

CRHXIX
Cahier des charges pour le projet scientifique sur
« La correspondance de Frédéric Le Play (1806-1882) : une source pour l’histoire
des sciences sociales en Europe (LEPLAY) »
Version 2 (qui diffère très peu de la version 1)

Première version réalisée par Lucie Slavik, M2 TNAH, ENC
dans le cadre de son stage filé au CRHXIX, mai-juillet 2020
Rédaction : juin 2020

I. Contexte

1. Le CRHXIX

Le Centre de recherche d’histoire du XIXe siècle (CRHXIX), fondé par le Professeur Louis Girard (1911-2003), relève des deux universités de Paris-1 et Sorbonne-Université (à savoir l’ex Paris-4)¹. Il se compose de plus de 150 personnes, que cela soit des professeurs des universités, de maîtres de conférences, de divers professeurs agrégés, d’allocataires, moniteurs, ATER (Attaché temporaire d’enseignement et de recherche), d’IATOS (ingénieurs, administratifs, techniciens, ouvriers de service), de chercheurs associés et doctorants². Il met à la disposition des chercheurs sa bibliothèque riche de près de 8000 ouvrages³.

Le CRHXIX développe actuellement ses recherches autour de quatre axes principaux⁴, à savoir en premier lieu, « Le “resserrement des sociétés” : circulations globales et pratiques locales au XIXe siècle », en second lieu « Du moral au social : pratiques et théories de l’enquête. Autour des archives du mouvement leplaysien », puis sur « Citoyennetés, sûretés, sécurités, souverainetés », et enfin « Images, imaginaires et écriture de l’histoire. »

2. Le projet Le Play

C’est donc dans ce cadre que s’inscrit notre projet scientifique sur « La correspondance de Frédéric Le Play (1806-1882) : une source pour l’histoire des sciences sociales en Europe », « porté par Sorbonne Université et le Centre d’histoire du XIXe siècle (EA 3550)⁵ ».

-
- 1 *Présentation du Centre*, Site web de l’Université Paris-1 Panthéon Sorbonne, URL : <https://www.pantheonsorbonne.fr/unites-de-recherche/crhxix/aboutthecenter/>, (visité le 18 juin 2020)
 - 2 *Présentation du Centre, l’équipe*, Site web de l’Université Paris-1 Panthéon Sorbonne, URL : <https://www.pantheonsorbonne.fr/unites-de-recherche/crhxix/about-us/faculty/>, (visité le 18 juin 2020)
 - 3 *Présentation du Centre, la bibliothèque*, Site web de l’Université Paris-1 Panthéon Sorbonne, URL : <https://www.pantheonsorbonne.fr/unites-de-recherche/crhxix/about-us/thelibrary/>, (visité le 18 juin 2020)
 - 4 *Présentation du Centre, les activités de recherche du Centre*, Site web de l’Université Paris-1 Panthéon Sorbonne, URL : <https://www.pantheonsorbonne.fr/unites-de-recherche/crhxix/about-us/research-activities/>, (visité le 17 juin 2020)
 - 5 Note rédigée par Monsieur Matthieu Brejon de Lavergnée et adressée à Lucie Slavik, 17 octobre 2019.

3. Acteurs

L'équipe de recherche de ce projet autour de « *La correspondance de Frédéric Le Play (1806-1882) : une source pour l'histoire des sciences sociales en Europe (LEPLAY)* » est « dirigée par Matthieu Brejon de Lavergnée, maître de conférences HDR, et constituée de Philippe Boutry, PR ; Eric Anceau, MCF HDR ; Rémy Hême de Lacotte, MCF ; Marie-Laure Massei-Chamayou, MCF ; Sophie Lhermitte, ingénieur d'études⁶. »

De nombreux étudiants, doctorants et stagiaires ont participé de près ou de loin à l'avancement du projet. L'on notera particulièrement Margaux Faure, qui a effectué de nombreuses relectures de transcriptions et qui a réalisé elle-même des transcriptions, ainsi qu'Edouard Coquet. Tous deux ont participé aux journées de formation à l'École Nationale des Chartes (ENC) autour de Transkribus.

4. Soutiens financiers

« Le projet a reçu un financement du GIS Collex-Persée dans le cadre des appels à projets lancé en 2018 pour soutenir la numérisation de corpus pour la recherche. Il est aussi soutenu par le Labex Ecrire une histoire nouvelle de l'Europe (EHNE)⁷. »

Les formations à l'ENC ont été financées par un prêt Collex.

Pour mener à terme ce projet dans les meilleures conditions, nous aurions besoin de financements plus importants, dans l'idéal un prêt Collex.

II. Objectifs du projet scientifique : numérisation et édition numérique de la correspondance de Frédéric Le Play

1. Description du corpus à numériser⁸

Le corpus de lettres à numériser représente 2091 lettres échangées entre Frédéric Le Play et 94 correspondants entre 1837 et 1882. Le corpus est conservé dans trois fonds principaux : fonds Le Play de la Société d'économie et de sciences sociales conservée au CEDIAS-Musée social ; fonds Le Play de la Bibliothèque de l'Institut de France ; fonds Le Play de l'association du château de Ligoure (Haute-Vienne).

A côté de ces trois massifs, la correspondance est particulièrement éclatée dans de nombreuses bibliothèques et centres d'archives, tant en France (Archives nationales, Bibliothèque nationale de France, bibliothèque Thiers, etc.) qu'à l'étranger (Angleterre, Belgique, Italie, Russie, etc.).

Le travail de numérisation a déjà été engagé, en particulier dans les fonds les plus importants.

Les fonds complémentaires sont les ouvrages de Le Play, les enquêtes publiées, les revues du mouvement. Une large partie est disponible sur Gallica (*La Réforme sociale en France*, 1864 ; *Les Ouvriers européens*, 6 volumes entre 1877 et 1879 ; revue *La Réforme sociale*, 1881-1934, etc.)

⁶ M. Brejon de Lavergnée, *Idem*.

⁷ M. Brejon de Lavergnée, *Ibidem*.

⁸ L'intégralité de cette partie est tirée de la note déjà citée, rédigée par Monsieur Matthieu Brejon de Lavergnée et adressée à Lucie Slavik, 17 octobre 2019.

Un premier inventaire sommaire a été réalisé en 2006 et publié dans *Les Etudes sociales* (n° 142-144, p. 231-247). L'inventaire précis pièce à pièce reste à réaliser.

2. Objectifs scientifiques⁹

Notre projet scientifique consiste à éclairer les conditions d'élaboration des enquêtes qui se multiplient en Europe à partir des années 1840 (Buret, Villermé, Blanqui, Engels...) et qui s'interrogent sur les mutations du monde du travail et la naissance d'un paupérisme de masse, ce que l'on appelait alors la « question sociale ». Or ces enquêtes publiées gommant toute la **phase préparatoire**, pourtant riche d'enjeux aussi bien épistémologiques qu'idéologiques. Les correspondances représentent une source inexploitée pour étudier cet aspect. En travaillant sur les pratiques savantes d'un groupe d'enquêteurs réunis au sein d'une école « sociologique » relativement homogène, nous pensons pouvoir contribuer de manière pluridisciplinaire à l'histoire des sciences sociales. L'« école » leplaysienne jouit d'un regain d'intérêt historiographique mais sa mémoire demeure relativement effacée par l'école durkheimienne concurrente. Notre projet de recherche promet donc des **résultats scientifiques neufs**¹⁰.

Il s'agit donc de **donner accès** à ces fonds dispersés et parfois peu accessibles en les numérisant et en les mettant en ligne, et de transcrire les lettres pour produire une édition électronique du corpus.

III. Moyens employés et mise en pratique des objectifs

En vue de la mise en ligne, plusieurs étapes sont à suivre. Tout d'abord, il faut s'assurer de la matière première que sont les **manuscrits** : avons-nous toutes les numérisations en notre possession, leur qualité est-elle suffisante ? De plus, les transcriptions ont-elles été réalisées ? Sinon, quelle procédure suivre pour réaliser ce travail ?

Ceci fait, nous nous intéresseront plus précisément à la mise en ligne. Pour cela, il s'agit d'avoir une **vue précise des objectifs poursuivis par la présente édition**. Pourquoi cette mise en ligne, à qui s'adresse-t-elle, à quels besoins et quelles requêtes doit-elle répondre ? Pour cela, nous établirons des « *Users Stories* ¹¹ » pour répondre au mieux aux besoins des utilisateurs. Grâce au fait de se

9 *Idem*. Les éléments soulignés en gras l'ont été par la stagiaire

10 À ce sujet, voir *Présentation du Centre, les activités de recherche du Centre*, Site web de l'Université Paris-1 Panthéon Sorbonne, URL : <https://www.pantheonsorbonne.fr/unites-de-recherche/crhxix/about-us/research-activities/area2/>, (visité le 18 juin 2020)

« L'objet d'un axe *Du moral au social : pratiques et théories de l'enquête. Autour des archives du mouvement leplaysien* est triple : **rendre accessible** un corpus d'enquête original et encore sous-exploité **par une saisie numérique et une édition en ligne** ; **appréhender, dans les procédures d'enquête, le lien entre société et morales du XIX^e siècle** : Le Play lui-même relève, on le sait, pour partie du positivisme comtien et pour partie du traditionalisme bonaldien, mais ses disciples et imitateurs obéissent à des logiques diverses au point de former pour certains un courant dissident de la science sociale ; **analyser enfin le legs de l'enquête leplaysienne** aux sciences sociales en gestation au tournant des XIX^e et XX^e siècle, sociologie en premier lieu, mais aussi psychologie sociale et histoire des mentalités collectives.

Cet axe, centré sur l'Europe et les États-Unis des années 1850-1920, aurait une dimension éditoriale, historique et historiographique ; il pourrait réunir, autour de l'archive leplaysienne, les **contributions de plusieurs disciplines**, histoire, sociologie, droit, philosophie, économie, statistique. »

NB ; Les éléments soulignés en gras l'ont été par la stagiaire

11 *Récit utilisateur*, dans *Wikipedia*, Page Version ID 6474167, 2020, URL : https://fr.wikipedia.org/wiki/R%C3%A9cit_utilisateur, (visité le 18 juin 2020).

projeter dans l'utilisation du site, nous pourrions établir ainsi un encodage plus pertinent, car on n'encode pas pour encoder mais dans le but d'exploiter cet encodage de façon pratique.

Une fois que nous nous serons fait une idée précise du site attendu, nous pourrions donc **encoder**. L'heure sera donc à la partie TEI, avec l'établissement d'un schéma et d'une ODD (tous ces termes sont expliqués ci-dessous).

Quand l'encodage sera réalisé, il sera temps d'**externaliser le travail de développement**.

Une question sous-jacente à toutes ces étapes est celle de la **pérennité des données**. À titre personnel, je ne suis pas sûre d'avoir le temps de la traiter, mais il faudra le faire pour le projet.

Enfin, il faudra considérer si l'on ne peut pas **inscrire le projet dans d'autres projets** pour en faciliter la réalisation.

- Soit on obtient un prêt Collex et la question ne se pose pas.
- Soit on fait un partenariat avec le Labex Obvil, pour profiter de leurs compétences techniques et de leur soutien financier.
- Soit on tente un partenariat avec l'IRHT (cf. courriel de Monsieur Clérice 17 juin 2020).

1. Manuscrits et transcriptions

a. Numérisations

Comme souligné plus haut, un travail de numérisation a déjà été engagé. À ce propos, si cela n'a pas déjà été réalisé, il serait bon de faire un point sur ces numérisations.

Par rapport au dossier qui a été transmis à la stagiaire lors de son stage en télétravail¹², les numérisations sont de qualité variable et certaines lettres ont été simplement prises en photos et sont d'une qualité médiocre : elles ne pourront donc pas être utilisées pour la future mise en ligne ni même pour l'entraînement du modèle dans Transkribus, qui nécessite une qualité correcte.

> Je viens d'écrire un courriel à Monsieur Rémy Hême de Lacotte pour lui demander s'il possède un point sur les numérisations qui ont été réalisées et celles qui restent à faire.

Selon sa réponse, je ferai le point moi-même (s'il me reste du temps, et selon les conseils de mes tuteurs) ou je reprendrai le sien pour le mettre ici.

> Le fichier en question vient de m'être transmis. Reconsidérer donc cette partie en fonction. (25 juin 2020)

Il faut établir :

- les numérisations de correspondance que nous avons en notre possession.
- les photos floues ou les numérisations de mauvaise qualité qu'il faudra refaire.
- quelles photocopies de manuscrits nous avons en notre possession (à voir au CRHXIX sur place) et en demander les numérisations pour le site.
- les manuscrits dont nous n'avons jamais eu la trace et dont il faut faire la demande de numérisation.
- la correspondance de Le Play qu'il y aurait sur Gallica et dont il faudrait demander la numérisation des manuscrits en regardant où se situent les sources. (J'ai un doute sur ce point).

¹² Le dossier a été transmis par *WeTransfer* à la stagiaire qui était en télétravail du fait du confinement dû au Covid-

19. Tout ne lui a pas été transmis, mais une grande partie.

À partir de cela, nous pourrions **estimer le coût des numérisations**, selon le prix par page. À titre d'exemple, aux Archives Nationales, le coût par page A4 est de 0,30 euros pour une reproduction en noir et blanc, et de 0,60 euros pour une reproduction en couleur, avec des prix plus avantageux selon le nombre de pages demandées¹³.

En vue de l'édition numérique de correspondance, il faut privilégier les numérisations en couleur pour le design du site. La majorité, si ce n'est l'intégralité des sites d'édition numérique de correspondance possèdent des numérisations en couleur. Le but de mettre en ligne le manuscrit est de coller au plus près de la réalité matérielle. Donc **privilégions la couleur**.

À ce propos, nous possédons, notamment pour les fonds Peruzzi et Loyson, des numérisations en noir et blanc, qui sont d'ailleurs de très bonnes qualité. Voir si à long terme on les remplace par des numérisations en couleur (ce qui aura un coût à la fois de numérisation mais aussi de manipulation des documents si on utilise Transkribus). À première vue, je dirais de les garder et donc de ne pas refaire des numérisations couleur, afin de ne pas compliquer le projet qui est déjà assez important. Tout dépend si nous bénéficions par la suite d'un soutien financier plus important. Dans ce cas-là, nous pourrions reconsidérer la question.

b. Transkribus

Une fois que la matière première, à savoir les manuscrits, est assurée, il s'agit d'effectuer les transcriptions.

Transcriptions manuelles

Nombre de manuscrits ont déjà été transcrits manuellement, comme souligné plus haut, que cela soit par des étudiants, des doctorants, des stagiaires. (*VOIR COMBIEN / qq centaines*)

Or, la transcription s'avère être parfois un exercice difficile : les hommes du XIX^e siècle disposent de certains codes d'écriture ou d'abréviations¹⁴ qui ne nous sont pas familières. Par ailleurs, nombre de mots sont malaisés à lire, du fait des différentes écritures. Je pense par exemple à celle de Jules Baroche¹⁵.

Même avec un œil exercé, et des mains habituées à taper rapidement, il faut compter au moins cinq minutes par page pour la transcription, sans compter la relecture ;

Ce qui fait pour la transcription uniquement, 12 pages en une heure, donc près de 200 heures de transcriptions pour 2000 lettres.

Ainsi, sur un corpus aussi important, comportant plus de deux-mille lettres, la question se pose d'automatiser les transcriptions en utilisant les moyens technologiques dont nous disposons aujourd'hui notamment avec l'apprentissage machine¹⁶. C'est ainsi que nous avons fait le choix de nous tourner vers Transkribus.

13 *Tarifs reproduction documents*, Site web des Archives Nationales, URL : http://www.archives-nationales.culture.gouv.fr/documents/10157/11409/2013_12_tarifs_reproduction_document.pdf/7f853b31-f601-4d6a-a673-0ba85d2bf115, (visité le 18 juin).

14 Comme par exemple pour la date qui est souvent chiffrée différemment. Ainsi, ils écrivent 7bre pour septembre etc.

15 Voir le pdf SIM MS 6062, les premières pages. Jules Baroche par exemple, lorsque deux « s » se suivent, allonge le premier. (p.7 du pdf).

16 Voir à ce sujet *Apprentissage automatique*, dans *Wikipedia*, Page Version ID 31947, 2020, https://fr.wikipedia.org/wiki/Apprentissage_automatique, (visité le 19 juin 2020)

Transcriptions automatisées

Transkribus est un outil est un logiciel de transcription collaborative¹⁷. Il permet de transcrire manuellement des manuscrits, mais également d'entraîner un modèle qui permettra à Transkribus de reconnaître par la suite l'écriture et d'effectuer lui-même les transcriptions.

La stagiaire actuelle a donc chargé une partie des manuscrits dans Transkribus, avec les transcriptions qui avaient déjà été réalisées, afin d'entraîner un modèle pour l'écriture de Le Play. Après une dizaine de jours où elle s'est attelée à ce travail et à rentrer environ 20 000 mots, c'est à dire, le nombre de mots conseillés pour entraîner un modèle, le résultat ne s'est pas avéré être optimal. Voir à ce sujet le récapitulatif qui a été fait dans le fichier Point_Transkribus.odt. Sur le dernier modèle entraîné, on obtient un taux d'erreur (CER, c'est à dire Character Error Rate) de 3,52 % pour le Training Set, divisé en 50 époques¹⁸, et de 6,09 % pour le Test Set, ce qui est en théorie déjà un bon résultat. En effet, comme nous le renseigne le Wiki, « les résultats avec un taux d'erreur (CER) de 10 % ou moins peuvent être utilisés pour les transcriptions automatisées. Les résultats avec un taux d'erreur de 20 à 30 % sont suffisants pour travailler avec la recherche de mots-clés. ¹⁹» Nous sommes donc sur la bonne voie.

Néanmoins, ceci nécessite une relecture attentive, car nous en sommes en moyenne à une faute tous les cinq mots au moins dans la transcription automatique.

Deux possibilités s'ouvrent désormais : soit on entraîne encore le modèle, pour obtenir un meilleur taux de réussite, en faisant bien attention à prendre une écriture de Le Play qui couvre bien toute sa vie, car elle évolue avec le temps, (ce qui d'ailleurs nécessite d'avoir toutes les numérisations en mains) et ceci avec le risque de n'avoir plus de lettres à transcrire une fois que l'on aura fini d'entraîner le modèle ; soit on laisse ainsi, mais il faudra effectuer une relecture et une correction qui risque de prendre presque autant de temps que la simple transcription : en effet, si l'on doit corriger une faute tous les cinq mots, le fait de rectifier quelque chose d'écrit prend je pense autant de temps que de bien transcrire directement et à la main. À voir.

Quoiqu'il en soit, même si Transkribus n'est pas utilisé pour la transcription automatisée, avec le modèle, l'on peut toujours l'employer pour la transcription manuelle. En effet, Transkribus a le grand avantage de pouvoir être utilisé par plusieurs personnes du même projet à la fois. Ces dernières peuvent transcrire et leurs transcriptions resteront disponibles et ouvertes à toutes les personnes du groupe, ce qui permet de ne pas perdre les transcriptions dans l'ordinateur de l'un ou de l'autre. Par ailleurs, on peut ainsi mieux coller au texte. On y trouve donc nombre d'avantages, même si l'on ne pousse pas la machine jusqu'au bout de ses capacités. Transkribus est donc un excellent outil pour la transcription en équipe.

De toutes façons, Le Play a une correspondance dont le nombre de lettres varie selon les correspondants, et si la majorité des lettres dont nous disposons sont écrites de sa main, beaucoup

17 Pour en savoir plus, voir *Accueil*, Site web de Transkribus, URL : <https://transkribus.eu/Transkribus/>, (visité le 19 juin 2020).935516

18 « le "Nombre d'époques" fait référence à la fréquence à laquelle les données de formation sont évaluées. Si vous augmentez le nombre d'époques, le processus de formation sera plus long ... » in Wiki, Transkribus, URL : https://transkribus.eu/wiki/images/8/84/Comment_entra%C3%Aener_un_Mod%C3%A8le_dans_Transkribus.pdf, (visité le 19 juin 2020).

19 *Idem*.

sont également des lettres passives. Certaines sont nombreuses, comme celles par exemple de Napoléon-Joseph Bonaparte²⁰, mais d'autres sont rares, comme celles de Louis-Joseph Buffet²¹.

Reste enfin la question de l'export des données, une fois que les transcriptions ont été réalisées, pour les encoder en vue de l'édition numérique de correspondance qui est tout l'objet de notre projet.

2. Un édition numérique accessible à tous

a. Besoin des utilisateurs

Le but de ce projet, nous l'avons dit plus haut, est de rendre accessible un savoir. Il s'agit véritablement de donner accès aux fonds dispersés de Le Play, afin de favoriser les recherches dans ce domaine et accroître ce savoir sociologique. Pour cela, il est important de savoir ce que veulent les utilisateurs, quel public nous ciblons, quelles personnes seront intéressées par cette édition numérique et quels seront leurs besoins.

C'est tout l'intérêt des « *Users Stories* » ou « récits utilisateurs » de nous éclairer sur ce point. À nous de nous projeter dans la démarche des futurs utilisateurs de notre plateforme :

Je suis un étudiant qui ne connaît Le Play que de nom. En allant sur ce site, je voudrais voir une rapide **présentation du personnage** et voir les personnes qui étaient en lien avec lui. Pour cela, il serait intéressant d'avoir un **index des correspondants** de Le Play. Pour mieux mesurer la profondeur de ces liens, une approche quantitative du nombre de lettres serait intéressante : il me serait bon de savoir **combien de lettres** Le Play a échangées avec tel ou tel correspondant.

Je suis un chercheur en sociologie, je consacre une de mes études à Le Play et je voudrais étudier la genèse de son ouvrage sur *La Réforme Sociale*. Pour cela, il serait intéressant de pouvoir faire une **recherche par mots-clés** et voir dans quelles lettres Le Play évoque son **projet de livre** ou se cite lui-même.

Je suis un chercheur en histoire et je voudrais mettre en lumière le rôle des enquêtes dans l'approche sociologique de Le Play. Pour cela, il me serait intéressant aussi de pouvoir faire une recherche avec le nombre de fois où apparaît le **mot enquête, monographie ou tout autre terme approchant** dans ses lettres. Par ailleurs, dans ce même but, il serait intéressant pour moi de cerner si les lettres de Le Play mentionnent ses **voyages à l'étranger**, quels pays il a eu l'occasion de visiter et d'étudier, de même pour les **régions de France**, et le **nombre de ces occurrences** par année. Ceci nécessite une **indexation des noms de pays et de lieux**.

Je suis un chercheur en histoire religieuse. Je voudrais voir dans quelle mesure Le Play s'inscrit dans les diverses phases de retour au catholicisme chez les intellectuels français à la fin du XIXe siècle et au début du XXe siècle, et même s'il n'en serait pas un des précurseurs. Pour cela, il me serait intéressant de voir ses références à la religion en général, et plus particulièrement au catholicisme mais aussi à toutes ces nouvelles églises, telle que l'église néo-gallicane indépendante fondée par le père Hyacinthe Loyson²² avec lequel il entretint pendant un moment une

20 On dénombre 98 lettres de Napoléon-Joseph Bonaparte à Frédéric Le Play.

21 Louis-Joseph Buffet n'adresse que deux lettres à Le Play.

22 *Hyacinthe Loyson*, dans *Wikipedia*, Page Version ID 935516, 2020, URL : https://fr.wikipedia.org/wiki/Hyacinthe_Loyson, (visité le 25 juin 2020).

correspondance abondante jusqu'à leur rupture, ou encore ses liens avec les saint-simoniens²³. Pour cela, un **index relativement précis de concepts religieux ou encore d'églises citées** serait intéressant, sans pour autant se perdre dans une indexation trop poussée ou bien trop large, avec des mots d'ordre trop général. Il serait par exemple inutile d'indexer des mots qui reviennent de toutes façons sans cesse dans sa correspondance, autant la lire intégralement.

Je suis un chercheur anglais et je m'intéresse à l'œuvre de Le Play dans les expositions universelles qui ont eu lieu dans mon pays dans le courant du XIXe siècle. Une recension des lettres actives et passives de Le Play autour des années 1850 m'intéresserait. Je pourrais y voir dans quelle mesure la préparation de l'Exposition était un souci pour Le Play et s'il en parlait dans ses lettres. Pour cela, il faudrait pouvoir mener une **recherche avancée (à facettes)**, qui prenne en compte les correspondants et les années, les fonds de provenance des lettres si possible.

Je suis un chercheur en histoire des idées politiques. Je voudrais étudier la réception des idées et principes de 1789 dans la sociologie de Le Play. Toutes les lettres mentionnant les principes de 1789, ou les concepts de démocratie, liberté etc. m'intéresse. Une **indexation** de ces mots est donc nécessaire.

Je fais une étude sur les relations épistolaires entre les intellectuels français et étrangers. Le Play pourrait être un bon exemple à prendre pour mon étude. Une recherche sur la nationalité de ses correspondants me serait utile. Si cette fonctionnalité était prise en compte dans les recherches à facettes, cela m'aiderait grandement. (VOIR si je supprime cette User Story. Ce n'est pas forcément primordial).

Dans le but de publier un article dans une revue, je fais une recherche sur l'univers mental de Le Play et ses correspondants. Pour cela, je voudrais avoir un **index des ouvrages et des personnages cités** dans la correspondance de Le Play, avec la référence (lettre de qui à qui, quel fonds, lien vers la page internet où se trouve la transcription de la lettre). L'index serait le plus large possible, citant tous les ouvrages évoqués, de n'importe quelle époque (contemporaine ou antérieure au XIXe siècle), de même pour les personnages. Si un petit résumé était fait, une petite présentation du personnage et ses dates clés, et même ses ouvrages majeurs, ce serait l'idéal. Si c'est trop ambitieux, un renvoi vers data.bnf²⁴ (**interopérabilité**) serait déjà un bon début. De même pour les ouvrages. Un lien vers l'un de ces sites ou vers un article de référence disponible en ligne (pourquoi pas sur l'*Encyclopædia Universalis* disponible en ligne, si ces pratiques sont autorisées par le droit) serait déjà une bonne chose, à défaut d'un article rédigé spécialement pour le site.

Je suis étudiant à la Sorbonne et j'aimerais enrichir mes connaissances sur Le Play. Je ne connais pas grand-chose de sa vie et de son œuvre, mais je veux pourtant me plonger dans sa correspondance. Pour cela, des **notes** écrites tout au long de sa correspondance, pour me remettre dans le contexte et m'expliquer de quoi il retourne, m'aideraient grandement à la compréhension de sa correspondance.

Je fais une étude sur le père Hyacinthe Loyson et l'influence de Le Play sur ses prêches lors des Avents à Notre-Dame. Je cherche donc toutes les lettres échangées entre Le Play et Loyson particulièrement entre 1864 (si cette correspondance est existante) et 1869. Il me faudrait donc avoir la possibilité de faire une recherche avancée, par correspondants et année. Les mots clés tels que

23 François VATIN, « TRAVAIL SOCIOLOGIE DU », Site web *Encyclopædia Universalis*, URL : <http://www.universalis-edu.com/janus.bis-sorbonne.fr/encyclopedie/sociologie-du-travail/> (visité le 25 juin 2020).

24 Lien vers Wikipedia non souhaité.

Concile Vatican I ou infailibilité pontificale me seraient également intéressants.

À COMPLÉTER si nécessaire.

Ces mises en perspectives peuvent donc nous aider à comprendre les besoins des utilisateurs et nous indiquer en conséquence comment il nous faudrait encoder et concevoir notre futur site.

b. Arborescence et architecture du site

Sources d'inspiration

Selon les besoins des utilisateurs et pour le succès de notre projet, l'arborescence du site est un élément primordial. Or, l'on ne part pas de rien. Plusieurs sites d'édition numérique de correspondance existent déjà. À nous de nous en inspirer et de les adapter à nos besoins.

Deux sites ont retenu plus particulièrement notre attention. C'est d'une part le site d'édition électronique de la correspondance de Gustave Flaubert²⁵, réalisée par le Centre Flaubert de l'université de Rouen, composante du laboratoire CÉRÉdI (Centre d'Études et de Recherche « Éditer-Interpréter »), et s'inscrivant dans le programme de l'IRIHS (Institut de Recherche Interdisciplinaire Homme et Société), soutenue par Huma-Num, cette « très grande infrastructure visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales ²⁶ » et son Consortium Cahier²⁷. Cette édition a particulièrement plu à Madame Sophie Lhermitte. On y retrouve sur la page d'accueil, le nombre de lettres consultables avec le récapitulatif des lettres écrites par Flaubert et celles qu'il a reçues. Est renseignée également une présentation des lettres par ordre chronologique (à la fois par période et par année), par correspondant, par lieu de rédaction et par lieu de conservation. On y trouve également un bouton qui dirige vers les thèmes des lettres, un index des noms propres, des cartes du voyage en Orient, et des informations sur l'ancienne édition Louis Conard. Enfin, s'y trouvent toutes sortes de renseignements sur l'édition, tels que les principes de transcription, une présentation de l'équipe, des partenaires et des soutiens. On peut également y faire des recherches, simples ou multicritères, en sélectionnant si c'est une correspondance active ou passive, quel correspondant, quelle période, quel lieu de rédaction nous intéresse, dans quel lieu de conservation se trouve le manuscrit, et enfin, nous pouvons renseigner combien de lettres nous voulons voir apparaître par page, si l'affichage comprend l'incipit, si l'on veut qu'elles soient triées par date.

Autant de critères intéressants qui peuvent nous être utiles dans la propre conception de notre édition numérique. Pour ma part, je pense qu'il faut retenir la recherche par correspondant, les lieux de rédaction, éventuellement les lieux de conservation.

Quant au deuxième site qui a été retenu par Monsieur Rémy Hème de Lacotte, c'est une édition numérique de la correspondance de Marcel Proust²⁸. Comme souligné, « L'édition des lettres de Proust a été l'objet, depuis plus de huit décennies, de recherches intenses qui convergent aujourd'hui dans le projet *Corr-Proust*, mené dans le cadre de la collaboration franco-américaine du

25 Accueil, Site web Centre Flaubert, URL : <https://flaubert.univ-rouen.fr/correspondance/edition/>, (visité le 19 juin 2020).

26 Accueil, Site web d'Huma-Num, URL : <https://humanum.hypotheses.org/6089>, (visité le 19 juin 2020).

27 Le Consortium, Site web OpenEdition, URL : <https://cahier.hypotheses.org/le-consortium> (visité le 19 juin 2020).

28 Accueil, Site web Corr-Proust, URL : <http://proust.elan-numerique.fr/>, (visité le 8 juin 2020).

Consortium « Proust21 ».²⁹ »

« Pour chaque lettre ([...]au hasard : <http://proust.elan-numerique.fr/letter/02844>), [on a] la vue numérisée (si elle existe), la transcription, le texte, les notes et les informations. [Il nous] semble n'avoir rien vu d'aussi précis jusque-là.

[On peut] aussi [retenir] l'argumentaire de la page d'accueil sur la pertinence de l'édition numérique : <http://proust.elan-numerique.fr/>. [...] "Elle permet des recherches plus rapides, plus nombreuses et plus précises à travers un corpus balisé en fonction de critères de recherche multiple : dates, lieux, personnes, personnages, titres, citations, informations codicologiques..." Supprimons, pour Le Play, "personnages" et peut-être "citations" et nous aurons déjà une bonne base de départ sur ce qu'il nous faut coder³⁰. » Il y a en plus la possibilité de se connecter.

Nous pouvons donc partir de ces deux exemples pour voir ce qu'il y a à adapter par rapport à nos besoins.

Arborescence ou architecture du site ?

Avant de commencer, il est important de poser une distinction entre ces deux concepts. En effet, pour un site, on parle à la fois d'architecture et d'arborescence. Si toutes deux ont des points communs, il est bon de voir également ce qui les distingue :

Ainsi, l'architecture SEO définit l'organisation du site pour les robots des moteurs de recherche (Google, par exemple), il s'agit donc d'un document uniquement destiné au référencement naturel, tandis que l'arborescence définit l'organisation du site pour les internautes, il s'agit donc d'un document lié à l'expérience utilisateur³¹.

Il sera bon de bien définir l'architecture du site, surtout si l'on se rend compte que celui-ci n'apparaît pas dans les premières pages de recherche Google. Pour cela, il sera peut-être nécessaire de recourir à un prestataire afin d'avoir un bon référencement naturel et donc de nombreuses visites.

Quelques principes pour optimiser le site

En attendant, l'on pourra toujours s'attacher à avoir une arborescence des plus claires et également optimisée pour le SEO. En effet, « l'arborescence [du] site internet aura un impact considérable sur son référencement naturel, que ce soit au niveau global ou dans la hiérarchie des pages [du] site. Les moteurs de recherche analysent les liens entre les pages [du] site pour en comprendre la structure générale, et déterminer quelles sont les pages clés. Plus une page est « profonde », moins elle sera prise en compte³². »

Ainsi, « l'organisation et la hiérarchisation des pages d'un site a des impacts souvent sous-estimés. Il joue aussi bien sur l'expérience utilisateur, sa navigation, que sur le référencement naturel (SEO)³³. » Il est donc important de bien penser le site afin de lui assurer un plus grand rayonnement.

29 *Le Projet*, Site web Corr-Proust, URL : <http://proust.elan-numerique.fr/>, (visité le 8 juin 2020).

30 Courriel de Monsieur Rémy Hème de Lacotte à Lucie Slavik, 8 juin 2020.

31 *Quelles différences entre architecture et arborescence de site web ?*, Site web de UpByWeb, URL : <https://upbyweb.com/seo/differences-architecture-arborescence/>, (visité le 25 juin 2020).

32 Construire l'arborescence de son site web, Site web La Fabrique du Net, URL : <https://www.lafabriquedunet.fr/conseils/conception-site-web/arborescence-site-web/>, (visité le 25 juin 2020).

33 *Optimiser l'architecture d'un site web pour le SEO*, Site web La Fabrique du Net, URL : <https://www.lafabriquedunet.fr/seo/articles/optimiser-architecture-site-web-seo/>, (visité le 25 juin 2020).

Quelques principes : la page d'accueil est « en général la page la plus forte d'un site : plus une page est éloignée de la page d'accueil, plus il est difficile pour elle d'être correctement référencée³⁴. »

Par ailleurs il est important de bien cibler les mots-clés que l'on utilisera, toujours dans l'idée du SEO³⁵. De même pour la rédaction de contenu, il serait bon de voir dans quelle mesure nous devons l'adapter pour être mieux référencés³⁶.

Présentation de l'arborescence du site

ICI, PRÉSENTATION DE L'ARBORESCENCE DU SITE. (J'aime beaucoup les index de la correspondance de Flaubert, qui à mon sens ne sont pas assez clairement présents dans l'édition de Proust ou plutôt qui y sont présents de façons trop détournée, par l'aléa des clics et non sur un onglet visible à tout instant).

> VOIR LES DOCUMENTS EN ANNEXE

> IL SERAIT BON DE REVOIR L'ARBORESCENCE DU SITE SELON LES CONSEILS DONNÉS AUX PAGES 21-22 DU GUIDE MÉTHODOLOGIQUE POUR L'ÉDITION NUMÉRIQUE DE CORRESPONDANCE PAR LE CONSORTIUM CAHIER.

À COMPLÉTER.

c. Encodage TEI

> Voir les termes à indexer.

d. Schéma et ODD

> Tutoriels pour TEI souhaités par le CRHXIX.

IV. Quelle suite ?

1. Former l'équipe à Transkribus. Tutoriels qui ont été réalisés dans ce but.

a. Entraînement du modèle LP

b. Entraînement d'autres modèles

c. Ou de toutes façons, transcriptions directement sur Transkribus

Pour les étudiants qui auraient à faire des transcriptions, ce serait un plus dans leur formation. Leur laisser le choix si l'initiation est trop dure à mettre en œuvre.

2. Former l'équipe à TEI

a. Formation prévue à l'ENC, financée par Collex.

b. Comment mettre en œuvre la suite de l'encodage ? Comment automatiser ? Avec python ? XSLT ? Depuis transkribus ?

³⁴ *Idem*.

³⁵ À ce sujet, voir l'article *Pourquoi et comment bien cibler les mots-clés de son site web*, Site Web anthedesign, URL : <https://www.lafabriquedunet.fr/seo/articles/comment-cibler-mots-cles/>, (visité le 25 juin 2020).

³⁶ *Comment rédiger pour le web*, Site Web anthedesign, URL : <https://www.anthedesign.fr/redaction-web/redaction-de-contenu-web/>, (visité le 25 juin 2020).

3. Externaliser le travail de développement
4. Voir la question de la pérennité des données.
5. Élargissement possible à d'autres projets.

C'est une issue possible, en considérant les échecs des projets numériques.

REVOIR LE PLAN SELON CE QUE J'AI ÉCRIT DANS MON CAHIER DE STAGE.

Proposition du Labex OBVIL pour nous aider dans notre projet et l'intégrer à Elicom. → En reparler.