Rapport de fin de stage
CRHXIX

Lucie Slavik
4 mai 2020 – 31 juillet 2020
3 jours par semaine.
TOTAL DE 31,5 JOURS travaillés (jours fériés et congés exclus)

I – Synthèse des tâches accomplies

A- Les premiers jours de stage ont été consacrés à la **prise en main du projet.** J'ai rempli des tableaux excel pour mieux cerner les correspondances en ma possession, leur qualité de numérisation, lesquelles il fallait traiter en priorité.
Ces tableaux ont plutôt une utilité personnelle.
Il vaut mieux se référer dans l'ensemble à celui qui a déjà été fait par Monsieur Matthieu Brejon de Lavergnée.

B- Prise en main de **Transkribus** :

~ Importation de la correspondance que j'ai choisi de traiter en priorité.
Ici, j'ai rencontré le problème (toujours à régler) de certains PDF trop lourds, notamment les SIM, et qu'il faudra diviser. Ils sont trop importants en général pour le faire gratuitement sur internet.
Il existerait certains outils pour le faire.

~ Copier/coller des transcriptions des étudiants et stagiaires
Je n'ai pas toujours pu avoir la vigilance nécessaire pour vérifier la qualité des transcriptions, car il fallait que je m'attache à la partie technique.
Il faut encore déterminer si on transcrit en ajoutant les accents ou non.

~ Entraînement d'un modèle pour la reconnaissance de l'écriture de Le Play.
Je suis arrivée à plus de 20 000 mots.
Le taux de réussite est assez satisfaisant, mais il serait bon de continuer l'entraînement.

~ Réalisation de tutoriels pour la prise en main de Transkribus par l'équipe du CRHXIX.

~ Relecture des transcriptions avant l'export
Fait que pour quelques pages. C'est encore un long chantier.

~ Export en format TXT, TEI.
Fait que pour quelques pages. A continuer.

C- Ébauche d'un **cahier des charges**
Écriture d'un cahier des charges en vue d'un futur site pour la mise en ligne de l'édition numérique de correspondance.

~ Rédaction d'un rapport qui fait le point sur le projet. Ce n'est pas un réel cahier des charges. Beaucoup de choses sont à modifier.

~ Difficultés rencontrées dans la rédaction de ***Users Stories***

~ **Architecture du site** faite dans sa globalité. Beaucoup de choses sont à revoir.
Sources d'inspiration : le site de la correspondance de Flaubert
https://flaubert.univ-rouen.fr/correspondance/edition/
Celui de la correspondance de Proust http://proust.elan-numerique.fr/
Madame Sophie Lhermitte prendra la suite sur ce point.
> Réalisation du site par la suite à confier plutôt à un prestataire (à confirmer).
Une bonne adresse à connaître, mais peut-être non adaptée à notre projet :
https://www.limonadeandco.fr/

~ Réflexions à peine ébauchée à continuer sur le **référencement naturel** (SEO)

D- Encodage **TEI**

~ Réflexions faites sur le schéma de balises à établir en fonction des attentes du site, avec :

* index des noms de personnes (à la fois historiques et contemporaines)
* index des noms de lieux
* index des ouvrages
* index des événements cités
*index des termes sociologiques leplaysiens, réalisé *a priori* avec l'aide du Professeur Antoine Savoye, spécialiste de Le Play, ainsi que l'équipe du CRHXIX.

~ réalisation de fiches pour la normalisation de ces noms indexés en TEI.

~ réalisation d'un tutoriel pour aider l'équipe du CRHXIX dans l'encodage TEI. Transmission d'un tutoriel pour lier les fichiers XML-TEI à l'ODD.

~ réalisation de l'ODD (One Document Does it all)
https://tei-c.org/guidelines/customization/getting-started-with-p5-odds/

II – Guide pour la suite

A- Chargement des données, transcriptions Transkribus

~ Continuer à importer sur Transkribus

* les manuscrits numérisés
* leurs transcriptions déjà réalisées

en commençant par ceux de la main de Le Play

~ Importer les autres manuscrits non transcrits et les transcrire sur Transkribus.

~ Continuer ainsi l'entraînement du modèle. Une fois qu'il est satisfaisant :

~ Transcrire automatiquement avec Transkribus les lettres avec le modèle Le Play

~ Pour les autres écritures, transcrire sur Transkribus directement. Si c'est un grand corpus, faire éventuellement un autre modèle.

A la fin, EXPORTER toutes les données. Sinon elles restent simplement dans le serveur de Transkribus et nous ne voyons pas le fruit de notre travail. Il faut que nos données soient dans nos propres bases de données.

B- Encodage

~ Nous avons deux choix :

1. exporter en TEI et transformer avec XSLT les fichiers tei de Transkribus en fichiers tei répondant plus précisément à nos attentes.

Pour cela, on peut utiliser les tags sur Transkribus avant l'export, ce qui nous avance pour l'encodage tei.

J'ai fait le test. Le fichier tei exporté étant très sale, et les noms des balises étant souvent impropres et invalides sur oXygen, j'ai préféré choisir la deuxième solution.

La première solution est en soi toujours envisageable. Il faudrait simplement se pencher plus longtemps sur la question. Un de ses avantages notamment est pour le style (italiques, souligné etc.). On le signale directement sur Transkribus et après on n'a plus besoin de se référer au manuscrit. Cependant, pour beaucoup de choses, en général, je trouve qu'il vaut mieux toujours

travailler avec le manuscrit en regard. Cela aide pour la précision ; même si l'on fera toujours des erreurs, mieux vaut en faire le moins possible.

2. exporter depuis Transkribus soit en format TEI, soit tout simplement en format TXT et créer nous-mêmes les fichiers XML-TEI. Les encoder selon le schéma de balises prédéfini et les lier à l'ODD.

J'ai trouvé cette solution plus simple pour les débutants que nous sommes.

~ Étapes à suivre

1. Sur Transkribus, relire les transcriptions, vérifier si les mots illisibles le sont vraiment, si certains mots n'ont pas été corrigés sans raison (ex: septembre au lieu de 9bre). Selon le choix qui aura été fait, ajouter ou enlever les accentuations et les majuscules aux noms communs.
2. Selon le corpus que l'on souhaite encoder, faire en premier un fichier XML-TEI, xxxx_base.xml, avec les parties qui seront communes à tous les fichiers de ce corpus (le teiHeader, avec la normalisation du nom du correspondant, les index etc.). Sur cette base, encoder une à une les lettres avec le fichier TXT ou TEI exporté de Transkribus.

Faire un dossier par fonds et correspondants. (ex. Fonds de la BnF = 1 dossier. Un sous-dossier par correspondant). A voir si c'est vraiment une bonne pratique.

C- Externalisation du travail de développement

Pour cela, reprendre et développer le cahier des charges.

On pourrait envisager de prendre un stagiaire de l'ENC l'année prochaine. Ce serait un travail intéressant pour le stagiaire. L'appel à stage a lieu vers janvier 2021.

D- Pérennité des données

Cette question sera probablement envisagée dans mon mémoire.