Remarques HTML Lamartine

Par Lucie Slavik
J2 Mercredi 20 mai, J3 mardi 26 mai, J4 mercredi 27 mai 2020

Remarques d'ordre général :
-Il y a plus de 5000 balises span pour George Sand pour le HTML alors que pour Lamartine, je vois qu'il n'y en a pas une seule, ce qui montre que tout est ok, l'OCR est bon en théorie.
-Pas de gap dans l'un ni dans l'autre… (est-ce pour la TEI?)
-Lamartine finit la plupart de ses lettres par « Adieu, mon cher ami… » 104 occurrences pour 97 lettres.
-L'OCR « recolle» un mot qui a été divisé par un tiret parce qu'en fin de ligne.
Un problème se pose lorsque le mot en fin de ligne a un tiret parce que cela fait partie d'une expression, comme à la ligne 4026 (p.365 dans le PDF) où l'expression « écris-moi » a été assemblée en un mot « écrismoi » dans l'html, ce qui n'est plus correct.
Dans Volume 1 correspondance Lamartine.

Questions :
-Lignes 1800 et suivantes dans le html de Lamartine : orthographe spéciale (ung pour un etc.) = doit-on normaliser pour le XML ? (balises orig/reg)
-Doit-on normaliser certains mots, par exemple l'orthographe de « poëte » que l'on retrouve à maintes reprises ?
- Comment fait-on pour l'index de noms de lieux et de personnes ? Est-ce que cela ne nous avancerait pas de le mettre déjà dans python ? Après, est-ce qu'on le fera manuellement pour repérer les noms de lieux et de personnes ou y a-t-il y moyen de le faire de façon automatisée ?
-Doit-on garder la division des années des grands titres ?
Sur la future plateforme, est-ce qu'il est prévu de faire une recherche par année. Doit-on pour cela laisser des titres spéciaux dans la TEI ?
J'ai compris qu'on voulait surtout extraire les lettres en tant que telles donc je les ai supprimés.

→ NETTOYAGE DE l'HTML :
1) Rendre l'HTML VALIDE et pouvoir l'indenter convenablement. Pour cela :
-Fermeture des balises meta <meta blablabla />
-Fermeture des balises
 (1 occurrence) et <hr/>(391 occurrences) par un tout rechercher <hr> /tout remplacer <hr/> sur oXygen XML Editor
-Parfois les années se sont perdues au fil du texte et ne sont pas sous balises. Il faut donc le faire (manuellement pour ma part)
exemple : ce qui est tout autre chose ANNÉE 1810. 259 que de le lire et de l'écrire.
Ou mélangées dans la balise de la signature ;

2) Résolution d'une perte de données :
Une partie de l'HTML (lignes 91 à 278 !) s'était perdu dans le copier/coller. Je dois je refaire la manipulation depuis Gallica pour cerner le problème et les récupérer. Cette perte de données s'est apparemment faite entre Sublime Text (vol1) et oXygen (VOL1)

3) Suppression de balises et de données inutiles
- A chaque page, le titre est rappelé : cela fait du bruit. Je vois comment supprimer ces choses inutiles.
→ par des REGEX dans oXygen XML Editor qui prennent directement entre les balises les chiffres qui mentionnent la page et les majuscules qui font référence au titre ou à l'année de correspondance.
A supprimer, toutes les occurrences de ces titres en haut de chaque page (à chaque en-tête)
30 CORRESPONDANCE DE LAMARTINE. → Le titre du livre qui est rappelé avec la page.
ANNÉE 1808. 31 → L'année de correspondance en cours, avec la page

→ Marqueurs
Récapitulatif sur les caractéristiques et MARQUEURS des lettres et de l'HTML de Lamartine, volume 1, d'après le cahier des charges :

MARQUEURS DE LA LETTRE par Camille

Début des lettres : Chiffre romain
adresse au destinataire en minuscules gras
lieu de résidence du destinataire (cette dernière information ne figure pas toujours)
lieu et date d'écriture de la lettre.

Fin de la lettre : signature irrégulière, quand elles est présente, c'est en petites capitales (ALPHONSE DE LAMARTINE ou A. DE L. ou A.L.)

- Correspondance active
- Distinction entre les lettres grâce à des chiffres romains
- Alinéas marquent des paragraphes
- Le destinataire est indiqué en gras et en minuscule. Au dessous à la ligne, en minuscule et en maigre, le lieu de destination de la lettre : « **A Monsieur Prosper Guichard de Bienassis**/A Bienassis, par Crémieu (Isère) » L'habitude de faire systématiquement figurer le lieu de destination de la lettre en deuxième ligne se perd à partir du tome 4.
- En principe date et lieux à droite sur le modèle (« Lieu, jj mois AAAA. »). Quelques exceptions où le lieu n'est pas mentionné. (tome 1 p ; 35) et où la date est imprécise (tome 4 p. 8)
- Structure de la lettre : La structure de la lettre n'est pas régulière. Les formules de politesse de début et de fin sont parfois détachées (tome 4, p. 25) mais dans la grande majorité des cas, elles sont intégrées au corps de la lettre (tome 4 p. 12, tome 1 p. 39, tome I, p. 55)

Remarques sur les marqueurs :
1) a) PREFACE
Supprimée
1) b) TITRES
-La correspondance est divisée selon les années par des titres. (Attention, ils ressemblent beaucoup aux titres des pages, sans comporter le numéro de page, c'est qui les distingue.
ANNÉES 1807 ET 1808
SUPPRIMÉS = ce qui compte, c'est d'extraire les lettres.

Arthur me dit qu'il vaut mieux les garder pour les métadonnées si jamais une lettre n'avait pas de date ! 2 juin 2020.

2) SOUS-TITRES

<hr /><hr /> CORRESPONDANCE DE LAMARTINE ANNEE 1807

→ Mis sous plusieurs balises

SUPPRIMÉS = ce qui compte, c'est d'extraire les lettres.

3) LETTRES

Chaque lettre commence sur une nouvelle page (pas en milieu de page, ni les unes à la suite des autres). Donc après une balise <hr/> ?!

A) DÉBUT DE LA LETTRE

a) Chiffre romain pour le début de la lettre

b) Destinataire

c)Adresse/ lieu d'écriture (parfois il est écrit « même adresse »

d) Lieu, date

B) CORPS DE LA LETTRE

a) il s'y glisse les **titres des pages** (numéro de page en chiffres arabes + titre en majuscule suivi d'un point pour la page de gauche 4 CORRESPONDANCE DE LAMARTINE. , indication de l'année en majuscule suivi de l'année en chiffres arabes, puis d'un point, et indication du numéro de page en chiffres arabes pour la page de droite ANNÉE 1807. 5)

Attention, parfois la page n'est pas indiquée. Il faut adapter les regex.

b) **Coquilles** à corriger : vais •que

Pas besoin de s'attarder sur les coquilles. Ce qui compte, c'est d'extraire les lettres. Remarque d'Arthur, 2 juin 2020.

c) Les changements de pages sont indiqués par la balise <hr/> (! voir si je la supprime) Je le fais pour une partie.

d) Les changements de paragraphes sont signalés par des balises

C) FIN DE LA LETTRE

a) La lettre se finit souvent(la plupart du temps?) par « Adieu blablabla »

b) Signature de Lamartine : En majuscules ALPHONSE DE LAMARTINE.

ALPHONSE DE LAMARTINE.
Ou ALPH. DE LAMARTINE.
ALPH. DE LAM.
ALPH. DE L.
ALPH. DE LAMART.
ALP. DE LAM.
ALP. DE L.
A. LAMARTINE
ALPHONSE LAMARTINE
ALPHONSE DE LAM.
ALPHONSE L.
AL. DE LAMARTINE.
ALP. DE LAMARTINE.
AL. DE L.
AL. DE LAM.
A. DE LAM.
AL. DE LAM.
A. DE L.

A. de L.
A. L.
A. DE LAM.
>ALONZO DE LAM. <
= 76 ? occurrences pour 97 lettres (XCVII)
Comment faire pour les autres lettres ?
c) Parfois des P.S. après la signature.
Signalés par P.-S. (4 occurrences)

Ou simple paragraphe en plus…

I A monsieur Prosper Guichard de Bienassis A Bienassis, par Crémieu (Isère). Milly, 24 septembre 1807. Mon cher ami, je vois que tu es un homme de parole, et je veux l'être aussi, car on m'a remis ta lettre hier à neuf heures […] de traverser nuitamment toute la maison de ton oncle remplie d'esprits follets et de revenants. <hr /> 4 CORRESPONDANCE DE LAMARTINE. Il y a huit jours que je suis arrivé
à Mâcon ; j'ai fait plus de la moitié du chemin à pied, avec mon petit paquet sur mon dos ; ainsi tu vois que mon voyage n'a guère été plus gai que le tien: je m'en allais tout le long de la route chantant comme un troubadour quelque vieille romance, j'en composais même tout en marchant ; lorsque je trouvais •quelque beau site, je m'asseyais et je le contemplais tout à loisir. […] voyage.. Je suis à présent à la campagne. J'ai chassé […]donnerai point. <hr /> ANNÉE 1807. 5 Je ne te […]désire, « quod sperat omis excidi, hoc, hoc soevius opprimet. » Tout le monde, […] et je t'écris
entre Gresset et Molière. Adieu, mon cher ami, écris-moi le plus tôt possible. J'espère que l'année prochaine nous verra plus liés que jamais, et que ta sincère amitié m'aidera à endormir mes peines présentes dans les songes d'un plus doux avenir. Je t'embrasse de tout mon coeur et suis ton plus sincère ami. ALPHONSE DE LAMARTINE.

Attention
Certaines lettres ont été écrites en deux fois. Comment traiter la chose ?
tienne. 31 octobre. Je te

Suppressions :
-de la ligne 24 à 124 (Rappel de la demande BnF, Préface (intéressante en soi, faut-il la mettre en note ? Ou la mettre quelque part dans le site?) de Valentine de Lamartine qui est à l'origine de cette édition.

-de la table des matières (l.4066 à 4123) qui récapitule les XCVII lettres du premier volume avec chaque destinataire : A monsieur Prosper Guichard de Bienassis
A monsieur Aymon de Virieu
Seuls deux destinataires pour ce premier volume.

-des balises inutiles mentionnant simplement un changement de page
<\/p><hr \/>
→ 289 matches

- des balises inutiles dues à un changement de ligne
-
2 matches
exemple : sont char- mants

- de tous les tirets suivis d'un espace qui coupent les mots en deux
23 matches

-des balises <hr/> quand elles divisent une phrase
40 matches

-des balises quand elles divisent une phrase.
58 matches avec regex.

-de chiffres inutiles et balises ajoutées
p>Adieu. ~~23~~
XCI A monsieur de Virieu Naples, mercredi 22 janvier 1812.
par : XCI A monsieur de Virieu

Remplacement, correction
- de El par Et 5 matches
- de Celte ou celte par Cette ou cette 17 matches
-de moimême par moi-même 2 matches
-de lu par tu 70 matches (espace|lu|espace)
-de tu par lu 11 matches
-de loi par toi 7 matches
-de écrismoi par écris-moi 5 matches
Il est plus facile de nettoyer à mon sens un texte avant sa mise en XML qu'après quant à son contenu.
Après on peut remarquer des erreurs et les corriger sur l'HTML (pour tout matcher d'un coup) puis tout remettre en XML avec Python (en bref, recommencer la manipulation).
A moins qu'il y ait possibilité de le faire directement dans python ?
Pour l'instant, se concentrer sur l'extraction et non sur la correction. Arthur, 2 juin 2020.

Ajout de balises
p>ALPH. DE LAM. LXII
il manque des balises pour séparer la signature des chiffres romains qui désignent une nouvelle lettre.
Je l'ajoute avec une regex. 65 matches.

Remarques de fin :
Avec quelle granularité corriger les fautes laissées par l'OCR de Gallica ? Pour Lamartine, j'avais une estimation de 100 % de réussite et en fait, beaucoup d'erreurs s'y sont glissées comme par exemple des confusions entre les l et les t. (tu océrisé en lu).

Comment s'y prendre pour nettoyer les textes ? Ce n'est pas mon soucis. Ce sera fait par la suite (ou pas fait, selon le nombre d'erreurs).
Selon le manuscrit (ou plutôt le livre dactylographié de correspondance) de base, les fautes ne sont pas les mêmes… (comme les titres qui se mêlent au corps du texte parce qu'il y en a un pour Lamartine ,ce qui n'est pas le cas pour Sand). Difficile de généraliser les corrections dans ce cas-là. Prendre le temps pour « perdre du temps » à penser la chose au début pour en gagner en fin de compte. C'est le cas pour les REGEX.
Automatiser les regex, le nettoyage de l'html ?
Dans quel mesure nettoyer les balises (hr, br…)
Il y a également du latin. Doit-on le mettre en italiques ? Beaucoup de mots, d'ouvrages sont indiqués en italiques dans les lettres de Lamartine.cf. Lettre 37.
A ce propos, y aura-t-il une ODD par auteur ou une ODD pour tous les auteurs ?
Y en a-t-il déjà une ?
→ Il faut une ODD pour tous les auteurs d'après Monsieur Clérice pour le CRHXIX.

→ Est-ce que c'est utile de supprimer certaines balises dans html (comme br) puisqu'on fait le tri dans python ?!

Mardi 2 juin 2020, J5

POUR PYTHON

- Les destinataires :
A monsieur Guichard de Bienassis, 22 matches (remplacement de Guiehard par Guichard)
A monsieur Prosper Guichard de Bienassis 2 matches
A monsieur$\n[^a-z]+(Guichard de Bienassis) 3 matches
Pour 37 lettres à lui:/

A monsieur$\n[^a-z]+(de Virieu) 2 matches (matche quand changement de ligne)
A monsieur$\n[^a-z]+(Aymon de Virieu) 2 matches
A monsieur Aymon de Virieu 33 matches
A monsieur de Virieu 4 matches
A$\n[^A-Z]+(monsieur Aymon de Virieu) 8 matches
A$\n[^A-Z]+(monsieur Guichard de Bienassis) 4 matches
A$\n[^A-Z]+(monsieur de Virieu) 1 match
A monsieur Prosper de Bienassis 1 match

=82 SUR 97

IX A monsieur Prosper. de Bienassis → J'enlève le point pour pouvoir ensuite mieux matcher dans les REGEX.

Attention, parfois il manque le chiffre romain qui indique le changement de lettre.
ALPHONSE DE LAMARTINE. A. monsieur Prosper Guichard de
Bienassis A Bienassis. Mâcon, 3 octobre 1807. Mon cher ami, il est minuit ; tout le monde dort
Je l'ajoute pour la lettre II.

Pour matcher malgré les changements de ligne
A monsieur$\n[^a-z]+(Guichard de Bienassis)