Protocole de codage TEI

Modéliser le changement : les voies du français

---

21 mai 2010

Révision 1.11

Université d'Ottawa (Ontario, Canada)
www.voies.uottawa.ca
(dir. F. Martineau)

Conception : France Martineau et Yves Charles Morin
Rédaction : Caroline Fauchon
Collaboration : Simon-Pier Labelle Hogue

Table des matières

1 Introduction

Dans le souci d'une diffusion simple et efficace des textes, le projet *Modéliser le changement : les voies du français* (MCVF) a opté pour un encodage international: le TEI (Text Encoding Initiative) reposant sur le format XML. Ce format international a l'avantage de décrire des données, de les structurer et de transmettre de l'information brute, ce qui permet de partager aisément les textes via Internet sans avoir l'inconvénient de l'incompatibilité des systèmes. Ce codage permet de noter les méta-informations des textes avant leur analyse morphosyntaxique.

Le présent document expose la liste des balises utilisées par le projet MCVF ainsi que leur contexte d'utilisation. Notre protocole de codage a pour source essentielle le document Principes directeurs pour l'encodage et l'échange de textes électroniques (2007) élaboré par le consortium de TEI et édité par Lou Bernard et Syd Bauman.[1] Comme source secondaire, nous avons eu recours au manuel d'encodage de la *Base de Français Médiéval* élaboré par Heiden, Guillot & Lavrentiev.[2] En règle générale, le projet MCVF utilise les mêmes balises que la BFM, pour faciliter l'échange des textes. Mais étant donné la spécificité du MCVF, quelques balises spécifiques à certains textes ont été ajoutées au corpus du MCVF et diffèrent donc des balises utilisées par la BFM.

1 Bernard, Lou & Syd Bauman. (2007) Guidelines for Electronic Text Encoding and Interchange, Version P5 ; <http://www.tei-c.org/Guidelines/P5/>

2 Heiden S., Guillot C., Lavrentiev A. (2005) Manuel d'encodage BFM / XML-TEI, Version 2.1, BFM - Base de Français Médiéval [En ligne] ; Lyon : UMR ICAR / ENS-LSH <http://bfm.ens-lsh.fr/IMG/pdf/Manuel_Encodage_TEI.pdf >

2 Principes de base du TEI

Le codage TEI permet de noter toutes les informations concernant la structure interne du texte, sa mise en page (celle du manuscrit ou bien celle de l'édition de référence), certaines interventions éditoriales (ajout ou suppression d'un mot par l'éditeur, correction, etc.) et, dans le cas d'un manuscrit, il permet de rendre compte de son aspect visuel (rature, insertion interlinéaire, etc.).

Ces informations sont indiquées par l'entremise de balises xml. Ces dernières sont délimitées par des chevrons et se présentent normalement en paires : une balise ouvrante <...> et une balise fermante </...> qui encadrent un ou plusieurs mots ou une section du texte.

(exemple de balises)

```
<title> Les femmes savantes</title>
<head>Deuxième chapitre</head>
<persName>Victor Hugo</persName>
```

Une balise peut s'insérer entre d'autres balises. Par exemple, plusieurs balises de paragraphes peuvent s'insérer entre les deux parties de la balise d'un chapitre, etc.

(exemple de balises)

```
 <div>
<p> Il était une fois....</p>
</div>
```

Les balises indiquant une frontière se présentent seules, déjà fermées. C'est le cas, entre autres, de la balise <lb/> qui indique le début d'une nouvelle ligne typographique. Il en va de même pour la balise des sauts de pages, <pb/>. Ces balises possèdent le caractère «/» en suffixe, ce qui démontre que la balise est 'vide'.

(exemple)

```
<lb/>Nous étions à l'étude, quand le proviseur entra, suivi
<lb/>d'un nouveau habillé en bourgeois et d'un garçon de
<lb/>classe qui portait un grand pupitre.
```

Chacune des balises peut contenir des informations supplémentaires, informations que l'on appelle *attributs*. Les attributs sont toujours dotés d'une valeur. Par exemple, une division peut comporter l'attribut *type* et une valeur "chapitre".

(exemple d'attribut)

```
<div type="chapitre">
```

Nous verrons les balises et leurs attributs de façon plus détaillée dans la prochaine section.

Une note importante au sujet des balises : quoique le protocole TEI permette une grande liberté au niveau de l'insertion des balises, nous avons décidé de ne jamais insérer une balise à l'intérieur d'un mot. Les fichiers TEI sont acheminés vers une analyse morphosyntaxique – une analyse qui est facilitée si les mots ne sont pas coupés par une balise.

3 Les éléments communs d'un document TEI

Certaines balises peuvent apparaître dans tous les types de documents TEI, et ce dans toutes les sections de la structure textuelle.

3.1 Prose : paragraphes <p> et lignes <lb/>

Les changements de paragraphes et de lignes sont des unités structurales fondamentales de tout texte en prose.

- **<p>** Encadre un paragraphe en prose.
- **<lb/>** Changement de ligne en prose. Toujours au début de la ligne et toujours fermée. Au début d'un paragraphe, <p> remplace <lb/> ; un changement de paragraphe comprend implicitement un changement de ligne. L'attribut n peut être ajouté si les lignes de l'édition sont numérotées.

(exemple)

```
<p> Une nuit, vers onze heures, ils furent réveillés par
<lb/>le bruit d'un cheval qui s'arrêta juste à la porte.
<lb/>La bonne ouvrit la lucarne du grenier et parlementa (...)
</p>
```

3.2 Poésie : groupes de lignes <lg> et lignes de vers <l>

Si le texte n'est pas en prose, il faut utiliser les balises spécifiques à la poésie.

- **<lg>** Encadre un groupe de lignes
- **<l>** Encadre une ligne de vers

(exemple)

```
<lg>
<l>Autour des corps, qu'une mort avancée</l>
<l>Par violance a privez d'un beau jour,</l>
<l>Les ombres vont, et font maint et maint tour,</l>
<l>Aimans encor leur dépoüille laissée. </l>
</lg>
```

3.3 Les listes <list> et leur contenu <item>

On peut retrouver des listes dans plusieurs sections d'un document TEI, notamment dans les tables des matières (voir section 5.1)

- **<list>** Élément sous lequel l'on organise des éléments en liste. Peut être accompagné de l'attribut *type* (généralement avec la valeur "table_des_matieres").
- **<item>** Composante d'une liste. Peut être accompagnée de l'attribut *n* pour numéroter les items. Par exemple, <item n="2"> indique qu'il s'agit du deuxième item dans la liste.
- **<label>** Étiquette associée à un item dans une liste. (À noter que si un item est accompagné d'une étiquette, tous les items de cette liste doivent l'être aussi.)
- **<head>** En-tête, par exemple d'une section ou d'une liste.
- **<headLabel>** Intitulé pour la colonne d'étiquettes dans une liste, le cas échéant.
- **<headItem>** Intitulé pour la colonne d'items dans une liste, le cas échéant.

(exemple)

```
<list>
<head>Fruits</head>
        <item  n="1">Pomme</item>
        <item  n="2">Orange</item>
        <item  n="3">Banane</item>
</list>
```

(exemple)

```
<list>
<headLabel>PRÉFÉRENCE</headLabel>
<headItem>FRUITS</headItem>
        <label>(1)</label><item><Pomme</item>
        <label>(2)</label><item>Orange</item>
</list>
```

3.4 Le format des dates

On peut retrouver de nombreuses balises de dates dans un document TEI : <date>, <originDate>, etc. Il y a plusieurs façons d'encoder une date. Le degré de certitude de cette dernière viendra modifier le choix de l'attribut et de la valeur de la balise:

- **<date when=" ">** On utilise l'attribut *when* lorsqu'on connaît la date, soit en partie ou en totalité. La valeur est entrée sous le format AAAA-MM-JJ. Si l'on ne connaît que l'année, on peut laisser tomber le mois et le jour. Si l'on connaît le mois, mais pas l'année, on peut remplacer l'année par deux tirets.
- **<date notBefore=" " notAfter=" ">** On utilise ces attributs lorsque l'on a une date approximative. En règle générale, les tranches de siècle sont de 25 ans (par exemple, un texte du début du XV^{e} siècle sera daté entre 1400 et 1425.)

```
                                                    (exemple d'une date exacte)
<date when="2008-01-01">Le premier janvier 2008</date>
```

```
                                                    (exemple d'une date incomplète)
<date when="1599">En l'an 1599</date>
```

```
                                                    (exemple d'une date incomplète)
<date when="--05">Au mois de mai</date>
```

```
                                                    (exemple d'une tranche de siècle)
<date notBefore="1400" notAfter="1425">Début du 15e siècle</date>
```

3.5 Les balises générales de structure

Tout fichier TEI bien formé comprend des méta-informations au sujet du document, ainsi que sur le document comme tel. Les méta-informations sont regroupées sous la balise de l'en-tête <teiHeader> (voir section 4), tandis que l'ensemble du texte se retrouve sous la balise de l'élément <text> (voir section 5).

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
        <teiHeader><!-- éléments subordonnés--> </teiHeader>
        <text><!-- éléments subordonnés--> </text>
</TEI>
```

4 L'en-tête

L'en-tête TEI sert principalement à décrire le document. Il fournit les informations descriptives et déclaratives précédant chaque texte conforme à la TEI et qui permettent de constituer une page de titre électronique. Il permet de préciser une vaste gamme de méta-informations au sujet du texte: la provenance du document, des données biographiques sur l'auteur, le scribe et, le cas échéant, l'imprimeur, etc. Il sert aussi de mémoire à l'équipe de recherche car l'on y retrouve des informations au sujet des procédures de codage et des modifications apportées au texte.

Ces informations se retrouvent dans un ensemble d'éléments regroupés sous <teiHeader>. L'en-tête comporte quatre sections :

1. **<fileDesc>** Description bibliographique du fichier
2. **<encodingDesc>** Description de l'encodage
3. **<profileDesc>** Description du profil (du texte, de l'auteur, etc.)
4. **<revisionDesc>** Description des révisions

Nous décrirons les balises spécifiques à chacune de ces sections un peu plus bas. De façon générale, la structure supérieure de l'en-tête TEI est comme suit:

```
<teiHeader>
    <fileDesc>
   <!-- éléments subordonnés-->
   </fileDesc>
   <encodingDesc>
   <!-- éléments subordonnés-->
   </encodingDesc>
   <profileDesc>
   <!-- éléments subordonnés-->
   </profileDesc>
   <revisionDesc>
   <!-- éléments subordonnés-->
   </revisionDesc>
</teiHeader>
```

Note : <!--...--> indique qu'il y a des balises à insérer à l'intérieur de l'élément.

Afin de faciliter la tâche des encodeurs du projet MCVF, nous avons créé un en-tête modèle (voir Header.squelette.xml). Les encodeurs n'ont qu'à entrer les informations dans les balises appropriées.

4.1 La description du fichier <fileDesc>

L'élément <fileDesc> contient une description bibliographique complète du fichier électronique (c'est-à-dire, du fichier TEI comme tel). Il comprend quatre éléments :

- **<titleStmt>** Mention de titre
- **<extent>** Importance matérielle
- **<publicationStmt>** Mention de publication
- **<sourceDesc>** Description de la source

4.1.1 La mention du titre <titleStmt>

La mention du titre regroupe des informations sur le titre de l'œuvre et les personnes responsables de son contenu intellectuel.

4.1.1.1 Le titre <title>

Le projet MCVF fait la distinction entre trois différentes sortes de titres pour les documents TEI : le titre, le sous-titre et l'abréviation.

- **<title type="main">** Le titre complet de l'oeuvre codée.
- **<title type="sub">** Le sous-titre de l'oeuvre, le cas échéant.
- **<title type="short">** L'abréviation MCVF du document. Il s'agit du nom de l'auteur et d'une abréviation du titre, séparés par un point. Ce titre raccourci servira à identifier le document une fois transféré dans *Philologic*. Si l'auteur est anonyme, on écrit Anonyme.Titre. S'il s'agit d'auteurs divers, on écrit Divers.Titre.

(Exemple avec sous-titre)

```
<title type="main">Chroniques</title>
<title type="sub">Premier livre</title>
<title type="short">Froissart.Chroniques</title>
```

(Exemple sans sous-titre)

```
<title type="main">Les Cent Nouvelles Nouvelles</title>
<title type="sub"/>
<title type="short">Anonyme.CNN</title>
```

4.1.1.2 L'auteur <author>

Cet élément sert à identifier l'auteur de l'oeuvre.

- **<author>** Élément sous lequel on regroupe les informations au sujet de l'auteur

- **<persName>** Élément sous lequel on regroupe le nom et le prénom de l'auteur
- **<surname>** Le nom de l'auteur
- **<forename>** Le prénom de l'auteur
- **<addName>** Élément sous lequel on retrouve le nom de plume de l'auteur, ou tout autre graphie non standard d'importance

(Exemple avec un auteur)

```
<author>
        <persName>
                <surname>Froissart</surname>
                <forename>Jean</forename>
        </persName>
</author>
```

(Exemple d'un texte anonyme)

```
<author>
        <persName>
                <surname>Anonyme</surname>
                <forename/>
        </persName>
</author>
```

(Exemple avec plusieurs auteurs)

```
<author>
        <persName>
                <surname>Lorris</surname>
                <forename>Guillaume</forename>
        </persName>
        <persName>
                <surname>Meung</surname>
                <forename>Jean</forename>
        </persName>
</author>
```

(Exemple avec deux graphies)

```
<author>
        <persName>
                <surname>Boileau</surname>
                <forename>Étienne</forename>
                <addName>
                        <surname>Boilique</surname>
                        <forename>Estienne</forename>
                </addName>
        </persName>
</author>
```

4.1.1.3 <editor> <principal> et <funder>

Dans le cadre du projet MCVF, les informations à insérer pour ces éléments sont prédéterminées. Les données ne changent pas d'un document à l'autre, elles restent toujours les mêmes (voir ci-bas).

- **<editor>** L'institution responsable de l'édition. Prédéterminé.
- **<principal>** Le nom du chercheur principal. Prédéterminé.
- **<funder>** L'institution responsable du financement. Prédéterminé.

```
<editor>Modéliser le changement, les voies du français GTRC/CRSH</editor>
<principal>
        <persName>France Martineau</persName>
</principal>
<funder>
        <orgName>Université d'Ottawa</orgName>
        <orgName>Conseil de recherche en sciences humaines du Canada</orgName>
</funder>
```

Pour des collaborations, nous suivons un modèle déterminé avec l'organisme avec qui nous collaborons.

4.1.1.4 La mention de responsabilité <respStmt>

Cet élément fournit une mention de responsabilité pour le responsable du contenu intellectuel du document.

- **<respStmt>** Élément sous lequel on regroupe les informations au sujet des responsabilités
- **<resp>** Décrit la nature de la responsabilité
- **<persName>** Le nom du responsable, le cas échéant
- **<orgName>** Le nom de l'institution ou de l'organisme responsable, le cas échéant
- **xml:id=** Attribut qui permet d'identifier le ou la responsable. Ce code est inséré à l'intérieur de la balise <persName> ou <orgName> et doit être unique. Dans le cas d'une personne, nous avons généralement recours au nom comme moyen d'identification. Dans le cas d'un organisme, nous utilisons une abréviation. Dans les deux cas, il faut placer le code entre guillemets anglais. Ce code est essentiel pour l'entrée des modifications apportées au document (voir la section 4.4)

(Exemple avec personnes et institution)

```
<respStmt>
        <resp>Codage</resp>
        <persName xml:id="Tailleur">Sandrine Tailleur</persName>
        <persName xml:id="Fauchon">Caroline Fauchon</persName>
        <orgName xml:id="BFM">Base de français médiéval</orgName>
 </respStmt>
```

La mention de responsabilité est le dernier élément de la mention du titre. Pour plus d'informations au sujet du <titleStmt>, consulter le site de TEI :
http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD2.

4.1.2 L'importance matérielle <extent>

L'importance matérielle décrit la taille approximative d'un document. Nous mesurons la taille par le nombre de mots contenus dans le texte.

Exemples

```
<extent>250 438</extent>
<extent>323</extent>
```

4.1.3 La mention de publication <publicationStmt>

L'élément <publicationStmt> regroupe les informations relatives à la publication ou à la diffusion du document.

- **<date>** Date de publication en ligne du document TEI, entrée selon le format : AAAA-MM-JJ.
- **<publisher>** Nom de l'institution responsable de l'édition du document TEI. Par défaut, on met les coordonnées du MCVF. Si le texte nous provient d'une autre institution, on met les coordonnées de cette institution.
- **<distributor>** Nom de l'institution responsable de la diffusion du document TEI. Par défaut, on met les coordonnées du MCVF. Si le texte nous provient d'une autre institution, on met les coordonnées de cette institution.
- **<authority>** Nom de l'institution responsable de la mise en ligne du document TEI. Si le texte nous provient d'une autre institution, l'autorité est le Projet MCVF.
- **<availability>** Informations sur la disponibilité du texte, comme des restrictions ou des copyright.

Exemple d'un texte du MCVF

```
    <publicationStmt>
            <date when="2008-01-01"/>
            <publisher>
                    <orgName>Modéliser le changement : les voies du français
            GTRC/CRSH</orgName>
                    <address>
                    <addrLine>60, rue Université</addrLine>
                    <addrLine>Ottawa, Ontario</addrLine>
                    <addrLine>K1N 6N5</addrLine>
                    <addrLine>Tél : 613-562-5797</addrLine>
                    <addrLine>Fax : 613-562-5981</addrLine>
                    <addrLine><email>fmartin@uottawa.ca</email></addrLine>
                    <addrLine>http://www.voies.uottawa.ca/</addrLine>
                    </address>
            </publisher>
            <distributor>
                    <orgName>Modéliser le changement : les voies du français
                    GTRC/CRSH</orgName>
                    <address>
                    <addrLine>60, rue Université</addrLine>
                    <addrLine>Ottawa, Ontario</addrLine>
                    <addrLine>K1N 6N5</addrLine>
                    <addrLine>Tél : 613-562-5797</addrLine>
                    <addrLine>Fax : 613-562-5981</addrLine>
                    <addrLine><email>fmartin@uottawa.ca</email></addrLine>
                    <addrLine>http://www.voies.uottawa.ca/</addrLine>
                    </address>
            </distributor>
      <availability status="restricted">
        <p>(c) Modéliser le changement : les voies du français (CRSH/GTRC, dirigé par F.
        Martineau)</p>
         <p>Les Corpus MCVF demeurent la propriété du projet Modéliser le changement : les
        voies du français. Les références à ces Corpus devraient apparaître ainsi : Corpus MCVF,
        Modéliser le changement : les voies du français (F. Martineau (dir.), U. d'Ottawa). Tous
        les documents du Corpus MCVF ne peuvent être divulgués à de tierces parties sans
        l'autorisation préalable de la Directrice du Projet </p>
       <p>Pour tout autre usage, contacter :
        <persName>France Martineau</persName>
        <orgName>Modéliser le changement : les voies du français, Université
        d'Ottawa</orgName>
        <email>fmartin@uottawa.ca</email></p>
      </availability>
    </publicationStmt>
```

Exemple d'un texte qui nous provient de la BFM

```
<publicationStmt>
                <date when="2008-01-01"/>
                <publisher>
                        <orgName>Projet BFM, UMR5191 ICAR, CNRS/ENS-LSH</orgName>
                        <address>
                        <addrLine>15, parvis René Descartes</addrLine>
                        <addrLine>69342 Lyon BP7000 Cedex 07</addrLine>
                        <addrLine>Tél : 04 37 37 63 10</addrLine>
                        <addrLine>Fax : 04 37 37 62 65</addrLine>
                        <addrLine><email>fmartin@uottawa.ca</email></addrLine>
                        <addrLine>bfm@ens-lsh.fr/</addrLine>
                        </address>
                </publisher>
                <distributor>
                        <orgName>Projet BFM, UMR5191 ICAR, CNRS/ENS-LSH</orgName>
                        <address>
                        <addrLine>15, parvis René Descartes</addrLine>
                        <addrLine>69342 Lyon BP7000 Cedex 07</addrLine>
                        <addrLine>Tél : 04 37 37 63 10</addrLine>
                        <addrLine>Fax : 04 37 37 62 65</addrLine>
                        <addrLine><email>fmartin@uottawa.ca</email></addrLine>
                        <addrLine>bfm@ens-lsh.fr/</addrLine>
                        </address>
                        </distributor>
                <authority>
                        <orgName>Modéliser le changement : les voies du français
                        GTRC/CRSH</orgName>
                        <address>
                        <addrLine>60, rue Université</addrLine>
                        <addrLine>Ottawa, Ontario</addrLine>
                        <addrLine>K1N 6N5</addrLine>
                        <addrLine>Tél : 613-562-5797</addrLine>
                        <addrLine>Fax : 613-562-5981</addrLine>
                        <addrLine><email>fmartin@uottawa.ca</email></addrLine>
                        <addrLine>http://www.voies.uottawa.ca/</addrLine>
                        </address>
                </authority>
                <availability status="restricted">
                        <p>c) 2002-2005, CNRS/ENS-LSH</p>
        <p>Conditions d'utilisation : usage dans le cadre de l'échange de textes entre les projets
        BFM (ENS LSH, UMR 5191 ICAR, Lyon, France) et GTRC "Modéliser le changement :
        Les voies du français" (Université d'Ottawa,Canada)</p>
       <p>Pour tout autre usage, contacter : <persName>Christiane Marchello-Nizia</persName>
        <orgName>Base de français médiéval</orgName>
        <email>marchell@linguist.jussieu.fr</email></p>
      </availability>
</publicationStmt>
```

4.1.4 La description de la source <sourceDesc>

L'élément <sourceDesc> sert à décrire la ou les sources d'où le fichier TEI est dérivé. Le contenu de la description de la source dépend de la source : s'il s'agit d'une source publiée, il faut générer une citation bibliographique ; s'il s'agit d'un texte non publié, il faut générer une description de manuscrit.

4.1.4.1 La liste des citations bibliographiques <listBibl>

Si le texte a été publié – sous presses ou sur Internet – l'on regroupe toutes les informations bibliographiques sous une liste de citations bibliographiques. Cette liste peut comprendre trois éléments :

- **<biblStruct type="Princeps" >** Élément sous lequel l'on regroupe les informations de la première édition du texte publié, le cas échéant (pour les textes du 16e et 17e siècles).
- **<biblStruct type="Source_publique" >** Élément sous lequel l'on regroupe les informations de la source publiée publique du texte.
- **<bibl type="Source_électronique" >** Élément sous lequel l'on regroupe les informations de la source électronique du texte, le cas échéant. (Par exemple, une image digitale sur Internet.)

```
<listBibl>
<biblStruct type="Princeps"><!-- éléments subordonnés--> </ biblStruct >
<biblStruct type="Source_publique"><!-- éléments subordonnés--> </ biblStruct >
<bibl  type="Source_électronique"><!-- éléments subordonnés--> </ bibl >
</listBibl>
```

4.1.4.2 La première édition <biblStruct type="Princeps" > et la source publique <biblStruct type="Source_publique" >

L'élément <biblStruct> contient une citation bibliographique structurée. Pour les textes imprimés à partir du 16e siècle, l'on regroupe les informations de la première édition sous le type *Princeps*. Pour les éditions contemporaines, l'on les regroupe sous le type *Source publique*. Sous le <biblStruct>, l'on regroupe les sous-éléments bibliographiques suivants et ce, dans cet ordre précis :

- **<biblStruct>** Élément sous lequel l'on regroupe les informations bibliographiques
- **<analytic>** Élément sous lequel l'on regroupe les informations au sujet d'une ressource (par exemple un poème ou un article de revue) publiée à l'intérieur d'une monographie et non publiée de façon indépendante
- **<monogr>** Élément sous lequel l'on regroupe les informations au sujet d'un objet publié en tant qu'objet indépendant (par exemple, une monographie)
- **<author>** Informations au sujet de l'auteur (voir 4.1.1.2 pour plus de détails) Peu être dédoublé s'il y a plus qu'un auteur.
- **<editor>** Informations au sujet de l'éditeur. Peu être dédoublé s'il y a plus qu'un éditeur.
- **<title>** Titre de la monographie

- **<imprint>** Élément sous lequel l'on regroupe les informations au sujet du lieu de publication
- **<pubPlace>** Lieu de publication
- **<publisher>** Élément sous lequel l'on regroupe les informations au sujet de la maison d'édition
- **<orgName>** Maison d'édition
- **<persName>** sous <publisher>, nom et prénom de l'imprimeur (pour Princeps seulement)
- **<date>** Date de publication
- **<biblScope>** Extension d'une référence bibliographique, (par défaut, on donne les pages)
- **<series>** Élément sous lequel l'on regroupe les informations au sujet de la série, le cas échéant

exemple d'un chapitre

```
<biblStruct type="Source_publique">
                <analytic>
                   <title>Journal militaire tenu par Nicolas Renaud d'Avène Des Méloizes,
                      Cher, Seigneur de Neuville, au Canada du 19 juillet 1756 au 30
                      octobre de la même année et du 8 mai 1759 au 21 novembre de la même
                      année</title>
                   <author>
                      <persName>
                         <forename>Nicolas Renaud d'Avène des</forename>
                         <surname>Méloizes</surname>
                      </persName>
                   </author>
                </analytic>
                <monogr>
                   <title>Rapport de l'archiviste de la province du Québec 1928-1929</title>
                   <editor>
                      <persName>
                         <forename>Pierre-Georges</forename>
                         <surname>Roy</surname>
                      </persName>
                   </editor>
                   <imprint>
                      <pubPlace>Québec</pubPlace>
                      <publisher><orgName>Rédempti Paradis</orgName></publisher>
                      <date when="1929">1929</date>
                   </imprint>
                   <biblScope>pp 5-86</biblScope>
                </monogr>
             </biblStruct>
```

```
<biblStruct type="Princeps">                                    exemple d'une monographie
            <monogr>
              <author>
                <persName>
                  <surname>Montaigne</surname>
                  <forename>Michel de</forename>
                </persName>
              </author>
              <editor>
                <persName>
                  <surname>De Gournay</surname>
                  <forename>Marie Le Jarsforename>
                </persName>
              </editor>
              <title>Les Essais</title>
              <title type="sub">Livre trois</title>
              <imprint>
                <pubPlace>Paris</pubPlace>
                <publisher>
                    <orgName>Abel L'Angelier</orgName>
                    <persName>
                     <surname>L'Angelier</surname>
                    <forename>Abel</forename>
                    </persName>
                </publisher>
                <date when="1595">1595</date>
              </imprint>
            </monogr>
  </biblStruct>
```

4.1.4.3 La source électronique <bibl type="source_électronique" >

Si nous nous sommes fiés à un document électronique pour faire la transcription (par exemple, une image digitale ou une transcription trouvée en ligne), nous notons les informations pertinentes sous <bibl> :

- **<orgName>** Nom de l'organisme qui gère le site web
- **<note type="Type_de_document">** Type de document électronique.
- **<ref target="http://www.exemple.ca">** Adresse URL du site web.
- **<note>** Autres informations pertinentes, le cas échéant.

exemple

```
<bibl type="Source_électronique">
<orgName>Gallica</orgName>
 <note type="Type_de_document">Image digitale</note>
<ref target="http://www.exemple.ca"> www.exemple.ca</ref>
</bibl>
```

4.1.4.4 La description d'un manuscrit <msDesc>

Si le document source n'a pas été publié, nous regroupons les informations sous l'élément <msDesc>, qui permet de décrire les informations liées au manuscrit.

4.1.4.5 L'identification du manuscrit <msIdentifier>

Nous identifions d'abord le manuscrit.

- **<msIdentifier>** Élément sous lequel l'on regroupe les informations requises pour identifier le manuscrit.
- **<settlement>** Lieu où se trouve le centre d'archives (habituellement la ville ou le pays)
- **<repository>** Nom du centre d'archives
- **<collection>** Nom de la collection dans laquelle se trouve le manuscrit
- **<idno>** Cote complète du manuscrit, telle que donnée par le centre d'archives
- **<msName>** Nom donné au manuscrit, le cas échéant

(exemple d'un manuscrit)

```
<sourceDesc>
    <biblStruct/>
     <msDesc>
      <msIdentifier>
         <settlement>Vatican</settlement>
         <repository>Bibliothèque du Vatican</repository>
        <repository> Fonds de la reine de Suède </repository>
         <idno> Reg. lat. 869</idno>
         <msName>Manuscrit de Rome</msName>
      </msIdentifier>
```

4.1.4.6 L'historique du manuscrit <history>

Une fois le manuscrit identifié, nous rassemblons les informations au sujet de l'historique de ce dernier.

- **<history>** Élément sous lequel l'on regroupe les informations sur l'historique du manuscrit
- **<origin>** Élément sous lequel l'on regroupe les informations sur la création du manuscrit
- **<originPlace>** Lieu de création du manuscrit
- **<originDate>** Date de création du manuscrit (voir section 3.4 pour le format des dates)

(exemple d'un manuscrit)

```
<sourceDesc>
    <biblStruct/>
     <msDesc>
      <msIdentifier><!-- éléments subordonnés--></msIdentifier>
       <history>
         <origin>
            <origPlace>Paris<origPlace>
            <origDate notBefore="1399" notAfter="1410"/>
         </origin>
       </history>
   </msDesc>
</sourceDesc>
```

La description de la source est le dernier élément de la description du fichier <fileDesc>. Pour de plus amples informations ou pour d'autres exemples, veuillez visiter le site web des Principes directeurs pour l'encodage TEI : http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/HD.html#HD2

4.2 La description de l'encodage <encodingDesc>

La section <encodingDesc> documente la relation entre un texte électronique et la ou les sources dont il est dérivé. Elle comprend deux grands éléments :

- **<projectDesc>** Description du but du projet d'encodage
- **<editorialDecl>** Déclaration des pratiques éditoriales

4.2.1 La description du projet d'encodage <projectDesc>

La description du projet décrit en détail le but visé par l'encodage d'un fichier électronique, ainsi que tout autre information pertinente sur la procédure de collecte ou d'assemblage[3]. Dans le cadre du MCVF, ces informations sont prédéterminées et doivent être entrées comme suit : (Note: ces informations se trouvent déjà dans l'en-tête modèle)

[3] Martineau, F., R. Diaconescu et P. Hirschbühler. 2007. « Le Corpus *Voies du français* : de l'élaboration à l'annotation », dans Kunstmann, Pierre & Stein, Achim (éds) *Le Nouveau Corpus d'Amsterdam*, p. 121-142. Actes de l'atelier de Lauterbad, 23-26 février 2006, Stuttgart: Steiner.

```
<projectDesc>
  <p>Projet : Modéliser le changement les voies du français GTRC/CRSH
  <lb/>Resp. :<persName>France Martineau</persName>, Département de
      français, Université d'Ottawa
    <address>
      <addrLine>60, rue Université</addrLine>
      <addrLine>Ottawa, Ontario</addrLine>
      <addrLine>K1N 6N5</addrLine>
      <addrLine>Tél : 613-562-5797</addrLine>
      <addrLine>Fax : 613-562-5981</addrLine>
      <addrLine><email>fmartin@uottawa.ca</email></addrLine>
      <addrLine>www.voies.uottawa.ca</addrLine></address></p>
</projectDesc>
```

4.2.2 La déclaration des pratiques éditoriales <editorialDecl>

Sous l'élément <editorialDecl>, l'on fournit des détails sur les pratiques et principes éditoriaux appliqués lors de l'encodage du texte. Dans le cadre du MCVF, ces informations sont prédéterminées et doivent être entrées comme suit: (Note: ces informations se trouvent déjà dans l'en-tête modèle)

```
<editorialDecl>
<p>Voir le manuel d'annotation TEI au http://www.voies.uottawa.ca</p>
<p>L'édition de référence a été respectée quant à l'orthographe, la ponctuation, la mise en page et la pagination. Cependant, la césure des mots en fin de lignes (ou de pages) a été supprimée, le mot est placé dans la ligne ou la page où il commence. </p>
<p>Les paragraphes sont notés par P pour la prose. Pour les vers, P rend="lg" indique que le paragraphe est constitué d'un groupe de vers. </p>
<p> Les débuts des vers sont encodés par LB. Cette balise peut également servir dans
les textes en prose à indiquer les débuts de lignes selon la pagination de l'édition. </p>
<p>Les changements de pages sont indiqués par la balise PB</p>
<p>Le corps du texte est obligatoirement compris dans le BODY. Le FRONT contenant le titre et/ou le prologue et le BACK (explicit) sont facultatifs et ne sont mis qu'au besoin. </p>
<p>Les différentes parties du texte (livre, chapitre...) sont encodées par DIV, auquel s'ajoute l'attribut TYPE qui spécifie sa nature. </p>
<p>Les paragraphes (P), lignes (LB) et pages (PB) de même de certaines divisions (DIV) sont ou peuvent être numérotés par l'attribut N s'ils le sont dans l'édition. </p>
<p>Quelques commentaires ou corrections ont été introduits à l'aide de la balise NOTE, CORR ou SIC. </p>
<p>Les discours directs et les citations ont été encodés par la balise Q. Cette
balise est mise à chaque fois que des guillemets sont utilisés dans l'édition. </p>
<p>L'élément FOREIGN indique un passage en latin qui est noté en italique dans l'édition, les occurrences non en italique sont signalées par l'attribut ROMAIN. </p>
<p>Les folios sont notés par la balise <tag>pb rend="folio"</tag></p>
<p>Les notes de bas de pages ont été retirées. </p>
 <p>Les paragraphes entièrement en latin ont été retirés, une balise P et une NOTE les signalent. </p>
<p>Note spécifique au texte ci-dessous: Aucune.</p>
</editorialDecl>
```

Si l'encodeur a une note à ajouter au sujet de l'encodage d'un texte en particulier, il peut l'ajouter au bas de la déclaration des pratiques d'encodage.

4.3 La description du profil <profileDesc>

La section <profileDesc> offre une description détaillée des aspects non bibliographiques d'un texte. Elle comprend cinq ensembles d'éléments:

- **<creation>** Création du document
- **<langUsage>** Langue utilisée dans le document
- **<textDesc>** Description du document
- **<textClass>** Classification du document
- **<particDesc>** Description des participants

4.3.1 La création du document <creation>

L'élément <creation> contient des informations concernant la création du texte. Ces informations sont essentielles, particulièrement pour les ouvrages publiés. L'année et le lieu de création d'un texte correspondent rarement à la date et au lieu de publication. Il faut donc s'assurer de faire la distinction entre les deux.

- **<creation>** Élément sous lequel l'on regroupe les informations sur la création du document
- **<date>** Date de création du document (à ne pas confondre avec la date de publication)
- **<placeName>** Lieu de création du document (à ne pas confondre avec le lieu de publication)

(exemple)

```
<creation>
   <date notBefore="1500" notAfter="1600">1500-1600</date>
   <placeName>Londres</placeName>
</creation>
```

4.3.2 La langue utilisée dans le document <langUsage>

L'élément <langUsage> contient des informations concernant la langue ou les langues employée(s) dans le document.

- **<language ident=" ">** Élément qui caractérise une seule langue employée dans le texte. L'attribut *ident* fournit un code ISO qui permet d'identifier la langue précise. On utilise le code *fr* pour le français moderne, *frm* pour le français moyen (1400-1600) et *fro* pour l'ancien français. Voir http://hapax.qc.ca/iso639-2.fr.htm pour la liste complète des codes ISO.

(exemple)

```
<langUsage>
   <language ident="frm">Français moyen</language>
</langUsage>
```

4.3.3 La description du document <textDesc>

L'élément <textDesc> fournit une brève description du texte. Il y a huit sous-éléments. Toutefois, dans le cadre du projet MCVF, le seul élément que l'on doit remplir est la composition du texte <constitution>. Les sept autres éléments peuvent rester vides.

- **<contitution>** Composition du texte. Deux seules réponses possibles : Texte intégral ou extrait.

(exemple)

```
<textDesc>
        <channel/>
         <constitution>Texte intégral</constitution>
        <derivation/>
        <domain/>
        <factuality/>
        <interaction/>
        <preparedness/>
        <purpose/>
</textDesc>
```

4.3.4 La classification du document <textClass>

La classification du document se fait par une liste de mots-clés ou d'expressions décrivant la nature ou le sujet du texte. Cette liste a été développée spécialement pour le projet MCVF afin de décrire le document source de la façon la plus détaillée possible.

- **<textClass>** Élément sous lequel l'on regroupe des informations décrivant la nature du texte
- **<keywords scheme="GTRC">** Élément sous lequel l'on insère une liste de mots-clés
- **<list>** Liste sous laquelle l'on regroupe une série de mots-clés
- **<item n=" ">** Mot-clé à insérer. Le type de mot-clé dépend de l'attribut *n* à l'intérieur de la balise. Par exemple <item n="domaine" > indique qu'il faut définir le domaine du document et <item n="genre" > indique qu'il faut définir le genre.

(exemple)

```
<textClass>
        <keywords scheme="GTRC">
        <list>
                <item n="domaine">Littéraire</item>
                <item n="genre">Nouvelle / Fabliau</item>
                <!—autres items de la liste-->
        </list>
        </keywords>
</textClass>
```

Il y a neuf types de mots-clés pour la classification du document. Nous les verrons en détails ci-dessous.

4.3.4.1 Le domaine <item n="domaine">

Le premier mot-clé à définir est le domaine du document. Tous les documents relèvent soit du domaine *Littéraire*, soit du domaine *Non littéraire*.

```
<item n="domaine">Littéraire</item>
Ou
<item n="domaine">Non littéraire</item>
```

4.3.4.2 Le genre <item n="genre">

Le genre du document dépend directement du domaine. Nous faisons la distinction entre les genres suivants:

Domaine	Littéraire	Non littéraire
Genre	Biographie Chanson de geste Chroniques Essai Hagiographie Nouvelle / Fabliau Poésie Roman	Annales Correspondance Document administratif Journal / Mémoire personnel Journal de compte Ouvrage didactique Relation de voyage

Lors de l'encodage, il faut s'assurer que le genre corresponde au domaine.

(exemple 1)

```
<item n="domaine">Non littéraire</item>
<item n="genre">Relation de voyage</item>
```

(exemple 2)

```
<item n="domaine">Littéraire</item>
<item n="genre">Hagiographie</item>
```

4.3.4.3 La région du document <item n="docRegion">

Les trois prochains mots-clés – docRegion1, docRegion2 et docRegion3 – décrivent le lieu de création du document. Le deuxième niveau de la région dépend du premier, et le troisième dépend du deuxième.

docRegion1	docRegion2	docRegion3
Angleterre		
Belgique (sauf Picardie)		
Picardie (incluant scripta picarde en Flandre)		
France (sauf Picardie)	Région inconnue	
	France Centre	Auvergne
		Berry
		Île-de-France
		Orléans
		Touraine
	France Est	Bourgogne
	France Nord	Bretagne
		Sarthes
		Normandie
	France Ouest	Anjou
		Bordeaux
		Champagne
		Charente

		Poitou
		Saintonge
		Vendée
Amérique française	Région inconnue	
	Acadie	
	Grands Lacs	
	Vallée du Saint-Laurent	

(exemple 1)

```
<item n="docRegion1">France</item>
<item n="docRegion2">France Nord</item>
<item n="docRegion3">Picardie</item>
```

(exemple 2)

```
<item n="docRegion1">France</item>
<item n="docRegion2">France Centre</item>
<item n="docRegion3">Orléans</item>
```

Quand il n'y a pas de données à entrer (par exemple : docRegion2 et docRegion3 pour la Belgique), nous laissons l'espace réservé à ce mot-clé vide.

(exemple)

```
<item n="docRegion1">Amérique française</item>
<item n="docRegion2">Acadie</item>
<item n="docRegion3"/>
```

4.3.4.4 La région de l'auteur <item n="auteurRegion">

Les trois prochains mots-clés – auteurRegion1, auteurRegion2 et auteurRegion3 – décrivent le lieu de naissance de l'auteur. La liste des régions est la même que celle de la région du document (voir section 4.3.3.3 ci-dessus).

(exemple)

```
<item n="auteurRegion1">France</item>
<item n="auteurRegion2">France Nord</item>
<item n="auteurRegion3">Picardie</item>
```

4.3.4.5 La région du coauteur <item n="auteur2Region">

Les trois prochains mots-clés – auteur2Region1, auteur2Region2 et auteur2Region3 – décrivent le lieu de naissance du coauteur du texte, le cas échéant. La liste des régions est la même que celle de la région du document (voir section 4.3.3.3 ci-dessus). Si le texte n'a qu'un auteur, on laisse ces balises vides.

4.3.4.6 Le sexe de l'auteur <item n="auteurSexe">

Nous indiquons si l'auteur est *Homme* ou *Femme.* Si nous ne connaissons pas l'auteur, et donc le sexe de l'auteur, nous mettons *Inconnu*.

(exemple)

```
<item n="auteurSexe">Homme</item>
```

4.3.4.7 Le sexe du coauteur <item n="auteur2Sexe">

Si le texte a été écrit par deux auteurs, on indique le sexe du coauteur : *Homme*, *Femme* ou *Inconnu*. On laisse la balise vide si le texte n'a qu'un seul auteur.

4.3.4.8 L'occupation de l'auteur <item n="occupation">

Lorsque possible, nous cherchons à identifier l'occupation de l'auteur. Il y a treize mots-clés possibles, liste que nous augmentons au fur des besoins. Cette distinction demeure très large et sert d'indication de la classe sociale :

Administrateur	Marchand	Religieux
Agriculteur	Militaire	Inconnue
Artisan	Milice	Divers
Érudit	Noble	
Habitant	Ouvrier	

(exemple)

```
<item n="occupation">Religieux</item>
```

4.3.4.9 L'occupation du coauteur <item n="occupation2">

Si le texte a plus qu'un auteur, on indique l'occupation du coauteur sous <item n="occupation2">, suivant la liste ci-haut. Si le texte n'a qu'un seul auteur, on laisse la balise vide.

4.3.4.10 La région du scribe <item n="scribeRegion">

Les trois prochains mots-clés – scribeRegion1, scribeRegion2 et scribeRegion3 – décrivent le lieu de naissance du scribe (pour les textes du Moyen Âge). La liste des régions est la même que celle de la région du document (voir section 4.3.3.3) Si l'on ne connaît pas le lieu de naissance du scribe ou s'il ne s'agit pas d'un texte du Moyen Âge, on peut laisser ces mots-clés vides.

(exemple)

```
<item n="scribeRegion1">France</item>
<item n="scribeRegion2">France Est</item>
<item n="scribeRegion3">Bourgogne</item>
```

4.3.4.11 La région de l'imprimeur <item n="imprimeurRegion">

Les derniers mots-clés – imprimeurRegion1, imprimeurRegion2 et imprimeurRegion3 – indiquent la région de la première maison d'édition, le cas échéant (pour les textes à partir du XVI^{e} siècle). La liste des régions est la même que celle de la région du document (voir section 4.3.3.3) S'il s'agit d'un texte sans imprimeur, on peut laisser ces mots-clés vides.

(exemple d'un texte sans imprimeur)

```
<item n="imprimeurRegion1"/>
<item n="imprimeurRegion2"/>
<item n="imprimeurRegion3"/>
```

4.3.4.12 Dernière note au sujet des mots-clés

Nous avons terminé notre survol de la liste des mots-clés. Soulignons qu'il est primordial de respecter scrupuleusement l'orthographe de ces derniers, car une seule erreur de frappe pourrait entraîner des

erreurs par mots-clés lorsque des recherches seront faites.

4.3.4.13 L'indice de Dees

Le dernier élément de la classification du document est l'indice de Dees; un indice attribué à certains textes du Moyen Âge par Anthonij Dees dans l'Atlas des formes linguistiques des textes littéraires de l'ancien français. Consulter le livre pour trouver la liste des textes concernés. Pour tous les autres textes, nous pouvons laisser cette balise vide.

```
                                                    (exemple d'un texte sans Indice de Dees)
        <classCode scheme="Indice de Dees"/>
```

4.3.5 La description des participants <particDesc>

La description des participants est la quatrième et dernière section de la description du profil du document. L'on y décrit les intervenants sur le texte. Dans le cadre du MCVF, ces intervenants sont l'auteur du document, le scribe (pour les documents du Moyen Âge) et l'imprimeur. Nous cherchons à donner le plus d'informations possibles à leur sujet, notamment en ce qui concerne leur lieu de naissance. (Notons que ces informations sont complémentaires à celles présentées dans la classification du document. Par exemple, si l'on a indiqué qu'un auteur était né en Acadie, c'est dans la description des participants que l'on indiquerait la ville et la date de sa naissance.)

- **<particDesc>** Élément sous lequel l'on regroupe les informations sur les intervenants.
- **<listPerson>** Élément sous lequel l'on regroupe une liste d'intervenants
- **<person role=" ">** Élément sous lequel l'on regroupe les informations d'un intervenant en particulier. Le rôle de l'intervenant dépend de l'attribut *n* à l'intérieur de la balise. Par exemple <person role="auteur"> indique qu'il s'agit d'informations sur l'auteur; <person role="auteur2"> pourr le coauteur (le cas échéant) ; <person role="scribe"> pour le scribe.
- **<persName>** Élément sous lequel on regroupe les informations sur le nom de l'intervenant (voir section 4.1.1.2 pour plus de détails)
- **<birth>** Élément sous lequel l'on regroupe les informations au sujet de la naissance de l'intervenant.
- **<date>** Date de naissance de l'intervenant
- **<placeName>** Lieu de naissance de l'intervenant

(exemple avec un seul auteur et un scribe inconnu)

```
<profileDesc>
        <creation><!-- éléments subordonnés--></creation>
        <textDesc><!-- éléments subordonnés--></textDesc>
        <textClass><!-- éléments subordonnés--></textClass>
        <particDesc>
          <listPerson>
            <person role="auteur">
              <persName>
                <surname>Froissart</surname>
                <forename>Jean</forename>
              </persName>
              <birth>
                <date notBefore="1333" notAfter="1337"/>
                <placeName>Picardie, France</placeName>
              </birth>
            </person>
            <person role="scribe">
              <persName>
                <surname>Inconnu</surname>
                <forename/>
              </persName>
              <birth/>
            </person>
          </listPerson>
        </particDesc>
</profileDesc>
```

La description des intervenants est le dernier élément de la description du profil <profileDesc>. Pour de plus amples informations ou pour d'autres exemples, veuillez visiter le site web des Principes directeurs pour l'encodage TEI : http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/HD.html#HD4

4.4 La description des révisions <revisionDesc>

La description des révisions est la dernière section de l'en-tête TEI. On y fournit un résumé de l'historique des révisions du fichier TEI comme tel.

- **<revisionDesc>** Élément sous lequel l'on regroupe les informations sur les révisions.
- **<change when=" " who="# ">** Résumé d'une modification ou d'une correction apportée au document. Chaque modification est datée à l'aide de l'attribut *when* accompagné de la date entre

guillemets anglais. La personne responsable de la modification est identifiée par l'attribut *who,* suivi de son code d'identification tel que déterminé dans la mention des responsabilités (voir la section 4.1.1.4 pour plus d'informations)

(exemple d'un historique très peu détaillé)

```
<revisionDesc>
   <change when="2008-02-15" who="#Fauchon">En-tête au complet</change>
   <change when="2007-07-12" who="#Tailleur">Vérification du codage</change>
   <change when="2005-10-24" who="#Brazeau">Codage TEI</change>
</revisionDesc>
```

L'historique des modifications peut être plus ou moins détaillé, selon les besoins de l'encodeur. Nous recommandons fortement d'être le plus précis possible, afin de faciliter le travail de vérification. Pour de plus amples informations ou pour d'autres exemples, veuillez visiter le site web des Principes directeurs pour l'encodage TEI : http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/HD.html#HD6

5 La structure générale d'un texte

Comme mentionné plus haut, la structure d'un fichier TEI doit comprendre un en-tête <teiHeader> et un texte. L'élément <text> encadre le texte en entier, c'est-à-dire le corps du texte et, s'il y a lieu, toute partie qui le précède (par exemple, le prologue) ou qui le suit (par exemple, l'explicit).

- **<text>** Élément sous lequel l'on retrouve le texte en entier
- **<front>** Parties préliminaires, le cas échéant
- **<body>** Le corps du texte
- **<group>** Corps d'un texte composite qui regroupe une suite de textes distincts
- **<back>** Parties postliminaires, le cas échéant

Nous décrirons les balises spécifiques à chacune de ces sections un peu plus bas. De façon générale, la structure générale, d'un texte est comme suit:

(exemple)

```
<TEI>
<teiHeader><!-- éléments subordonnés--></teiHeader>
    <text>
        <front><!-- éléments subordonnés--></front>
        <body><!-- éléments subordonnés--></body>
        <back<!-- éléments subordonnés--></back>
    </text>
</TEI>
```

5.1 Les parties préliminaires <front>

L'élément <front> contient tout ce qui est au début du document, avant le corps du texte : en-têtes, page titre, préface, table des matières, etc. Exception faite des pages titres, les parties préliminaires sont délimitées simplement par la balise de division , accompagnée de l'attribut *type*. Le tableau suivant indique la valeur de l'attribut pour les différents types de parties préliminaires:

Partie préliminaire	Balise et attribut
Incipit	
Préface	
Prologue	
Table des matières	

- **<front>** Élément sous lequel l'on retrouve les parties préliminaires
- ** ** Élément par lequel l'on délimite les parties préliminaires
- **<head>** En-tête, le cas échéant
- **<list>** Liste, notamment pour les tables des matières (voir section 3.3 pour plus d'informations)

(exemple)

```
<front>
<div type="prologue">
        <head>Les XV joies de mariage</head>
        <p> Pluseurs ont travaillé a moustrer par grans raisons
        <lb/> et auctorités que c'est plus grant felicité en terre a
        <lb/> homme de vivre... </p>
</div>
</front>
```

5.1.1 Les pages titres <titlePage>

Les pages titres, contrairement aux autres parties préliminaires, suivent un format particulier.

- **<titlePage>** Élément sous lequel l'on retrouve les informations sur la page titre
- **<docTitle>** Titre du document tel que donné sur la page titre. Peut être accompagné de l'attribut *type*, avec la valeur "main" pour le titre principal et "sub" pour le sous-titre.
- **<titlePart>** Section du titre d'un ouvrage telle qu'elle est indiquée sur la page de titre
- **<byline>** Mention de responsabilité principale pour une œuvre donnée sur sa page titre
- **<docAuthor>** Auteur du document tel que donné sur la page titre
- **<docEdition>** Mention d'édition telle que donnée sur la page titre
- **<docImprint>** Mention de dénomination commerciale d'éditeur (lieu et date de publication, nom de l'éditeur), telle que généralement donnée au bas de la page de titre
- **<docDate>** Date du document telle que donnée sur la page titre

(exemple)

```
<front>
<titlePage>
        <docTitle>Madame Bovary</docTitle>
        <byline>par
                <docAuthor>Gustave Flaubert</docAuthor>
        </byline>
</titlePage>
</front>
```

Pour de plus amples informations et pour d'autres exemples, consulter le site web des Principes directeurs pour l'encodage TEI : http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/DS.html#DSTITL

5.2 Les parties postliminaires <back>

L'élément <back> contient tout ce qui est à la fin du document, après le corps du texte : explicit, colophon, index, notes, etc. Comme pour les parties préliminaires, elles sont délimitées simplement par la balise de division , accompagnée de l'attribut *type*. Le tableau suivant indique la valeur de l'attribut pour les différents types de parties postliminaires.

Partie préliminaire	Balise et attribut
Annexe	
Colophon	
Explicit	
Index	

- **<back>** Élément sous lequel l'on retrouve les parties postliminaires
- ** ** Élément par lequel l'on délimite les parties postliminaires
- **<head>** En-tête, le cas échéant
- **<list>** Liste, notamment pour les index (voir section 3.3 pour plus d'informations)
- **<ref>** Page de référence pour les index

(exemple)

```
<back>
        <div type="explicit">
                <p>Icy finent les C. nouvelles nouvelles.</p>
        </div>
</back>
```

(exemple)

```
<back>
        <div type="index">
        <head>Index</head>
        <list type="index">
                <item>Agriculture<ref>200</ref></item>
                <item>Pisciculture<ref>395</ref></item>
        </list></div>
</back>
```

Pour de plus amples informations et pour d'autres exemples, consulter le site web des Principes directeurs pour l'encodage TEI : http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html#DSBACK

5.3 Les groupes de textes <group>

L'élément <group> contient le corps d'un texte composite qui regroupe une suite de textes distincts (ou des groupes de textes tels), lesquels sont considérés comme formant une unité dans un but quelconque, par exemple les œuvres complètes d'un auteur, une suite d'essais en prose, une collection de lettres, etc.

Un avantage d'utiliser l'élément <group> est la possibilité de pouvoir traiter individuellement chacun des textes en accordant à chacun un <front> et un <back>. Il faut toutefois mentionner que le protocole TEI ne permet pas d'utiliser l'élément <group> à l'intérieur d'une division . Par conséquent, l'élément <group> ne pourrait pas être utilisé pour un texte dont les chapitres seraient composés de groupes de textes (par exemples des lettres ou des poèmes) car ces derniers se trouveraient à l'intérieur de la balise .

(exemple d'un texte composite)

```
<text>
        <front><!-- parties préliminaires du texte composite--></front>
        <group>
                <text>
                <front><!-- parties préliminaires du 1er texte--></front>
                <body><!-- corps du 1er texte--></body>
                <back><!-- parties postiminaires du 1er texte--></back>
                </text>
                <text>
                <front><!-- parties préliminaires du 2e texte--></front>
                <body><!-- corps du 2e texte--></body>
                <back><!-- parties postliminaires du 2e texte--></back>
                </text>
        </group>
        <back><!-- parties postliminaires du texte composite--></back>
</text>
```

5.4 La structure du corps du texte <body>

Le corps du texte <body> peut s'enrichir de diverses balises selon la complexité du texte, ses divisions textuelles, ses marques typographiques et ses interventions éditoriales. La seule balise obligatoire dans le <body> est celle indiquant le paragraphe, <p>.

5.4.1 La division du texte

La majorité des textes sont présentés avec une division structurale : parties, chapitres, etc. Lors de l'encodage, il est important de respecter les divisions présentes dans le texte original.

- ** ** Subdivision d'une partie liminaire ou du corps d'un texte. Nous cherchons à toujours accompagner la division du texte de l'attribut *type* et d'une valeur afin de clarifier la structure du texte (voir tableau ci-bas). Nous cherchons aussi à inclure l'attribut *n* et une valeur numérique afin de suivre la hiérarchie des divisions.
- **<header>** En-tête de la division, le cas échéant
- **<trailer>** Informations à la fin de la division, le cas échéant
- **<p>** Paragraphe. Élément obligatoire de toute division.

Division	Balise et attribut
Chapitre	
Livre	
Partie	
Volume	

(exemple)

```
<body>
<div type="partie" n="1">
        <header>Partie 1</header>
                <div type="chapitre" n="1">
                <p><!-- texte de la partie 1, chapitre 1--></p></div>
                <div type="chapitre" n="2">
                <p><!-- texte de la partie 1, chapitre 2--></p></div>
</div>
<div type="partie" n="2">
        <header>Partie 2</header>
                <div type="chapitre" n="1">
                <p><!-- texte de la partie 2, chapitre 1--></p></div>
                <div type="chapitre" n="2">
                <p><!-- texte de la partie 2, chapitre 2--></p></div>
</div>
</body>
```

5.4.2 Le changement de page <pb/>

Le changement de page est indiqué par la balise <pb/>. Cette dernière est toujours fermée. En règle générale, nous cherchons à l'accompagner de l'attribut *n* et d'une valeur numérique.

(exemple)

```
<body>
<div type="chapitre" n="1">
<pb n="1"/>
      <p>Il y a quelques années qu'en visitant, ou, pour
      </lb>mieux dire, en furetant Notre-Dame (...) </p>
<pb n="2"/>
      <p>lui consacre ici l'auteur de ce livre (...)</p>
</div>
</body>
```

5.4.3 Les balises de bornage <milestone>

En plus des changements de lignes, de paragraphes, de pages et de divisions, le protocole TEI nous permet d'encoder d'autres points limites du document grâce à la balise de bornage <milestone>. Dans le cadre du projet MCVF, nous n'encodons que le changement de folio, le cas échéant. (Ce changement de folio est souvent indiqué dans le texte publié pour les éditions de documents médiévaux.) Le changement de folio est indiqué par l'attribut *unit* et la valeur "folio", ainsi que l'attribut *n* et la valeur numérique du folio. Cette valeur numérique dépend du format adopté par l'éditeur du texte en question. Par principe, nous numérotons les folios dans le fichier TEI de la même façon que l'édition du texte.

(exemples)

```
<lb/><milestone unit="folio" n="85"/> ont esté faictes et obtenues a
<lb/><milestone unit="folio" n="85r°"/> ont esté faictes et obtenues a
<lb/><milestone unit="folio" n="85c"/> ont esté faictes et obtenues a
<lb/><milestone unit="folio" n="85bis a"/> ont esté faictes et obtenues a
```

5.4.4 Le discours direct <q>

Dans un texte en discours indirect, le passage à un discours direct n'est pas noté par les guillemets, mais par la balise <q>. Cette dernière délimite le début et la fin d'un passage en discours direct.

(exemple)

```
<lb/>demande ce qu'elle bien scet: <q>Qui est-ce la?</q>
Et il respond: <q>C'est vostre mary...</q>
```

6 Problèmes d'encodage

6.1 Les problèmes relatifs aux manuscrits

Certaines particularités sont propres à l'encodage de manuscrits:

- **<gap>** Omission dans une transcription, que ce soit pour des raisons éditoriales ou parce que le matériel est illisible.
- **<unclear>** Mot ou passage qui ne peut être transcrit avec certitude parce qu'il est illisible
- **<space>** Espace vide significatif
- **<del>** Lettre, mot ou passage supprimé, marqué comme supprimé ou autrement indiqué comme superflu ou erroné dans le texte par un auteur, un copiste, un annotateur ou un correcteur
- **<add>** Lettres, mots ou phrases insérés dans le texte par un auteur, un copiste, un annotateur ou correcteur. Il est généralement suivi des attributs *resp* et *space*.

Pour les trois premiers éléments, nous pouvons insérer les attributs *extent* et *unit*, le premier avec une valeur numérique, le deuxième avec une valeur spatiale, afin de bien représenter l'étendue de la difficulté.

(exemple)

```
<p>Les ornières devinrent plus profondes.
On approchait <gap extent="2" unit="mots"/>  Le petit garçon (...)</p>
```

(exemple)

```
<p>C'était une ferme de bonne <unclear>apparence</unclear></p>
```

(exemple)

```
<p>Une jeune femme<space extent="4" unit="lignes"/>
<lb/>Des vêtements humides séchaient (...) </p>
```

(exemple)

```
<p>Une fois le <add>pansement</add> fait,
<lb/>le médecin fut <del>cordialement</del> invité par M. Rouault (...)</p>
```

6.2 Les interventions éditoriales

La transcription à partir d'éditions présente aussi des particularités:

- **<choice>** Accompagné des balises <sic> et <corr>, permet de montrer la forme corrigée (par le scripteur ou le transcripteur, tel qu'indiqué par l'attribut *resp*) et erronée d'un mot.
- **<corr>** Forme correcte d'un passage apparemment erroné dans le texte original. Si la forme erronée est donnée, elle peut être placée sous l'attribut *n*. Le diacritique " ̥" (U+031F) peut, au besoin, montrer l'ajout d'une lettre, en exposant, par le scripteur, lorsqu'intégré dans la balise <choice>.
- **<supplied>** Texte restitué par l'éditeur à la place d'une section de texte qui est illisible. Les lettres ou mots ajoutés par le codeur sont accompagnées du diacritique " ̥" (U+031F). Il est possible d'intégrer une valeur *reason* à l'attribut <supplied>. Si cet attribut peut contenir des espaces, il est par contre impossible d'y mettre des diacritiques. Les lettres ajoutées au mot par le diacritique " ̥" (U+031F) ne seront conséquemment pas intégrées à cet attribut.
- **<w>** Présente, en attribut *n*, la forme restituée d'un mot autrement incompréhensible, notamment les graphies fortement divergentes. Le mot à la graphie divergente se trouve entre les balises ouvrantes et fermantes.

(exemple)

```
<head> Le 12
<choice><sic>Novebre</sic><corr>novembre</corr></choice></head>
```

(exemple)

```
<p>(...)L'hôtel du Lion d'or tenu par Mme <corr>Lefrançois</corr> sur la place d'Armes
(...)</p>
```

(exemple)

```
<p>Est-ce qu'on peut faire <supplied>entendre raison</supplied> à des femmes
pareilles?(...)</p>
```

(exemple)

```
<p>Elle prévoit venir <w n="dans">dun</w> un véhicule à chevaux</p>
```

6.3 Les passages en langue étrangère <foreign>

S'il y a des mots ou des expressions d'une autre langue que le français dans le texte, nous les identifions en tant que tels. Nous utilisons la balise <foreign> qui identifie un mot ou une expression comme appartenant à une langue différente de celui du texte environnant.

Si l'on connaît la langue employée, on l'indique sous l'attribut *xml:id.* Pour une liste complète des valeurs possibles pour *xml :id,* consulter : http://hapax.qc.ca/iso639-2.fr.htm. Notons que, pour les textes médiévaux, dans lesquels l'on retrouve fréquemment des mots latins, l'élément <foreign> employé seul est suffisant.

Si le passage en langue étrangère est accompagné de marques typographiques particulières, nous pouvons les noter sous l'attribut *rend*.

(exemple)

```
<p>Anne mange un <foreign xml :id="jpn" rend="italique" >daifuku</foreign> tous les soirs.</p>
```

6.4 Les marques typographiques

Si un mot ou une expression est graphiquement distinct du texte environnant dans le texte d'origine, nous pouvons l'indiquer dans le fichier TEI par la balise de mise en évidence, <hi>. Cette dernière marque un changement dans la typographie.

Marque typographique	Balise
Italique	<hi rend="italique">
Souligné	<hi rend="underline">
Gras	<hi rend="gras">

Petites majuscules	<hi rend="pmaj">
Mot entre crochets [...]	<hi rend="crochets">
Italique entre crochets	<hi rend="crochets-ital">
Édition : point grammatical	<hi rend="pt"/>
Série de points	<hi rend="points"/>
Exposants	Voir 6.5

Une précaution au sujet des marques typographiques : nous cherchons avant tout à ne pas insérer une balise au milieu d'un mot. Si la typographie change au milieu d'un même mot, il est préférable d'ignorer les marques typographiques et de garder le mot entier.

(exemple)
```
<p>André n'avait jamais lu <hi rend="italique">Madame Bovary</hi>.</p>
```

(exemple)
```
<p>(...)et en voz plus <lb/> devotes prieres ma pouvre ame <hi rend="points"/></p>
```

De plus, en vue de faciliter l'analyse morphosyntaxique, le schéma permet l'utilisation de balises de ponctuation spécifiques :

- **<pc/>** Comme la balise de ponctuation permet plusieurs attributs parfois redondants, nous n'en avons retenu qu'un, *type*, qui prend soit la valeur "point", soit la valeur "virgule". La balise de ponctuation n'apparaissant pas dans les analyses, il est possible de la posée à côté d'un signe de ponctuation soit identique, soit différent.

(exemple d'un point)
```
<p> le reste pares tranquille mais <lb/>on ne ſy fie pas,<pc type="point"/> […]</p>
```

6.5 Les abréviations

Les abréviations sont marquées par la balise TEI <abbr>. Nous indiquons aussi les abréviations par le diacritique ₊ en dessous de la lettre marquant l'abréviation. Ce diacritique est entré en Unicode (U+031F), code tapé immédiatement après la lettre concernée. Voir la section Unicode pour plus d'informations.

- **Pour les textes médiévaux :** On note la résolution de l'abréviation en plaçant le diacritique " ₊" sous les lettres concernées
- **Pour les textes à partir du XVI^e^ siècle :** On code les éléments de l'abréviation (souvent des lettres en exposant) en plaçant le diacritique " ₊" sous les lettres concernées, et les abréviations atypiques peuvent être résolues.
- **Pour les éditions de textes :** On code les lettres manquantes rajoutées par l'éditeur par le diacritique " ₊" sous les lettres concernées, et les abréviations atypiques peuvent être résolues.
- À NOTER : Si un texte présente à la fois des résolutions d'abréviations et des résolutions de lettres manquantes, il faut deux diacritiques différents pour éviter toute confusion. On suggère le « combining inverted bridge below » (U+033A) pour les lettres manquantes.

(exemple d'une résolution d'abréviation - Médiéval)
« bõ » (l'abréviation de « bon » dans certains textes médiévaux) est entré « bo̪n ».

(exemple du codage des éléments d'une abréviation - XVI^e^)
« m^r^ » est entré « m̪r»

(exemple de l'ajout des lettres manquantes par un éditeur)
« semon» est entré « se̪rmon»

6.6 La présence de points non-linguistiques

Les éditions de textes médiévaux et les textes manuscrits peuvent faire usage du point dans un but autre que celui d'indiquer une fin de phrase. Il s'agit des points numériques, qui servent à mettre en évidence des nombres dans les manuscrits (par exemple .XVII.) et des points abréviatifs qui indiquent des abréviations. Pour éviter que ces points ne soient interprétés comme une frontière linguistique lors de l'analyse morphosyntaxique automatique, nous employons le point moyen (Unicode U+00B7) pour remplacer les points numériques et abréviatifs. Voir la section Unicode pour plus d'informations.

6.7 L'indentation

L'indentation d'un paragraphe n'est jamais marquée par les espaces ou la tabulation. Ces derniers ne sont pas pertinents pour l'encodage TEI. Afin de marquer l'indentation, nous optons plutôt pour l'attribut *rend*. Ce dernier peut être inséré dans n'importe quelle balise et permet, entre autres, de marquer

l'indentation de l'élément. Par exemple, en donnant un *rend* à une balise de paragraphe, le paragraphe en entier se verra attribué une indentation.

Notons que l'on devrait toujours marquer l'indentation dans le plus haut niveau de la hiérarchie structurale. Par exemple, au lieu d'attribuer une indentation à chaque ligne d'un poème, l'on devrait l'attribuer au groupe de vers <lg> ou à la division en entier.

- **rend="indent(1)"** Cet attribut, inséré dans une balise, marque l'indentation de l'élément au complet. La valeur numérique (par défaut, 1) doit être placée entre parenthèses. Si l'indentation est plus marquée, on peut changer la valeur numérique pour 2, 3, etc. À l'inverse, l'on peut aussi donner des valeurs négatives (-1, -2, etc)
- **rend="first-indent(1)"** Cet attribut, inséré dans une balise, marque l'indentation de la première ligne de l'élément. Par exemple, s'il est inséré dans une balise de paragraphe, elle indique l'indentation de la première ligne de ce dernier. La valeur numérique dépend de l'importance de l'indentation.
- **rend="indent(-1) first-indent(3)"** Cet attribut, inséré dans une balise, marque l'indentation de l'élément au complet ET de la première ligne, au cas où elles seraient différentes. Les valeurs numériques dépendent de l'importance de l'indentation.

(exemple)

```
<p rend="indent(2) first-indent(-1)">Il y a aujourd'hui trois cent quarante-huit ans
<lb/>six mois et dix-neuf jours (...)</p>
```

6.8 La césure de mot

Pour la majorité des textes manuscrits et une grande partie des textes publiés, la césure de mot en fin de ligne pose problème. Pour l'analyse morphosyntaxique, il est important de garder les mots complets. Les parties de mots coupés doivent être regroupées pour former une seule unité graphique.

Les deux parties du mot coupé sont regroupées en remontant la deuxième moitié à la fin de la ligne précédente. La marque de césure est identifiée, pour l'instant, par « ÷ », ce qui permet d'indiquer sans ambiguïté que le mot était coupé dans la version originale et à quel endroit il était coupé. Si le mot coupé était suivi d'une ponctuation, nous portons cette ponctuation sur la ligne précédente également (pour éviter d'avoir une ligne qui commence par une ponctuation).

(exemple)

```
<lb/>L'opération, du reste, s'est prati
<lb/>quée comme par enchantement
```

Ainsi :

Devient:

(exemple)

<lb/>L'opération, du reste, s'est prati÷quée
<lb/>comme par enchantement

Une fois le fichier TEI terminé, nous pouvons utiliser des fonctions de recherche et de remplacement globales pour transformer ces raccourcis. La marque de césure sera identifiée par le diacritique Unicode de soulignement « Combining low line » (U+0332). Ce diacritique paraît à l'écran comme un soulignement du caractère précédent le code, donc de la lettre précédant la césure.

(exemple)

<lb/>L'opération, du reste, s'est pratiquée
<lb/>comme par enchantement

(Pour plus d'informations sur Unicode, voir la Section 7)

6.9 Les mots coupés ou soudés

Pour les textes manuscrits, nous retrouvons souvent des mots coupés (par exemple: « enfant » écrit « en fant ») ou des mots soudés (par exemple « l'amour » écrit « lamour »). Ces graphies ne sont pas marquées par des balises. À l'instar des césures de mot en fin de ligne, les mots coupés utilisent le signe « ÷ », qui est par la suite modifié pour un « Combining low line » (U+0332). Cependant, avenant le cas d'un mot coupé qui devrait, selon l'orthographe moderne, présenter un trait d'union, nous employons le signe « = ». Pour leur part, les mots soudés présentent une barre verticale qui propose une scission entre les deux termes.

(exemples de mots coupés)

« en fant » est entré « en÷fant » et devient « enfant »
« grand mère» est entré « grand=mère »

(exemples de mots soudés)

« lamour » est entré « l|amour »
« cejour » est entré « ce|jour »

6.10 Les passages écrits par une autre main

Avec les manuscrits, il se peut que l'on trouve des passages écrits d'une main autre que celle du scripteur principal. Le cas échéant, on indique les passages de mains avec la balise **<handShift/>.** On l'utilise pour marquer les frontières d'une section de texte écrite par une nouvelle main. Si l'on sait à qui appartient la main, on l'indique à l'aide de l'attribut **scribe=**. Il convient de noter que la balise <handShift/> doit être fermée. Il faut donc indiquer le retour à la main la répétition de la balise et un changement de l'attribut *scribe*. Finalement, l'attribut *scribe* ne permet pas les espaces. Il faut donc utiliser le tiret bas « _ » pour unir le prénom et le nom du nouveau scripteur

(exemple)

```
<p>Texte rédigé par le scripteur principal</p>

<handShift scribe="Prénom_Nom"/>
<p>Texte rédigé par une autre main</p>

<p>Retour au texte rédigé par le scripteur principal</p>
```

Le changement de main, lorsqu'il n'est fait que pour un ajout ou une délétion, peut être intégré à même la balise par l'attribut *hand*. Cet attribut, relativement rare, ne peut cependant pas être utilisé dans les balises de structures , <p>, <l>, etc.

7 Unicode

7.1 Principes de base

Tous les caractères spéciaux qui se trouvent dans les documents sources doivent être transposés dans le fichier TEI. Afin d'éviter un surplus de balises TEI et afin de faciliter la tâche des encodeurs, tous ces caractères sont entrés à partir de Unicode, une norme informatique qui vise à donner à tout caractère de n'importe quel système de langue ou d'écriture un code et un identificateur.

Chacun des caractères spéciaux dans les textes sources a donc un code Unicode qui lui est propre. Nous n'avons qu'à entrer le code (de quatre chiffres) correspondant et le caractère sera inséré dans le fichier TEI. Pour les logiciels opérant sous la plateforme Mac OS X, il suffit d'appuyer la touche *option* et de taper le code. Pour la plateforme Mac X 10.2 et toutes les versions subséquentes, il y a une palette de

caractère Unicode qui permet de choisir et d'insérer les caractères Unicode avec la souris (Menu Pomme – Edition – Caractères spéciaux).

Pour ceux qui travaillent sur la plateforme Windows, il faut appuyer la touche *Alt* et, tout en la maintenant abaissée, appuyer la touche d'addition du clavier numérique, suivie du code. Comme ce raccourci ne fonctionne pas pour tous les logiciels, il est possible de télécharger un outil qui permet l'entrée directe des codes (par exemple, *Unicode Input* : http://www.fileformat.info/tool/unicodeinput/)

Au cas où le logiciel d'édition TEI rendrait la saisie directe des caractères spéciaux difficile au clavier, nous proposons d'utiliser une séquence spéciale qui n'a pas de chance d'apparaître dans le texte normal (tel que « aE » pour le caractère« ä »). Une fois le fichier TEI complété, il suffit d'utiliser des fonctions de recherche et de remplacement globales pour transformer ces raccourcis en des caractères corrects.

7.2 Où trouver les codes

Comme il y a plus de 100 000 caractères Unicode à ce jour, il est impossible de fournir une liste détaillée de tous les codes dans ce protocole. L'on peut consulter des listes plus ou moins complètes sur les sites web suivants: http://en.wikipedia.org/wiki/List_of_Unicode_characters et http://unicode.org/charts/ (voir les tableaux de caractères latins).

Nous invitons les encodeurs à consulter ces sources externes pour trouver les codes dont ils ont besoin. La palette de caractères Unicode de la plateforme Mac est aussi un outil excellent pour trouver les caractères spéciaux et leurs codes.

7.3 Le format de transformation Unicode (UTF)

Le format de transformation Unicode est le **UTF-16**. Il pourrait y avoir des erreurs de caractères dans les formats UCS-2 ou UTF-8. Dans le fichier TEI comme tel, nous indiquons le format de transformation dans le haut du document, avant toutes les balises TEI.

(exemple)

```
<?xml version="1.0" encoding="UTF-16"?>

<TEI xmlns="http://www.tei-c.org/ns/1.0">
<!-- éléments subordonnés-->
</TEI>
```

Annexe I : les équivalences PhiloLogic

Afin de faciliter la recherche des utilisateurs dans le module PhiloLogic, une série d'équivalences ont été codées en Perl.

(extrait du code Perl)

```
'a',
"(a|\xc3\xa0|\xc3\xa1|\xc3\xa2|\xc3\xa3|\xc3\xa4|\xc3\x82|\xc3\x41|\xc3\x80|\xc3\x84|\x
61\xcc\xb2|\x61\xcc\xb8|\x61\xcc\x9f|\x41\xcc\xb2|\x41\xcc\xb8|\x41\xcc\x9f|\xc3\xa1\
xcc\xb2|\xc3\xa1\xcc\xb8|\xc3\xa1\xcc\x9f|\xc3\xa2\xcc\xb2|\xc3\xa2\xcc\xb8|\xc3\xa2\
xcc\x9f|\xc3\xa3\xcc\xb2|\xc3\xa3\xcc\xb8|\xc3\xa3\xcc\x9f|\xc3\xa4\xcc\xb2|\xc3\xa4\x
cc\xb8|\xc3\xa4\xcc\x9f|\xc3\x82\xcc\xb2|\xc3\x82\xcc\xb8|\xc3\x82\xcc\x9f|\xc3\x80\x
cc\xb2|\xc3\x80\xcc\xb8|\xc3\x80\xcc\x9f|\xc3\x84\xcc\xb2|\xc3\x84\xcc\xb8|\xc3\x84\x
cc\x9f)",
```

Les lettres, indiquées en hexadécimal, forment un réseau relationnel entre celles qui ont des diacritiques (accent, trémas, U+0332, U+0338 et U+031F) et celles qui n'en ont pas. Ainsi, que le scripteur ait, ou non, utilisé un accent, ou si une lettre a été ajoutée et présente un diacritique, le moteur de recherche trouvera le terme entré dans le champ prévu à cet effet.

Voici une liste des équivalences qui ont été codées, en sus des diacritiques U+0332, U+0338 et U+031F de chaque caractère, qui n'ont pas été indiqués par soucis de concision :

Marque typographique		Balise	
a	a, à, á, ä, â, ã, À, Á, Â, Ä, A	o	o, ò, ó, ô, õ, Ò, Ó, Ô, Õ, Ö
c	c, ç, s, C, Ç, S, ſ	q	k, c, q, K, C, Q
e	e, è, é, ê, ë, È, É, Ê, Ë	r	r, ŕ
i	i, ì, í, î, ï, j, y, ÿ, I, Ì, Í, Î, Ï, J, Y, Ÿ	s	Voir « c »
j	Voir « i »	u	u, ù, ú, û, ü, v, w, ẅ, Ù, Ú, Û, Ü, V, W, Ẅ
k	k, c, q, K, C, Q	v	Voir « u »
l	l, 1	w	Voir « w »
n	n, ñ, N, Ñ	y	Voir « i »

Les caractères « = », « - » et « | » marquent les césures de mot. Il n'y a, de ce fait, aucune différence entre les recherches « grand-mère » et « grand mère ». L'utilisateur peut, donc utiliser une orthographe davantage moderne, ce qui facilite la navigation. Néanmoins, on doit s'assurer, dans le codage des

textes, de n'avoir aucun diacritique sur ces trois caractères. Une séquence « =÷ », avant un changement pour le « Combining low line », en fin de transcription, devra par conséquent être changée pour « ÷= ».