Structuration des données et des documents : balisage XML

TEI – Modélisation de la source et des interventions éditoriales : un premier cas d'étude

J.B. Camps – mercredi 15 décembre 2015
M2 *Technologies numériques appliquées à l'histoire*

Avant de réfléchir à l'encodage

- Quelles sont les particularités de ces documents ?

- Dans quelles perspectives (éditoriale et d'exploitation) cette édition et cet encodage se placent-ils ?

En conséquent : quels sont les besoins que l'on peut attendre en termes de modélisation et d'encodage des documents ?

Le projet

Le français à la mesure d'un continent : un patrimoine en partage, dir. France Martineau, Département de français, Université d'Ottawa (2011-2018),
en ligne : http://continent.uottawa.ca/

objectifs :

« **évaluer l'impact réel des contacts linguistiques** et culturels dans les communautés multiculturelles et d'examiner les **conditions de maintien du français et des autres langues en situation de contact.** Il alimente la réflexion sur les enjeux contemporains du Canada et de la Francophonie en matière de diversité culturelle et d'épanouissement des minorités linguistiques. »

Les correspondances et le corpus FRAN

- Méthode qui « allie des **perspectives historique, linguistique et sociologique** et des **approches dialectologique, variationnelle et diachronique, à partir de corpus textuels et oraux**. Elle prend en compte la **complexité des compétences linguistiques des individus et des réseaux** dans lesquels ils évoluent » (*site*)

- *Envoyer et recevoir: lettres et correspondances dans les diasporas francophones*, dir. Yves Frenette, John Willis et Marcel Martel, Québec, 2006. Voir aussi les ouvrages de la coll. *Les Voies du français* (http://continent.uottawa.ca/fr/publications/livres/). Série de colloques et ouvrages passés ou en préparation.

- Série de Corpus, dont le **Corpus FRAN** (publ. lancée en sept. 2014) « premier **corpus panfrancophone sur l'Amérique française** réunissant dans un ensemble unifié des **corpus écrits historiques et des corpus oraux modernes** qui constituera la référence fondamentale pour l'**histoire linguistique et culturelle des francophones du Canada et des États-Unis** » http://continent.uottawa.ca/corpus-et-ressources-electroniques/corpus/corpus-interrogeable-fran/

Les sous-corpus

- **Famille Papineau** (années 1804-1830 ; Hull, Montréal, Petite-Nation – Québec)

- **Mélanie Hébert** (années 1871-1876 ; Plaquemine, Bâton Rouge – Louisiane)

  « Melanie Hebert, a grandchild of Raphael Hebert and Marie Odile Landry, wrote her relatives from the Academy of St. Basil in Plaquemine, Louisiana, from 1870 to 1874 » (cf. Hebert (RAPHAEL And Family) Papers (Mss. 4769), Inventory Compiled by Germain J. Bienvenu, en ligne : http://www.lib.lsu.edu/special/findaid/4769_inv.pdf, p. 4 et 7).

- **Argyle** (ou « Par-en-Bas », 1868-1891 ; West Pubnico, municipalité d'Argyle, Acadie/ Nouvelle-Écosse)

  cf. http://www.museeacadien.ca/french/index.htm et la liste des fonds, http://www.museeacadien.ca/french/archives/fonds/index.htm.

Perspective éditoriale

- perspective historique (migrations, réseaux, ...) ;
- perspective linguistique et sociolinguistique (linguistique variationnelle ; de contact ; oralité/écrit, ...) ;
- perspective d'histoire de l'écrit.

Perspective d'exploitation des fichiers

- Envisager deux affichages d'édition (proche de la source et normalisé) ;

- utilisation avec un logiciel permettant des recherches dans le corpus (cooccurrences, segments répétés, …) ; utilisation du logiciel *Philologic*.

Première réflexion

À partir des lettres vues ensemble et de celle reçue, réfléchir à :

- particularité du genre de la lettre (structuration logique, ...) ;
- particularités linguistiques, paléographiques, etc. ;
- matérialité des sources.

Réfléchir notamment en termes de :
- structuration logique ;
- régularisations, normalisations, etc. ;
- représentation de la structure matérielle et des phénomènes liés.

I. Réflexion préalable sur la structuration de ces lettres et de leur contenu

1. Structure logique des lettres

Structure logique des lettres

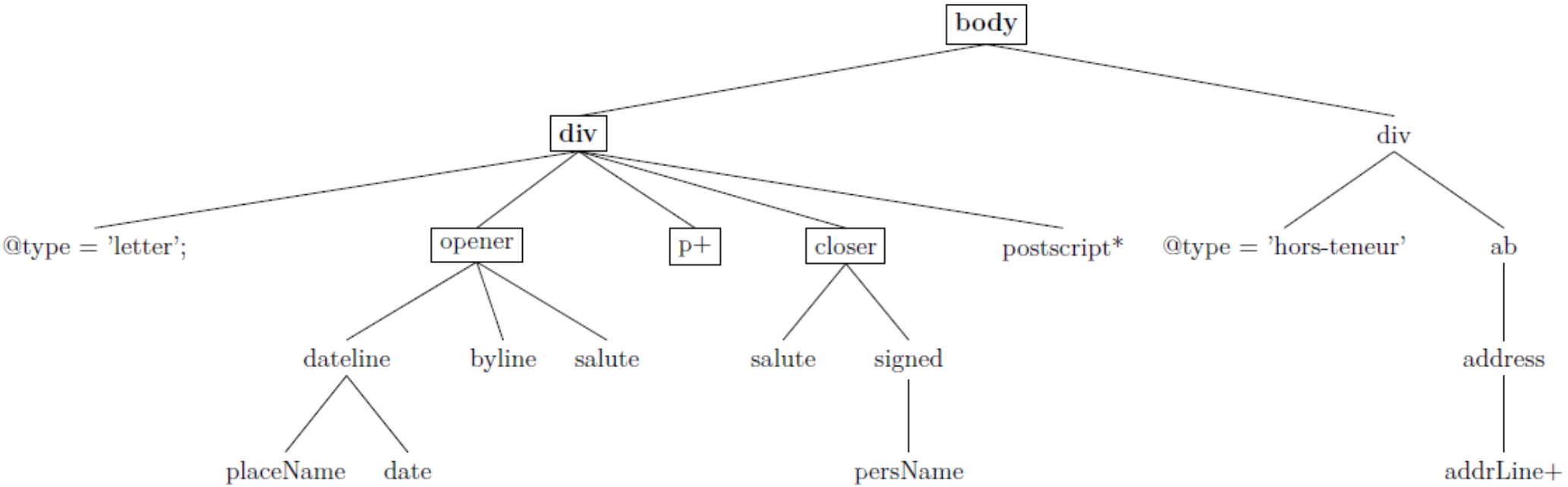
Teneur de la lettre

- **Mentions ouvrant la lettre** (*cadre formel initial*)
  - lieu et date (*formule de date initiale*, dont *date de lieu* et *date de temps*) ;
  - nom de l'expéditeur (*suscription*) ;
  - salutations (*adresse* ; *salut*).
- **Texte de la lettre** (paragraphes de prose)
- **Mentions fermant la lettre** (*cadre formel final*)
  - salutations (*salutation finale*) ;
  - signature (*souscription* ; *signature*).
- **Post-scriptum.**

Mentions « hors-teneur »

- adresse postale ;
- …

Structure logique des lettres



```
<body>
  <div type="letter">
    <opener>
      <dateline>
        <placeName>Hull</placeName>, <date when="1822-10-17"> 17 octobre 1822</date>
      </dateline>
      <!--<byline></byline> (suscription, s'il s'en trouve une)-->
      <salute>mon cher <persName>Benjamin</persName></salute>
      <!-- salute pourrait aussi contenir une éventuelle adresse du type «M. B.
Papineau»  -->
    </opener>
    <p>
    <!-- Corps de la lettre - autant de paragraphes que nécessaire -->
    </p>
    <closer>
      <salute> adieu mes amitiés a angelle, ambrasse les petits enfants pour moi et
        Crois moi sincerement</salute>
      <signed> ton pere affectionné <persName>jh. papineau</persName></signed>
    </closer>
    <!--<postscript><p></p></postscript> post-scriptum éventuel-->
  </div>
  <!--<div type="hors-teneur"> (s'il y a lieu)
        <ab>
          <address>
            <addrLine></addrLine>
          </address>
        </ab>
       </div>-->
</body>
```

N.B.

- impossible de typer (@type) les éléments <salute> pour préciser s'il s'agit de l'adresse ou du salut proprement dit ;

- le cas des mentions hors-teneur (à distinguer des additions ultérieures) ne connaît pas de solution unique en TEI, ni d'élément spécialisé. On constate une variété de solutions mises en œuvre.

Édition de correspondances : schéma ENC

(http://developpements.enc.sorbonne.fr/diple/schema/correspondance)

<nota>

Contient les mentions hors teneur, marginales ou portées au dos du dernier feuillet soit l'adresse ou diverses mentions ajoutées lors de la réception de la dépêche (nom de l'expéditeur, date de réception, sujet).

Leurs places et types peuvent être renseignés de manière obligatoirement normalisée afin de pouvoir dresser plus facilement des typologies, ces informations peuvent ne pas apparaître au sein de l'édition. La description des mentions hors teneur étant souvent faite en texte libre, celle-ci doit être inséré directement dans un élément @desc.

```
<nota type="adresse" desc="adresse au dos">A Monsieur, Monsieur le chancelier.</nota>
```

Attributs	@place ? @type ? @desc ?
Contenu	(*text()*)*
Usage	div type="transcription"

@type

Valeur	[*"adresse"* | *"reception"* | *"note"*]

@desc

Valeur	*text()*

Description d'une mention hors teneur de manière plus libre qu'au travers des attributs @place et @type.

Quelle solution pour ces mentions ?

En les maintenant avec le corps de la lettre

- <seg> (segment arbitraire)
- <add> (additions)
  - @type
  - @place
  - @hand (si besoin)

  Peu satisfaisant d'un point de vue sémantique

En les séparant (les mettant au même niveau)

-

Respecte mieux la distinction teneur/hors-teneur de la diplomatique.

2. Régularisation

2. Régularisation

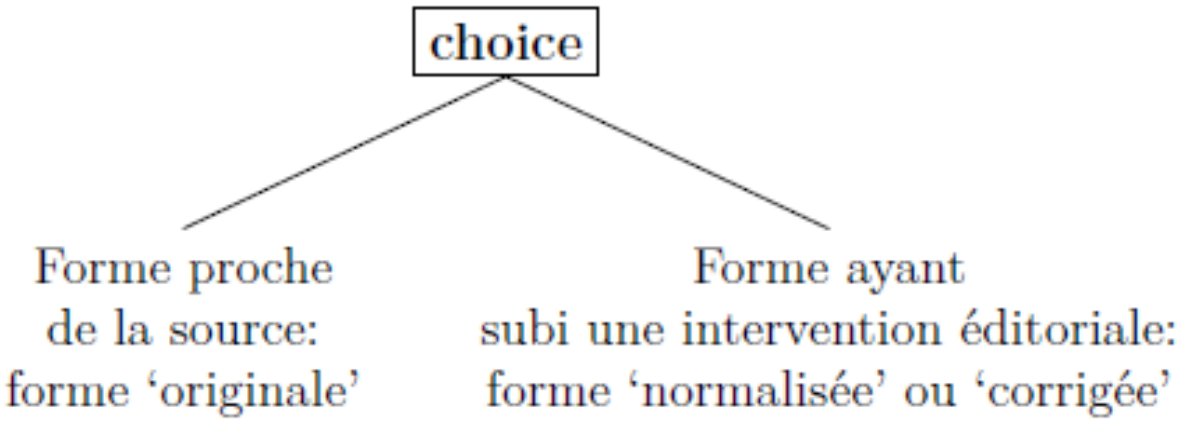
1. **graphiques**
   a. paléographiques
      i. allographes
      1. abréviations
   b. orthographiques (mais par rapport à quelle norme ?)
2. **segmentation de la chaîne graphique** (**séparation des mots**)
   a. mots séparés à réunir
      i. séparés en pleine ligne
      ii. séparés par une césure en fin de ligne
   b. mots réunis à séparer
      i. cas général
      1. cas particulier de l'élision
3. **Usages typographiques, diacritiques et ponctuation**
   a. emploi des majuscules
   b. accents, cédilles, tréma
   c. ponctuation
4. **Corrections** (par rapport à quelle définition d'erreur?)
   Omissions, répétitions, substitutions de mots ; variations substantielles (sémantiques) ; mais édition d'originaux…

Une représentation double avec <choice>

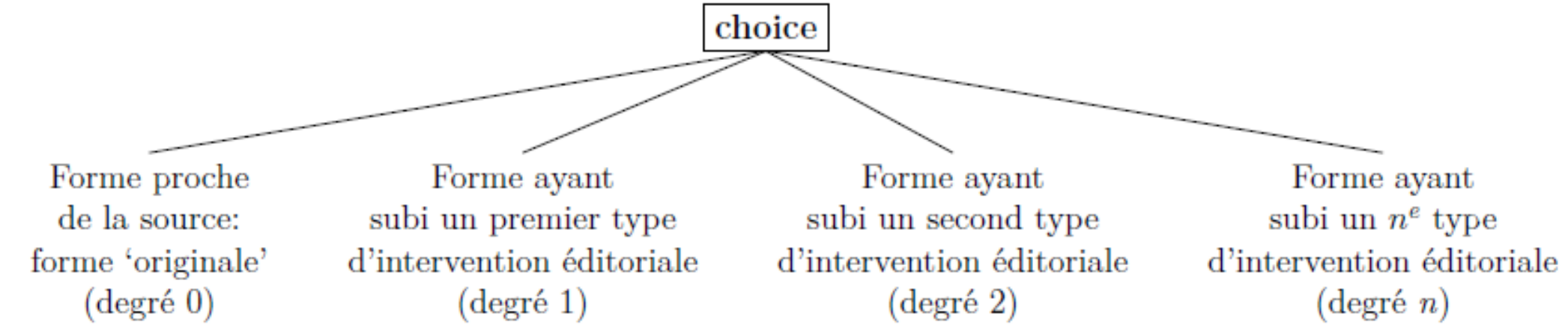
Permet la représentation simultanée de deux états du texte

Peut contenir tous les éléments du model.choicePart :

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



(Deux états ou plus)



Distinguer :

- nature/objectif de l'intervention (correction, régularisation,...) ;
- responsabilité, source, justification de cette intervention (att.global.responsibility, att.editLike).

Correction… ou régularisation des graphies ?

reg + orig

```
adieu mes amitiés a angelle,
<choice>
  <reg>embrasse</reg>
  <orig>ambrasse</orig>
</choice>
les petits enfants pour moi
```

Autre possibilité : réserver sic et corr aux corrections que vous apporteriez à la transcription fournie

« quoique les affaires de mr wrigth ne soient pas [fiais] »
(voir reprod.)

```
<choice>
  <sic>[fiais]</sic>
  <corr>finis</corr>
</choice>
```

Utiliser att.responsibility pour préciser l'origine et le degré de certitude de la correction

```
<corr cert="high" resp="#Untel">finis</corr>
```

N.B. : @resp contient un pointeur, donc, il faut spécifier dans le header

```
<titleStmt>
    <!-- [...] -->
    <respStmt>
      <resp>Encodé par</resp>
      <name xml:id="Untel">Untel</name>
    </respStmt>
</titleStmt>
```

Abréviations

« 17 oc^{bre} 1822 »

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

N.B. : en TEI Lite, il est a priori impossible de marquer les lettres ajoutées lors de la résolution de l'abréviation (et inutile si on ne propose pas une transcription proprement diplomatique)

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

Traitement des allographes

En perfaite ſanté je vous ſouaite boucoup de plaisir dans vos vaquance

Traitement des allographes

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

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

Traitement des allographes

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

Et dans le corps du texte :

```
&s-long;
```

La solution <g>

Module gaiji: Character and glyph documentation

(N.B. : attention, pas disponible en TEI-lite)

Dans le header :

```
<encodingDesc>
   <charDecl>
      <glyph xml:id="s-long">
         <glyphName>LATIN SMALL LETTER LONG S EXTENDING BELOW THE
LINE</glyphName> <!-- Nom donné au glyphe -->
         <charProp> <!-- Description de ses propriétés -->
            <localName>entity</localName>
            <value>s-long</value>
         </charProp>
         <mapping type="modern">s</mapping>
         <mapping type="simplified">&#383;</mapping>
      </glyph>
   </charDecl>
</encodingDesc>
```

Dans le corps du document :

```
<g ref="#s-long">s</g>
```

Ou, en se simplifiant la vie : `<!ENTITY s-long '<g ref="#s-long">s</g>' >`

et `&s-long;`

Cf. TEI G-P5, « 5 Non-standard Characters and Glyphs »

Séparations de mots

Séparation de mots unis dans la source

« ou même **enville** »

Séparation de mots donnant lieu à une élision

« **Leun** en vers L'autre »

Union de mots séparés dans la source

« Leun **en vers** L'autre »

Cas particulier de la fin de ligne

mademois- // **elle**

Cas cumulatifs

« pourmoi **de-** //**montreal** »

Séparations de mots

Séparation de mots unis dans la source

« ou même enville »

<w rend="aggl">en</w><w>ville</w>

Séparation de mots donnant lieu à une élision

« Leun en vers L'autre »

<w rend="elision">L</w><w>eun</w>

Union de mots séparés dans la source

<w>en vers</w> L'autre

ou, en codant en dur l'espace fautif (*NB a priori impossible en TEI-Lite, mais a ses mérites*) : <w>en<space/>vers</w>

Cas particulier de la césure

mademois- // elle

<w>mademois<pc type="hyphen">-</pc><lb/>elle</w>

ou : mademois<lb rend="hyphen"/>elle

Séparations de mots

Cas cumulatifs

« pourmoi **de-** //**montreal** »

```
<w rend="aggl">de</w><lb rend="hyphen"/><w>montreal</w>
```

Autres usages typographiques

Majuscules

Emploi possible de <**reg**> **et** <**orig**>, et d'entités,
[NB : il devrait être envisageable de ne pas encoder cette normalisation et de la faire ensuite par expressions régulières (en tenant compte du balisage des noms propres et de la ponctuation)].

Accents

Emploi possible également de <**reg**> **et** <**orig**>, ou <**reg**> seul, en utilisant des accents combinatoires.

Ponctuation

S'il est nécessaire de distinguer ponctuation moderne et ancienne, des éléments pc typés peuvent être utilisés.
Sinon, on peut simplement distinguer l'ajout de ponctuation moderne en utilisant **supplied**.

3. Éléments de représentation de la source

3. Éléments de représentation de la source

1. **Disposition du texte dans la source** (et mise en page)
   a. changements de page, colonne, ligne…
2. **Éléments propres au support matériel / à la lisibilité**
   a. trous, tâches, etc. sur le support
   b. mots illisibles, effacés,…
3. **Interventions sur la source** (plus graphique que matériel)
   a. type : suppressions, ratures, corrections, ajouts, gloses et annotations…
   b. responsabilité : par le scribe du document, par un lecteur postérieur,…

Changements de feuillets et de ligne

<pb n="1"/> (changement de page) et <lb/> (changement de ligne)

Difficulté de lecture

quoique les affaires de mr wrigth ne soient pas [fiais] :
pas <unclear reason="illegible">fiais</unclear>

Long passage impossible à transcrire pour diverses raisons

<**gap**> en spécifiant

- **la raison**, @reason
- **les dimensions**, att.dimensions (@unit, @quantity)

Voir aussi « 11.3.3.2 Use of the <gap>, <del>, <damage>, <unclear>, and <supplied> Elements in Combination »,
http://www.tei-c.org/release/doc/tei-p5-doc/en/html/PH.html#PHCOMB

Interventions dans la source

Mot raturés dans la source par le scribe

« Et votre papa ~~arangeras~~ vous pourvoiras »

Et votre papa <del rend="strikethrough">arangeras</del> vous pourvoiras

Mots ajoutés dans la source par le scribe

Mr

jai vu Souligni : jai vu <add place="above">Mr</add> Souligni

N.B. : il existe une possibilité de grouper les deux quand cela est nécessaire (mais a priori pas en TEI-Lite)

```
<subst>
  <del rend="strikethrough">arangeras</del>
  <add>vous pourvoiras</add>
</subst>
```

Spécifier les mains ?

attribut **@hand** pour les éléments **add** et **del**

Doit renvoyer à une description des mains dans le header (mais ne fait pas partie de la personnalisation TEI-Lite)

```
<profileDesc>
   <handNotes>
      <handNote xml:id="main1">Main
principale du document, écriture du début du XIXe
siècle</handNote>
      <handNote xml:id="main2">Corrections
d'une main postérieure, différente de la main
principale</handNote>
   </handNotes>
</profileDesc>
```

Noms de personnes et de lieux

```
Nom :
   <placeName>Hull</placeName>
   mon cher <persName>Benjamin</persName>

Date :
<date when="1822-10-17"> 17 octobre 1822 </date>
```

T.P. : débuter l'encodage d'une lettre,
vérifier la validité des solutions retenues,
chercher à identifier des cas qui
n'auraient pas été pris en compte

2. Métadonnées du document : l'entête teiHeader des correspondances

Réflexion à mener sur les métadonnées :

- quelles données concernant ces lettres serait-il souhaitable de normaliser et d'intégrer au header ?
- quelle réalisation technique peut-on envisager ?