

## Protocole de codage TEI adopté pour les textes du GTRC *Les voies du français*

### 1. Introduction : Choix d'encodage des textes du *GTRC: les voies du français*

Dans le souci d'une diffusion simple et efficace des textes, le projet du *GTRC les voies du français* 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 morpho-syntaxique.

Le présent document expose la liste des balises utilisées par le *GTRC: les voies du français* ainsi que leur contexte d'utilisation. Notre protocole de codage a pour source essentielle le Manuel élaboré par Heiden, Guillot & Lavrentiev (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](http://bfm.ens-lsh.fr/IMG/pdf/Manuel_Encodage_TEI.pdf)). En règle générale, le GTRC utilise les mêmes balises que la BFM, pour faciliter l'échange des textes. Mais vu la spécificité du GTRC, quelques balises spécifiques à certains textes ont dû cependant être ajoutées au corpus du GTRC et diffèrent donc des balises utilisées par la BFM.

### 2. Principes de base du codage

- Le codage permet de noter toutes les informations typographiques (italique, gras, couleurs, par exemple) ajoutées à un texte, 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.).
- Le format des documents codés est en texte brut.
- Les balises sont toujours délimitées par des chevrons < >.
- Les balises se présentent normalement en paires: une balise ouvrante <...> et une balise fermante </...> qui encadrent un ou plusieurs mots ou une section du texte. La balise fermante possède le caractère «/» en préfixe. 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 de chapitre.
- Les balises indiquant une frontière, se présentent seules, déjà fermées. C'est le cas des changements de lignes <lb/>, indiquant la frontière entre deux lignes, c'est également le cas des sauts de pages <pb/>. Ces balises possèdent le caractère «/» en suffixe.
- Chacune des balises peut contenir des informations supplémentaires, appelées attributs, et une valeur. Ainsi, une division <div> peut comporter l'attribut **type** et sa valeur = "**chapitre**". La balise se présentera ainsi: <div type="chapitre">...</div>.
- Seules les balises marquant le début et la fin du texte sont obligatoires selon les normes du TEI. La personne chargée du codage utilise les balises supplémentaires en fonction des exigences de chacun des textes.

### 3. Les étapes de codage des textes

Du texte manuscrit ou édité à la version électronique, trois étapes de travail sont nécessaires :

- Saisie numérique du texte (édition imprimée) ou transcription du manuscrit.
- Relecture et vérification de la numérisation ou de la transcription. Le texte numérisé reproduit fidèlement les éditions quant aux coupures de mots et aux choix éditoriaux. Cependant, les notes de bas de page et l'apparat critique ne sont pas reproduits. Les manuscrits sont transcrits avec la plus grande fidélité.
- Balisage du texte qui comprend la vérification du balisage et sa validation finale. Le TEI prévoit une méthode d'identification formelle pour chaque personne intervenant dans la constitution du document électronique.
- L'**author** définit celui qui a écrit le texte original.
- L'**editor** correspond à l'éditeur scientifique.
- Le **transcriber** est la personne chargée de transcrire le texte manuscrit ou de saisir numériquement le texte édité.
- L'**encoder** est le responsable du codage et de sa vérification, de même que de la relecture de la numérisation ou de la transcription.

On utilise ces titres à l'intérieur même du balisage afin d'identifier tous les commentaires, interventions éditoriales ou corrections effectués au cours du travail d'élaboration du document.

### 4. Description des balises utilisées dans les corpus du *GTRC: Les voies du français*

#### 4.1 - La structure générale d'un texte

- **<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 (prologue) ou qui le suit (appendice, explicit). Cette balise est obligatoire dans le codage en TEI.
- **<body>** contient le corps du texte. Cette balise est également obligatoire. Il doit contenir au minimum un paragraphe **<p>**.
- **<front>** délimite toutes informations précédant le corps du texte (prologue, sommaire, titre de l'œuvre).
- **<back>** balise les informations suivant le corps du texte (explicit, appendice, index).

Exemple :

```
<text>
<front>.....[voir ci-dessous].....</front>
<body>
<p>.....</p>
</body>
<back>.....[voir ci-dessous].....</back>
</text>
```

#### 4.2 - La structure du **<front>**

À l'intérieur du **<text>** :

- **<front>** contient tout type d'information précédant le corps du texte (prologue, sommaire, préface, table des matières, titre de l'œuvre, incipit, dédicace, résumé, etc.).

- une division **<div>** est toujours présente lorsque le **<front>** concerne le texte en entier **<body>**.
  - **type** spécifie le type de division (= "prologue, sommaire, titre")
  - **<p>** est le paragraphe contenant le titre de l'œuvre

Exemple :

```
<text>
<front>
<div type="titre">
<p>titre</p>
</div>
</front>
<body>.....</body>
</text>
```

À l'intérieur d'un **<group>**

- **<front>** peut également contenir des informations préliminaires (en-tête) concernant les textes indépendants regroupés, notamment les lettres (voir ci-dessous 4.4). Il contient un:
  - **<head>** comprenant l'en-tête
  - **<date>** indique le lieu et la date de la lettre
  - **<rs type="source">** pour indiquer la source (rs signifie *referencing string*)

Exemple :

```
<group>
<text n="1">
<front>
<head>
<date>Bourges, 21 juin 1438.</date>
<rs type="source">Orig. Bibl. nat., Fr. 20620, fol. 2.</source>
</head>
</front>
<body>...</body>
</text>
```

### 4.3 - La structure d'un **<back>**

- **<back>** inclut les informations suivant le texte (explicit, colophon, etc.). Il encadre une
  - **<div>** (division) et **type** spécifie le type de division (ex. = "explicit")
  - **<p>** est le paragraphe.

Exemple :

```
<back>
<div type="explicit">
<p> texte de l'explicit</p>
</div>
</back>
```

#### 4.4 - La structure pour les groupes de textes <group>

Dans le cas où une édition imprimée regroupe un ensemble de textes indépendants, par exemple un recueil de lettres, l'élément **<group>** permet de conserver le caractère individuel de chacun de ces textes regroupés dans un seul document électronique. La structure peut être potentiellement la suivante:

```
<text> [Le texte global du volume, qui inclut des textes indépendants]
<front>....</front> [au besoin, partie précédant l'ensemble des textes]
<group>
<text n="1"> [Le texte numéroté, s'articule comme un texte indépendant, avec
son <front> (voir 4.2), son <body>, son <back>]
<front>....</front>
<body>....</body>
</text> [Fin du premier texte]
<text n="2">.....</text> [Deuxième texte indépendant]
</group>
</text>
```

L'avantage d'utiliser l'élément **<group>** est de pouvoir traiter individuellement chacun des textes, comme les lettres, et de pouvoir inclure pour chacun d'eux un **<front>**, comprenant par exemple des informations supplémentaires sur la rédaction de la lettre (lieu et date d'écriture, destinataire ou source) (voir ci-dessus *La structure du <front>*).

Par contre, la structure de certains textes édités comprenant des lettres ne permet pas toujours l'emploi de l'élément **<group>**, par exemple, lorsque le texte contient des divisions en chapitres (**<div>**) contenant des lettres, (le TEI ne permet pas d'utiliser l'élément **<group>** à l'intérieur d'une **<div>**). Ces textes comprennent donc une division **<div>** suivie d'un **<head>** contenant la date.

Exemple :

```
<body>
  <div n="21" type="lettre">
    <head>
      <hi rend="gras">Pierre de Cussac dit Cardinal à Edward II. Février
1324.</hi>
    </head>
    <p>...</p>
  </div>
```

#### 4.5 - La structure du corps du texte <body>

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

##### 4.5.1 La division du texte <div>

- **<div>** indique une division du texte.

- **<div>** peut contenir un attribut et sa valeur pour indiquer de quel **type** de division il s'agit:

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

auquel on peut également ajouter la numérotation du livre ou du chapitre grâce à l'attribut **n**.

```
<div type="livre" n="1">
<div type="chapitre" n="5">
```

- **<div>** doit contenir nécessairement un paragraphe **<p>**.

Exemple :

```
<div type="livre" n="1">
<div type="chapitre" n="5">
<p>texte du paragraphe</p>
</div>
</div>
```

- **<div>** peut également contenir un

- **<head>** signalant un titre ou un en-tête
- **<trailer>** qui comprend des informations situées à la fin de la division.

```
<div type="chapitre" n="5">
<head><p>titre du chapitre</p></head>
<p>texte du paragraphe</p>
</div>
```

- La balise **<div type="pseudo-div">...</div>** indique qu'une division a lieu dans un texte, annoncée par un titre, mais qu'elle ne correspond pas à une division logique.

#### 4.5.2 La numérotation des pages <pb/>

- **<pb/>** indique un changement de page. Elle se présente toujours de manière fermée.

- **<pb/>** peut contenir l'attribut **n** et sa valeur **<pb n="2"/>**

- Les pages blanches ou entièrement occupées par un tableau ou une image (non reproduites dans l'édition numérique) doivent être indiquées.

- Il ne doit pas y avoir d'espace entre la balise du numéro de page et le début du texte.

Exemple :

```
<p>texte du paragraphe.....
<pb n="23"/>
<pb n="24"/>
....suite du texte du paragraphe</p>
```

**Note:** L'image se retrouve à la page 23.

- Une balise **<pb rend="folio" n=""/>** a été utilisée pour rendre compte des changements de folios qui sont signalés par l'éditeur scientifique. Cette balise est

unique au GTRC: *Les voies du français*<sup>1</sup>. Cette balise s'ajoute à la balise **<pb n="1"/>**.

### 4.5.3 Les paragraphes <p>

Dans les textes en prose:

- **<p>** et **</p>** encadrent les paragraphes.

Exemple :

```
<p>texte du paragraphe</p>
```

Dans les textes en vers:

- **<p>** et **</p>** encadrent les groupes de vers. La première moitié de la balise prend obligatoirement l'attribut
  - **rend** et sa valeur (laisse, strophe, etc.) ou gv (groupe de vers) dans le cas d'un groupe de vers difficilement qualifiable.

exemple

```
<p rend="gv">
<lb/> Que que il parloient ensi
<lb/> li rois fors de la chanbre issi
<lb/> ou il ot fet longue demore,
<lb n="652"/> que dormi ot jusqu'a ceste ore.....</p>
```

### 4.5.4 Les changements de lignes <lb/>

- Les **<lb/>** peuvent être insérés automatiquement à l'aide de la macro dans Word.
- Cette balise est indispensable, autant dans les textes en prose que dans les textes en vers.
- La balise se place au début de la ligne et est toujours fermée.
- Au début d'un nouveau paragraphe, **<p>** remplace **<lb/>**. Les deux balises ne peuvent se combiner.
- L'attribut **n** peut être ajouté si les lignes de l'édition sont numérotées (**<lb n="3"/>**)<sup>2</sup>.

```
<text>
<body>
<p rend="gv">
<lb/>.....
<lb/>.....
....
<pb n="2"/>
<lb/>.....
.....
</p>
</body>
</text>
```

<sup>1</sup> Les contraintes du TEI empêchent d'utiliser l'attribut **type="folio"** qui aurait été plus approprié, car l'attribut **rend**, ici utilisé, est normalement employé pour rendre les marques typographiques.

<sup>2</sup> Nous suivons en cela l'usage de la BFM, qui diffère de celle du TEI, qui n'emploie que l'élément **<l>**.

#### 4.5.5 Présence de points non linguistiques <num> et <abbr>

Les éditions de textes médiévaux ou les textes manuscrits peuvent faire usage du point dans un but autre que celui d'indiquer une fin de phrase. Pour éviter que ces points ne soient interprétés comme une frontière linguistique (fin de phrase) lors de l'analyse morpho-syntaxique automatique, un encodage spécifique a dû s'appliquer à ces cas :

- Point numérique **<num>** :

Ces points sont utilisés conformément à l'usage des manuscrits pour mettre en évidence des nombres, entre des points.

- **<num>** indique les chiffres encadrés par les points

Note : Dans le projet *GTRC : les voies du français*, la balise **<num>** est également employée dans le cas d'un chiffre suivi par un point, mais non précédé de celui-ci.

- Points abrégatifs **<abbr>** :

- **<abbr>** indique les abréviations suivies d'un point.

Exemple :

```
<p>.....et li hom ..... paiera <num>.II.</num> <abbr>s.</abbr> le  
jour.....</p>
```

Note: *L'homme paiera deux* (numéral entre points) *sols* (abrégé par s.).

#### 4.5.6 Corrections et interventions éditoriales <corr>

- **<corr>** indique les corrections faites par l'éditeur scientifique et signalées dans l'édition de manière particulière (italique, crochets, etc.)

- **resp** signale le responsable de la correction, ici l'éditeur (**= "editor"**)

- **sic** mentionne le texte original erroné. Pour un ajout, il reste vide (**sic=""**)

- **rend** indique la marque typographique notant la correction.

Exemple :

L'éditeur ajoute un ou des mots dans le texte.

```
<corr resp="editor" sic="" rend="crochets">.....</corr>
```

L'éditeur rétablit une lettre manquante, ici le n.

```
<corr resp="editor" sic="e" rend="e[n]">en</corr>
```

#### 4.5.7 Proposition de correction de l'encodeur <sic>

Le texte de l'édition n'est jamais modifié, mais l'encodeur peut indiquer une erreur flagrante de l'édition dans l'attribut de la balise sans toutefois toucher au texte lui-même (ex. répétition d'un même mot).

- **<sic>** signifie qu'il y a erreur et introduit l'attribut.

- **resp** indique son responsable, ici, l'**encoder**.

- **corr** indique la correction proposée.

Une **<note>** permet, si nécessaire, d'ajouter un commentaire supplémentaire. (voir 4.5.10)

Exemple :

**<sic resp="encoder" corr="">...</sic>** Suppression d'un mot  
**<sic resp="encoder" corr="monseigneur">monseigneur</sic>** Correction de la répétition de la syllabe

#### 4.5.8 Passages en langue étrangère <foreign>

- **<foreign>** indique un changement de langue.
- **lang** permet de définir cette langue.
- **rend** mentionne la typographie utilisée

- Note : Le latin fréquemment utilisé dans les textes médiévaux n'est pas spécifié par l'attribut **lang**, seul l'élément **<foreign>** est suffisant. De même, comme les mots latins sont majoritairement écrits en italique, cette spécificité typographique n'a pas été indiquée, par contre, la valeur **rend="romain"** a été ajoutée lorsque les mots latins n'étaient pas notés en italique.

- Dans le projet *GTRC : les voies du français*, la balise **<foreign>** est toujours accompagnée de l'attribut **lang** parce qu'on retrouve dans les textes de la Nouvelle-France tout autant l'anglais et les langues amérindiennes que le latin.

Exemple :

**<foreign>**mot latin**</foreign>**  
**<foreign lang="" rend="marque typographique">...</foreign>**

Pour la notation des valeurs de **lang**, se reporter à la liste fournie à l'adresse <http://www.loc.gov/standards/iso639-2/englangn.html>

#### 4.5.9 Marques typographiques <hi>

- **<hi>** note un changement de la typographie
- **rend** indique la typographie utilisée.

**<hi rend="italique">...</hi>**  
**<hi rend="gras">...</hi>**  
**<hi rend="pmaj">...</hi>** pour PETITE(S) MAJUSCULE(S)  
**<hi rend="maj">...</hi>** pour MAJUSCULE. Peu utilisé; les majuscules suffisent.  
**<hi rend="exp">... </hi>** pour <sup>exposant</sup>  
**<hi rend="crochets">... </hi>** pour [....]  
**<hi rend="crochets-ital">...</hi>** pour les italiques mis entre crochets [*italique*]  
**<hi rend="susp">... </hi>** pour les trois points de suspension  
**<hi rend="points">... </hi>** pour plusieurs points .....

Exemple :

**<p>** En la **<hi rend="pmaj">**Tierce partie**</hi>** et en la deharenriere....

#### 4.5.10 Insertions de notes de l'encodeur <note>

- **<note>** contient des notes ou des commentaires,
- **resp** mentionne le responsable de la note, ici, l'**encoder**.

Exemple :

**<note resp="encoder">.....</note>**

.....Va ton chemin, il ne te fault aultre confession, car <pb n="61"/> tu es en tresbon estat. - **<note resp="encoder">**Des guillemets ouvrants sont placés ici, sans qu'il y ait les fermants**</note>** Ha dya! dist l'yvroigne....

#### 4.5.11 Le discours direct et indirect <q>

Dans un texte en discours indirect, le passage à un discours direct est noté par **<q>** et remplacé de ce fait les guillemets.

- **<q>** permet d'indiquer le début et la fin d'un passage en discours direct.

exemple

.....demande ce qu'elle bien scet:**<q>**Qui est **<pb n="28"/>**ce la? **</q>** Et il respond: **<q>**C'est vostre mary....**</q>**.....

#### 4.5.12 Les listes <list> et leur contenu <item>

Un texte peut présenter une liste d'items, comme par exemple une table des matières contenue dans le **<front>**.

- **<list>** contient des items

- **type** permet de spécifier de quel genre de liste il s'agit.

- **<item>** est un des éléments de la **<list>**.

- **n** permet de numéroter les items.

Exemple :

```
<list type="table_des_matières">
  <item n="1">
    <lb n="5"/>
    <p>Compté par Monseigneur le Duc.</p>
    <p> La premiere nouvelle traicte d'un qui trouva façon...</p>
  </item>
  <item n="2".....</item>
</list>
```

## 4.6 Problèmes relatifs aux manuscrits indiqués dans les éditions

Les deux éléments suivants conviennent autant aux textes édités qu'aux manuscrits eux-mêmes.

### 4.6.1 Lacunes <gap>

- **<gap>** indique les lacunes du manuscrit ou encore le texte supprimé rendu complètement illisible. (Pour le texte supprimé lisible, voir **<del>**). Les attributs suivants sont susceptibles d'apparaître:

- **resp** identifie le responsable, l'éditeur (**= "editor"**).

- **rend** note la marque typographique utilisée dans l'édition.

- **reason** inclut la description de cette lacune  
(ex. reason="manuscrit endommagé" ou ="vers omis")
- **extent** définit la longueur de la lacune (="1 line"; ="2 words")

Exemple :

.....<gap resp="editor" reason="manuscrit endommagé"/>.....

#### 4.6.2 Passages du manuscrit illisibles ou difficiles à lire <unclear>

L'éditeur scientifique note (surtout dans les éditions diplomatiques) un passage difficile à lire et dont il n'est pas certain de sa lecture. Les passages supprimés partiellement lisibles entrent dans cette catégorie. On l'indique de cette manière:

- **<unclear>** englobe la partie du texte difficile à lire
  - **resp** indique la responsabilité
  - **rend** note la typographie utilisée
  - **reason** raison du passage peu clair (illegible, ambiguous)

exemple

<p>.....Pierre Amable de <unclear resp="encoder">bonne saullcaize</unclear></p>

## 5. L'encodage des corpus manuscrits

La représentation des manuscrits exige un codage plus précis afin de noter fidèlement chacun des aspects du support manuscrit. Le guide du TEI consacre un chapitre entier (chapitre 18) au codage de ces sources. Le TEI expose de nombreuses variantes qui encodent parfois un phénomène similaire. Dans un souci d'alléger et de simplifier le codage, les corpus manuscrits du GTRC présentent un nombre réduit de types de balises.

### 5.1.1 Correction du manuscrit de la part de l'auteur/scribe <corr>

Comme aucune correction éditoriale n'a été effectuée lors de la transcription des manuscrits, l'élément **<corr>** ne concerne que les corrections dues à l'auteur/scribe. La notation de la balise est identique à celle utilisée pour les sources éditées.

Exemple :

<corr resp="author" sic="">

### 5.1.2 Ajout de la part de l'auteur/scribe <add>

- **<add>** note les mots ou lettres insérés par l'auteur/scribe
  - son attribut **place** indique l'endroit de l'insertion. Elle peut être:
    - **"inline"** pour une insertion dans un espace laissé sur la ligne.
    - **"interlinear"** pour une insertion entre les lignes
    - **"supralinear"** pour l'ajout au-dessus de la ligne
    - **"infralinear"** pour l'insertion sous la ligne
    - **"left"** pour l'addition faite dans la marge de gauche
    - **"right"** pour l'addition faite dans la marge de droite

- ="**top**" pour l'ajout fait dans le haut de la marge
- ="**bottom**" ou "**foot**" pour l'insertion faite dans la marge du bas
- ="**opposite**" pour l'addition écrite sur la page opposée
- ="**verso**" pour l'insertion faite au verso
- ="**previous**" pour une insertion sur la page précédente
- ="**next**" pour l'addition faite sur la page suivante
- ="**end**" pour un ajout fait en fin de document
- ="**mixed**" pour une addition s'étalant sur plusieurs endroit décrits ci-dessous.

Exemple :

....macoutumer a te **<add place="supralinear">le</add>** dire....

### 5.1.3 Ajout de la part du transcripteur <supplied>

- **<supplied>** permet d'ajouter les mots ou lettres qui sont illisibles ou manquantes
  - La responsabilité **resp** en revient normalement au transcripteur (= "**transcriber**")
  - **reason=" "** indique la cause physique (support endommagé) ou intellectuelle (omission de la part du scribe) de cet ajout.

Exemple :

....nos coutumes qui devrai**<supplied resp="transcriber">e</supplied>**nt naturellement....

### 5.1.4 Suppression d'un mot ou d'un passage de la part de l'auteur/scribe <del>

- l'élément **<del>** permet de conserver le texte supprimé par l'auteur/scribe, mais qui reste visible pour le transcripteur de sorte qu'il puisse retranscrire le texte supprimé. Si le texte est illisible, utiliser **<gap>**.
  - **rend** indique la manière utilisée pour la suppression. Ses valeurs peuvent être:
    - ="**subpunction**" pour les point de suppression sous la lettre
    - ="**overstrike**" pour les ratures
    - ="**erasure**" pour les mots effacés ou grattés.
    - **etc.**

Exemple :

**<p>**...et municipales de ses fidels **<del rend="overstrike"> sujet</del>** et loyaux sujets canadiens**</p>**

### 5.1.5 Les abréviations non résolues <abbr>

Les abréviations des manuscrits ne sont jamais résolues. Dans le cas des textes publiés, la décision prise par l'éditeur est respectée. Dans le cas d'une abréviation non développée, celle-ci est encadrée par :

- l'élément **<abbr>** note l'abréviation non résolue.
  - **type** indique le genre d'abréviation (typologie) qui peut se définir comme:
    - ="**suspension**" seule la fin du mot est abrégée
    - ="**contraction**" les lettres au milieu du mot sont abrégées

- ="breviatif" symbole ou marque d'abréviation
- ="superscription" l'abréviation inclue des lettres au-dessus de la ligne
- ="acronym" comprend seulement les lettres initiales

Exemple :

```
<abbr type="superscription">Do</abbr>
```

### 5.1.6 La résolution des abréviations <expan>

- <expan> indique que l'abréviation a été résolue. On peut le préciser par l'ajout de
  - **abbr=""** qui indique la forme abrégée
  - **type**, pour indiquer le genre d'abréviation (typologie) qui peut se définir comme:
    - ="suspension" seule la fin du mot est abrégée
    - ="contraction" les lettres au milieu du mot sont abrégées
    - ="breviatif" symbole ou marque d'abréviation
    - ="superscription" l'abréviation inclut des lettres au-dessus de la ligne
    - ="acronym" comprend seulement les lettres initiales

Les éléments <expan> et <abbr> peuvent coexister dans un même document. <abbr> est employé pour indiquer les abréviations dont la résolution est difficile et donc non résolue, ou pour celles suivies d'un point (fréquemment utilisé pour les prénoms) (Voir 4.5.5.). Pour les transcriptions à partir de manuscrits, <abbr> est préféré pour rester le plus fidèle possible au texte original.

Exemple :

```
<abbr>Hyp.</abbr> <expan abbr="Sr">Sir</expan> Georges Dupré
```

### 5.1.7 Les espaces significatifs <space>

Les manuscrits présentent parfois des espaces laissés blancs de longueur assez considérable, en vue parfois d'y insérer des informations.

- L'élément <space> indique ces espaces et son attribut **extent** indique leur longueur.

Exemple :

```
vous recevrez la somme de <space extent="6 lettres"/> que Monsieur...
```

N.B. Un espace a été réservé pour noter la somme et n'a pas été rempli.

### 5.1.8 Les changements de main <HandShift>

- L'élément <HandShift> note le passage d'une main d'écriture à une autre et identifie ces mains par ces attributs :
  - **old**, indiquant entre les guillemets la personne qui écrivait jusqu'alors.
  - **new**, notant la nouvelle main.

Exemple :

<handShift old="unknown" new="Papineau">.....</HandShift>

### 5.1.9 Le cas de la correspondance : l'inscription de l'adresse <address>

L'adresse postale est notée en fin de document (après les salutations et la signature), peu importe l'endroit où elle est notée dans le manuscrit, et elle est signifiée grâce à la balise <address> :

- <address> et </address> encadrent le texte.
- Les changements de lignes sont indiqués par <addrLine> </addrLine> plutôt que par <lb/>.

Exemple :

```
<address>
<addrLine>Madame</addrLine>
  <addrLine>R. Chevrier papineau</addrLine>
    <addrLine>St Hyacinthe</addrLine>
</address>
```

## 6. L'En-tête TEI

L'en-tête TEI doit être ajouté au début de tous les documents. Les informations qu'il contient permettent aux chercheurs d'avoir facilement accès aux données bibliographiques essentielles et doivent servir de mémoire à l'équipe de recherche. Il contient les références bibliographique, le genre littéraire (vers, prose ou mixte), les numéros des pages ou de folios contenus dans le fichier, le nom du fichier, le nom de la personne à qui appartient le texte (cette information est importante quand on parle d'échange), le nom de la personne qui a saisi le texte, le nom de la personne qui a révisé le texte, le nom de la personne qui a ajouté les codes de balisage, etc.

### En-tête TEI

CODAGE	SIGNIFICATION
<TEI.2>	En-tête d'un texte
<teiHeader type='text'>	Description bibliographique du fichier électronique
<fileDesc>	
<titleStmt>	« Title statement »
<title>_____</title>	- Titre de l'ouvrage : (Nom du fichier : ce nom devrait contenir datation et un nom facile à retenir (Exemple pour <i>Les quinze joyes de mariage</i> : 13.9-QJM)
<principal>_____</principal>	- Principal: nom du chercheur principal responsable de la création de la version électronique
<possesseur>_____</possesseur>	- Nom de la personne à qui appartient le texte
<respStmt>	- RespStmt: « statement of responsibility »

```

    <resp>transcription</resp>
<name>____</name>
<date>____</date>
<resp>correction_transcription</resp>
<name>____</name>
<date>____</date>
    <resp>numérisation</resp>
<name>____</name>
<date>____</date>
<resp>correction_numérisation</resp>
<name>____</name>
<date>____</date>
    <resp>codage</resp>
<name>____</name>
<date>____</date>
    <resp>correction_codage</resp>
<name>____</name>
<date>____</date>
    </respStmt>
  </titleStmt>
<extent>__ mots</extent>
<extent>__ pages</extent>
<extent>__ folios</extent>

<sourceDesc>
  <biblStruct >
    <monogr>
      <title>____</title>
      <title type='sub'>____</title>.
      <author>____</author>
      <editor>____</editor>
      <imprint>
        <pubPlace>____</pubPlace>
        <publisher>____</publisher>
        <date>____</date>
      </imprint>
    </monogr>
  </biblStruct>
</sourceDesc>
  
```

la personne responsable du contenu intellectuel d'une transcription, vérification, etc.  
 -----Ici, nom des responsables et date de:  
 -----transcription  
 -----correction de la transcription  
 -----numérisation  
 -----correction de la numérisation  
 -----codage (ou balisage)  
 -----correction du codage

« Type et longueur du texte (optionnel) »  
 Longueur du texte électronique  
 Nombre de pages ou de folios (pour manuscrits)  
 contenus dans le fichier

---

Source du fichier électronique (Référence bibliographique)

Description de la source selon qu'elle soit :  
 Une monographie

```
<biblStruct >
  <analytic>
    <author>_____</author>
    <title>_____</title>
    <title type='sub'>_____</title>.
  </analytic>
  <monogr>
    <title>_____</title>
    <title type='sub'>_____</title>.
    <author>_____</author>
    <imprint>
      <biblScope type="vol:pp">_: __-
_____</biblScope>
      <pubPlace>_____</pubPlace>
      <publisher>_____</publisher>
      <date>_____</date>
    </imprint>
  </monogr>
</sourceDesc>
```

Un article

```
<biblStruct >
  <manuscrit>
    <lieu>_____</lieu>
    <fond>_____</fond>
    <cote>_____</cote>
    <folios'>_____</folio>
  </manuscrit>
</biblStruct>
</sourceDesc>
</fileDesc>
```

Un manuscrit

```
<encodingDesc>

  <projectDesc>
    <p>_____</p>
  </projectDesc>
  <samplingDecl>
    <p>_____</p>
  </samplingDecl>
  <editorialDecl ">
    <p>_____</p>
  </editorialDecl>
</encodingDesc>

<profileDesc>
```

---

Fin du FILEDESC

---

- EncodingDesc : les méthodes et principes éditoriaux gouvernant la transcription, la numérisation et le codage du texte
- but du codage de ce fichier en particulier
- SamplingDecl: description de la logique et de la méthodologie utilisées pour choisir les textes et échantillons du corpus (ex.: 2000 mots pris au début du texte)
- Editorial Decl: principes éditoriaux et méthodes spéciales qui s'appliquent au codage de ce texte en particulier (ou plus général)
- « Profile Description »

```
<creation>
  <date>_____</date>
</creation>
<langUsage>
  <language>_____</language>
</langUsage>
<condi_auteur>
  <métier>_____</métier>
```

- info sur la création du texte : date : Le siècle, avec précision du quart de siècle, ex. 13.1 (premier quart du 13<sup>e</sup> siècle)
- dialecte du scripteur
- Condition sociale de l'auteur
- Le métier de l'auteur si connu

<éducation> ____ </éducation>	Le degré d'éducation de l'auteur si connu
</condi-auteur>	
<catRef = "____"/>	CatRef: typologie du texte (Prose/Vers)
</profileDesc>	
<revisionDesc>	Sert à noter les changements apportés texte électronique (ordre décroissant)
<change n="____">	
<date> ____ </date>	
<respStmt>	
<resp>encoder</resp>	
<name>	
____ </name>	
</respStmt>	
</change>	
</revisionDesc>	
</teiHeader>	Fin de L'EN-TÊTE

## 7. PROCÉDURE DE TRANSCRIPTION ET DE CODAGE – OTTAWA

### 7.1 Étapes de travail sur un texte

- A- Saisie numérique du texte d'une édition imprimée ou transcription d'un manuscrit
- B- Relecture et vérification de la numérisation ou de la transcription. Insertion des balises TEI.
- C- Balisage du texte qui comprend la vérification du balisage et sa validation finale.

### 7.2 Protocole de transcription : transcription de textes manuscrits ou saisie numérique de textes édités

#### 7.2.1 Conseils généraux :

- \*\*Garder en tête les objectifs des transcriptions; elles servent à des études linguistiques et à une éventuelle diffusion sur le Web.
- \*\* Être constant et uniforme dans ses décisions tout au long de la transcription d'un texte.\*\*

#### 7.2.2 Création du fichier WORD :

Généralement, nous créons un seul fichier par document (journal, mémoires, registres de comptes), sauf pour la correspondance, où nous créons un fichier différent pour chacune des lettres. Il y a des exceptions, comme celle du journal de Madame Bégon.

#### 7.2.3 Titre du fichier de transcription :

#### 7.2.4 Règles générales de la transcription – aspect formel et mise en page

- A- Conserver l'allure générale du texte (manuscrit ou version éditée)
- B- Respecter les fins de lignes : retour (↵) dans Word **et** code TEI applicable <lb/>

- C- Respecter les paragraphes : un espace (ligne) supplémentaire dans Word **et** code TEI applicable <p>
- D- Respecter les fins de pages : saut de page dans Word **et** code TEI applicable <pb n="xx"/>
- E- Espaces et alinéas : Reproduire dans Word l'allure générale de la page manuscrite ou de la page éditée
- F- Respecter l'alignement du texte : Reproduire l'allure générale dans Word, par exemple, texte à gauche, à droite, centré, etc.
- G- Espaces entre les mots : Dans le cas de la transcription à partir d'un manuscrit, n'indiquer, au mieux de notre jugement, que les espaces ORTHOGRAPHIQUES et laisser tomber les espaces jugés calligraphiques.

## 7.2.5 Règles générales de la transcription – texte et orthographe

- A- NE JAMAIS CORRIGER LE TEXTE
- B- Textes édités: ne garder que le texte, éliminer les notes de bas de page, les commentaires de l'éditeur, préfaces, postfaces, etc.
- C- Textes manuscrits: TOUT garder, sauf les notes des centres d'archives comme les étampes, no. de document, etc.
- D- Indiquer les majuscules: respecter exactement la version originale.
- E- Indiquer les "s" longs (ſ), s'il y en a : Dans Word. Sélectionnez Insertion ⇒ Caractères spéciaux... ⇒ Symbol ⇒ Sélectionnez ſ ⇒ Correction automatique ⇒ Le remplacer par la touche de son choix, dont vous n'aurez pas besoin lors de la transcription (\$, par exemple) ⇒ Cliquer sur Ajouter ⇒ OK. Chaque fois que vous taperez "\$", c'est "ſ" qui apparaîtra.
- F- **Ne pas** résoudre les abréviations : Les indiquer telles qu'elles se retrouvent dans le manuscrit, de même que dans la version éditée (même si parfois les éditions résolvent les abréviations). Attention, si l'édition a résolu l'abréviation, suivre l'édition. Les indiquer par le code TEI applicable <abbr>.
- G- **Ne pas** indiquer les différents "e". Les seules marques calligraphiques de ce genre dont nous devons tenir compte sont les "s" longs.
- H- **Ne pas** indiquer les traits en fins de lignes afin de ne pas surcharger le texte, SAUF si nous jugeons que c'est un trait d'union orthographique servant à séparer un mot (encore une fois, être uniforme dans nos décisions); **dans ce cas, il faut l'indiquer par le signe " ÷ " à la fin de la ligne.**
- I- Ratures et exposants : Indiquer les mots raturés et les mots écrits en exposants dans Word à l'aide de la palette de mise en forme, par exemple : M<sup>lle</sup>, monsieur, jamais<sup>toujours</sup>, etc. Indiquer également le code TEI applicable <hi rend="xxx">.
- J- Renvois et écriture en marges : Lorsque le manuscrit contient des renvois à l'intérieur du texte, on écrit le texte du renvoi à la suite du texte principal dans Word (on préserve la suite sémantique) **et** on indique le renvoi par le code TEI applicable (<add place=XXX> on écrit le texte de l'addition </add>).
- K- Correspondance : Il faut inclure l'adresse. On la trouve souvent, dans le cas des manuscrits, sur l'enveloppe ou à l'endos de la page. Elle est également parfois indiquée dans les textes édités. Écrire l'adresse en fin de document, après les salutations, la signature, etc., et la noter avec le code TEI applicable (<address> et </address>).
- L- Doutes : Lorsqu'une décision est prise par le transcripateur mais qu'un doute persiste, ou lorsqu'une note est incluse dans l'édition

mentionnant que l'éditeur n'était pas certain d'un passage, indiquer ce passage avec le code TEI applicable (<unclear resp="editor" (ou "transcrire")> texte </unclear>)

M- Passages illisibles : Lorsqu'un passage est complètement illisible dans le manuscrit (on ne peut même pas deviner le texte) ou que l'éditeur a indiqué un passage qu'il a trouvé lui-même illisible, on l'indique avec le code TEI applicable (<gap resp="editor" (ou "transcrire") extent="1 line"> espace correspondant environ à celui du passage illisible </gap>)

## 7.2.6 Modèle d'en-tête TEI utilisé

Les assistants peuvent utiliser le squelette de cet en-tête TEI en utilisant la fonction copier-coller, mais ils ne doivent pas oublier de modifier les informations propres à chacun des textes (cette information est surlignée en gris) et de supprimer les commentaires en rouge.

### Exemple

```
<TEIheader>
<titre>NF-LET-CONT-6B-MarieCharlotteContrecoeurLaPerrière07-24</titre>
<possesseur>France Martineau</possesseur>
<numérisation>
<name>Jennifer Dionne</name>
<date>08-05</date>
<projet>GTRC Modéliser le changement les voies du français, Corpus Nouvelle-
France dirigé par F. Martineau</projet>
</numérisation>
<transcription>
<name>Sandrine Tailleur</name>
<date>20-02-05</date>
mettre la date du début de la transcription
<projet>GTRC Modéliser le changement les voies du français Corpus Nouvelle-
France dirigé par F. Martineau</projet>
</transcription>
<correction_transcription>
<name>none</name>
<date>none</date>
ici, comme cela n'a pas été fait, mettre none
<projet>GTRC Modéliser le changement les voies du français Corpus Nouvelle-
France dirigé par F. Martineau</projet>
</correction_transcription>
<codage>
<name>Sandrine Tailleur</name>
<date>4-03-05</date>
ici mettre le nom de la personne qui code TEI même partiel (donc la personne qui
transcrit) et la date de début de ce TEI
<projet>GTRC Modéliser le changement les voies du français Corpus Nouvelle-
France dirigé par F. Martineau</projet>
</codage>
<correction_codage>
<name>none</name>
<date>none</date>
```

```

ici, comme cela n'a pas été fait, mettre none
</correction_codage>
<extent>5 200 mots</extent>
<extent>0 page</extent>
<extent>4 folios</extent>
Ici, le nombre de mots est compté à la fin (du texte seulement, pas de l'en-tête).
Le nombre de pages correspond à l'édition publiée si elle est utilisée comme base
Le nombre de folios correspond au nombre de feuillets pour un document manuscrit.
Dans les balises TEI, on utilise celle de pages.
Ici, c'est la référence du document publié :
<sourceDesc>
<biblioStruct>
<monogr>
<title>Papiers Contrecoeur et autres documents concernant le conflit anglo-français
sur l'Ohio de 1745 à 1756</title>
<title type="sub">none</title>
<author>Marie-Charlotte Contrecoeur LaPerrière</author>
<editor>Fernand Grenier</editor>
<imprint>
<pubPlace>Québec</pubPlace>
<publisher>Les Presses de l'Université Laval </publisher>
<date>1952</date>
<collection>Publications des Archives du Séminaire de Québec</collection>
</imprint>
</monogr>
</biblioStruct>
</sourceDesc>
<samplingDecl>pages 325-326</samplingDecl>
<editorialDecl>

<p>Le texte de référence a été respecté scrupuleusement quant à l'orthographe, la
ponctuation, la mise en page et la pagination. Les notes de bas de pages dans les
documents publiés ont été retirées. L'élément FOREIGN indique un passage dans
une autre langue.</p>
</editorialDecl>
<profileDesc>
<creation>
<date>1765-01-05</date>
(date où le document a été écrit. S'il n'y en a pas, mettre au moins le siècle)
<lieu>Montréal</lieu>
(lieu où le texte a été écrit ; si on ne sait pas, on écrit ??)
<représentationSociale>militaire</représentationSociale>
Choix : Administration ; Marchand ; Artisans ; Religieux ; Militaires ; Habitants ;
Femmes ou Filles ou Fils de X selon classe de métier)
<représentationDialectale>Québec</représentationDialectale>
Choix : Acadie ; France ; Québec ; Ontario ; Louisiane (si on ne sait pas, on écrit ??)
<type_de_document>correspondance</type_de_document>
Choix : Correspondance ; Journal ; Annales ou récit ; Livre de comptes ; Journal
intime
<scripteur>Marie-Charlotte Contrecoeur La Perrière</scripteur>
<destinataire>père (Claude-Pierre Pécaudy de Contrecoeur)</destinataire>
Ici, écrire le nom du destinataire si connu, ainsi que sa relation avec le scripteur
</creation>
</profileDesc>

```

</TEIheader>

<b>1 INTRODUCTION : CHOIX D'ENCODAGE DES TEXTES DU GTRC: LES VOIES DU FRANÇAIS.....</b>	<b>1</b>
<b>2 PRINCIPES DE BASE DU CODAGE.....</b>	<b>1</b>
<b>3 LES ÉTAPES DE L'ENCODAGE DES TEXTES.....</b>	<b>2</b>
<b>4 DESCRIPTION DES BALISES UTILISÉES DANS LES CORPUS DU GTRC: LES VOIES DU FRANÇAIS .....</b>	<b>2</b>
4.1 - LA STRUCTURE GÉNÉRALE D'UN TEXTE .....	2
4.2 - LA STRUCTURE DU <FRONT> .....	2
4.3 - LA STRUCTURE D'UN <BACK> .....	3
4.4 - LA STRUCTURE POUR LES GROUPES DE TEXTES <GROUP> .....	3
4.5 - LA STRUCTURE DU CORPS DU TEXTE <BODY> .....	4
<b>4.5.1</b> La division du texte <div> .....	4
<b>4.5.2</b> La numérotation des pages. <pb/> .....	5
<b>4.5.3</b> Les paragraphes <p> .....	5
<b>4.5.4</b> Les changements de lignes <lb/>.....	6
<b>4.5.5</b> Présence de point non linguistique <num> et <abbr> .....	6
<b>4.5.6</b> Correction et intervention éditoriales <corr> .....	7
<b>4.5.7</b> Proposition de correction de l'encodeur <sic> .....	7
<b>4.5.8</b> Passages en langue étrangère <foreign> .....	8
<b>4.5.9</b> Marques typographiques <hi> .....	8
<b>4.5.10</b> Insertions de notes de l'encodeur <note> .....	8
<b>4.5.11</b> Le discours direct et indirect <q>.....	9
<b>4.5.12</b> Les listes <list> et leur contenu <item>.....	9
4.6 - PROBLÈMES RELATIFS AUX MANUSCRITS INDIQUÉS DANS LES ÉDITIONS .....	9
<b>4.6.1</b> Lacunes <gap>.....	9
<b>4.6.2</b> Passages du manuscrit illisibles ou difficiles à lire <unclear>.....	10
<b>5 L'ENCODAGE DES CORPUS MANUSCRITS .....</b>	<b>10</b>
<b>5.1.1</b> Correction du manuscrit de la part de l'auteur/scribe <corr> .....	10
<b>5.1.2</b> Ajout de la part de l'auteur/scribe <add> .....	10
<b>5.1.3</b> Ajout de la part du transcripteur <supplied> .....	11
<b>5.1.4</b> Suppression d'un mot ou d'un passage de la part de l'auteur/scribe <del> .....	11
<b>5.1.5</b> Les abréviations non résolues <abbr>.....	11
<b>5.1.6</b> La résolution des abréviations <expan> .....	12
<b>5.1.7</b> Les espaces significatifs <space> .....	12
<b>5.1.8</b> Les changements de main <HandShift> .....	12
<b>5.1.9</b> Le cas de la correspondance : l'inscription de l'adresse <address>.....	13
<b>6 L'EN-TÊTE TEI .....</b>	<b>13</b>
<b>7 PROCÉDURE DE TRANSCRIPTION ET DE CODAGE – OTTAWA .....</b>	<b>16</b>
7.1 ÉTAPES DE TRAVAIL SUR UN TEXTE .....	16
7.2 PROTOCOLE DE TRANSCRIPTION : TRANSCRIPTION DE TEXTES MANUSCRITS OU SAISI NUMÉRIQUE DE TEXTES ÉDITÉS .....	16
<b>7.2.1</b> Conseils généraux :.....	16
<b>7.2.2</b> Création du fichier WORD : .....	16
<b>7.2.3</b> Titre du fichier de transcription :.....	16
<b>7.2.4</b> Règles générales de la transcription – aspect formel et mise en page .....	16

<b>7.2.5</b>	<i>Règles générales de la transcription – texte et orthographe.....</i>	<b>17</b>
<b>7.2.6</b>	<i>Modèle d’en-tête TEI utilisé.....</i>	<b>18</b>