Rapport de fin de stage
Labex Obvil

Lucie Slavik, stagiaire M2 TNAH ENC
19 mai 2020 – 29 juillet 2020
Stage filé, deux jours par semaine
TOTAL de 20 JOURS

Résumé du travail accompli

Lors de ce stage, j'ai travaillé sur le projet ELICOM porté par Obvil, sous la direction d'Arthur Provenier.
J'ai été chargée de l'extraction des correspondances et de la correction des fichiers XML générés. C'était plus une correction technique (correction des balises) qu'orthographique.

Si l'on considère le cahier des charges en regard de ce résumé, voici les correspondances que j'ai extraites avec du code python, m'inspirant d'un squelette déjà réalisé par Arthur et que j'ai adapté à chacun des auteurs :

- LAMARTINE. Rédaction du code python, extraction et correction du premier volume de correspondance. Total de **97 fichier**s extraits et corrigés.

J'ai commencé à extraire l'HTML du deuxième volume et à le nettoyer.

- LAMENNAIS. Rédaction du code python, extraction du premier volume (comprend trois tomes) et correction partielle. J'ai corrigé 80 fichiers. Il faut encore corriger 124 fichiers, à savoir les fichiers164 à 195, 2 à 9, 17 à 99. Total de **204 fichiers**. Certains sont nommés a, b, c car ils ont dû être divisés *a posteriori.*

- PROUDHON. Rédaction du code python, extraction du premier volume et correction des lettres. Total de **87 fichiers** extraits et corrigés.

Des rapports ont été faits au fur et à mesure pour indiquer le travail réalisé.
Les fichiers XML ont été mis sur github.
Reste à transmettre certains fichiers où sont recensés les REGEX mais j'aimerais les retravailler avant.

Au total, 388 lettres ont été extraites, dont 264 fichiers ont été corrigés durant ces 20 jours de stage.

Procédures pour y arriver

Pour chaque correspondance, j'ai procédé de la même manière (l'ordre ci-dessous est à titre indicatif) :

A) HTML
1) Extraction de l'OCR en HTML sur Gallica
2) Rendre l'HTML valide sur oXygen et bien indenté
3) Nettoyage de l'HTML avec des REGEX pour supprimer les numéros de pages, en-têtes de pages, certaines balises inutiles (<hr/>
) etc. C'était donc un nettoyage technique si l'on peut dire. En parallèle, j'ai également corrigé certaines fautes d'océrisation afin de tout matcher directement dans l'HTML ce qui est plus simple que de le faire après un fichier XML après l'autre.
Exemple de fautes : quille au lieu de quitte ; celte au lieu de cette.

Ajout de balises manquantes, par exemple pour séparer la signature de Lamartine des chiffres romains marquant le début d'une nouvelle lettre, empêchant ainsi une bonne extraction des lettres.
4) Repérage des destinataires et mise en place de REGEX pour les matcher.
5) Repérage des marqueurs de la lettre pour l'extraire en python

B) XML
1) Écriture du code python en s'inspirant du squelette commun *extraction-elicom.py*
2) Création d'un dossier dump pour récupérer tous les fichiers XML.
3) Ecriture des REGEX python
4) Essais et extraction finale
5) Correction des balises
Normalisation des dates et des auteurs
Remplacement des balises <p> par des balises <l> quand il y a des vers, ce qui nécessite de consulter le pdf en regard (beaucoup le cas pour Lamartine, un peu moins pour Lamennais).
Création de balises <l>
Suppression des notes restées dans le corps du texte malgré les regex, en raison d'une mauvaise océrisation des chiffres des notes de bas de page. Cela nécessite également d'avoir le pdf de la correspondance en regard.
6) Mise en ligne sur github https://github.com/OBVIL/elicom/tree/master/extraction_cor_stage2020 (vérifier que tout a bien été transmis à OBVIL)

Remarques sur le stage

Le stage a été en quelque sorte un défi étant donné qu'il a été entièrement réalisé en télétravail. Je tiens à remercier à cette occasion mon tuteur qui s'est rendu très disponible et m'a consacré beaucoup de temps pour répondre à mes nombreuses questions et m'indiquer le travail à faire.

La majeure difficulté a été pour moi la rédaction du code python. J'ai beaucoup apprécié par la suite le fait de voir le résultat de mes efforts une fois que les fichiers XML étaient extraits. J'ai donc appris à mieux maîtriser python, notamment via le terminal pour voir au fur et à mesure les effets de chaque bout de code. J'ai beaucoup apprécié cela.
Pour la personne qui prendra la suite, il faudra d'ailleurs améliorer le code python pour Lamartine afin de mieux séparer les balises de la dateline (date et lieu) car cela est facilement possible, nous l'avons fait pour les autres correspondants.

J'ai également apprécié l'utilisation de github, même si je n'ai pas toujours su l'exploiter au mieux. À ce propos, je regrette les nommages ambigus de certains fichiers sur github (dump corrigé ou non corrigé). J'espère améliorer ce point à l'avenir.

Une autre difficulté a été le choix des balises et l'unité entre les différents corpus. Tous n'ont pas les mêmes structures et caractéristiques et il était difficile d'unifier tout cela sans ODD et d'avoir une vue d'ensemble sur cet immense corpus. Plus vite l'ODD sera réalisée, plus facilement le problème d'unité sera réglé car nous seront contraints par un schéma.

Dans l'ensemble, ce stage m'a vraiment aidé à consolider mes connaissances en XML et m'a par là aidé pour mon autre stage au CRHXIX. J'ai beaucoup apprécié le fait de travailler sur des correspondances du XIX^{e} siècle et je me réjouis par avance de voir le fruit de ce travail d'équipe sur ELICOM.