`msdescription`
`transcr`
`(gaiji)`
Guidelines, chap.10-11 (et 5)

Structuration des données et des documents : balisage XML

La TEI et les manuscrits

J.B. Camps – mercredi 29 novembre 2017

M2 *Technologies numériques appliquées à l'histoire et Humanités numériques*

Programme

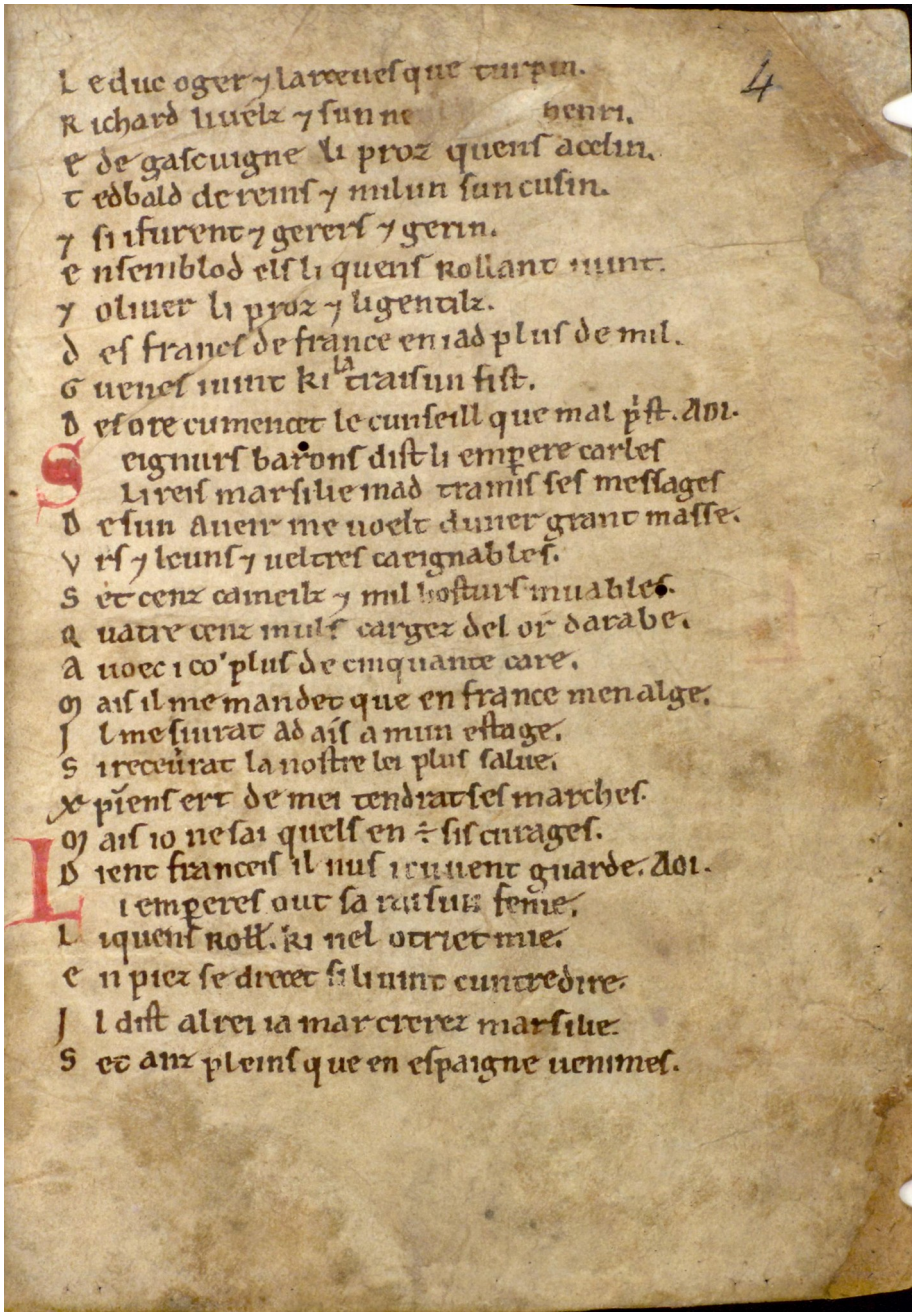
IV. La TEI et les manuscrits : La transcription des sources

V. La TEI et les manuscrits : notices des manuscrits

VI. Édition critique et apparat

Cas d'étude et travaux pratiques :
Les manuscrits médiévaux romans : les manuscrits épiques et le *Roland* d'Oxford

Le *Roland* d'Oxford (Oxford, Bodleian Library, Digby 23)



- Plus ancien témoin conservé de chanson de geste (et de *Roland*) ;
- anglo-normand ;
- c. 1130-1140 ;
- version « archaïque » du texte (fin XIe).

Objet des T.D. :

1. édition représentant le système graphique de ce témoin (et donnant une version régularisée) ;
2. description du manuscrit.

IV. Représentation des sources primaires

Sujets : Lien facsimile/transcription ; transcriptions imitatives et éditions à visée paléographique

Guidelines, « 11 Representation of Primary Sources »

Modules `transcr` (*Transcription of primary sources*) et `gaiji` (*Characters, Glyphs, and Writing Modes*)

J.B. Camps – Mercredi 29 novembre 2017

M2 *Technologies numériques appliquées à l'histoire*
M2 *Humanités numériques*

Édition à visée paléographique

- Représentation du **système graphique** ;
- **Analyse paléographique et linguistique** traditionnelle et computationnelle ;
- Développement de méthodes d'alignement, d'identification de la mise en page, d'acquisition du texte (OCR), etc.

Voir par ex., Peter Robinson et Elizabeth Solopova, «Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue », *The Canterbury Tales Project Occasional Papers*, (1993), http://server30087.uk2net.com/canterburytalesproject.com/pubs/op1-transguide.pdfDominique Stutzmann, « Paléographie statistique pour décrire, identifier, dater. . . Normaliser pour coopérer et aller plus loin ? », dans *Kodikologie und Paläographie im digitalen Zeitalter*, dir. Fischer Franz, Christiane Fritz et Georg Vogeler, Norderstedt, (Schri en des Instituts r Dokumentologie und Editorik), https://halshs.archives-ouvertes.fr/halshs-00596970/.

Unicode et fontes pour médiévistes

MUFI (Medieval Unicode Font Initiative)

http://www.mufi.info/

Œuvrer pour la création de caractères Unicode répondant aux besoins des médiévistes, et élaborer une norme qui puisse être reprise par des créateurs de fontes (voir la liste, http://www.mufi.info/fonts/).

Quelques exemples :

- Junicode (http://junicode.sourceforge.net/) :

S ı lapucele cūmɑnde e ıol otrıe / D ıſt beliſſɑnt eıo me tíent pur garíe

- Palemonas MUFI (http://www.mufi.info/fonts/) :

S ı lapucele cūmɑnde e ıol otrıe / D ıſt beliſſɑnt eıo me tíent pur garíe

Et d'autres encore…

Gestion des fontes sous Linux

- via le *Visionneur de polices*
- via la ligne de commande

Installation manuelle : *dans* `/usr/share/fonts` *ou* `~/.fonts` *ou* `~/.local/share/fonts`

```
sudo fc-cache -fv
fc-list
```

Pour Junicode, via un *package*

```
sudo apt-get install fonts-junicode
```

TP

- Installer Junicode ;
- Réaliser un première encodage « courant » du texte des fol. 4 et 4v de *Roland*

2.1 Correspondance fac-similé et transcription

ex. Le *Didascalicon* d'Hugues de Saint-Victor (*Thélème*, 100), en ligne :
http://theleme.enc.sorbonne.fr/dossiers/notice100.php

Mise en correspondance entre facsimile et transcriptions

Nécessité d'accomplir **deux opérations :**

1. Définir et délimiter des zones de l'image, d'une part, et des zones de texte, de l'autre ;
2. Mettre en correspondance les seconds avec les premiers

Délimitation des zones du facsimile : <facsimile/>

<facsimile/> « contient une représentation d'une source écrite quelconque sous la forme d'un ensemble d'images plutôt que sous la forme d'un texte transcrit ou encodé. »
(ou <sourceDoc/> pour une édition génétique)

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader/>
   <facsimile> <!-- ... --> </facsimile> <!-- Élément répétable si besoin -->
   <text/>
</TEI>
```

<facsimile/>

Éléments

`<surfaceGrp/>` : regroupe des éléments surface (ex., recto et verso d'un même feuillet)

`<surface/>` : définit une surface (ex. une page)

`<zone/>` : définit une zone à l'intérieur de l'élément `surface` parent

`<graphic/>` avec @url pour définir les emplacements des images

Attributs

att.coordinated (pour <surface/> et <zone/>)

@ulx (upper left x), abscisse du coin supérieur gauche (ex. 5)

@uly (upper left y), ordonnée du coin supérieur gauche (ex. 140)

@lrx (lower right x), abscisse du coin inférieur droit

@lry (lower right y), ordonnée du coin inférieur droit

ainsi que

@points, pour définir une zone non rectangulaire exprimée comme une série (séparée par des espaces) de coordonnées de points, sous la forme x,y x',y' x'',y'' … (ex. 220,100 300,210 170,250 123,234)

NB : les coordonnées sont généralement exprimées en pixel. Elles sont facilement obtenues avec un logiciel de retouche d'image, comme **Gimp** ou de manière plus automatisée, par un algorithme d'analyse d'image

Ex. : Le *Didascalicon* d'Hugues de Saint-Victor (*Thélème*, 100)

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader/>
   <facsimile>
      <surface>
         <graphic url="fax.jpg" width="693px" height="1100px"
xml:id="fax"/>
         <zone xml:id="zone_1" ulx="213" uly="3" lrx="342" lry="67"/>
         <zone xml:id="zone_2" ulx="29" uly="31" lrx="214" lry="67"/>
         <zone xml:id="zone_3" ulx="81" uly="65" lrx="335" lry="91"/>
         <zone xml:id="zone_4" ulx="80" uly="89" lrx="342" lry="117"/>
      </surface>
   </facsimile>
   <text/>
</TEI>
```

(exemple tiré de l'édition d'une page du *Didascalicon* d'Hugues de Saint-Victor (Paris, Bibliothèque Mazarine, 717, 93v° ; voir http://theleme.enc.sorbonne.fr/dossiers/notice100.php)

En pratique

Ouvrir avec *Gimp* le facsimile du *Didascalion* et vérifier que les coordonnées exprimées soient correctes

`att.global.facs` et `@facs`

classe d'attribut s'ajoutant aux attributs globaux, (i.e., pouvant être portés par *tous les éléments* de la TEI, dont les marqueurs <pb/>, <cb/>, <lb/>, ... mais aussi tous les autres !

`att.global.facs` « attributs utilisables pour les éléments correspondant à tout ou partie d'une image, parce qu'ils contiennent une représentation alternative de cette image, généralement mais pas nécessairement, une transcription »

`@facs` « (fac-similé) pointe directement vers une image ou vers une partie d'une image correspondant au contenu de l'élément. »

N.B. :

Il y a une certaine logique à faire porter l'attribut @facs, renvoyant à des portions du facsimile, à des éléments renvoyant à la structure physique du ms. ou à la représentation matérielle du texte, à savoir :

<lb/> (début de ligne)
<cb/> (début de colonne)
<pb/> (début de page)
<gb/> (début de cahier)
<milestone/> (début de quelque chose)
(voir *Guidelines*, « 3.10.3 Milestone Elements »)

mais on peut également utiliser selon les besoins tout type d'élément, ou créer des segments arbitraires (avec `<seg>` par exemple)

Ex. 1 : exemple minimal, tiré des *Guidelines*

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <!--...-->
   </teiHeader>
   <text>
      <pb facs="page1.png"/>
      <!-- text contained on page 1 is encoded here -->
      <pb facs="page2.png"/>
      <!-- text contained on page 2 is encoded here -->
   </text>
</TEI>
```

Ex. : Le *Didascalicon* d'Hugues de Saint-Victor (*Thélème*, 100)

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader/>
  <facsimile>
    <surface>
      <graphic url="fax.jpg" width="693px" height="1100px" xml:id="fax"/>
      <zone xml:id="zone_1" ulx="213" uly="3" lrx="342" lry="67"/>
      <zone xml:id="zone_2" ulx="29" uly="31" lrx="214" lry="67"/>
      <zone xml:id="zone_3" ulx="81" uly="65" lrx="335" lry="91"/>
      <zone xml:id="zone_4" ulx="80" uly="89" lrx="342" lry="117"/>
    </surface>
  </facsimile>
  <text>
    <body>
      <div type="transcription" facs="#fax">
        <p>
          <seg facs="#zone_1">De tribus generibus lectorum. </seg>
          <seg facs="#zone_2">Satis ut puto aperte </seg>
          <seg facs="#zone_3">demonstratum e<expan>st</expan> provectis et
            aliquid </seg>
          <seg facs="#zone_4">amplius de se
            p<expan>ro</expan>mittentib<expan>us</expan>
            n<expan>on</expan> idem e<expan>ss</expan>e </seg>
          <!– etc. –>
        </p>
      </div>
    </body>
  </text>
</TEI>
```

(exemple tiré de l'édition d'une page du *Didascalicon* d'Hugues de Saint-Victor (Paris, Bibliothèque Mazarine, 717, 93v° ; voir http://theleme.enc.sorbonne.fr/dossiers/notice100.php)

Une autre possibilité : la mise en correspondance via linkGrp

```
<TEI>
  <teiHeader/>
  <facsimile>
    <surface>
      <graphic url="fax.jpg" width="693px" height="1100px" xml:id="fax"/>
      <zone xml:id="zone__1" ulx="213" uly="3" lrx="342" lry="67"/>
      <zone xml:id="zone__2" ulx="29" uly="31" lrx="214" lry="67"/>
    </surface>
  </facsimile>
  <text>
    <body>
      <div type="transcription" facs="#fax">
        <p>
          <seg xml:id="seg__001">De tribus generibus lectorum. </seg>
          <seg xml:id="seg__002">Satis ut puto aperte </seg>
        </p>
      </div>
      <linkGrp>
        <link targets="#zone__1 #seg__001"/>
        <link targets="#zone__2 #seg__002"/>
      </linkGrp>
    </body>
  </text>
</TEI>
```

T.P. Roland

1. Créer un élément <facsimile/> pour le fol. du ms. de la *Chanson de Roland*, et y définir des groupes de surface, des surfaces, et des zones ;
2. Mettre en correspondance les sections de la transcription avec les zones définies. Raccrocher de préférence ces indications à des éléments déjà présents, sans créer de segments arbitraires.

Alignement texte/image et ROC du manuscrit

L'alignement texte et image peut tant servir à l'établissement d'édition à visée paléographique ou linguistique, que permettre la fourniture de données utiles pour perfectionner la reconnaissance optique des caractères (anglais : *Optical character recognition*) pour le manuscrit

- Projet ORIFLAMMS, voir https://oriflamms.hypotheses.org/ et « Oriflamms Alignment Software », cf. A. Lavrentiev, Y. Leydier, D. Stutzmann, « Specifying a TEI-XML Based Format for Aligning Text to Image at Character Level », Symposium on Cultural Heritage Markup, 2015, en ligne : http://www.balisage.net/Proceedings/vol16/print/Lavrentiev01/BalisageVol16-Lavrentiev01.html
- Transkribus, https://transkribus.eu/Transkribus/.

Détection des lignes

Ces outils intègrent généralement des algorithmes d'analyse de la mise en page et de détection des zones de texte. Par exemple, Ocropy (https://github.com/tmbdev/ocropy)

Étapes principales

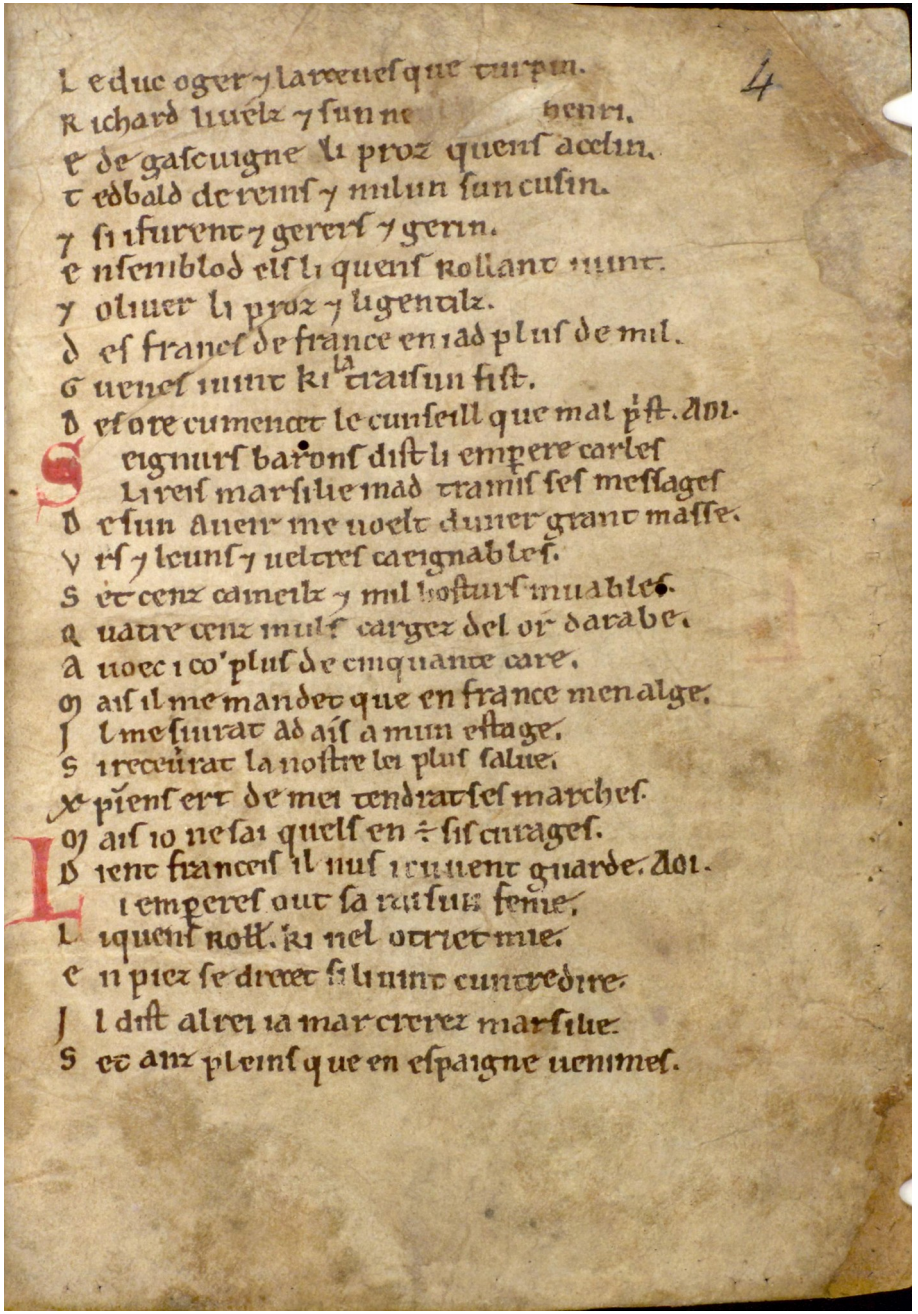
1 Acquisition des images (de préférence tif avec une résolution d'au moins 300 DPI) ;
2 Binarisation ;
3 détection des colonnes et des lignes ;
4 acquisition du texte
5 si erreurs, corriger et réentraîner le modèle
…
6 extraction des données textuelles, éventuellement de l'alignement avec l'image

2.2 TEI et éditions allographétiques

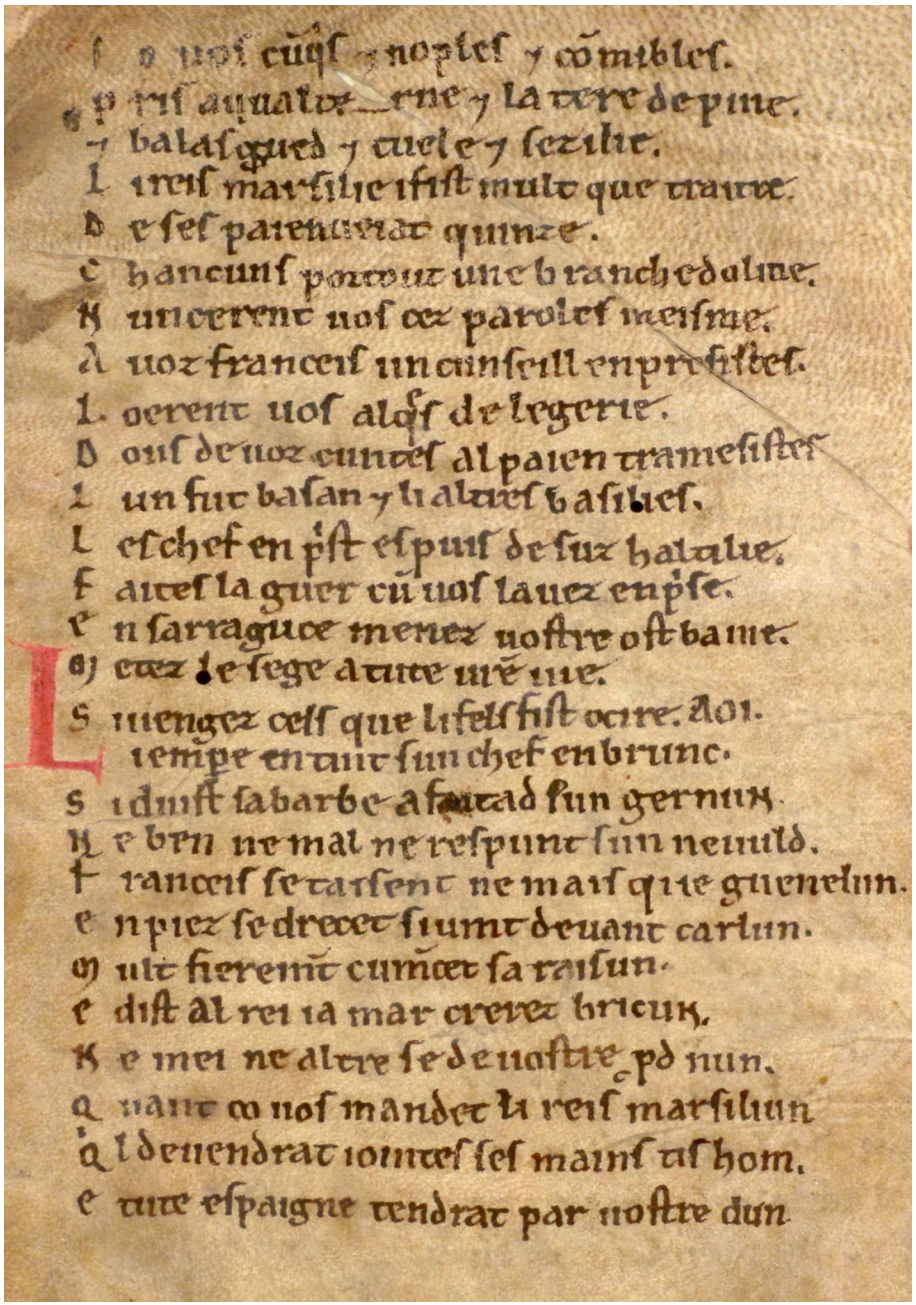
Numérisation du ms.

Oxford, Bodleian Library, Digby 23

http://image.ox.ac.uk/show?collection=bodleian&manuscript=msdigby23b



L educ oger y lareuesque turpin.
R ichard liuelz y sun ne benri.
E de gascuigne li proz quens acelin.
T edbald de reins y milun sun cusin.
Y si furent y gerers y gerin.
E nsemblod els li quens rollant uint.
Y oliuer li proz y ligentilz.
D es franc de france en mad plus de mil.
C uenes uint ki la traisun fist.
D e ore cumencet le cunseill que mal pist. Aoi.
S eignurs barons dist li empere carles
L i reis marsilie mad tramis ses messages
D e sun auer me uolt duner grant masse.
V ers y leuns y uelteres carignables.
S et cenz camelz y mil hosturs muables.
Q uatre cenz mulz cargez del or darabe.
A uoec i co plus de cinquante care.
O ais il me mandet que en france menalge.
I l me siurat ad ais a mun estage.
S i receurat la nostre lei plus salue.
X pïens ert de mei tendrat ses marches.
O ais io ne sai quels en ⁊ sis curages.
D ient francois il nus i cuuent guarde. Aoi.
I empere ot sa raisun fenie.
L iquens rollz. ki nel otriet mie.
E n piez se dreecet si li uint cuntredire.
I l dist al rei ia mar creirez marsilie.
S et anz pleins que en espaigne uenimes.



⁊ o uos cũqs ⁊ noptes ⁊ cõmibles.
P ris aquaive—rne ⁊ la tere de pine.
⁊ balasgued ⁊ cuele ⁊ sezilie.
I reis marsilie ifist mult que traitre.
D e ses parentet quinze.
C hanguns portout une branche d olive.
K uncerent uos cez paroles meisme.
A uez franceis un cunseill enpristet.
I. oerent uos alos de legerie.
D ous de uez cuntes al paien trametistes
I un fuc basan y li altres basilies.
I es chef en pist espuis de sue batalie.
F aites la guer cũ uos lauez enpse.
E n sarragoce menez uostre ost baime.
O ez le sege a tute mẽ uie.
S uenget cels que li fels fist ocire. Aoi.
I empe en tint sun chef en brunc.
S i duist sa barbe a fraitet sun gernun.
K e bien ne mal ne respunt sun neuld.
F ranceis se taisent ne mais que guenelun.
E n piez se dreecet si uint deuant carlun.
O ilt fierement cumcet sa raisun.
E dist al rei ia mar creirez brieus.
K e mei ne altre se de uostre pd nun.
Q uant co uos mandet li reis marsiliun
I l deuendrat ioinz ses mains tis hom.
E tute espaigne tendrat par uostre dun

Modélisation

Quels éléments veut-on représenter ?

Modélisation

Quels éléments veut-on représenter ?

Système graphique : allographes, abréviations, segmentation, ponctuation ancienne…

Interventions scribales : ajouts, suppressions, corrections, …

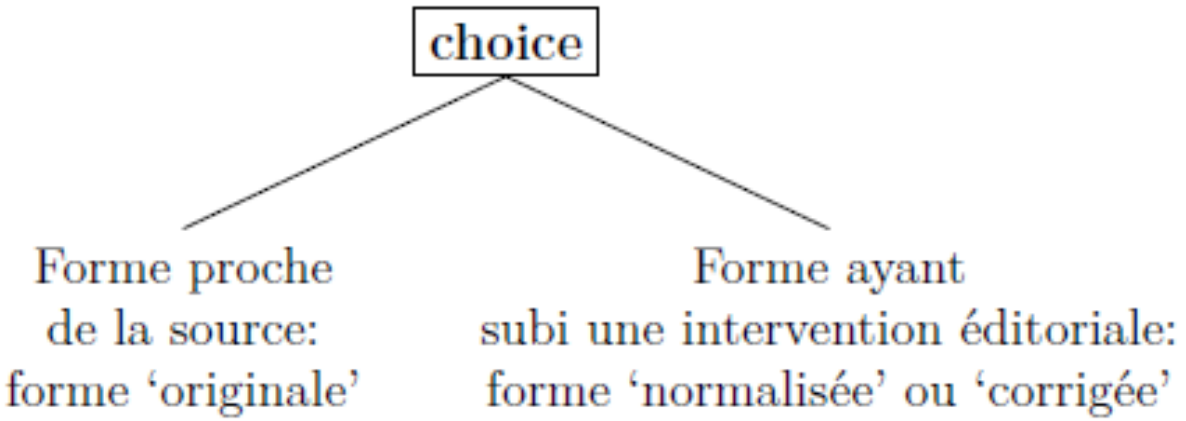
Interventions éditoriales : normalisations, corrections, difficultés de transcription, …

2.2.1 une représentation double avec ‹choice›

Permet la représentation simultanée de deux états du texte

Peut contenir tous les éléments du model.choicePart :

sic/corr ; reg/orig ; unclear ; abbr/expan ; ex, am et seg



Abréviations

« 17 oc^{bre} 1822 »

Solution utilisable avec TEI-Lite

```
17
<choice>
  <abbr>oc<hi rend="sup">bre</hi></abbr>
  <expan>octobre</expan>
</choice>
1822
```

Utilisation de ex pour marquer les lettres restituées

```
17
<choice>
  <abbr>oc<hi rend="sup">bre</hi></abbr>
  <expan>oc<ex>to</ex>bre</expan>
</choice>
1822
```

Solution sans choice (plus limitée) :

```
17 oc<ex>to</ex>bre 1822
```

NB : am peut être utilisé si des lettres figurent dans l'abréviation et pas dans sa résolution. Ex. ss. pour sancti

Traitement des allographes

```
<choice>
        <reg>s</reg>
        <orig>ſ</orig>
</choice>
```

```
<choice>
        <reg>s</reg>
        <orig>&#383;</orig>
</choice>
```

Quelques limites de cette première solution

- les *Guidelines* tendent à recommander un emploi de `expan`, `reg`, …, au niveau du mot (expan « should usually be a complete word or phrase ») ;
- détournement du sens de caractères Unicode

  par. ex., si l'on utilise U+0303 COMBINING TILDE (̃) pour la tilde abréviative médiévale, alors que son sens est « IPA: nasalization / Vietnamese tone mark »

2.2.2 Définir des caractères (module `gaiji`)

En toute bonne rigueur, lorsque l'on utilise des caractères qui n'ont pas de correspondant exact dans Unicode, que l'on détourne des glyphes de fontes Unicode*, ou que l'on veut décrire spécifiquement des variantes de caractères Unicode existants, il importe de clarifier cela en utilisant des éléments du module gaiji, cf. *Guidelines*, « 5 Characters, Glyphs, and Writing Modes »), pour soit annoter un caractère existant (précision à apporter), soit créer un nouveau caractère.

⚠ Ne pas confondre *caractère* (symbole abstrait) et *glyphe* (une représentation physique donnée, une réalisation d'un caractère)

* (par exemple, le caractère U+0303 COMBINING TILDE, qui représente la nasalisation dans l'API, pour représenter la tilde abréviative médiévale).

Définir des caractères (module `gaiji`)

Deux éléments :

Dans le teiHeader

`charDecl` pour définir ou annoter des caractères dans le teiHeader/encodingDesc (`model.encodingDescPart`), à l'aide d'éléments

- **`char`** : définit un nouveau caractère
- **`glyph`** : précise une réalisation (un glyphe) d'un caractère existant

Dans le corps du document

`g` (et `@ref`) pour faire référence à un glyphe dans le corps du document (`model.gLike`)

`char` / `glyph`

- **`charName`** ou **`glyphName`** pour préciser le nom d'un caractère ou d'un glyphe ;
- **`gloss`** et **`desc`**
- **`charProp`** (character property) permet de décrire des propriétés du caractère (paires nom/valeur), avec `unicodeName`/`localName` et `value` ;
- **`mapping`** (character mapping) « contient un ou plusieurs caractères reliés par certains aspects (spécifiés par l'attribut type) au glyphe ou au caractère défini dans l'élément parent »
- **`figure`**
- **`note`**

Exemple de définition de caractère (*Guidelines*)

```
<charDecl>
  <desc>Description des caractères et glyphes : Exemples tirés des Guidelines</desc>
    <char xml:id="aenl">
      <charName>LATIN LETTER ENLARGED SMALL A</charName>
      <charProp>
        <localName>entity</localName>
        <value>aenl</value>
      </charProp>
      <mapping type="standard">a</mapping>
    </char>
< !-- Autres définitions -->
</charDecl>
```

Exemple d'annotation (*Guidelines*)

```
<charDecl>
    <desc>Description des caractères et glyphes : Exemples tirés des Guidelines</desc>
    <!-- ... -->
    <glyph xml:id="r1">
      <glyphName>LATIN SMALL LETTER R WITH ONE FUNNY STROKE</glyphName>
      <charProp>
        <localName>entity</localName>
        <value>r1</value>
      </charProp>
      <figure>
        <graphic url="r1img.png"/>
      </figure>
    </glyph>
    <!-- ... -->
</charDecl>
```

Répertoires de définitions de caractères

James Cummings, « The **ENRICH Gaiji Bank** », *Manuscriptorium : Building Virtual Research Environment for the Sphere of Historical Resources*, en ligne : http://v2.manuscriptorium.com/apps/gbank/gbank_table.php

Définir des caractères (module `gaiji`)

Utiliser les caractères dans le corps du document :

```
<p>Attention, ce <g ref="#r1">r</g> a une
forme bizarre. Et ce <g ref="#aenl">a</g>
    n'existait pas dans Unicode. </p>
```

N.B. : la TEI définit un élément <c> (charactère), de la même série que <w> (*word*) ou <s> (*sentence*), qui peut aussi être utilisé pour spécifier le rendu d'un caractère précis.

```
<c rend="lettrine">A</c>ttention
```

2.2.3 utilisation d'une DTD et d'entités

La déclaration de DTD

DTD externe

```
<!DOCTYPE nomÉlémentRacine SYSTEM "madtd.dtd">
```

DTD interne

```
<!DOCTYPE nomÉlémentRacine [
   <!ELEMENT nomÉlémentRacine (#PCDATA)>
   Définition des autres éléments...
] >
```

DTD externe avec sous-ensembles internes

```
<!DOCTYPE document SYSTEM "madtd.dtd" [
   <!ELEMENT elementNonDecritDansmadtd (#PCDATA)>
]
>
```

Rappel de cours : les entités

```
<!ENTITY entité "contenu">
<!ENTITY ENC "École nationale des chartes">
```

&ENC; sera développé par le parseur en « École nationale des chartes »

Entités

Peuvent contenir des éléments TEI :

```
<!DOCTYPE TEI [
  <!ENTITY s-long
'<choice><reg>s</reg><orig>&#383;</orig
></choice>' >
]>
```

Des difficultés de modélisation qui subsistent

- Problématique de la définition des variantes de forme ;
- Portée abréviative et son niveau.

Un exemple d'encodage du début du fragment de Mende

(Voir les fichiers)

T.P.

Mise en place

6) Ajouter une déclaration DOCTYPE, interne ou externe, au document

Définition des entités et transcription

7) Créer les entités nécessaires à la transcription du fragment (allographes, signes abréviatifs,…), et les utiliser où elles sont nécessaires ;

8) Utiliser, là où ils sont nécessaires les éléments pour représenter les lacunes matérielles, les difficultés de lecture, les notes, la séparation des mots…

Bibliographie

En complément des références données au cours de la présentation.

Texte de référence

TEI Consortium, *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html>, particulièrement

« 10 Manuscript Description »

« 11. Representation of Primary Sources »

Éditions à visée paléographique et analyse statistique

A. Lavrentiev, Y. Leydier, D. Stutzmann, « Specifying a TEI-XML Based Format for Aligning Text to Image at Character Level », *Symposium on Cultural Heritage Markup*, 2015, en ligne : http://www.balisage.net/Proceedings/vol16/print/Lavrentiev01/BalisageVol16-Lavrentiev01.html

Dominique Stutzmann, « Paléographie statistique pour décrire, identifier, dater… Normaliser pour coopérer et aller plus loin ? », dans *Kodikologie und Paläographie im digitalen Zeitalter 2 / Codicology and Palaeography in the Digital Age 2*, dir. Franz Fischer *et al.*, 2010, p. 247-277, en ligne : http://kups.ub.uni-koeln.de/4353/1/15_stutzmann.pdf.