Manuel d'encodage XML-TEI des textes de la Base de Français Médiéval

Normes utilisées à la date du 1er février 2005

Serge Heiden (slh@ens-lsh.fr)
Céline Guillot (cguillot@ens-lsh.fr)
Alexei Lavrentiev (Alexei.Lavrentev@ens-lsh.fr)

CNRS / ENS-LSH, UMR 5191 ICAR

(VERSION 1.0- JUILLET 2002)
(VERSION 2.0- JANVIER 2003)
(VERSION 2.1 – FEVRIER 2005)

Ce document de travail est élaboré dans le cadre des opérations de relecture et de balisage en XML-TEI des textes de la BFM. Il sert de référence à l'équipe des relecteurs/encodeurs de la BFM et pour des échanges de textes avec nos partenaires.

SOMMAIRE

Introduction

Ce document présente l'ensemble des méta-informations susceptibles d'être intégrées dans un texte en vue de sa gestion dans la base de textes et de son traitement automatique par Weblex. Ces méta-informations sont représentées explicitement dans le texte sous la forme de balises XML. Le nom et la structuration de ces balises correspondent à un sous-ensemble des recommandations de la TEI.
En dehors des balises délimitant le début et la fin du texte **aucune balise n'est obligatoire**.
Le relecteur/encodeur[1] ajoute les seules méta-informations dont il dispose et qu'il croit nécessaires à l'exploitation future du texte. Ce document ne présente donc pas un « format » de texte particulier mais bien un moyen de communication formel entre le chercheur, ses partenaires et les outils d'exploitation permettant d'*expliciter* les notions du texte nécessaires à leur traitement (la structure interne du texte, la mise en page de l'édition de référence, les interventions éditoriales…). La représentation de ces notions se fait au moyen de balises insérées dans le texte.
Pour un encodage plus complet on pourra consulter la documentation de la TEI située à l'url http://www.tei-c.org/Guidelines2/index.html. De manière générale il faudra consulter les TEI Guidelines comme complément aux imprécisions de ce manuel.

1 Généralités

1.1 Principes de base

1.1.1 Définition de l'unité textuelle

Les textes de la Base de Français Médiéval sont une représentation d'**éditions de référence imprimées**. Cependant l'unité physique d'un volume imprimé ne correspond pas toujours à une unité sémantique (ou « intellectuelle »), alors que cette dernière est de première importance pour l'étude linguistique ou littéraire d'un texte. En effet, certaines œuvres littéraires sont éditées en plusieurs volumes (le *Roman de Thèbes* par exemple). Dans d'autres cas, un volume physique comporte plusieurs œuvres d'un même auteur (l'édition des *Lais* de Marie de France par exemple) ou de plusieurs auteurs (le *Roman de la Rose* commencé par Guillaume de Lorris et terminé plus tard par Jean de Meun).
Face à cette situation, nous avons énoncé un ensemble de principes qui définissent les limites d'un texte dans la Base de Français Médiéval. Dans le cas où une œuvre est éditée en plusieurs volumes, chacun de ces volumes donnera lieu à la création d'un fichier séparé. Mais pour l'exploitation de la base, l'unité textuelle au sens d'unité sémantique est recomposée grâce aux références contenues dans les en-têtes TEI. Dans le cas où un volume physique comprend un ensemble d'œuvres d'un même auteur, du même genre et pour lesquelles on ne possède pas de données distinctes de datation (ce qui est le cas des *Lais* de Marie de France), le contenu du volume physique est représenté par un seul fichier. Mais à l'intérieur de ce fichier chaque œuvre correspond à une division structurelle délimitée au moyen de balises TEI (cf. ci-dessous) et peut donc être récupérée pour une analyse plus fine. Dans le cas enfin où un

[1] La personne qui vérifie la conformité du texte avec l'édition de référence et qui ajoute les balises au texte.

volume physique rassemble des œuvres d'auteurs différents et/ou écrites à des dates distinctes, chacune de ces œuvres sera représentée dans un fichier séparé[2].
Dans son état actuel la représentation des textes dans la BFM se limite au texte proprement dit. Sont donc écartés la page de garde, la préface et les documents liminaires, ainsi que les notes de l'éditeur quelle que soit leur place, le glossaire, etc. Les informations bibliographiques portées par la page de garde sont intégrées dans les rubriques correspondantes des en-têtes TEI. On maintient le titre de l'œuvre dans le texte uniquement dans le cas où ce titre figure sur la première page du texte imprimé.

1.1.2 Principes de balisage

- les documents à baliser sont au format texte brut (il n'y a pas de police ou de style particulier à utiliser) ;
- un document est composé d'éléments, qui sont délimités par des balises ;
- chaque balise est délimitée par des chevrons (**<, >**) ;
- on distingue les éléments qui contiennent un autre type d'élément (ex : le chapitre contient une portion du texte et il est susceptible d'être divisé en paragraphes...), et ceux qui indiquent uniquement une frontière (ex : le saut de ligne marque la frontière entre deux lignes) :
    - dans le premier cas, la balise qui a un contenu (ou une portée) se place au début de son contenu et le termine par une balise fermante qui possède un caractère « / » en préfixe (ex : ** **...** **) ;
    - dans le second cas, la balise qui n'a pas de contenu est unique et possède un caractère « / » en suffixe (ex : **<pb/>**).

Exemple :

```
<div type="chapitre" n="1">
contenu du premier chapitre
</div>
```

Glose : on appelle élément la partie du document qui commence avec la balise ouvrante ** ** et se termine avec la balise fermante ** **

- toute balise peut posséder plusieurs propriétés en plus de son nom, sous la forme d'une succession de relations **nom-attribut="valeur-attribut"** situées entre les chevrons (ex : **<pb n="2">**, où l'attribut **n** encode le numéro de la page qui débute) ;
- l'ordre dans lequel on indique les attributs est libre ;
- on indiquera le nom et la valeur de l'attribut si on possède cette information ;
- le nombre d'espaces ou de tabulations situés entre deux mots ou balises dans le corps du texte et dans les valeurs d'attributs n'a pas d'interprétation particulière ;
- de même pour les sauts de ligne et leur nombre ;
- les textes doivent être enregistrés en format texte brut (avec un encodage des caractères Windows ou Unicode).

[2] Dans le cas unique du *Roman de la Rose* dont le premier volume de l'édition rassemble le début du texte, composé par Guillaume de Lorris entre 1225 et 1230, et la suite de ce texte, composée par Jean de Meun entre 1269 et 1278, nous avons choisi de séparer dans deux fichiers distincts ces deux parties.

1.2 Rôle des différents acteurs participant à l'édition du texte informatisé

Le document numérique étant constitué à partir d'une édition de référence imprimée, il est nécessaire d'indiquer au fil du texte les interventions de l'éditeur scientifique mises en évidence par des procédés typographiques. Pour passer du document imprimé à la version électronique du texte, trois étapes se succèdent :

1) la numérisation du texte de l'édition de référence ;
2) la vérification que cette version est conforme au texte de l'édition de référence, et le prébalisage du texte ;
3) l'enrichissement du balisage, la vérification formelle de la structure du document XML et la validation finale.

La TEI prévoit la description formelle de chaque personne intervenant dans la constitution du document. Chaque correction, intervention éditoriale ou commentaire doit porter la mention de la personne qui en est responsable. On utilise pour ce faire l'attribut **`resp`** auquel on adjoint l'une des quatre valeurs suivantes :

- **`"editor"`**, qui correspond à l'éditeur scientifique ;
- **`"digitizer"`**, qui correspond à la personne qui numérise le texte de l'édition de référence ;
- **`"proofreader"`**, qui correspond à la personne qui relit et fait le prébalisage du texte (si plusieurs personnes sont chargées de ce travail on leur attribue un numéro, par exemple : "proofreader1", etc.) ;
- **`"encoder"`**, qui correspond à la personne qui réalise la vérification formelle et la validation finales.

2 Description formelle du balisage des textes de la BFM

2.1 Délimitation du corps du texte et des éléments qui lui sont externes (prologue…)

- l'élément **`<text>`** regroupe à la fois le corps du texte et tous les éléments qui lui sont externes
- l'élément **`<body>`** marque le début et la fin du corps du texte proprement dit
- l'élément **`<front>`** marque les informations liminaires : prologue, sommaire…, et surtout le titre de l'œuvre
- l'élément **`<back>`** marque les informations supplémentaires : appendice, index…, et surtout l'explicit de l'oeuvre

NOTA BENE :

- en plus de son encadrement par l'élément **`<body>`**, le contenu du texte doit toujours se trouver dans au moins un élément de structuration, par exemple dans un élément paragraphe **`<p>`** ;
- dans le même ordre d'idées, un élément **`<front>`** devra toujours encadrer au moins un paragraphe (l'élément **`<p>`**) et une division **`<div>`**, cette division contenant éventuellement un **`<head>`** (cf. 2.2.1 et 2.2.3) ;

- même chose pour l'élément **<back>**
  - dans le cas où le **<front>** encadre le titre, la balise ** ** a un attribut **type="titre"** et la balise **<p>** n'a pas d'attribut ;
  - dans le cas d'un explicit, la balise ** ** a un attribut **type="explicit"** et la balise **<p>** n'a pas d'attribut.

La structure potentielle d'un texte est donc la suivante :

```
<text>
<front>
<div type="titre">
<p>
titre
</p>
</div>
</front>
<body>
contenu du texte
</body>
<back>
<div type="explicit">
<p>
explicit
</p>
</div>
</back>
</text>
```

Seuls les éléments **<text>** et **<body>** sont obligatoires dans un texte encodé en TEI.

Exemple :

```
<text>
<body>
<p>
Buona pulcella fut Eulalia,
Bel auret corps, bellezour anima.
Uoldrent la ueintre li Deo inimi,
Uoldrent la faire diaule seruir.
Elle no.nt eskoltet les mals conselliers,
Qu'elle Deo raneiet chi maent sus en ciel.
…
Tuit oram que por nos degnet preier
Qued auuisset de nos Christus mercit
Post la mort et a lui nos laist uenir
Par souue clementia.
</p>
</body>
</text>
```

2.2 Structure du corps du texte

2.2.1 Délimitation des parties du texte : livres, chapitres, sections, sous-sections…

- l'élément ** ** marque n'importe quelle division du texte ;
- un élément ** ** contient au moins un élément d'un niveau de structuration inférieur : soit un ** ** de niveau inférieur, soit un **<p>** (cf. 2.2.3 pour plus de détails) ;
- l'élément ** ** peut être renseigné avec l'attribut **type** pour indiquer le type de la division (chapitre, section…) et par un attribut **n** pour indiquer son numéro éventuel ;

Exemple :

```
…
Et pour vous informer du temps dont ay eu congnoissance
dudit seigneur, dont faictes demande, m'est force de
commancer avant le temps que je veinse en son service; et
puis, par ordre, je suyvray mon propos jusques à l'heure
que je devins son serviteur, et continueray jusques à son
trespas.
</p>
</div>
<div type="livre" n="1">
<div type="chapitre" n="1">
<p>
Au saillir de mon enfance et en l'aage de povoir monter
à cheval, fus amené à Lisle devers le duc Charles de
Bourgoigne, lors appellé conte de Charroloys, lequel me
print en son service, et fut l'an mil quatre cens soixante
quatre … </p>
…
</div>
</div>
```

Notation des titres de livre, chapitre…

- chaque élément ** ** peut s'ouvrir avec un premier élément **<head>** contenant le titre ou l'entête de la division et se clore par un élément **<trailer>** contenant des informations présentes en fin de division.

Exemple :

```
<div type="chapitre" n="1">
<head>
```

```
Ci commence li premiers chapitres qui parole de l'office as baillis.
</head>
Tout soit il ainsi qu'il n'ait pas en nous toutes les
graces qui doivent estre en homme qui s'entremet de baillie,
pour ce ne lerons nous pas a traitier premiers en cest
chapitre de l'estat et de l'office as baillis, et dirons briement
une partie des vertus qu'il doivent avoir, et comment il se
doivent maintenir, si que cil qui s'entremetront de l'office
i puissent prendre aucune essample...
</div>
```

NOTA BENE :

Il peut arriver qu'on rencontre dans un texte un titre qui ne correspond à aucune division logique (et qui n'est pas numéroté). On considèrera alors qu'on a affaire à une pseudo-division qu'on balisera au moyen de la balise ** … **. Le titre se trouvera comme dans les cas précédents dans son **<head>**.

2.2.2 Numérotation des pages

- l'élément **<pb/>** marque les sauts de page :
  - il a un attribut **n** qui permet d'encoder le numéro de la page qui s'ouvre[3] ;
  - il a aussi un attribut **ed** qui permet éventuellement de préciser de quelle édition provient la pagination (en cas d'annotation simultanée de la pagination de plusieurs éditions).

Exemple :

```
…
«Or i voist donc, fait ele, car se il demain ne deust
revenir il n'i alast hui par ma volenté.» Et il monte
et la damoisele ausi, <pb n="2"/> si se partent de laienz sanz
autre congié, et sanz plus de compaignie, fors
solement dui escuier qui avec la damoisele
estoient venuz. Et quant il sont issu de Kamaalot
…
```

NOTA BENE :

Il arrive que des illustrations s'intercalent dans le texte. Si elles occupent la totalité d'une page, il est nécessaire d'insérer un saut de page à l'endroit où elles se trouvent dans le texte (y compris si elles ne sont pas conservées dans la version numérique du texte).

2.2.3 Délimitation des unités inférieures

Textes en prose

[3] Cet attribut est généralement ajouté par des outils de numérotation automatique lors de la vérification finale du balisage.

- l'élément **<p>** marque les paragraphes ;
- l'élément **<lb/>** peut être utilisé pour marquer les sauts de ligne, mais le plus souvent les sauts de ligne de l'édition ne sont pas pris en compte lors de la numérisation de textes en prose.

Exemple :

```
…
<p>
«Or i voist donc, fait ele, car se il demain ne deust
revenir il n'i alast hui par ma volenté.» Et il monte
<pb n="2"/>
et la damoisele ausi, si se partent de laienz sanz
autre congié, et sanz plus de compaignie, fors
solement dui escuier qui avec la damoisele
estoient venuz. Et quant il sont issu de Kamaalot
</p>
…
```

Textes en vers

- l'élément **<p>** marque les groupes de vers : laisses, strophes, refrains…
  - il a obligatoirement un attribut **rend** auquel on peut donner plusieurs valeurs : laisse, strophe…[4] Si la qualification des groupes de vers d'un ouvrage est difficile, l'attribut **rend** a la valeur **"gv"** (groupe de vers)[5].
- l'élément **<lb/>** marque les débuts de vers, y compris quand le vers est incomplet.
  - cet élément est placé au début (et non à la fin) de chaque vers pour faciliter la mise en forme des textes au moyen de feuilles de styles ;
  - il a un attribut **n** qui permet d'encoder éventuellement le numéro du vers (cf. élément **<pb/>**).

Exemple :

```
<text>
<body>
<p rend="gv">
<lb/>Buona pulcella fut Eulalia,
<lb/>Bel auret corps, bellezour anima.
…
<lb/>Post la mort et a lui nos laist uenir
<lb/>Par souue clementia.
</p>
</body>
</text>
```

[4] Voir l'Annexe 3.5 pour la « liste d'autorité » des valeurs de cet attribut.

[5] Nous avons choisi de ne pas utiliser dans ce cas l'élément **<lg>** recommandé par la TEI, parce qu'il impose qu'on balise chaque vers au moyen de l'élément **<l>** qui entraîne par ailleurs la présence de contraintes indésirables, notamment en cas de balisage des unités syntaxiques et/ou du discours direct.

Théâtre

- l'élément **<sp>** marque les prises de parole :
  - son attribut **who** permet de renseigner le nom du locuteur ;
- l'élément **<stage>** marque les didascalies.

Exemple :

```
<sp who="Pathelin">
Encor ne le dis je pas pour me
vanter, mais n' a, au territoire
ou nous tenons nostre auditoire,
homme plus saige fors le maire.
</sp>
<sp who="Guillemette">
Aussy a il leu le grimaire
et aprins a clerc longue piece.
</sp>
```

2.3 Indications dans le corps du texte

2.3.1 Délimitation explicite des lexies

L'objectif premier de la BFM est d'être utilisée dans des recherches linguistiques au moyen d'outils de requête et d'analyse, tels que le logiciel Weblex. L'intégration de textes dans Weblex prévoit notamment un balisage automatique et une indexation des mots (lexies) et des phrases. Cette procédure se base sur des critères formels : les lexies sont séparées par des espaces ou autres « séparateurs » (par exemple l'apostrophe), les phrases sont délimitées par des marques de ponctuation forte (point, point d'exclamation, point d'interrogation).
Dans les cas où la présence d'un « séparateur » formel ne correspond pas à une frontière linguistique (par exemple le point après une abréviation ne signifie pas une fin de phrase), il convient de procéder à un balisage explicite.

2.3.1.1 Nombres

Certaines éditions utilisent les points pour la mise en évidence des chiffres, en suivant la pratique des manuscrits médiévaux.

- l'élément **<num>** marque les chiffres mis en évidence par les points.

Exemple :

```
<p n="1"> Sachiez que <num>.M.</num> et <num>.C.</num> et quatre
vinz et <num>.XVII.</num> anz aprés l'incarnation Nostre Sengnor
Jesu Crist, …
</p>
```

2.3.1.2 Abréviations

- l'élément **<abbr>** marque les abréviations suivies d'un point.

Exemple :

```
<lb/>Je vaudroie bien avoir mis
<lb/>En amender vostre pesance
<lb/><num>.C.</num> <abbr>s.</abbr>, ke ceste desevrance
<pb n="204"/>
<lb/>Me fait plus mal que jou n' os dire."
```

Glose : Ici, le mot *sol* (ancienne unité monétaire) est noté dans l'édition par une abréviation *s*.

2.3.1.3 Mots composés

Dans l'orthographe du français moderne, certains lexèmes sont notés en plusieurs mots graphiques (*aujourd'hui*, *parce que*). Dans la perspective diachronique, il est cependant difficile de définir *a priori* à quel moment précis une locution devient un mot unique. Il a donc été décidé de s'en tenir à une définition formelle de l'unité-mot, telle qu'elle est représentée par des caractères séparateurs dans l'édition critique.

2.3.2 Corrections et interventions éditoriales

Les éditions de textes en ancien français contiennent souvent des marques typographiques (italiques, caractères gras, majuscules, crochets, etc.) qui servent à mettre en évidence des passages dans une langue étrangère, un changement de manuscrit, une coquille, une intervention éditoriale, etc. Les pratiques varient selon les éditions. Il convient donc d'analyser l'usage des marques typographiques dans chaque édition et de baliser les passages mis en évidence conformément à leur nature avec les éléments qu'offre la TEI.

2.3.2.1 Corrections de l'éditeur scientifique

- l'élément **<corr>** sert à marquer les corrections effectuées par l'éditeur scientifique et qui sont indiquées de façon explicite dans le corps du texte de l'édition ;
  - son attribut **resp** indique le responsable de la correction, c'est-à-dire l'éditeur scientifique ("editor") ;
  - son attibut **sic** permet d'indiquer le texte original remplacé par la correction. Il sera vide (**sic=""**) en cas d'ajout (en cas, par exemple, d'une préposition

manquante). Il sera absent si l'éditeur scientifique n'indique pas le texte original sur la même page ;
- son attribut **rend** permet d'indiquer la marque typographique utilisée dans l'édition pour mettre la correction en évidence. La valeur de cet attribut dépend de la « portée » de la correction ;
- son attribut **cert** permet d'indiquer éventuellement la certitude de la correction.

NOTA BENE :

Pour faciliter le traitement ultérieur des textes (et en particulier l'étiquetage morphosyntaxique), nous avons distingué les corrections qui touchent une partie d'un mot et celles qui concernent un mot entier, voire plusieurs mots. Dans tous les cas, l'élément **<corr>** contient au moins un mot entier.

Exemple de balisage d'une correction d'une partie de mot :

```
<lb n="2"/>Ki son sens aüse et <corr resp="editor" sic="travalle"
rend="trava[i]lle">travaille</corr>
```

Glose : Ici, d'après l'éditeur scientifique, une lettre *i* est omise par erreur de copiste dans la graphie *travaille*. L'éditeur l'a donc ajoutée entre crochets. Le contenu de l'élément est alors la « bonne » forme, d'après l'éditeur. La forme (erronée) du manuscrit est fournie par l'attribut **sic**, et la forme de l'édition est reproduite « telle quelle » dans l'attribut **rend**.

Exemple de balisage d'une correction concernant plusieurs mots :

```
<lb n="6068"/>Quant li palefrois biaus et gens
<lb n="6069"/><corr resp="editor" sic="" rend="crochets">Fu venus la
pucele i monte.</corr>
<lb n="6070"/>Li maistre cambrelens le conte
```

Glose : Ici, d'après l'éditeur scientifique, un vers est omis dans le manuscrit de base. Ce vers est rétabli d'après un autre manuscrit et placé entre crochets. Ces crochets sont remplacés par la balise **<corr>** dont l'attribut **sic** est « vide » (ce qui signifie qu'il s'agit d'un ajout) et l'attribut **rend** signale la marque typographique utilisée. Pour les valeurs de cet attribut, il convient de se référer à la liste d'autorité présentée dans l'Annexe 3.4.

2.3.2.2 Propositions de correction du relecteur/encodeur

- L'élément **<sic>** peut être utilisé par le relecteur ou l'encodeur pour proposer une correction d'une erreur flagrante de l'édition. La règle fondamentale de la représentation numérique des textes de la BFM est que le texte numérisé doit être la copie exacte du texte papier de l'édition de référence. La correction est proposée dans un attribut de la balise, et le contenu de l'élément reste une reproduction fidèle de l'édition de référence. Le texte de l'édition **ne saurait donc être modifié** ;
  - l'attribut **resp** indique quelle est la personne responsable de la correction (**"proofreader"** ou **"encoder"**) ;

- l'attibut **corr** contient la correction proposée. Il sera vide (**corr=""**) en cas de suppression (en cas, par exemple, d'un mot répété), et le contenu de l'élément sera vide en cas d'ajout ;
- l'attribut **cert** permet éventuellement au relecteur d'indiquer s'il est sûr ou non de la correction (**cert="yes"** ou **"no"**) ;
- le relecteur peut utiliser l'élément **<note>** pour ajouter un commentaire (cf. 2.3.4 ci-dessous).

Exemple :

```
<sic resp="encoder" corr="Ço">Co</sic> que li plus halz fist plus
bas peüst desfaire;
```

Glose : Ici le démonstratif est noté avec un *C* sans cédille dans l'édition. Cependant, il y a dans l'édition d'autres occurrences du pronom qui comportent la cédille. L'encodeur, qui a constaté cette incohérence, a décidé de proposer une correction en utilisant l'attribut **corr** de l'élément **<sic>** (la forme utilisée dans le contenu de l'élément est celle de l'édition).

2.3.2.3 Lacunes du manuscrit indiquées par l'éditeur scientifique

- l'élément **<gap/>** permet d'indiquer les lacunes constatées par l'éditeur scientifique dans le texte ;
  - son attribut **resp** permet d'indiquer la personne qui a constaté la lacune, c'est-à-dire l'éditeur scientifique (**resp="editor"**) ;
  - son attribut **rend** permet d'indiquer éventuellement la marque typographique utilisée dans l'édition ("…", par exemple). Pour les valeurs de cet attribut, il convient de se référer à la liste d'autorité présentée dans l'Annexe 3.4 ;
  - son attribut **desc** permet éventuellement de décrire la nature de la lacune (par exemple, **"manuscrit endommagé"**, **"vers omis"**). Il convient de se référer aux notes de l'éditeur scientifique pour renseigner cet attribut ;
  - son attribut **extent** permet éventuellement d'indiquer l'ordre de grandeur de la lacune. On peut utiliser des valeurs comme **"1 line"**, **"0,5 line"**, etc.

Exemples :

```
<lb n="721"/>L'autres gaians, qui rostissoit
<lb n="719"/><gap resp="editor"/>
<lb n="720"/>Et aveuc son poivre faisoit.
```

```
Dedens vont, regardent les <gap resp="editor"/>
<lb n="6068"/>Afaitent les, metent
```

2.3.2.4 Passages difficiles à lire dans le manuscrit

- l'élément **<unclear>** permet de marquer des passages qui ne sont pas clairs dans le manuscrit source et que l'éditeur scientifique met en évidence à l'aide de marques typographiques. Son usage est limité normalement aux éditions diplomatiques ;
  - son attribut **resp** sert à indiquer la personne qui a mis en évidence le passage incertain, c'est-à-dire l'éditeur scientifique (**resp="editor"**) ;
  - son attribut **rend** permet éventuellement d'indiquer la marque typographique utilisée dans l'édition. Pour les valeurs de cet attribut, il convient de se référer à la liste d'autorité présentée dans l'Annexe 3.4 ;
  - son attribut **reason** permet éventuellement d'indiquer la raison pour laquelle le passage est considéré comme n'étant pas clair (par exemple, **"illegible"** ou **"ambiguous"**) ;

Exemple :

```
Que les prisons touz uos r<unclear resp="editor">en</unclear>drai. <lb>
```

2.3.3 Autres types de passages mis en évidence

2.3.3.1 Passages en langue étrangère

- L'élément **<foreign>** marque les passages écrits dans une langue différente de celle du texte.
  - si c'est du latin, ce qui est le cas le plus fréquent, la balise **<foreign>** est suffisante. S'il s'agit d'une autre langue, il convient d'ajouter l'attribut **lang** dont la valeur permet de préciser quelle la langue est utilisée ;
  - l'attribut **rend** permet d'indiquer quelle marque typographique est employée. Pour les valeurs de cet attribut, il convient de se référer à la liste d'autorité présentée dans l'Annexe 3.4.

Exemple :

```
…
- Et tout li haut homme, et clerc et lai et petit et grant,
demenerent si grant goie a l'esmovoir que onques encore si faite
goie ne si fais estoires ne fu veus ne oïs; et si fisent li pelerin
monter as castiaus des nes tous les prestres et les clers qui
canterent <foreign> Veni creator spiritus </foreign>. Et trestout
et grant et petit plorerent de pec et de le grant goie qu'i eurent
…
```

2.3.3.2 Mises en évidence typographiques dont la signification n'est pas claire

- L'élément **<hi>** marque les passages imprimés dans une typographie différente de celle qu'on trouve habituellement dans le texte et dont on n'a pas d'interprétation particulière :

- la marque typographique de l'italique est encodée au moyen de l'élément **<hi rend="ital">…</hi>** ;
- la marque typographique de mots en majuscules est encodée au moyen de l'élément **<hi rend="maj">…</hi>** ;
- la marque typographique de mots en petites majuscules est encodée au moyen de l'élément **<hi rend="pmaj">…</hi>** ;
- la marque typographique de mots en exposant est encodée au moyen de l'élément **<hi rend="exp">…</hi>**.
- la marque typographique de mots en indice est encodée au moyen de l'élément **<hi rend="ind">…</hi>**[6].

Exemple : ;

jj^{c} → jj**<hi rend="exp">**c**</hi>**

2.3.4 Notes et commentaires du relecteur/encodeur

Nous avons fait le choix de ne pas conserver dans notre version numérique du texte les notes de l'éditeur scientifique (dans lesquelles il donne en particulier des variantes du texte).

- l'élément **<note>** marque les annotations et les commentaires du relecteur/encodeur.
  - son attribut **resp** doit indiquer quelle est la personne responsable de la note : le relecteur (**"proofreader"**), l'encodeur (**"encoder"**)…

Exemple :

```
<lb n="423"/>A Com cist cheualiers qui ci siet. <note
resp="proofreader">Deux lettres majuscules initiales</note>
<lb n="423"/>Qu'il ne respont ne un neel.
```

Glose : l'éditeur indique la présence de deux majuscules en début de ligne.

2.3.5 Passage au discours direct ou indirect

- l'élément **<q>** marque les passages au discours direct ou indirect :
  - son attribut **who** permet de marquer le nom du locuteur ;
  - son attribut **type** permet éventuellement d'indiquer s'il s'agit de paromes prononcées ou de pansée ;
  - son attribut **direct** permet éventuellement d'indiquer si le passage est au discours direct ou indirect (les valeurs acceptées sont "y", "n" et "unspecified");
  - un élément **<q>** peut en contenir un autre, si un personnage cite les paroles d'un autre.

[6] Cf. Annexe 3.4.

Exemple :

```
Et quant Melyan voit ces letres si dist a Galaad :
<q who="Melyan">Frans chevaliers por Dieu lessiez moi entrer en cele
a senestre, car en cele porrai je esprover ma force, et connoistre
s'il avra ja en moi proesce ne hardement por quoi je doie avoir los
de chevalerie.</q>
- <q who="Galaad">S'il vos pleust</q>, fait Galaad,
<q who="Galaad">je m'en entrasse en cele a senestre, car si com je
pens je m'en getasse mielz que vos.</q>
```

2.3.6 Citations

- l'élément **<quote>** marque les passages cités ou les mentions (du type : indication du nom qui est écrit sur un siège…).

Exemple :

```
…
Si troevent le perron qui estoit venuz a rive et issuz
hors de l'eve, et estoit de marbre vermeil, et ou perron
estoit une espee fichiee qui mout estoit bele et riche
par semblant, et en estoit li ponz d'une pierre
precieuse ouvrez a letres d'or mout soutilment, et li baron
resgardoient les letres qui disoient : <quote>Ja nus ne
m'ostera de ci se cil non a qui costé je pendrai, et cil sera
li mieldres chevaliers del monde </quote>. Et quant li rois voit
ces letres si dist a Lancelot : « Biau sire ceste espee est vostre
par bon droit, car je sai bien que vos estes li mieldres
chevaliers dou monde… »
…
```

2.3.7 Annotations linguistiques

- l'élément **<w>** encadre toutes les indications à caractère linguistique portant sur un lexème du texte (indications de nature morphologique, syntaxique, sémantique, pragmatique, etc.)
  - son attribut **type** permet de préciser la nature ou la valeur de ces indications et l'attribut **lemma** permet d'introduire le lemme.

Exemple :

```
<lb n="1"/>Puis que ma dame de Chanpaigne
<lb n="2"/>vialt que romans a feire anpraigne,
<lb n="3"/>je l'anprendrai molt volentiers
<lb n="4"/>come <w type="8">cil</w> qui est suens antiers
```

```
<lb n="5"/>de quanqu'il puet el monde feire
```

3 Annexes

3.1 Index des balises et des attributs

3.2 Un exemple de prose : extrait des Conquête de Constantinople *de Villehardouin (avec en-tête renseigné)*

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<!DOCTYPE TEI.2 PUBLIC "-//TEI P4//DTD Main DTD Driver File//EN"
"/dtd/sgml/tei/tei2.dtd" [
<!ENTITY % TEI.XML "INCLUDE" >
<!ENTITY % TEI.general "INCLUDE" >
<!ENTITY % TEI.corpus  "INCLUDE" >
<!ENTITY % TEI.analysis "INCLUDE" >
<!ENTITY % TEI.certainty  "INCLUDE" >
<!ENTITY % TEI.figures  "INCLUDE" >
<!ENTITY % TEI.linking  "INCLUDE" > ]>
<TEI.2>
  <teiHeader lang="fr" type="text" date.created="2004-01-19">
    <fileDesc>
      <titleStmt>
        <title type="normal">Conquête de Constantinople</title>
        <title type="formal" n="1">villehardouin</title>
        <title type="reference">villehardouin1</title>
        <title type="gmd">transcription électronique</title>
        <author>Villehardouin</author>
        <editor role="editor">Équipe diachronie et bases textuelles
d'ancien et moyen français - UMR5191 ICAR, CNRS/ENS-LSH</editor>
        <respStmt>
          <resp>Encodage TEI et validation SGML/XML</resp>
          <name type="person">Céline Guillot</name>
          <resp>Encodage TEI et validation SGML/XML</resp>
          <name type="person">Serge Heiden</name>
        </respStmt>
      </titleStmt>
      <extent>- Taille approximative du fichier encodé en TEI non compressé
: 107908 octets
- Texte brut : 339 lignes, 18112 mots, 96123 caractères
(ces statistiques élémentaires ont été calculées avec l'outil Unix 'wc')
- Numéros de pages saisies : de 2 à 210.
      </extent>
      <publicationStmt>
        <distributor><name>UMR5191 ICAR, CNRS/ENS-LSH</name>
<address>
<addrLine>15, parvis René Descartes</addrLine>
<addrLine>69342 Lyon BP7000 Cedex 07</addrLine>
<addrLine>Tél 04 37 37 63 10</addrLine>
<addrLine>Mèl : Celine.Guillot@ens-lsh.fr</addrLine>
</address>
        </distributor>
        <availability status="restricted">
          <p>(c) 2002, CNRS/ENS-LSH.
<hi>Conditions d'utilisation</hi> : usage dans le cadre du transfert
des textes de la BFM (Base de Français Médiéval) de l'UMR 5191
ICAR (Lyon) à l'UMR 7118 Analyse et Traitement Informatique de la Langue
Française (Nancy),
pour leur intégration dans la base Frantext.
Pour d'autres conditions d'usage, contacter :
      <address>
<addrLine><name type="person">Christiane Marchello-Nizia</name></addrLine>
<addrLine>Mèl : marchell@linguist.jussieu.fr</addrLine></address>
    </p>
        </availability>
```

```
        <idno type="Cote Bibliothèque ENS-LSH">COR3-VIL</idno>
      </publicationStmt>
      <seriesStmt>
        <title>Les Classiques de l'Histoire de France au Moyen Âge,
18</title>
        <idno>18</idno>
      </seriesStmt>
      <sourceDesc>
        <biblFull>
          <titleStmt>
            <title>VILLEHARDOUIN : LA : CONQUÊTE DE CONSTANTINOPLE : ÉDITÉE
ET TRADUITE PAR : EDMOND FARAL : MEMBRE DE L'INSTITUT : ADMINISTRATEUR DU
COLLÈGE DE FRANCE : TOME Ier : (1199-1203) : DEUXIÈME ÉDITION REVUE ET
CORRIGÉE</title>
            <respStmt id="editor">
              <resp>éditeur scientifique</resp>
              <name type="person">E. Faral</name>
            </respStmt>
          </titleStmt>
          <editionStmt>
            <edition/>
          </editionStmt>
          <extent>209 p. soit NA</extent>
          <publicationStmt>
            <publisher>Belles Lettres</publisher>
            <pubPlace>Paris</pubPlace>
            <date>1938-1939</date>
            <idno type="ISBN">n-a</idno>
          </publicationStmt>
        </biblFull>
      <p>
          <name id='latin' type='lang'/>
        </p>
      </sourceDesc>
    </fileDesc>
    <encodingDesc>
      <projectDesc>
        <p>Projet : BFM - Base de Français Médiéval
Resp : <name type="person">Christiane Marchello-Nizia</name>
Équipe : diachronie et bases textuelles d'ancien et de moyen français
Laboratoire : UMR5191 ICAR
Institution : CNRS/ENS-LSH, Lyon
<address>
<addrLine>15, parvis René Descartes</addrLine>
<addrLine>69342 Lyon BP7000 Cedex</addrLine>
<addrLine>Tél 04 37 37 63 10</addrLine>
</address></p>
      </projectDesc>
      <editorialDecl>
        <p><hi>Introduction</hi>
Les principes éditoriaux de la BFM seront publiés dans le recueil des actes
du colloque
"Ancien et moyen français sur le Web : enjeux méthodologiques, 4 et 5
octobre 2002, Ottawa", sous le titre :
HEIDEN Serge, GUILLOT Céline, "Capitalisation des savoirs par le web : une
application de la TEI pour l'encodage
et l'exploitation des textes de la Base de Français Médiéval", Benjamins
éds.</p>
        <p><hi>Corrections</hi> : le texte a été relu 2 fois, une fois pour
le texte et
```

```
une fois pour le texte et les balises. Il a été validé en SGML avec l'outil
NSGMLS et la DTD TEI P4X
et en XML avec l'outil RXP et la DTD TEI P4X.</p>
        <p><hi>Normalisation</hi> : l'édition de référence a été respectée
au maximum
pour les éléments suivants :
- l'orthographe ;
- la ponctuation ;
- la mise en page ;
- la pagination.</p>
        <p><hi>Citations</hi> : le discours direct et les citations ne sont
pas encodés.</p>
        <p><hi>Notes</hi> : les notes de bas de page ou de fin ont été
retirées.</p>
        <p><hi>Césures</hi> : les césures de mots ont été réduites. En cas
de césure sur
une limite de page, le mot est placé dans la page où il commence.</p>
        <p><hi>Notes éditoriales</hi> : le travail éditorial, souvent
marqué par des crochets
dans l'édition de référence, a été réduit. Les éléments CORR, SIC et GAP
servent à encoder ces informations.</p>
        <p><hi>Segmentation</hi> : le titre du texte se trouve dans un
élément FRONT ainsi
que le prologue (s'il existe). Le corps du texte est dans le BODY.
Les explicits (s'ils existent) sont placés dans un élément BACK.
Les livres, les chapitres, les lais, les chansons, les miracles et les
serments
sont encodés par des éléments DIV dont l'attribut TYPE précise la nature.
Les paragraphes en prose sont encodés par l'élément P.
Dans les textes en vers les strophes et les laisses sont encodés par
l'élément P rend="gv".
Les paragraphes et les strophes sont numérotés quand ils le sont dans
l'édition de référence, à l'aide de l'attribut N.
Les débuts de vers sont encodées par l'élément LB.
L'attribut N des éléments LB encode le numéro du vers.
Les sauts de page sont encodés par l'élément PB.</p>
        <p><hi>Interprétation</hi> : L'élément TEXT n'encode que le texte
du (des) manuscrit(s)
proprement dit(s), les introduction, sommaire, glossaire, index, table des
matières, etc. sont absents de l'édition électronique.
Un élément TEXT correspond toujours à un volume physique unique. Quand
l'oeuvre
est éditée en plusieurs volumes, chaque volume correspond donc à un élément
TEXT. Un fichier ne contient qu'un seul élément TEXT.
Les toponymes, les noms de personnes et de groupes ne
sont pas encodés. Les passages en langue étrangère sont encodés avec
l'élément FOREIGN (le latin essentiellement).</p>
        <p><hi>Décomptes</hi> : la section tagsDecl de l'en-tête TEI ne
tient pas compte des éléments de l'en-tête TEI.</p>
      </editorialDecl>
      <tagsDecl>
        <tagUsage gi="body" occurs="1"/>
        <tagUsage gi="div" occurs="28"/>
        <tagUsage gi="head" occurs="28"/>
        <tagUsage gi="lb" occurs="3"/>
        <tagUsage gi="note" occurs="2"/>
        <tagUsage gi="num" occurs="104"/>
        <tagUsage gi="p" occurs="231"/>
        <tagUsage gi="pb" occurs="209"/>
        <tagUsage gi="text" occurs="1"/>
      </tagsDecl>
```

```
    </encodingDesc>
    <profileDesc>
      <creation><name type="author">Villehardouin</name>
     <date value="13" rend="ancien">début XIIIème, entre 1199 et
1213</date></creation>
      <langUsage>
        <language id="fr" usage="100">Le texte est entièrement écrit en
prose en ancien français</language>
      </langUsage>
      <textDesc n="chronique">
        <channel mode="w">print</channel>
        <constitution type="single"></constitution>
        <derivation type="original"></derivation>
        <domain>historique</domain>
        <factuality>mixed</factuality>
        <interaction type="none"></interaction>
        <preparedness type="prepared"></preparedness>
        <purpose type="dialecte">non défini</purpose>
        <purpose type="forme">prose</purpose>
        <purpose></purpose>
      </textDesc>
    </profileDesc>
  </teiHeader>
<text>
<pb n="2"/>
<body>
<div part="N" n="1" type="section">
<head rend="crochets-maj">Les origines de la croisade</head>
<div part="N" n="1" type="sous-section">
<head rend="crochets-pmaj">La prédication</head>
<p rend="sous-titre">(1198)</p>
<p n="1"> Sachiez que <num>.M.</num> et <num>.C.</num> et quatre vinz et
<num>.XVII.</num> anz aprés l'incarnation Nostre Sengnor Jesu Crist, al
tens Innocent, apostoille de Rome, et Phelippe, roy de France, et Ricchart,
roy d'Engleterre, ot un saint home en France, qui ot nom Folques de Nuilli
(cil Nuillis si est entre Ligni sor Marne e Paris) et il ere prestres et
tenoit la parroiche de la ville. Et cil Folques dont je vos di comença a
parler de Dieu par France et par les autres terres entor; et Nostre Sires
fist maintes miracles por lui. </p>
<p n="2"> Sachiez que la renomee de cel saint home ala tant qu'ele vint a
l'apostoille de Rome Innocent; et l'apostoille envoia en France, et manda
al prodome
<pb n="3"/>
<pb n="4"/>
que il preechast des croiz par s'autorité; et aprés i envoia un suen
chardonal, maistre Perron de Chappes, croisié, et manda par lui le pardon
tel con je vos dirai: tuit cil qui se croisseroient et feroient le servise
Deu un an en l'ost seroient quite de toz les pechiez que il avoient faitz,
dont il seroient confés. Porce que cil pardons fu issi granz, si s'en
esmurent mult li cuer des gens, et mult s'encroisierent porce que li
pardons ere si granz.
</p>
</div>

…

</body>
</text>
</tei.2>
```

3.3 *Un exemple de poésie : extrait des* Lais *de Marie de France (en-tête non renseigné)*

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<!DOCTYPE TEI.2 PUBLIC "-//TEI P4//DTD Main DTD Driver File//EN"
"/dtd/sgml/tei/tei2.dtd" [<!ENTITY % TEI.XML "INCLUDE" ><!ENTITY %
TEI.general "INCLUDE" ><!ENTITY % TEI.corpus  "INCLUDE" ><!ENTITY %
TEI.analysis "INCLUDE" ><!ENTITY % TEI.certainty  "INCLUDE" ><!ENTITY %
TEI.figures  "INCLUDE" ><!ENTITY % TEI.linking  "INCLUDE" > ]>
<TEI.2>
<teiHeader type='text'>
<fileDesc>
<titleStmt>
<title/>
</titleStmt>
<publicationStmt>
<p/>
</publicationStmt>
<sourceDesc>
<p>
          <name id='latin' type='lang'/>
        </p>
</sourceDesc>
</fileDesc>
</teiHeader>
<text>
<front>
<div type='prologue'>
<pb n='1'/>
<head>PROLOGUE</head>
<p rend='gv'>
<lb n='1'/><hi rend="pmaj">Ki</hi> Deus ad duné escïence
<lb n='2'/>E de parler bone eloquence
<lb n='3'/>Ne s'en deit taisir ne celer,
<lb n='4'/>Ainz se deit voluntiers mustrer.
<lb n='5'/>Quant uns granz biens est mult oïz,
<lb n='6'/>Dunc a primes est il fluriz,
<lb n='7'/>E quant loëz est de plusurs,
<lb n='8'/>Dunc ad espandues ses flurs.
<lb n='9'/>Custume fu as ancïens,
<lb n='10'/>Ceo testimoine Precïens,
<lb n='11'/>Es livres ke jadis feseient,
<lb n='12'/>Assez oscurement diseient
<lb n='13'/>Pur ceus ki a venir esteient
<lb n='14'/>E ki aprendre les deveient,
<lb n='15'/>K'i peüssent gloser la lettre
<lb n='16'/>E de lur sen le surplus mettre.
<lb n='17'/>Li philesophe le saveient,
<lb n='18'/>Par eus meïsmes entendeient,
<lb n='19'/>Cum plus trespassereit li tens,
…
</div>
</body>
</text>
</tei.2>
```

3.4 Liste d'autorité pour les valeurs de l'attribut rend des éléments suivants : <corr>, <hi>, <gap> et <foreign>[7]

Valeur	Description	Exemple
ital	italiques	*creator*
gras	gras	**creator**
maj	majuscules	ENEAS
pmaj	petites majuscules	RENAUD DE BEAUJEU
exp	exposant	XX^{C}
ind	indice	XX_{C}
crochets[8]	cas où il y a plusieurs mots à l'intérieur des crochets	[et nen estoit leus de deffendre. Tote ert la vile mise en cendre].
susp	3 points de suspension	…
points	plus de 3 points de suspension	……
crochets-ital	italiques entre crochets	[*69v*]
crochets-pmaj	Petites majuscules entre crochets	[LA PREDICATION]
crochets-susp	3 points de suspension entre crochets	[…]

NOTA BENE

En cas de correction de l'éditeur scientifique qui rajoute soit un mot, soit un ou plusieurs caractère(s) à l'intérieur d'un mot, et qui le note dans l'édition au moyen de crochets, on utilise l'élément <corr> dont l'attribut rend restitue la chaîne de caractères telle qu'elle se présente dans l'édition.

Exemple :

```
<corr resp="editor" sic="travalle" rend="trava[i]lle">travaille</corr>
```

Pour des raisons techniques il serait difficile d'utiliser la valeur habituelle "crochets" dans ce cas. Si nous le faisions, nous serions soit conduits à éclater le mot en plusieurs unités, ce qui pourrait perturber l'exploitation automatique ultérieure de la base, soit incapables de repérer la position des crochets situés à l'intérieur d'un mot.
Par extension et pour faciliter le balisage automatique des textes de notre base, ce principe de notation a été adopté dans les cas où les crochets enserrent un seul mot.

[7] A distinguer de la liste d'autorité des valeurs de l'attribut rend de l'élément <p>.
[8] voir le NOTA BENE.

3.5 Liste d'autorité pour les valeurs de l'attribut rend de l'élément <p> pour les textes en vers[9]

Valeur	Description
strophe	strophe
laisse	laisse
couplet	couplet
gv	groupe de vers sans dénomination particulière

De façon générale, la valeur sera déterminée à partir du terme qu'utilise l'éditeur scientifique. En cas d'absence, on emploiera la valeur "gv".

[9] Pour les textes en prose, l'attribut rend n'est pas utilisé avec l'élément <p>.