

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

Serge Heiden (*slh@ens-lyon.fr*), Céline Guillot (*Celine.Guillot@ens-lyon.fr*), Alexei Lavrentiev (*Alexei.Lavrentev@ens-lyon.fr*) et Lauranne Bertrand (*laurannebp@gmail.com*)

PROJET BFM, UMR5191 ICAR, CNRS/ENS DE LYON

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. Ce document est publié librement sur le web à destination de la communauté scientifique dans le cadre de la licence Creative Commons « Paternité-Pas d'Utilisation Commerciale-Partage des Conditions Initiales à l'Identique 2.0 France ». En accord avec cette licence, si vous utilisez ce document dans vos travaux, vous êtes prié de mentionner sa référence (projet BFM, titre, auteurs).



(VERSION 4.0 - AOUT 2010)

Table des matières

1	Introduction	3
2	Généralités	3
2.1	Principes de base	3
2.1.1	Définition de l'unité textuelle	3
2.1.2	Principes de balisage	4
2.2	Rôle des différents acteurs participant à l'édition du texte informatisé	4
3	Description formelle du balisage des textes de la BFM	5
3.1	Délimitation du corps du texte et des éléments qui lui sont externes (prologue...)	5
3.2	Structure du corps du texte	6
3.2.1	Délimitation des parties du texte : livres, chapitres, sections, sous-sections...	6
3.2.1.1	Notation des titres de livre, chapitre...	7
3.2.1.2	Indication de date dans des documents juridiques ou historiques	7
3.2.2	Numérotation des pages	8
3.2.3	Délimitation des unités inférieures	8
3.2.3.1	Textes en prose	8
3.2.3.2	Textes en vers	9
3.2.3.3	Théâtre	9
3.2.4	Autres références	10
3.3	Indications dans le corps du texte	10
3.3.1	Délimitation explicite des lexies	10
3.3.1.1	Nombres	11
3.3.1.2	Abréviations	11
3.3.1.3	Mots composés	12
3.3.2	Corrections et interventions éditoriales	12
3.3.2.1	Corrections de l'éditeur scientifique ou du relecteur	12

3.3.2.2	Lacunes du manuscrit indiquées par l'éditeur scientifique et non comblées	14
3.3.2.3	Passages difficiles à lire dans le manuscrit	14
3.3.3	Autres types de passages mis en évidence	15
3.3.3.1	Passages en langue étrangère	15
3.3.3.2	Mises en évidence typographiques dont la signification n'est pas claire	15
3.3.4	Notes et commentaires du relecteur/encodeur	15
3.3.5	Passage au discours direct ou indirect	16
3.3.6	Citations	16
4	Annotations linguistiques	17
4.1	Annotations ponctuelles	17
4.2	Tokenisation (balisage des lexies et des phrases)	17
4.2.1	Lexies	18
4.2.2	Phrases	19
4.3	Etiquetage morphosyntaxique	19
5	Annexes	19
5.1	Un exemple de prose :	19
5.2	Un exemple de théâtre :	19
5.3	Un exemple de vers :	19
5.4	Liste d'autorité pour les valeurs de l'attribut rend des éléments suivants : corr, supplied, hi, gap et foreign A distinguer de la liste d'autorité des valeurs de l'attribut rend de l'élément p.	20
5.5	Liste d'autorité pour les valeurs de l'attribut type de l'élément ab pour les textes en vers Pour les textes en prose, l'attribut rend n'est pas utilisé avec l'élément p.	20
	Index	21

1 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/encodeur¹ 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'*explicit*er 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/Guidelines/P5/>. De manière générale, il faudra consulter les TEI Guidelines comme complément aux imprécisions de ce manuel.

2 Généralités

2.1 Principes de base

2.1.1 Définition de l'unité textuelle

Les textes de la Base de Français Médiéval sont une représentation d'**éditions de référence** qui peuvent être corrigées en cas d'erreurs manifestes par nos relecteurs, spécialistes de français médiéval. 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 volume physique rassemble des œuvres d'auteurs différents et/ou écrites à des dates distinctes, chacune de ces œuvres sera représentée dans un fichier séparé².

Dans son état actuel la représentation des textes dans la BFM se limite au texte proprement dit. Soit 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.

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

²Dans 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.

2 GÉNÉRALITÉS

2.1.2 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 : `<div> ... </div>`) ;
 - 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">
  <p>contenu du premier chapitre</p>
</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).

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

Le document numérique étant constitué à partir d'une édition de référence imprimée ou numérique, 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, plusieurs étapes se succèdent :

- 1) la numérisation du texte de l'édition de référence ;

3 DESCRIPTION FORMELLE DU BALISAGE DES TEXTES DE LA BFM

- 2) le « nettoyage » et un pré-balisage du texte numérisé au format Doc Word (Encodage1)
- 3) la vérification que cette version est conforme au texte de l'édition de référence, l'analyse de l'usage des marques typographiques dans l'édition, le repérage de coquilles éventuelles dans l'édition et la vérification du pré-balisage (Relecture1) ;
- 4) la conversion du Doc Word au format XML-TEI avec un balisage automatique (Encodage2) ;
- 5) la vérification et l'enrichissement du balisage et une nouvelle vérification du contenu textuel (Relecture2),
- 6) la vérification formelle de la structure du document XML et la validation finale (Encodage3).

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` qui renvoie à l'identifiant de la personne ou de son rôle dans la chaîne de la numérisation. La description des différentes responsabilités dans la chaîne de la numérisation et les identifiants des intervenants se trouvent dans l'en-tête TEI (voir le *Manuel de description des textes pour la BFM* pour plus de détails).

Exemple :

```
<note resp="#proofreader2">Erreur probable dans l'édition</note>
```

L'attribut `@resp` de cette note renvoie à l'identifiant `@xml:id="proofreader2"` d'un élément `<resp>` dans l'en-tête TEI.

3 Description formelle du balisage des textes de la BFM

3.1 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'œuvre

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>` ou un élément « block anonyme » `<ab>` si le texte n'est pas organisé en paragraphes ;
- 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. 3.2.1 et 3.2.3) ;
- même chose pour l'élément `<back>`
 - dans le cas où le `<front>` encadre uniquement le titre, la balise `<div>` a un attribut `@type="titre"` et la balise `<p>` n'a pas d'attribut ;
 - dans le cas du balisage d'un *explicit*, la balise `<div>` a un attribut `@type="explicit"` et la balise `<p>` n'a pas d'attribut.

3 DESCRIPTION FORMELLE DU BALISAGE DES TEXTES DE LA BFM

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

Exemple :

```
<text>
  <front>
    <div type="titre">
      <p>titre</p>
    </div>
  </front>
  <body>
    <ab type="groupe de vers">contenu du texte</ab>
  </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 (en dehors de l'en-tête).

Exemple :

```
<text>
  <body>
    <ab>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.</ab>
  </body>
</text>
```

3.2 Structure du corps du texte

3.2.1 Délimitation des parties du texte : livres, chapitres, sections, sous-sections...

- l'élément <div> marque n'importe quelle division du texte ;
- un élément <div> contient au moins un élément d'un niveau de structuration inférieur : soit un <div> de niveau inférieur, soit un <p> (cf. 3.2.3 pour plus de détails) ;
- l'élément <div> 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 ;

3 DESCRIPTION FORMELLE DU BALISAGE DES TEXTES DE LA BFM

Exemple :

```
<div>
  <p>... Et pour vous informer du temps dont ay eu congnoissance
    dudit seigneur, dont faictes demande, m'est force de
    commencer 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>
```

3.2.1.1 Notation des titres de livre, chapitre...

Chaque élément `<div>` peut s'ouvrir avec un premier élément `<head>` contenant le titre ou l'en-tête de la division et se clôturer 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>
  <p>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...</p>
</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>`.

3.2.1.2 Indication de date dans des documents juridiques ou historiques

Certains documents juridiques peuvent contenir une indication de la date et du lieu de la réalisation du document. Ces indications sont balisées à l'aide de l'élément `<dateline>` qui peut être placé à l'intérieur d'une `<div>`, avant ou après `<head>` ou à la fin.

Exemple :

```
<div type="document" n="I">
  <dateline>1418, 4 septembre, Mont Saint-Michel</dateline>
  <head xml:lang="fr">Vidimus par Laurent le Grant, sénéchal du
Mont-Saint-Michel, [...]</head>
  <p>À tous ceulx qui ces lettres verront Laurens le Grant [...]</p>
</div>
```

3.2.2 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 ;
 - le texte de la page, dont le numéro est indiqué par un attribut de la balise `<pb/>` se trouve après cette balise. Par conséquent, un premier « saut de page » se trouve avant le début du texte ;
 - 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) ;
 - si une césure de mot se trouve en fin de page, il faut supprimer la césure et placer la balise `<pb/>` après la fin du mot.

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 et si cette page est numérotée, il est nécessaire d'insérer un saut de page à l'endroit où elles se trouvent dans le texte (même si elles ne sont pas conservées dans la version numérique du texte).

3.2.3 Délimitation des unités inférieures

3.2.3.1 Textes en prose

- 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 ;
- s'il reste une trace de césure de fin de ligne au milieu d'un mot, il faut la supprimer et éventuellement placer la balise `<lb/>` après la fin du mot, au début de la ligne suivante.

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

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 laienc 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>
```

3.2.3.2 Textes en vers

- l'élément `<ab>` marque les groupes de vers : laisses, strophes, refrains...
 - il a obligatoirement un attribut `@type` auquel on peut donner plusieurs valeurs : "laisse", "strophe"...⁴ Si la qualification des groupes de vers d'un ouvrage est difficile, l'attribut `@type` 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 ; un premier « saut de vers » est placé donc avant le premier vers ;
 - il a un attribut `@n` qui permet d'encoder éventuellement le numéro du vers (cf. élément `<pb/>`). Cet attribut est normalement ajouté au stade de la vérification finale à l'aide d'un outil de numérotation automatique.

Exemple :

```
<text>  
<body>  
  <ab type="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.  
  </ab>  
</body>  
</text>
```

3.2.3.3 Théâtre

- l'élément `<sp>` marque les prises de parole :
 - son attribut `@who` permet de renseigner le nom du locuteur sous une forme normalisée ;
 - cet élément contient éventuellement un élément `<speaker>` (cf. ci-dessous) et comporte obligatoirement un ou plusieurs éléments `<p>` ou `<ab>` selon que le texte est en prose ou en vers.

⁴ Voir l'Annexe 5.5 pour la « liste d'autorité » des valeurs de cet attribut.

⁵ Nous avons choisi de ne pas utiliser dans ce cas l'élément `<l>` 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.

3 DESCRIPTION FORMELLE DU BALISAGE DES TEXTES DE LA BFM

- l'élément `<speaker>` marque l'indication du locuteur dans le texte source.
- l'élément `<stage>` marque les didascalies.

Exemple :

```
<sp>
  <speaker>Li desciples</speaker>
  <stage>dit a sun maistre :</stage>
  <p>Maistre, jeo te pri que tu me dies ceo que jeo te
    demanderai a l'honor de Deu e al pruffit de sainte
    iglise.</p>
</sp>
```

3.2.4 Autres références

Outre les numéros de pages et de vers, l'édition peut fournir d'autres marques de référence, comme par exemple les numéros de folios dans le manuscrit de base

- l'élément `<milestone/>` pourra être utilisé pour représenter ces informations.
 - son attribut `@unit` (obligatoire) sert de préciser l'unité de référence (par exemple, "folio" ou "column" pour indiquer la foliotation du manuscrit de base) ;
 - son attribut `@n` portera le numéro.

Exemple :

```
<text>
  <body>
    <p>... par la simplece <milestone unit="column" n="160c"/> que
      il i voit i espoire il tant de bien qu'il li
    <pb ed="Pauphilet1923" n="3"/>
      plest mout qu'il le face chevalier.</p>
  </body>
</text>
```

3.3 Indications dans le corps du texte

3.3.1 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.

3 DESCRIPTION FORMELLE DU BALISAGE DES TEXTES DE LA BFM

NOTA BENE : Dans le cadre des opérations d'enrichissement linguistique, la délimitation explicite exhaustive des lexies et des phrases (ou la *tokenisation*) est effectuée sur l'ensemble des textes (cf. section 4.2 ci-dessous). Les balises présentées dans la présente section servent à préparer les textes à la tokenisation automatique.

3.3.1.1 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 après l'incarnation Nostre  
Sengnor Jesu Crist, ...</p>
```

Si le nombre n'est pas entouré de points, le balisage n'est pas obligatoire.

3.3.1.2 Abréviations

- l'élément `<abbr>` marque les abréviations suivies d'un point ;
- l'élément `<expan>` marque les résolutions des abréviations (mots entiers) ;
- l'élément `<ex>` marque les caractères ajoutés à la place d'une abréviation.

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*.

Exemple :

```
<expan>Par</expan> foi, dient li baron, li mieudres <ex>con</ex>siaus  
q<ex>ue</ex> nous vous sachiens donner, ce est q<ex>ue</ex>  
vous la laissiez aler,
```

Glose : Ici, les résolutions des abréviations sont marquées explicitement à l'aide des balises `<expan>` (mot entier *Par*) et `<ex>` (quelques caractères dans *consiaus* et dans *que*).

3.3.1.3 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.

3.3.2 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.

NOTA BENE : La technique de balisage des corrections a changé considérablement dans la version P5 de la TEI (lancée officiellement en novembre 2007). Ce manuel d'encodage tient compte de ces modifications.

3.3.2.1 Corrections de l'éditeur scientifique ou du relecteur

Deux types de corrections éditoriales sont à baliser

- 1) corrections de l'éditeur scientifique par rapport à son manuscrit de base **signalées par des marques typographiques dans le corps du texte**. Le plus souvent, il s'agit de lettres ou de mots placés entre crochets ;
- 2) corrections du relecteur en cas d'erreur manifeste dans l'édition de référence (coquille typographique ou erreur flagrante de lecture du manuscrit).

Dans tous les cas le balisage doit permettre à l'utilisateur d'accéder au texte corrigé aussi bien qu'au texte initial.

Le plus souvent, les corrections visibles dans le corps du texte d'une édition se limitent à l'ajout de quelques lettres ou de quelques mots qui manquent dans le manuscrit de base. Une seule balise suffit dans ces cas pour indiquer la correction :

- l'élément `<supplied>` : sert à baliser les lettres ou les mots insérés par l'éditeur à la place de lacunes ou de coquilles de son manuscrit de base ;
 - son attribut `@rend` permet d'indiquer la marque typographique utilisée dans l'édition pour mettre la correction en évidence. Par défaut, on considère que la marque typographique utilisée dans ce cas est une paire de crochets ; ce même attribut permet d'indiquer l'étendu de la correction (cf. NotaBene ci-dessous).

Dans des cas plus complexes (et notamment en cas de correction proposée par un relecteur), il convient d'utiliser un schéma plus conséquent :

- l'élément `<choice>` sert à baliser la zone de texte concernée par la correction, il permet de faire un choix de lecture « avant » ou « après » la correction. Il contient obligatoirement les deux éléments suivants :
- l'élément `<corr>` contient le texte corrigé ;
 - son attribut `@resp` renvoie à l'identifiant du responsable de la correction selon l'en-tête TEI, c'est-à-dire l'éditeur scientifique ("`#editor`") ou le relecteur ("`#proofreader1`" ou "`#proofreader2`") ;

3 DESCRIPTION FORMELLE DU BALISAGE DES TEXTES DE LA BFM

- son attribut @rend permet d'indiquer la marque typographique utilisée dans l'édition pour mettre la correction en évidence ;
- son attribut @cert permet d'indiquer éventuellement la certitude de la correction.
- l'élément <sic> contient le texte avant la correction, apparemment erroné dans le manuscrit de base (correction de l'éditeur scientifique) ou dans l'édition (correction du relecteur) ;
 - si la correction consiste en un ajout d'un ou de plusieurs mots, cet élément sera vide : <sic/>
 - le relecteur peut utiliser l'élément <note> pour ajouter un commentaire (cf. 3.3.4 ci-dessous).

NOTA BENE : Pour faciliter le traitement ultérieur des textes (et en particulier la segmentation lexicale et l'étiquetage morphosyntaxique), nous avons distingué les corrections qui touchent une partie d'un mot de celles qui concernent un mot entier, voire plusieurs mots. Dans tous les cas, les éléments <corr> et <sic> contiennent au moins un mot entier. Si la correction consiste en un ajout d'un ou de quelques caractères omis, ceux-ci doivent être balisés à l'aide de <supplied> muni d'un attribut @rend="word_part" (si cet attribut est absent, on considère qu'au moins un mot entier est ajouté).

Exemple :

```
<lb n="2"/>Ki son sens aüse et  
trava<supplied rend="word_part">i</supplied>lle
```

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.

Exemple de balisage d'une correction concernant plusieurs mots :

```
<lb n="6068"/>Quant li palefrois biaus et gens  
<lb n="6069"/>  
<supplied resp="#editor" rend="ital">Fu venus la pucele i  
monte.</supplied>  
<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 <supplied> dont 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 5.4.

Exemple d'une correction proposée par un relecteur ou un encodeur :

```
<choice>  
  <corr resp="#encoder2">Ço</corr>  
  <sic>Co</sic>  
</choice> 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 ce manque d'homogénéité, a décidé de proposer une correction.

3.3.2.2 Lacunes du manuscrit indiquées par l'éditeur scientifique et non comblées

- 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 5.4 ;
 - son attribut `@reason` 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.

Exemple :

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

Exemple :

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

3.3.2.3 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 5.4 ;
 - 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.
```

3.3.3 Autres types de passages mis en évidence

3.3.3.1 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 `@xml: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 5.4.

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 xml:lang="lat"> Veni creator spiritus
</foreign>.
Et trestout et grant et petit plorerent de pec et de le grant
goie qu'i eurent ...
```

3.3.3.2 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>`
 - la marque typographique de mots en majuscules est encodée au moyen de l'élément `<hi>`
 - la marque typographique de mots en petites majuscules est encodée au moyen de l'élément `<hi>`
 - la marque typographique de mots en exposant est encodée au moyen de l'élément `<hi>`
 - la marque typographique de mots en indice est encodée au moyen de l'élément `<hi>` . . .⁶ .

Exemple :

```
jjc -> jj<hi rend="exp">c</hi>
```

3.3.4 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.

⁶ Cf. Annexe 5.4.

3 DESCRIPTION FORMELLE DU BALISAGE DES TEXTES DE LA BFM

- 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.

3.3.5 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 paroles prononcées ou de pensées ;
 - 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.

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, fait Galaad, je m'en entrasse en  
cele a senestre, car si com je pens je m'en getasse mielz que  
vos.</q>
```

3.3.6 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 sutilment, 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 mielres 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 mielres chevaliers dou monde... »
...
```

4 Annotations linguistiques

4.1 Annotations ponctuelles

- 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.
- l'élément `<c>` peut être utilisé exceptionnellement pour annoter les caractères particuliers qui ne correspondent à aucun caractère standard de la table ISO-8859-1 (ASCII).
 - son attribut `@function` permet de préciser éventuellement la fonction justifiant l'usage de ce caractère particulier et l'attribut `@rend` permet de donner les indications sur la forme de ce caractère (par exemple, son nom conventionnel).

Dans la base non étiquetée cet élément est utilisé pour mémoriser les annotations antérieures à l'intégration d'un texte à la base (en cas d'échanges) et pour baliser explicitement certaines lexies dont la reconnaissance automatique poserait un problème.

L'exemple ci-dessous présente l'usage de l'attribut `@type` pour conserver l'identifiant d'un lemme dans un texte du corpus reçu dans le cadre d'un échange.

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
```

4.2 Tokenisation (balisage des lexies et des phrases)

La tokenisation des textes de la BFM a pour but d'assurer le bon fonctionnement des outils d'étiquetage morphosyntaxique. Elle consiste en un balisage explicite des lexies et des phrases, qui permet par ailleurs d'augmenter la fiabilité des résultats des requêtes et des analyses textométriques.

4 ANNOTATIONS LINGUISTIQUES

La tokenisation des textes est effectuée grâce à une procédure automatique basée sur l'analyse des caractères séparateurs et des balises XML présentes dans le texte.

4.2.1 Lexies

Toutes les occurrences-mots sont balisées au moyen de l'élément `<w>`. De façon générale, les limites de mots sont définies par la présence d'un caractère séparateur ou de l'une des balises qui sont, par définition, externes à un mot.

Caractères séparateurs :

```
[espace] " . ! ? , ; " " " « » ( ) [ ]
```

La présence de l'un de ces caractères provoque la fermeture de la balise `<w>`. Les entités xml/html (`´`; etc.) sont autorisées et ne sont pas résolues lors de la tokenisation.

Quelques caractères d'un mot peuvent être placés entre parenthèses ou crochets. Normalement, cela signale une intervention éditoriale, et le balisage correspondant (`<corr>`, cf. section 3.3.2 ci-dessus) doit être utilisé. En cas de tokenisation de textes où les interventions éditoriales n'ont pas été balisées, les parenthèses ou crochets sont autorisés à l'intérieur d'un mot.

L'apostrophe est placée à l'intérieur de la balise `<w>`.

Exemple :

```
<w>d' </w>
<w>iceol </w>
```

NOTA BENE : Certaines éditions de textes très anciens (*Serments de Strasbourg, Sainte Eulalie, Vie de Saint Léger, Passion de Clermont...*) utilisent le « point haut » pour marquer l'enclise (comme dans les textes provençaux) ou d'autres phénomènes phonétiques combinatoires (*no-l, a-ddextris*).

S'il s'agit d'une simple enclise d'un mot grammatical à un autre mot grammatical, ces segments sont considérés comme des formes uniques (*no-s*). S'il s'agit d'autres phénomènes, où l'autonomie des morphèmes est préservée, un balisage séparé est pratiqué, le point haut étant incorporé à la forme qui le suit : (*me-ttrestoz*).

Les marques de ponctuation sont balisées .

Exemple :

```
<w type="pon"> , </w>
```

Les abréviations et les chiffres romains entourés de points sont balisés et respectivement. Les points sont placés à l'intérieur de ces balises, s'ils ne jouent pas le rôle de ponctuation de phrase.

Exemple :

```
<w type="num"> .LX. </w>
<w type="abbr"> s. </w>
```

NOTA BENE : Les balises `<num>` et `<abbr>` remplacent dans un texte tokenisé les balises `<num>` et `<abbr>` respectivement (cf. section 3.3.1 ci-dessus).

4.2.2 Phrases

Les phrases « orthographiques » sont balisées à l'aide de l'élément `<s>`. Les frontières des phrases sont déterminées par la présence d'une ponctuation forte ou par une balise de fin d'unité structurale de niveau supérieur à une phrase (, ,) ou d'une frontière (initiale ou finale) d'un discours direct (ou).

Caractères de ponctuation forte :

```
.!?
```

4.3 Etiquetage morphosyntaxique

L'attribut `@type` de l'élément `<w>` est utilisé actuellement pour porter l'étiquette morphosyntaxique. Si plusieurs jeux d'étiquettes sont utilisés simultanément, les valeurs sont séparées par un point-virgule.

L'attribut `@lemma` peut être utilisé pour fournir le lemme de la forme étiquetée :

Exemple :

```
<w part="N" type="nil;nco;nc" lemma="escrire">escrit</w>
```

5 Annexes

5.1 Un exemple de prose :

[à mettre à jour]

5.2 Un exemple de théâtre :

[à mettre à jour]

5.3 Un exemple de vers :

[à mettre à jour]

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

Valeur	Description	Exemple
"chevrons"		<ce>
"chevrons-susp"	3 points de suspension entre chevrons	<...>
"crochets" ⁸	plusieurs mots à l'intérieur des crochets	[Tote ert la vile mise en cendre].
"crochets-ital"	italiques entre crochets	[69v]
"crochets-pmaj"	Petites majuscules entre crochets	[LA PRÉDICTION]
"crochets-susp"	3 points de suspension entre crochets	[...]
"exp"	exposant	XX ^C
"gras"	gras	creator
"ind"	indice	XX _C
"ital"	italiques	<i>creator</i>
"maj"	majuscules	ENEAS
"parentheses"	parenthèses	culpa(l)bles
"pmaj"	petites majuscules	RENAUD DE BEAUJEU
"points"	plus de 3 points de suspension
"susp"	3 points de suspension	...

5.5 Liste d'autorité pour les valeurs de l'attribut @type de l'élément <ab> pour les textes en vers⁹

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".

⁷ A distinguer de la liste d'autorité des valeurs de l'attribut rend de l'élément <p>.

⁸ voir le NOTA BENE.

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

Index

Éléments

- ab, 5, 9, 20
- abbr, 11, 18
- back, 5
- body, 5, 6
- c, 17
- choice, 12
- corr, 12, 13, 18, 20
- dateline, 7
- div, 5–7
- ex, 11
- expan, 11
- foreign, 15, 20
- front, 5
- gap, 14, 20
- head, 5, 7
- hi, 15, 20
- l, 9
- lb, 8, 9
- lg, 9
- milestone, 10
- note, 13, 15
- num, 11, 18
- p, 5, 6, 8, 9, 20
- pb, 8, 9
- q, 16
- quote, 16
- resp, 5
- s, 19
- sic, 13
- sp, 9
- speaker, 9, 10
- stage, 10
- supplied, 12, 13, 20
- text, 5, 6
- trailer, 7
- unclear, 14
- w, 17–19

Attributs

- cert, 13
- direct, 16
- ed, 8
- extent, 14
- function, 17
- lemma, 17, 19
- n, 4, 6, 8–10
- reason, 14
- rend, 12–15, 17, 20
- resp, 5, 12, 14, 16
- type, 5, 6, 9, 16, 17, 19, 20

- unit, 10
- who, 9, 16
- xml:id, 5
- xml:lang, 15

Valeurs

- #editor, 12
- #proofreader1, 12
- #proofreader2, 12
- 0,5 line, 14
- 1 line, 14
- ambiguous, 14
- chevrons, 20
- chevrons-susp, 20
- column, 10
- couplet, 20
- crochets, 20
- crochets-ital, 20
- crochets-pmaj, 20
- crochets-susp, 20
- editor, 14
- encoder, 16
- exp, 20
- explicit, 5
- folio, 10
- gras, 20
- gv, 9, 20
- illegible, 14
- ind, 20
- ital, 20
- laisse, 9, 20
- maj, 20
- manuscrit endommagé, 14
- n, 16
- parentheses, 20
- pmaj, 20
- points, 20
- proofreader, 16
- proofreader2, 5
- strophe, 9, 20
- susp, 20
- titre, 5
- unspecified, 16
- vers omis, 14
- word_part, 13
- y, 16