Propositions pour l'édition des sources primaires CCFM*

Nicolas Mazziotta, Université de Liège
nicolas.mazziotta@ulg.ac.be

5-6 octobre 2006, ENS Lyon

Nous reprenons ici la position que nous défendons concernant la transcription des sources primaires. Nous avons pu en appliquer la philosophie dans la thèse que nous rédigeons sous la direction de Marie-Guy Boutier – qui consiste en l'étude des relations entre ponctuation et syntaxe dans un corpus de chartes originales rédigées en français à Liège entre 1236 et 1291 –, ainsi qu'au cours de la construction des éditions du projet *Khartês* [1]. Nous avons veillé à rendre la proposition conforme à XML.

Ce document est encore à l'état de brouillon et ne se suffit pas encore à lui-même. La lecture préalable de notre proposition de recommandation concernant l'annotation syntaxique est préférable (Mazziotta 2006).

Enfin, les conventions ci-dessous sont prévues pour permettre l'encodage de manuscrits isolés. La numérisation d'éditions effectuées sur base de plusieurs manuscrits pose des problèmes différents et spécifiques qui compliquent fortement le problème.

Nous avons structuré ce brouillon en deux parties : la première résume brièvement la philosophie suivie pour élaborer notre proposition (1) ; la seconde est consacrée à l'exposé technique des règles d'encodage que nous préconisons (2).

1 Problématique et principes généraux

Abordons tout d'abord la manière de concevoir une édition électronique (1.1) avant d'exposer les principes d'encodage qui nous paraissent convenir (1.2) .

1.1 Multiplicité des regards, multiplicité des éditions

Dans son entreprise, l'éditeur fait en sorte que son travail soit profitable à une catégorie spécifique de chercheurs. Nous commencerons par présenter brièvement cette responsabilité (1.1.1).

L'évolution technologique permet à présent à l'éditeur de donner au lecteur le choix parmi plusieurs vues d'un même texte. D'un point de vue technique, les manières de procéder ne sont pas toutes équivalentes, et nous avons dû nous positionner clairement (1.1.2).

1.1.1 Responsabilités de l'éditeur : description et élaboration

Le rôle primordial d'une édition est de « donner un texte à lire » à un public spécifique. L'éditeur scientifique doit faire des choix, parmi lesquels on distingue les opérations de *réduction* et les opérations d'*élaboration*. La réduction consiste en la sélection des aspects matériels du document dont l'éditeur veut rendre compte : il s'agit d'éliminer ce dont il ne compte pas faire la *description* (par exemple, pour les éditeurs de textes médiévaux, il est courant de réduire l'opposition originale entre les lettres accentuées et celles qui ne le sont pas). L'élaboration est par contre un *enrichissement* : le texte est accompagné d'indications supplémentaires qui visent à le rendre plus accessible (par exemple, l'éditeur peut résoudre les abréviations ou donner au texte une ponctuation moderne).

* La composition de la présente contribution a été entièrement réalisée à l'aide de logiciels libres.

1. Édition et étude de chartes rédigées en français en Wallonie avant 1300 ; voir la présentation synthétique dans Mazziotta (2004).

Quel que soit le choix posé, aucun éditeur n'échappe à ces décisions. Cependant, il se peut qu'il ait, pour une raison ou une autre, envie de prévoir la lecture de son édition par plusieurs types de chercheurs. Dans ces conditions, l'éditeur doit rendre possible l'accès à un nombre choisi de « vues » d'un même texte. Par exemple, pour ne prendre que les exemples les plus classiques, il est possible de faire une édition diplomatique et une édition critique d'un même texte. Si le lecteur est intéressé par la ponctuation ou la répartition des signes diacritiques, la première édition lui sera plus utile qu'elle ne le sera pour un syntacticien, qui préfèrera employer une édition plus accessible au lecteur moderne. L'édition diplomatique réduisant moins, elle met l'accent sur la description, alors que l'édition critique traditionnelle est orientée vers l'élaboration.

C'est au cœur du débat si présent dans la *new philology*, sur le conflit entre la lisibilité et la fidélité de l'édition, que nous nous trouvons ici.

1.1.2 Choix possibles et position adoptée

XML a rendu possible l'intégration de multiples informations au texte de l'édition, pour permettre ensuite au lecteur de choisir celles auxquelles il veut avoir accès. La technique permet ainsi l'élaboration de différentes « vues » d'un même document.

Lors de l'encodage, il est possible de séparer d'emblée ces vues ou de les fusionner. Voyons en quoi consistent les deux options avant de nous positionner.

a. Séparation des vues. L'éditeur construit plusieurs éditions parallèles. C'est l'option actuellement choisie par le projet *BFM-Manuscrits* (Lavrentiev 2005 : §2), sur les traces du projet *Menota* (Menota 1.1 : chap. 3) ; par exemple [2] :

```
<text>
 <front>
 </front>
 <body>
  <div>
   <p>
    <w>
     <lemma>chevalier</lemma>
     <norm>chevaliers</norm>
     <dipl>ch<expan>eualie</expan>rs</dipl>
     <facs>
      <mdv_abbr>ch&apos;&rrot;</mdv_abbr>s
     </facs>
    </w>
    <punct>
     <norm></norm>
     <dipl>;</dipl>
     <facs>&punctelev;</facs>
    </punct>
   </p>
  </div>
 </body>
</text>
```

L'élément `facs` contient une représentation où l'élaboration est faible, mais où l'éditeur s'est concentré sur la description. `norm` représente la position inverse. Quant à `dipl`, il manifeste une attitude intermédiaire.

Cette solution, qui est totalement conforme à XML, est très facile à gérer. On obtient facilement la représentation souhaitée (vue) en employant des feuilles de style XSLT (Clark 1999).

b. Fusion des vues. L'éditeur construit une seule édition, « farcie » d'une multitude d'informations combinant la description et l'élaboration des données. Nous appelons ce type d'édition *édition monolithique*.

C'est cette méthode qu'emploie Martin-Dietrich Gleßgen (2005 : 97) dans son entreprise d'édition des chartes lorraines :

```
<txt>
 <pub>
  <div n=1><maj>C</maj>onue chose soit a-toz</div>
 </pub>
 <exp>
  <div n=2>
   q<abr>ue</abr> li abes <abr>et</abr> li chapitles
   de Salinvas /. at laissié a Wirion <zw/> <abr>et</abr>
   Huillon, les dous freres de Gev<abr>er</abr>lise /.
   <abr>et</abr> a lor oirs [...]
  </div>
 </exp>
</txt>
```

2. D'après (Lavrentiev 2005 : 3).

On voit que l'éditeur a superposé les informations de description et d'élaboration. Par exemple, la ponctuation moderne (élaboration) est insérée parallèlement à la ponctuation ancienne (précédée d'une barre oblique).

C'est aussi cette même méthode que nous suivons dans le cadre du projet *Khartês*. Prenons comme base à l'exposé de nos méthodes le texte de l'adresse [3] de cette charte :

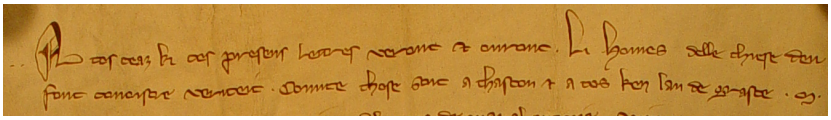

FIG. 1 – Document 1277-05-04 : deux premières lignes

Nous encodons cette adresse comme suit :

```
<l id="_1821"/>
<ponct id="_532">**</ponct>
<w id="_1"><c id="_1823" t="1">A</c></w>
<w id="_2">to$</w>
<w id="_3">cez</w>
<w id="_4">ki</w>
<w id="_5">ce$</w>
<w id="_6">pre$en$</w>
<w id="_7">lettre$</w>
<w id="_8">veront</w>
<w id="_9">[et]</w>
<w id="_10">oront</w>
<mponct id="_589">,</mponct>
<ponct id="_533">*</ponct>
```

Il est visible que la pratique d'encodage monolithique, mêlant description et élaboration ne fait pas toujours appel à XML [4] : ainsi, notre encodage distingue bien deux sortes de ponctuation à l'aide d'XML (éléments `mponct` et `ponct`, respectivement pour la ponctuation moderne et la ponctuation ancienne), mais ne marque pas par du XML que les ‹ $ › devront être remplacés par des ‹ s › lors de la conversion en une « vue » de type *édition critique*.

D'un point de vue technique, la conversion vers une visualisation particulière est plus difficile à gérer que dans le cas d'un encodage séparant les vues. Il nécessite le recours constant à des motifs d'expressions régulières parfois compliqués – ce qui est difficile à implémenter en XSLT [5]. Par ailleurs, la validation (au sens technique) est quasiment impossible quand les données textuelles contiennent des chaînes de caractères qui jouent un rôle similaire au balisage.

c. Position adoptée. La première des deux attitudes, qui consiste à diffracter les éditions est assez proche de la démarche traditionnelle, qui présente, par exemple, une page de gauche en transcription diplomatique, accompagnée, sur la page de droite, d'une édition critique [6]. Les désavantages principaux que ce choix éditorial implique sont essentiellement (Mazziotta 2004 : 796-797) :

- le constant aller-retour entre les différentes éditions en cas de correction ou d'amélioration de celles-ci (complique le travail s'il est évolutif) ;
- la multiplication *ad infinitum* du nombre d'éditions distinctes, en fonction des intérêts des lecteurs potentiels ;
- la diffraction conceptuelle du texte.

L'intérêt principal de la seconde des deux attitudes est de maintenir l'unité conceptuelle du texte. Il nous semble qu'il serait donc intéressant de reprendre les principes de l'édition monolithique, mais en les adaptant le plus possible à XML, de manière à avoir accès aux outils permettant de traiter ce type de documents (XSLT, SAX et XSchema).

3. Mention du destinataire, à la connaissance de qui l'acte écrit porte la teneur de la transaction qu'il consigne

4. L'exemple tiré du travail de Martin-Dietrich Gleßgen est en outre du SGML bien formé et non du XML.

5. XSLT2 gère cependant les expressions régulières, voir Kay 2006 : §15.

6. Voir, par exemple, le recueil Careri *et al.* 2001 : 106.

1.2 Principes d'encodage

L'intégration d'un nombre important de données au texte mène immanquablement à sa surcharge (1.2.1), à laquelle il faut tenter de remédier. Nous proposons de le faire en considérant le caractère comme l'unité de base de la description (1.2.2). Ces conventions générales laisseront ensuite toute latitude aux éditeurs pour décider de la teneur des informations qu'ils souhaitent introduire dans leur corpus (1.2.3).

1.2.1 Danger de surcharge

Dans le cas de la construction de l'édition[7], il paraît obligatoire d'encoder une partie de l'analyse directement dans le texte. Cependant, l'insertion d'annotations sur les caractères mène à une surcharge importante du texte :

```
<w id="_6">
 pre
 <elaboration turn_into="s">&#x17f;</elaboration>
 en
 <elaboration turn_into="s">&#x17f;</elaboration>
</w>
```

Si on veut mélanger des annotations concernant la description et l'élaboration, on est obligé de les imbriquer ; par exemple, pour le premier mot de l'extrait étudié :

```
<w id="_1">
 <description
  size="huge"
  weight="bold"
  decoration="some">
  <elaboration
   uppercase="yes">
   a
  </elaboration>
 </description>
</w>
```

Il faut donc trouver une solution à cette surcharge.

1.2.2 Le caractère comme individu

Voyons comment intégrer au texte les informations voulues en combinant aux caractères unicodes le principe bien connu de *pointeur*.

a. Pointeurs. Le principe est le même que celui exposé dans Mazziotta 2006 (2-3) : chaque caractère devant être décrit ou élaboré peut être délimité et identifié. L'ensemble des informations contenues dans d'autres attributs ou éléments peut donc être reporté hors du texte (*stand-off*). L'encodage du texte en sera substantiellement allégé :

```
<w id="_6">
 pre
 <c id="_c1">&#x17f;</c>
 en
 <c id="_c2">&#x17f;</c>
</w>
```

L'intérêt du pointeur augmente d'autant plus que le nombre d'informations grandit. Ainsi, pour le premier mot de notre extrait :

```
<w id="_1">
 <c id="_c1">a</c>
</w>
```

b. Balises vides. Il peut arriver que l'analyse doive être insérée à un endroit du texte sans pouvoir s'appuyer sur des données caractères. Dans ce cas, nous proposons d'employer une balise vide.

c. Unicode. Par convention, pour réduire l'atomisation, un caractère sera représenté en unicode si c'est possible. Il faut avant tout respecter la *sémantique de caractère* (ne pas choisir les signes parce qu'ils ressemblent à ce qu'on voit, mais parce qu'ils ont la même valeur).

On peut considérer que les caractères mimétiques du manuscrit sont des abréviations synthétisant un ensemble d'informations qui pourraient être exprimées de manière plus analytique. Le signe représenté de la sorte *doit* correspondre à une unité de description ou d'élaboration.

7. Contrairement à la construction d'analyse de second rang, qui s'élabore à partir d'une édition déjà existante, cf. Mazziotta 2006.

1.2.3 Des déclarations modulables

Nous aurons ici le même raisonnement que pour l'annotation syntaxique (Mazziotta 2006 : 3) : pour que la recherche évolue, il est indispensable que chaque chercheur puisse choisir lui-même les données qu'il désire transcrire et l'élaboration qu'il désire y apporter. Chacun doit être autorisé à choisir ses propres limites.

Nous préconisons donc la définition d'un *système* de déclaration des annotations, et non d'un jeu d'étiquettes fixe.

2 Proposition de recommandation

En repartant de l'exemple que nous avons donné de notre manière d'encoder les documents, nous avons tenté de le rendre conforme à XML.

Une fois définie la base d'encodage des caractères (2.1), nous donnons ci-dessous les conventions qui permettent de regrouper systématiquement (2.2) les moyens d'en enrichir la description (2.3) et de décrire leur élaboration éditoriale (2.4). Nous abordons enfin brièvement le délicat problème du traitement des abréviations (2.5).

2.1 Base d'encodage du texte

2.1.1 Définition non formalisée

La transcription est placée dans un élément `div`. Cet élément contient un attribut `type`, qui la met en relation avec une analyse [8]. Cette analyse détermine en outre la base de caractères unicodes de l'édition : soit la base est diplomatique, soit elle est plus moderne. Les deux choix sont possibles, mais sont mutuellement exclusifs : si une édition est diplomatique, elle conserve cette propriété du début jusqu'à la fin.

Tout caractère dont on veut préciser la description ou l'élaboration peut être, conformément à ce que recommande la TEI (TEI P5 0.4.2 : §15.1), délimité par un élément `c` et identifié par un `@id`.

Tout emplacement peut être représenté par un élément `anchor` vide (TEI P5 0.4.2 : §10.3.2) identifié de la même manière.

Quant aux abréviations, on peut soit noter leur forme dans l'élément `abbr`, soit leur résolution, dans l'élément `expan` (TEI P5 0.4.2 : §6.5.5)[9]. Il ne nous paraît pas nécessaire que l'élément `abbr` contienne du texte : il se peut que l'abréviation (par suspension) ne soit marquée par rien du tout. Le traitement des abréviations reste cependant problématique. Nous abordons la question ci-dessous [10].

La combinaison des caractères unicode et du balisage les délimitant et les identifiant sert de *base d'encodage* du texte.

2.1.2 Exemples

Si nous choisissons une base diplomatique, nous avons essentiellement besoin de trois types d'éléments : `c`, `anchor` et `abbr`.

```
<div type="Transcription diplomatique Khartês">
 <lb/>
 <c id="_c0">·</c>
 <c id="_c1">·</c>
 <w id="_1"><c id="_c2">A</c></w>
 <w id="_2">to<c id="_c3">&#x17f;</c></w>
 <!-- 017F LATIN SMALL LETTER LONG S -->
 <w id="_3">cez</w>
 <w id="_4">k<c id="_c4">&#x0131;</c></w>
 <!-- 0131 LATIN SMALL LETTER DOTLESS I -->
 <w id="_5">ce<c id="_c5">&#x17f;</c></w>
 <w id="_6">pre<c id="_c6">&#x17f;</c>en<c id="_c7">&#x17f;</c></w>
 <w id="_7">lettre<c id="_c8">&#x17f;</c></w>
 <w id="_8">veront</w>
 <w id="_9"><abbr id="_c3">&#x204a;</abbr></w>
 <!-- 204A TIRONIAN SIGN ET -->
 <w id="_10">oront</w>
 <anchor id="_a1"></anchor>
 <c id="_c4">·</c>
</div>
```

8. Voir 2.2.

9. Nous ne préconisons pas le regroupement de l'abréviation et de sa résolution dans le texte à l'aide de l'élément `choice` proposé par la TEI (TEI P5 0.4.2 : §6.5.5). Cette technique serait contraire au principe de l'édition monolithique.

10. Voir 2.5

En prenant une base modernisée, nous employons les mêmes éléments, à ceci près que `abbr` est remplacé par `expan` et que son contenu est révisé en conséquence.

```
<div type="Édition critique Khartês">
 <lb/>
 <anchor id="_a0"/>
 <w id="_1"><c id="_c1">A</c></w>
 <w id="_2">to<c id="_c2">s</c></w>
 <w id="_3">cez</w>
 <w id="_4">k<c id="_c3">i</c></w>
 <w id="_5">ce<c id="_c4">s</c></w>
 <w id="_6">pre<c id="_c5">s</c>en<c id="_c6">s</c></w>
 <w id="_7">lettre<c id="_c7">s</c></w>
 <w id="_8">veront</w>
 <w id="_9"><expan id="_c8">et</expan></w>
 <w id="_10">oront</w>
 <c id="_c9">,</c>
 <anchor id="_a1"/>
</div>
```

Comme on le voit, certains caractères ont été délimités et identifiés, de manière à permettre l'application du système de pointeurs. La base d'encodage ne permet pas de faire une description exhaustive, mais est complétée par une analyse supplémentaire.

Nous soulignons l'importance des choix posés implicitement par l'éditeur – qu'il pourrait, si nécessaire, exprimer au travers d'une introduction en langue naturelle ou par le remplissage d'une fiche technique formalisée. Il serait tout à fait envisageable de prendre une base d'encodage diplomatique, mais de ne pas y reporter l'opposition entre ‹ s › et ‹ ſ ›.

2.2 Localisation et format général

À l'instar de ce que nous avons proposé pour l'encodage de l'analyse syntaxique, la description et l'élaboration éditoriale des caractères identifiés par un `@id` sont regroupées à l'extérieur du texte, dans un élément nommé `graphData`.

L'attribut `@charbase` de cet élément détermine si le format de départ choisi est la description (`D`) ou l'élaboration (`E`). On définit ainsi explicitement la valeur des caractères unicodes employés. L'attribut `@type` identifie le type d'édition et permet aux encodages de texte de référer à l'analyse.

L'élément `graphData` contient un élément `graphDescr` et un élément `graphElab`, qui expriment respectivement la description des caractères [11] et leur élaboration [12].

```
<graphData charbase="D" type="Transcription diplomatique Khartês">
 <graphDescr>...</graphDescr>
 <graphElab>...</graphElab>
</graphData>
```

L'attribut `@type` du `div` contenant l'encodage de base fait ainsi référence explicitement à un ensemble d'analyses extérieures, dont nous allons détailler les principes.

2.3 Description des caractères

Voyons tout d'abord comment nous concevons l'encodage des analyses avant de nous concentrer sur la manière de définir les conventions propres à chaque éditeur.

2.3.1 Encodage de l'analyse

L'annotation se fait manière similaire à celle que nous avons proposée pour annoter les structures syntaxiques (Mazziotta 2006 : 7-8).

a. Définition non formalisée. La description des caractères est encodée dans l'élément `graphDescr`. Cet élément contient un ensemble d'éléments `graphDTag` définis par un `@type`, qui correspond à la *nature* de l'information encodée. Ces éléments contiennent à leur tour des éléments nommés `graphDValue`, qui donnent par leur `@type` la *valeur* de l'information encodée. Chacun de ces éléments contient quant à lui un ensemble de `link`, qui pointent vers les caractères concernés par l'analyse.

11. Voir 2.3.
12. Voir 2.4.

b. Application. En prenant une base diplomatique, il est possible d'enrichir la description que la base de codage unicode donne à l'aide d'informations supplémentaires concernant, par exemple, la taille, la graisse ou la décoration des lettres.

```
<graphDescr>
 <graphDTag type="character.size">
  <graphDValue type="huge">
   <link target="_c1"/>
  </graphDValue>
 </graphDTag>
 <graphDTag type="character.weight">
  <graphDValue type="bold">
   <link target="_c1"/>
  </graphDValue>
 </graphDTag>
 <graphDTag type="character.decoration">
  <graphDValue type="some">
   <link target="_c1"/>
  </graphDValue>
 </graphDTag>
</graphDescr>
```

Si l'on part d'une base modernisée, le mécanisme est le même, mais le contenu de la description des caractères en sera souvent augmenté, les oppositions qui étaient marquées par un choix de caractères unicodes spécifiques dans la base diplomatique n'étant plus présentes.

```
<graphDescr>
 <graphDTag type="character.size">
  <graphDValue type="huge">
   <link target="_c1"/>
  </graphDValue>
 </graphDTag>
 <graphDTag type="character.weight">
  <graphDValue type="bold">
   <link target="_c1"/>
  </graphDValue>
 </graphDTag>
 <graphDTag type="character.decoration">
  <graphDValue type="some">
   <link target="_c1"/>
  </graphDValue>
 </graphDTag>
 <graphDTag type="character.shape">
  <graphDValue type="long s">
    <link target="_c1"/>
    <link target="_c4"/>
    <link target="_c5"/>
    <link target="_c6"/>
    <link target="_c7"/>
  </graphDValue>
  <graphDValue type="dotless i">
    <link target="_c3"/>
  </graphDValue>
  <graphDValue type=".">
    <link target="_a0"/>
    <link target="_a0"/>
    <link target="_a1"/>
  </graphDValue>
  <graphDValue type="et">
    <link target="_c8"/>
  </graphDValue>
 </graphDTag>
</graphDescr>
```

À nouveau, nous insistons sur le fait que ces éléments de descriptions sont entièrement facultatifs. L'éditeur pourrait très bien choisir de partir d'une base modernisée et de ne réintroduire l'opposition ‹ s › *vs* ‹ ſ › à aucun endroit. Dans ce cas, il pourrait faire l'économie des éléments `c` qui entourent les caractères concernés dans la base d'encodage.

2.3.2 Déclaration

Pour que les pratiques descriptives puissent être partagées tout en restant souples, il serait intéressant de se baser sur les mêmes principes que ceux qui ont été proposés pour l'analyse syntaxique (Mazziotta 2006 : 9).

a. Définition non formalisée. Le contenu des éléments `graphDescr` est défini explicitement par une déclaration dans un élément `graphDescrDecl`. Cet élément comprend

un ensemble d'éléments `graphDTagDecl` dont le `@type` correspond à celui des éléments `graphDTag` employés dans l'analyse. Chaque élément `graphDTagDecl` contient à son tour les éléments `graphDValueDecl` dont le `@type` correspond à celui des éléments `graphDValue` de l'analyse. Les données des éléments `graphDValueDecl` explicitent (en langue naturelle ou dans n'importe quel formalisme jugé adéquat) la valeur sémantique de l'annotation dont il est question.

b. Application. Appliqué à l'exemple que nous avons choisi, la déclaration aurait éventuellement la forme suivante :

```
<graphDescrDecl>
 <graphDTagDecl type="character.size">
  <graphDValueDecl type="large">
   taille supérieure à la ligne
  </graphDValueDecl>
  <graphDValueDecl type="huge">
   taille atteignant deux lignes ou plus
  </graphDValueDecl>
 </graphDTagDecl>
 <graphDTagDecl type="character.weight">
  <graphDValueDecl type="bold">
   gras
  </graphDValueDecl>
 </graphDTagDecl>
 <graphDTagDecl type="character.decoration">
  <graphDValueDecl type="some">
   peu décoré
  </graphDValueDecl>
  <graphDValueDecl type="lot">
   très décorée
  </graphDValueDecl>
 </graphDTagDecl>
 <graphDTagDecl type="character.shape">
  <graphDValueDecl type="dotless i">
   <!-- LATIN SMALL LETTER DOTLESS I -->
   &#x0131;
  </graphDValueDecl>
  <graphDValueDecl type="long s">
   <!-- LATIN SMALL LETTER LONG S -->
   &#x017F;
  </graphDValueDecl>
  <graphDValueDecl type="et">
   <!-- TIRONIAN SIGN ET -->
   &#x204A;
  </graphDValueDecl>
  <graphDValueDecl type="·">
   <!-- MIDDLE DOT -->
   &#x00B7;
  </graphDValueDecl>
 </graphDTagDecl>
</graphDescrDecl>
```

On voit que nous avons choisi, quand cela était possible, d'identifier la forme des caractères par le code unicode correspondant (description formalisée). Par contre, la définition des valeurs possibles pour décrire la taille des caractères a été faite en français. La rigueur et le formalisme des descriptions est laissé à l'appréciation des utilisateurs.

2.4 Élaboration des caractères

Les règles d'encodage des pratiques d'élaboration sont similaires à celles employées pour la description des caractères. Seuls les noms des balises employées changent. Les exemples qui suivent sont clairs.

2.4.1 Encodage de l'analyse

Si l'on part d'une base d'encodage diplomatique, l'élaboration est explicitée en suivant la même méthode que celle qui a servi à expliciter les informations de description.

```
<graphElab>
 <graphETag type="character.shape">
  <graphEValue type="i">
    <link target="_c4"/>
  </graphEValue>
  <graphEValue type="s">
    <link target="_c3"/>
    <link target="_c5"/>
    <link target="_c6"/>
    <link target="_c7"/>
    <link target="_c8"/>
  </graphEValue>
  <graphEValue type=",">
```

```
    <link target="_a1"/>
   </graphEValue>
  </graphETag>
  <graphETag type="character.case">
   <graphEValue type="uppercase">
    <link target="_c2"/>
   </graphEValue>
  </graphETag>
  <graphETag type="abbreviation">
   <graphEValue type="et">
    <link target="_c3"/>
   </graphEValue>
  </graphETag>
</graphElab>
```

La méthode permet donc d'expliciter tout changement dans la forme d'une lettre, que ce soit pour transformer un ‹ ı › en un ‹ i › ou un ‹ ſ › en ‹ s ›, mais aussi pour changer un ‹ u › en ‹ v › ou le contraire.

Si la base d'encodage est moderne, il se peut que l'éditeur n'ait rien à ajouter, comme c'est le cas ici.

2.4.2 Déclaration

La déclaration accompagnant cette analyse est ainsi :

```
<graphElabDecl>
 <graphETagDecl type="character.shape">
  <graphEValueDecl type="i">
   <!-- LATIN SMALL LETTER I -->
   &#x0069;
  </graphEValueDecl>
  <graphEValueDecl type="s">
   <!-- LATIN SMALL LETTER S -->
   &#x0073;
  </graphEValueDecl>
  <graphEValueDecl type=",">
   <!-- COMMA -->
   &#x002C;
  </graphEValueDecl>
 </graphETagDecl>
 <graphETagDecl type="character.case">
  <graphEValueDecl type="uppercase">
   majuscule
  </graphEValueDecl>
 </graphETagDecl>
 <graphETagDecl type="abbreviation">
  <graphEValueDecl type="et">
   <w>et</w>
  </graphEValueDecl>
 </graphETagDecl>
</graphElabDecl>
```

2.5 Note sur les abréviations

Comme le fait remarquer Alexei Lavrentiev, le traitement des abréviations est loin d'être simple :

> L'encodage des abréviations médiévales pose plusieurs problèmes méthodologiques. Le premier problème concerne les limites de l'abréviation. L'abréviation se limite-t-elle à la marque graphique (point, caractère suscrit ou caractère spécial, ou encore « marque zéro ») qui « remplace » les caractères de la forme pleine du mot ? Ou bien inclut-elle la ou les lettres sur lesquelles (ou à côté desquelles) la marque est placée ? Ou bien faut-il considérer le mot entier comme une abréviation ? En plus, certaines marques d'abréviation sont pratiquement indissociables des lettres auxquelles elles se rattachent (par exemple, le « p barré » ou « p à queue ». (Lavrentiev 2005 : 8)

L'exemple que nous avons donné montre une manière d'encoder la résolution d'une abréviation qui correspond à un mot entier, mais comment faire lorsque l'on prend le parti de dire que seules certaines lettres sont remplacées ? Par exemple, dans la graphie du premier mot ci-dessous, le tilde remplace manifestement la chaîne ‹ ier › :

FIG. 2 – Document 1271-12-03 : 11

Nous proposons d'encoder le texte comme suit (ou bien en utilisant le caractère unicode du tilde) :

```
<w>
 cheual<abbr id="_a1"/>s
</w>
```

et de décrire l'abréviation comme ceci :

```
<graphData>
 <graphDescr>
  <graphDTag type="abbreviation">
   <graphDValue type="titulus">
    <link target="_a1"/>
   </graphDValue>
  </graphDTag>
 </graphDescr>
 <graphElab>
  <graphETag type="abbreviation">
   <graphEValue type="ier">
    <link target="_a1"/>
   </graphEValue>
  </graphETag>
 </graphElab>
</graphData>
```

Les informations concernant la forme et la résolution de l'abréviation sont encodées distinctement, conformément aux principes que nous avons exposés.

Cette technique peut être facilement exploitée dans le cas de graphies comme ‹ excoïement ›, résolue en ‹ exco*mini*ement ›, « excommunication » (Document 1288 : 10), où l'on voit clairement que l'abréviation remplace deux lettres disjointes, localisées de part et d'autre du ‹ ı ›.

```
<w>
 exco<abbr id="_a1"/>&#x69;<abbr id="_a2"/>ement
</w>
```

On peut employer l'élément `linkGrp` pour indiquer que plusieurs caractères correspondent à une seule abréviation :

```
<graphData>
 <graphDescr>
  <graphDTag type="abbreviation">
   <graphDValue type="titulus">
    <linkGrp>
     <link target="_a1"/>
     <link target="_a2"/>
    </linkGrp>
   </graphDValue>
  </graphDTag>
 </graphDescr>
 <graphElab>
  <graphETag type="abbreviation">
   <graphEValue type="m">
    <link target="_a1"/>
   </graphEValue>
  </graphETag>
  <graphETag type="abbreviation">
   <graphEValue type="n">
    <link target="_a2"/>
   </graphEValue>
  </graphETag>
 </graphElab>
</graphData>
```

La question reste néanmoins ouverte et nous ne prétendons pas donner ici une solution universelle.

3 Conclusion

Nous sommes à l'heure où il faut penser la construction des corpus pour qu'ils tirent parti de tout apport conceptuel offert par la technique offre pour respecter le plus possible la nature de notre travail.

Même si ce que nous proposons est difficile à mettre en œuvre, il nous semble essentiel d'en conserver la philosophie : l'édition monolithique est plus évolutive et conceptuellement plus « propre » que la structure éclatée.

Pratiquement, la technique du pointeur et le choix de la base d'encodage devrait permettre d'enrichir des éditions déjà numérisées sans les modifier autrement que par l'ajout de balises identifiant les caractères.

Du reste, il est cependant impossible de se passer d'un format de travail, tant our augmenter l'ergonomie que pour améliorer les performances.

4 Annexe : ensembles d'annotations complets

Nous donnons ci-dessous l'ensemble des annotations correspondant au petit passage donné en exemple.

4.1 Base d'encodage diplomatique

```
<graphData charbase="D" type="Transcription diplomatique Khartês">
 <graphDescr>
  <graphDTag type="character.size">
   <graphDValue type="huge">
    <link target="_c1"/>
   </graphDValue>
  </graphDTag>
  <graphDTag type="character.weight">
   <graphDValue type="bold">
    <link target="_c1"/>
   </graphDValue>
  </graphDTag>
  <graphDTag type="character.decoration">
   <graphDValue type="some">
    <link target="_c1"/>
   </graphDValue>
  </graphDTag>
 </graphDescr>
 <graphElab>
  <graphETag type="character.shape">
   <graphEValue type="i">
     <link target="_c4"/>
   </graphEValue>
   <graphEValue type="s">
     <link target="_c3"/>
     <link target="_c5"/>
     <link target="_c6"/>
     <link target="_c7"/>
     <link target="_c8"/>
   </graphEValue>
   <graphEValue type=",">
     <link target="_a1"/>
   </graphEValue>
  </graphETag>
  <graphETag type="character.case">
   <graphEValue type="uppercase">
     <link target="_c2"/>
   </graphEValue>
  </graphETag>
  <graphETag type="abbreviation">
   <graphEValue type="et">
     <link target="_c3"/>
   </graphEValue>
  </graphETag>
 </graphElab>
</graphData>
```

4.2 Base d'encodage moderne

```
<graphData charbase="E" type="Édition critique Khartês">
 <graphDescr>
  <graphDTag type="character.size">
   <graphDValue type="huge">
    <link target="_c1"/>
   </graphDValue>
```

```
7    </graphDTag>
8    <graphDTag type="character.weight">
9     <graphDValue type="bold">
10     <link target="_c1"/>
11    </graphDValue>
12   </graphDTag>
13   <graphDTag type="character.decoration">
14    <graphDValue type="some">
15     <link target="_c1"/>
16    </graphDValue>
17   </graphDTag>
18   <graphDTag type="character.shape">
19    <graphDValue type="long s">
20      <link target="_c1"/>
21      <link target="_c4"/>
22      <link target="_c5"/>
23      <link target="_c6"/>
24      <link target="_c7"/>
25    </graphDValue>
26    <graphDValue type="dotless i">
27      <link target="_c3"/>
28    </graphDValue>
29    <graphDValue type="·">
30      <link target="_a0"/>
31      <link target="_a0"/>
32      <link target="_a1"/>
33    </graphDValue>
34    <graphDValue type="et">
35      <link target="_c8"/>
36    </graphDValue>
37   </graphDTag>
38  </graphDescr>
39  <graphElab></graphElab>
40 </graphData>
```

Références

Travaux cités

Careri, Maria, Fery-Hue, Françoise, Gasparri, Françoise, Hasenor, Geneviève, Labory, Gillette, Lefèvre, Sylvie, Leurquin, Anne-Françoise et Christine, Ruby (2001). *Album de manuscrits français du XIII^e siècle. Mise en page et mise en texte*, Roma : Viella.

Clark, James, éd. (1999). *XSL Transformations (XSLT). Version 1.0. W3C Recommendation 16 November 1999*, http ://www.w3.org/TR/1999/REC-xslt-19991116.

Gleßgen, Martin-Dietrich (2005). « Editorische, lexikologische und graphematische Erschließung altfranzösischer Urkundtexte mit Hilfe von TUSTEP. Stand der Arbeiten », dans Gärtner, Kurt et Holtus, Günter, éds, *Überlieferungs- und Aneignungsprozesse im 13. und 14. Jahrhundert auf dem Gebiet der werstmitteldeutschen und ostfranzösischen Urkunden und Litteratursprachen*, Trier : Kliomedia (Trierer Historische Forschungen 59) : 91-107.

Haugen, Odd Einar, éd. (2004). *The Menota handbook : Guidelines for the electronic encoding of Medieval Nordic primary sources. Version 1.1*, Bergen : Medieval Nordic Text Archive, http ://www.hit.uib.no/menota/guidelines/.

Kay, Michael, éd. (2006). *XSL Transformations (XSLT). Version 2.0. W3C Candidate Recommendation 8 June 2006*, http ://www.w3.org/TR/2006/CR-xslt20-20060608/.

Lavrentiev, Alexei (2005). *Manuel d'encodage XML-TEI étendu des transcriptions de manuscrits dans le projet BFM-Manuscrits. Normes utilisées à la date du 14 décembre 2005*, http ://bfm.ens-lsh.fr/IMG/pdf/BFM-Mss_Encodage-XML.pdf, au 10 mars 2006.

Mazziotta, Nicolas (2004). « Le texte dans tous ses états. Philosophie d'encodage du projet Khartês », dans Purnelle, Gerald, Fairon, Cédrick et Dister, Anne, éds, *Le poids des mots*, Louvain : Presses universitaires de Louvain : 793-803.

Mazziotta, Nicolas (2006). « Proposition de recommandation pour l'annotation syntaxique des corpus CCFM. Avec logiciels perl », proposition transmise aux membres du CCFM pour les journées des 5 et 6 octobre 2006 à l'ENS de Lyon, non publié.

Menota 1.1 = Haugen 2004.

Sperberg-McQueen, C. M., Burnard, Lou, Bauman, Syd, DeRose, Steven et Rahtz, Sebastian (2005). *TEI P5. Guidelines for Electronic Text Encoding and Interchange. Version 0.4.1*, http ://www.tei-c.org/P4X/.

TEI P5 0.4.2 = Sperberg-McQueen *et al.* 2005.

Documents d'archives

Document (1271-12-03), *3 décembre*, Archives de l'État à Liège (Cathédrale Saint-Lambert à Liège), main K.

Document (1277-05-04), *4 mai*, Archives de l'État à Liège (Collégiale Saint-Martin à Liège), main A.

Document (1288), *reste de la date illisible*, Archives de l'État à Liège (Couvent de Robermont à Robermont), main inconnue.