

ÉCOLE NATIONALE DES CHARTES

---

**Lucie Slavik**

*Licenciée ès histoire*

L'édition numérique de correspondance,  
à travers deux applications sur des  
corpus du XIX<sup>e</sup> siècle : la  
correspondance de Frédéric Le Play  
(CRHXIX) et ELICOM (Labex  
OBVIL)

Mémoire pour le diplôme  
« Technologies numériques appliquées à l'histoire »

2020



# Résumé

Ce mémoire a été réalisé en vue de l'obtention du diplôme de Master 2 « Technologies numériques appliquées à l'histoire » de l'École nationale des chartes. Il a été rédigé suite à la réalisation d'un stage d'environ trois mois au sein de deux institutions dépendant de la Sorbonne, le Centre de Recherche d'Histoire du XIX<sup>e</sup> siècle, sur le projet d'édition numérique de la correspondance du sociologue Frédéric Le Play (1806-1882) d'une part, et le Labex OBVIL sur le projet ELICOM d'autre part. Ce travail n'est ni un mémoire de recherche, ni un rapport de stage. Il vise à apporter une analyse critique des enjeux, stratégies et résultats de ces projets qui ont en commun le siècle sur lequel ils se penchent, le type de sources, la volonté de valorisation et d'accessibilité, l'utilisation de certains outils numériques, mais qui diffèrent aussi sous certains rapports, et qui s'inscrivent dans le cadre des humanités numériques.

**Mots-clefs :** apprentissage machine ; architecture et arborescence de site ; Centre de Recherche d'Histoire du XIX<sup>e</sup> siècle ; édition numérique de correspondance ; ELICOM ; Frédéric Le Play ; Gallica ; HTML ; HTR ; index ; Labex Obvil ; numérisation ; OCR ; ODD ; Python ; Relax NG ; SEO ; sociologie ; XML-TEI ; XSLT

**Informations bibliographiques :** Lucie Slavik, *L'édition numérique de correspondance, à travers deux applications sur des corpus du XIX<sup>e</sup> siècle : la correspondance de Frédéric Le Play (CRHXIX) et ELICOM (Labex OBVIL)*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Thibault Clérice et Arthur Provenier, École nationale des chartes, 2020.



# Remerciements

Mes remerciements s'adressent en premier lieu à Monsieur Thibault Clérice, responsable du Master « Technologies Numériques appliquées à l'Histoire » et mon tuteur, qui a pris de son temps pour me conseiller et répondre inlassablement à toutes mes questions, et m'a particulièrement aiguillée pour mener à bien le projet Le Play. Je remercie également le corps enseignant du Master, et tout particulièrement Madame Ariane Pinche pour s'être rendue disponible pour répondre à certaines questions spécifiques, ainsi que tous ceux qui ont bien voulu répondre à mes questions, notamment Madame Alix Chagué.

Je me dois de remercier tout particulièrement Monsieur Matthieu Brejon de Lavergnée pour m'avoir proposé de réaliser mon stage de fin d'études au Centre de Recherche d'Histoire du XIX<sup>e</sup> siècle (CRHXIX) et pour avoir assuré le partenariat avec le Labex OBVIL.

Un grand merci à l'équipe du Labex OBVIL, pour avoir accepté d'assurer l'encadrement technique de mon stage, et spécialement à mon tuteur technique, Monsieur Arthur Provenier, qui a permis la réussite de mon stage en télétravail et m'a soutenue de ses conseils et avis.

Merci à l'équipe du CRHXIX, tout particulièrement à Monsieur Rémy Hême de Lacotte et Madame Sophie Lhermitte, pour leur accueil chaleureux et leur soutien tout au long de ce stage à distance, ainsi qu'à Monsieur Antoine Savoye pour ses éclairages sur Frédéric Le Play.

Enfin, je me dois de remercier ma famille, tout particulièrement mes parents ainsi que Jacinthe, Marie et Damien pour leur soutien durant cette année.

Merci à mes camarades de promotion, toujours à l'écoute et soucieux d'aider. Merci à mes amis, tout particulièrement Jacinthe L.C. et Ménehould.



# Liste des sigles et abréviations

- A.D. : Archives Départementales
- AFNOR : Association française de normalisation
- AIRE : Association Interdisciplinaire de Recherches sur l'Epistolaire
- ANR : Agence nationale de la recherche
- BIF : Bibliothèque de l'Institut de France
- BNF : Bibliothèque Nationale de France
- CAHIER : Corpus d'Auteurs pour les Humanités, Informatisation, Édition, Recherche
- CÉRÉdI : Centre d'Études et de Recherche « Éditer-Interpréter »
- CNRS : Centre national de la recherche scientifique
- CR : Chargé de recherche
- CRHXIX : Centre de Recherche d'Histoire du XIX<sup>e</sup> siècle
- DR : Directeur de recherche
- EADH : *European association for digital humanities*
- EHNE : Ecrire une histoire nouvelle de l'Europe
- ELICOM : Éditer, Lire des Correspondances Multidisciplinaires
- EHESS : École des hautes études en sciences sociales
- ENC : École Nationale des Chartes
- ENS : École Normale Supérieure
- HDR : Habilitation à diriger des recherches
- IA : Intelligence artificielle
- INHA : Institut National d'Histoire de l'Art
- MCF : Maître de conférences
- OBVIL : Observatoire de la vie littéraire
- READ : *Recognition and Enrichment of Archival Documents*
- TAL : Traitement Automatique des Langues

- SESS : Société d'économie et de science sociales
- TGIR : Très grandes infrastructures de recherche
- UPS : Union de Paix Sociale

★

- API : *Application Programming Interface*
- BL : *Base line*
- CER : *Character Error Rate*
- CSS : *Cascading Style Sheets*
- DMP : *Data Management Plan*
- HTR : *Handwritten Text Recognition*
- ID : *IDentifier*
- ISO : *International Organization for Standardization*
- JPEG : *Joint Photographic Experts Group*
- ML : *Machine Learning*
- ODD : *One Document Does it all*
- OCR : *Optical character recognition*
- PDF : *Portable Document Format*
- PNG : *Portable Network Graphics*
- SEO : *Search Engine Optimization*
- SERP : *Search engine result page*
- SIG : *Special Interest Group*
- SOAP : *Simple Object Access Protocol*
- TEI : *Text Encoding Initiative*
- TIFF : *Tagged Image File Format*
- TR : *Text region*
- URL : *Uniform Resource Locator*
- US : *User Story*
- VIAF : *Virtual International Authority File*
- XML : *eXtensible Markup Language*
- XSLT : *eXtensible Stylesheet Language Transformations*

# Introduction



*« Pratiquer l'édition numérique signifie prendre en compte ce lien étroit entre la technique et la culture. »*

Michaël E. Sinatra  
*Pratiques de l'édition numérique*



A la fin du XX<sup>e</sup> siècle apparaît le terme de « révolution numérique », désignant les mutations profondes des sociétés dues à l'essor des nouvelles technologies numériques. Cette révolution atteint tous les niveaux de la société<sup>1</sup> et le monde universitaire et culturel connaît également des changements. Ainsi constate-t-on que depuis plusieurs années,

« les chercheurs en humanités et sciences sociales vivent une transformation radicale de leur travail. Ces érudits que l'on imagine volontiers enfouis sous des piles de livres, fouillant des masses de vieux papiers dans les archives, [...] passent aujourd'hui le plus clair de leur temps sur leur... ordinateur. Comme le médecin, l'avocat ou le journaliste, le chercheur contemporain a vécu en moins de vingt ans, c'est-à-dire même pas l'espace d'une génération, une dématérialisation pratiquement complète des conditions d'exercice de son métier<sup>2</sup>. »

L'on assiste à la naissance des humanités numériques ou *digital humanities*. En effet, comme le souligne le « Manifeste des *digital humanities*<sup>3</sup> », on remarque que « le tournant numérique pris par la société modifie et interroge les conditions de production et de diffusion des savoirs. »

Cette dématérialisation ne peut se faire sans l'aide de personnes qui améliorent sans cesse les outils de recherche et l'offre proposée en ligne, sur les bibliothèques numériques ou les sites abritant des savoirs, permettant ainsi leur diffusion, leur partage et leur valorisation. Une de ces applications significatives est l'édition numérique de correspondance. Notre sujet se trouve donc au cœur des humanités numériques.

Avant tout, il faut savoir que l'éditeur d'une édition papier est chargé du choix des contenus, de leur légitimation et de leur diffusion. L'édition numérique garde ces caractéristique et se double d'un

« ensemble complexe de pratiques qui vont bien au-delà du rôle que l'éditeur a eu dans le modèle de l'édition imprimée à partir du XVIII<sup>e</sup> siècle. [Elle] regroupe toutes les actions destinées à structurer, rendre accessible et visible un contenu sur le web<sup>4</sup>. »

Elle a donc recours à nombre de pratiques, de traitements et d'outils technologiques avancés pour assurer à la fois une édition de qualité quant au fond et à la forme, et la pérennité des données.

1. Nous ne traitons pas ici de la fracture numérique qui pose des limites à cette affirmation. La révolution numérique n'en est pas moins une réalité.

2. Pierre Mounier, « Les Humanités numériques, gadget ou progrès ? Enquête sur une guerre souterraine au sein de la recherche », *Revue du Crieur*, 2017/2, p. 144-159, URL : <https://www.cairn-int.info/revue-du-crieur-2017-2-page-144.htm> (visité le 02/09/2020)

3. Manifeste rédigé par les participants du THATCamp Paris de mai 2010. Voir Michaël E. Sinatra et Marcello Vitali-Rosati, *Pratiques de l'édition numérique*, Montréal, Les Presses de l'Université de Montréal, 2014

4. Michaël E. Sinatra et Marcello Vitali-Rosati, *Pratiques de l'édition numérique*, Presses de l'Université de Montréal, 2014, URL : [https://books.openEdition.org/pum/308](https://books.openedition.org/pum/308) (visité le 03/09/2020)

Au sein de l'édition numérique en général se trouve une édition plus spécifique qui représente à elle seule un monde à part : c'est l'édition numérique de correspondance.

La correspondance a elle-même ses caractéristiques : c'est un genre à la fois protéiforme - c'est à dire qu'elle est adressée à un public restreint ou étendu - réticulaire - elle se trouve dans un réseau de lettres et de correspondants - et elliptique - d'où l'importance d'une documentation conséquente<sup>5</sup>. L'édition numérique de correspondance doit donc prendre en compte toutes ces caractéristiques et tirer profit de l'aspect numérique pour enrichir les possibilités d'une édition classique.

Lors de notre stage, nous avons donc été confronté à toutes ces problématiques autour de l'édition numérique de correspondance. Notre réflexion a eu pour cadre deux institutions dépendant de la Sorbonne (Sorbonne Université et Paris I), à savoir le Labex OBVIL (Observatoire de la vie littéraire) et le Centre de Recherche d'Histoire du XIX<sup>e</sup> siècle (CRHXIX). Chacune des institutions menant son propre projet, nous avons donc travaillé sur deux projets distincts.

Au Labex OBVIL, nous avons pris part au projet de plateforme multifonctionnelle ELICOM, le nom ELICOM étant l'abréviation de Éditer, Lire des Correspondances Multidisciplinaires : l'appellation du projet fait entrevoir à lui seul ses buts et son ampleur. Il répond à trois objectifs : la collecte et l'enrichissement de correspondances épistolaires de différentes disciplines, la fouille transversale des données et métadonnées, et l'exploitation visuelle et statistique des résultats de cette fouille. Par ailleurs, les corpus de correspondances épistolaires sont multidisciplinaires, c'est à dire à la fois littéraires, philosophiques et scientifiques, et tous appartiennent au XIX<sup>e</sup> siècle. C'est donc un projet innovant et au carrefour de nombre de standards, technologies et outils.

Le CRHXIX quant à lui porte un projet d'édition numérique à caractère plus classique, autour de la correspondance du sociologue Frédéric Le Play (1806-1882). Nous sommes encore sur un corpus du XIX<sup>e</sup> siècle, mais qui nécessite d'être traité différemment. Au Labex OBVIL, nous partons de différentes éditions papier pour aller vers une édition numérique de grande ampleur. Au CRHXIX au contraire, nous partons des sources manuscrites pour réaliser une édition numérique faite de documents inédits. Cette différence conséquente entraîne un traitement autre et l'utilisation d'outils spécifiques pour les transcriptions.

Certes, à des besoins différents répondent l'emploi d'outils distincts et d'un traitement approprié à chacune des éditions. Néanmoins, nombre de problématiques sont communes à ces éditions numériques de correspondance, car nous restons dans un même sujet. Deux grands axes rejoignent à travers l'acquisition et le traitement des données.

Tout d'abord, nous sommes arrivés aujourd'hui à l'ère de l'intelligence artificielle

---

5. Richard Walter (dir.), *L'édition numérique de correspondances – guide méthodologique*, 2018, p.4  
URL : <https://cahier.hypotheses.org/guide-correspondance> (visité le 17/06/2020)

(IA)<sup>6</sup>, et autour de cette réalité gravitent nombre de technologies. Dans l'édition numérique de correspondance, l'on a ainsi recours à l'apprentissage machine ou *machine learning* dans la phase d'acquisition des données, que ce soit via la Reconnaissance optique de caractères dite OCR (*Optical Character Recognition*), ou même via la technologie plus avancée qu'est la Reconnaissance de l'écriture manuscrite dite HTR (*Handwritten Text Recognition*).

D'autre part, quant au traitement des données, l'édition numérique de correspondance utilise ultimement le standard XML-TEI, comme l'a souligné assez récemment l'étude du Consortium CAHIER dans son *Guide méthodologique* consacré à ce sujet<sup>7</sup>.

L'objectif de ce mémoire sera donc de voir, à travers ces deux projets qui portent sur des corpus d'un même siècle mais diffèrent parfois dans leur méthodologie et outils, les grandes tangentes et les possibilités qui caractérisent l'édition numérique de correspondance aujourd'hui.

Pour cela, il nous faudra tout d'abord nous pencher, dans une première partie, sur les projets en eux-mêmes, pour mieux comprendre leurs enjeux, leurs besoins, leurs objectifs, afin de pouvoir y répondre.

Or, l'on ne pourra y répondre sans avoir, au préalable, mené une réflexion conséquente sur ce qu'implique l'édition numérique de correspondance aujourd'hui. Il faudra s'interroger sur ses moyens et ses outils et prendre du recul face à ces problématiques pour mieux les appliquer à nos projets. Penser l'édition numérique de correspondance sera donc l'objet de notre deuxième partie.

Ces réflexions nous amèneront à leur mise en pratique qui se fera en deux temps. Tout d'abord, notre troisième partie sera consacrée aux moyens employés pour l'acquisition des données, au cœur de laquelle se trouve l'apprentissage machine, sans oublier d'autres technologies que nous développerons à cette occasion.

Une fois les données acquises, il s'agira d'assurer leur traitement, ce qui fera l'objet de notre quatrième partie, qui abordera la question des standards et formats employés pour une meilleure pérennité des données, notamment XML-TEI.

---

6. L'IA est un ensemble de techniques permettant à des machines d'accomplir des tâches et de résoudre des problèmes normalement réservés aux êtres humains.

7. Richard Walter (dir.), *L'édition numérique de correspondances – guide méthodologique*, 2018  
URL : <https://cahier.hypotheses.org/guide-correspondance> (visité le 17/06/2020)



## Première partie

Des projets portés par des institutions  
culturelles



# Chapitre 1

## Un contexte universitaire

Avant tout, il est à noter que les deux stages ont été réalisés de mai à juillet 2020<sup>1</sup>. Suite à la situation sanitaire, nous avons été amenée à réaliser l'intégralité de ces stages en télétravail.

Pour mieux comprendre les objectifs de ces projets, il est important de souligner tout d'abord dans quel cadre ils se sont accomplis. Or, un de leurs points communs est leur contexte universitaire.

### 1.1 Le Centre de Recherche et d'Histoire du XIX<sup>e</sup> siècle

#### 1.1.1 Une institution dédiée à la recherche autour du XIX<sup>e</sup> siècle

Situé dans les locaux de l'université parisienne de la Sorbonne, Paris I, au cœur du quartier latin, le Centre de Recherche et d'Histoire du XIX<sup>e</sup> siècle accueille en son sein les chercheurs qui ont décidé de consacrer leurs travaux au XIX<sup>e</sup> siècle. Il dispose pour cela d'une bibliothèque de près de huit-mille volumes<sup>2</sup>.

Fondé par le Professeur Louis Girard (1911-2003), il relève des deux universités de Paris-1 et Sorbonne-Université (anciennement Paris IV)<sup>3</sup>. Il se compose de plus de cent-cinquante personnes, à la fois des professeurs des universités, des maîtres de conférences, divers professeurs agrégés, allocataires, moniteurs, ATER (Attaché temporaire d'enseignement et de recherche), IATOS (ingénieurs, administratifs, techniciens, ouvriers de service), chercheurs associés et doctorants<sup>4</sup>. Centre dynamique, il mène de front de

---

1. Plus précisément du 4 mai au 31 juillet, trois jours par semaine pour le Centre d'Histoire du XIX<sup>e</sup> siècle, soit un total de trente-et-un jours et demi de travail, et du 19 mai au 29 juillet, soit un total de vingt jours de travail au Labex OBVIL.

2. *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 18/06/2020).

3. *Présentation du Centre*, Site web de l'Université Paris-1 Panthéon Sorbonne, URL :<https://www.pantheonsorbonne.fr/unites-de-recherche/crhxix/aboutthecenter/>, (visité le 18/06/2020).

4. *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 18/06/2020).

nombreux projets.

### 1.1.2 Un Centre dynamique

Le CRHXIX développe actuellement ses recherches autour de quatre axes principaux<sup>5</sup>, à savoir en premier lieu, « Le “resserrement des sociétés” : circulations globales et pratiques locales au XIX<sup>e</sup> siècle », en second lieu « Du moral au social : pratiques et théories de l’enquête. Autour des archives du mouvement leplaysien », puis sur la problématique « Citoyennetés, sûretés, sécurités, souverainetés », et enfin le thème « Images, imaginaires et écriture de l’histoire ».

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 CRHXIX.

### 1.1.3 Les acteurs du projet

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 » 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'études.

De nombreux étudiants, doctorants et stagiaires ont participé de près ou de loin à l'avancement du projet. L'on notera particulièrement la coopération de Madame Margaux Faure, qui a effectué de nombreuses relectures de transcriptions et qui a réalisé elle-même des transcriptions, ainsi que Monsieur Edouard Coquet. Tous deux ont participé aux journées de formation à l'École Nationale des Chartes (ENC) autour de Transkribus, afin de prendre en main les outils qui nous sont nécessaires pour l'avancement de notre projet sur l'édition numérique de la correspondance de Le Play, outils qui seront présentés plus particulièrement dans la troisième partie.

Lors de notre stage, nous avons été surtout amenée à collaborer avec le chef actuel du projet, Monsieur Rémy Hême de Lacotte, avec les avis de Monsieur Matthieu Brejon de Lavergnée à l'origine du projet, et l'ingénieur d'études Madame Sophie Lhermitte<sup>6</sup>.

### 1.1.4 Partenaires susceptibles d'être mobilisés

Parmi les partenaires du projet, mobilisés ou susceptibles de l'être, figurent notamment l'Institut universitaire de France (IUF), le Labex EHNE (Ecrire une histoire

---

5. *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/crhix/about-us/research-activities/>, (visité le 17/06/2020).

6. C'est avec eux que nous avons fait des points réguliers sur l'avancement de notre travail.

nouvelle de l'Europe), l'École des Chartes (ENC), le Sénat, l'Académie des Sciences morales et politiques, le Centre de sociologie des organisations-Sciences Po et la revue *Les Études Sociales*.

### 1.1.5 Soutiens financiers

Qui dit projet dit financement. 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 EHNE. Les formations à l'ENC ont été financées par un prêt Collex.

C'est donc dans ce contexte universitaire que s'inscrit notre projet d'édition numérique de la correspondance de Frédéric Le Play.

Qu'en est-il du contexte du projet ELICOM ?

## 1.2 Le Labex OBVIL

### 1.2.1 Un laboratoire d'excellence pour les humanités numériques

Le projet ELICOM (Éditer, Lire des Correspondances Multidisciplinaires) relève, lui aussi, de la Sorbonne. Il est porté par le Labex (laboratoire d'excellence) OBVIL (Observatoire de la vie littéraire), qui est affilié à Sorbonne Université.

Si le CRHXIX est un centre de recherche, le Labex OBVIL est quant à lui un laboratoire de recherche. Même si le terme centre ou laboratoire peut être considéré comme à peu près équivalent, ce dernier a une connotation plus scientifique, et il est par ailleurs classé laboratoire d'excellence. En effet, le Labex OBVIL est très investi dans les humanités numériques. Il regroupe les laboratoires de littérature de l'Université Paris Sorbonne et s'articule avec le Laboratoire Informatique de l'Université Pierre et Marie Curie<sup>7</sup>.

Le Labex OBVIL réunit près de quatre cents chercheurs (PR, DR, MCF, CR, Doctorants, post doctorants) répartis dans 24 projets de recherche. Fort d'une équipe de six ingénieurs et d'un soutien administratif à la recherche, délibérément transdisciplinaire, puisqu'il conçoit des outils de numérisation et d'édition, de fouille de texte, d'alignements de texte, de visualisation, le Labex OBVIL mène une recherche novatrice, concevant des humanités numériques littéraires, où littérature et informatique se façonnent mutuellement<sup>8</sup>.

Depuis 2016, il se consacre à des données massives (création de la Très Grande

---

7. *Laboratoire d'excellence OBVIL*, Site web de l'enseignement supérieur et de la recherche, URL : [https://cache.media.enseignementsup-recherche.gouv.fr/file/Fiches\\_Labex\\_2/63/8/OBVIL\\_207638.pdf](https://cache.media.enseignementsup-recherche.gouv.fr/file/Fiches_Labex_2/63/8/OBVIL_207638.pdf) (visité le 04/09/2020).

8. *Observatoire de la vie littéraire*, Site web de l'ANR, URL : <https://anr.fr/ProjetIA-11-LABX-0059> (visité le 04/09/2020).

bibliothèque<sup>9</sup>, 130 000 documents), avec le partenariat de la Bibliothèque nationale de France (BNF), et à la conception d'ontologies des corpus critiques et au *text mining*<sup>10</sup>.

### 1.2.2 De nombreux partenaires

Le Labex Obvil s'entoure pour mener à bien tous ses multiples projets. Il a de nombreux partenaires parmi lesquels on compte le Centre d'étude de la langue et des littératures françaises (CELLF), le Centre de recherche en littérature comparée (CRLC), le Centre de Recherches Interdisciplinaires sur les Mondes Ibériques Contemporains (CРИMIC), Civilisation et littérature d'Espagne et d'Amérique (CLEA), le Programme de Recherches Interdisciplinaires sur le Théâtre et les Pratiques scéniques (PRIPEPS), Voix Anglophones : littérature et esthétique (VALE), l' Equipe Littérature et Culture italiennes<sup>11</sup>.

C'est donc dans ce contexte particulièrement favorable de recherche scientifique en humanités numériques que s'inscrit un de ses nombreux projets : ELICOM.

### 1.2.3 Les acteurs du projet

L'équipe travaillant plus spécifiquement sur le projet ELICOM est sous la direction scientifique de Monsieur Glenn Roe - directeur du Labex OBVIL - avec la participation de Monsieur Arthur Provenier, ingénieur d'études, et de Madame Camille Koskas agrégée de lettres modernes en contrat post-doctoral.

Lors de ce stage, nous avons donc bénéficié de l'encadrement scientifique et technique du Labex OBVIL, et plus spécifiquement de Monsieur Arthur Provenier, qui a également proposé l'aide de ses conseils et avis pour le projet du CRHXIX.

C'est donc dans ce contexte universitaire de centre et laboratoire de recherche que se sont développés nos deux projets en humanités numériques. Tous deux sont des projets d'édition numérique de correspondance, sur des corpus du XIX<sup>e</sup> siècle. Malgré ces points communs, ils présentent néanmoins des caractéristiques et mises en pratique différentes. Il convient donc de mieux considérer, dans un deuxième chapitre, la nature de ces projets avant de voir leurs sources.

9. *API et jeux de données*, Site web de la BNF, URL : <http://api.bnf.fr/mise-disposition-de-la-tres-grande-bibliothèque-du-labex-obvil> (visité le 04/09/2020).

10. « Le *text mining* est l'ensemble des techniques et méthodes destinées au *traitement automatique* de données *textuelles en langage naturel*, disponibles sous forme informatique, en assez grande quantité, en vue d'en *dégager et structurer le contenu, les thèmes* dans une perspective d'analyse rapide (non littéraire), de découverte d'informations cachées ou de prise automatique de décision. Voir Stéphane Tufféry, *Data mining et statistique décisionnelle*, 4<sup>e</sup> édition, Paris, 2012, p. 2. »

11. *Accueil*, Site web de l'Observatoire de la vie littéraire, URL : <http://obvil.sorbonne-universite.site/obvil/presentation> (visité le 04/09/2020).

# Chapitre 2

## Deux projets ambitieux

### 2.1 L'édition numérique de la correspondance de Frédéric Le Play

Penchons-nous tout d'abord sur le projet porté par le CRHXIX, à savoir l'édition numérique de la correspondance du sociologue Frédéric Le Play (1806-1882). Il sera nécessaire avant tout de comprendre quel est cet homme pour mieux saisir pourquoi il est pertinent de publier sa correspondance aujourd'hui.

#### 2.1.1 Au service de l'histoire des sciences sociales

Ce projet d'édition numérique de la correspondance est au service de l'Histoire, et en particulier de l'histoire de la sociologie. L'on voit en ce sens combien, dans les humanités numériques, toutes ces technologies élaborées sont au service de la culture et des humanités. Comme nous l'avons vu plus haut, notre projet s'inscrit dans l'axe de recherche du CRHXIX intitulé « Du moral au social : pratiques et théories de l'enquête. Autour des archives du mouvement leplaysien » .

##### 2.1.1.1 Autour des archives du mouvement leplaysien

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 XIXe 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 XIXe et XXe 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. 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<sup>1</sup>.

Pour mieux en saisir toute la portée, il convient d'apprécier qui est Frédéric Le Play.

### 2.1.1.2 Le Play, un sociologue méconnu

Parmi les personnages qui ont eu leur rôle à jouer dans l'histoire, mais qui sont souvent trop oubliés, figure Frédéric Le Play. À nous de le redécouvrir à travers sa vie, son œuvre et son réseau, puisque tout l'objet de notre projet est de les mettre en lumière.

Né le 11 avril 1806 dans le Calvados, Frédéric Le Play passe son enfance et ses études entre la Normandie et Paris. À 19 ans, il intègre l'École Polytechnique, où il rencontre les saints-simoniens Michel Chevalier et Jean Reynaud<sup>2</sup>. C'est à 1830 que remonte sa « vocation sociale ».

Lors de sa carrière dans le corps des Mines, entre 1831 et 1856, il fait de nombreux voyages en Europe : Espagne, Belgique, Sud de la France, Italie, Grande-Bretagne, Scandinavie, Suisse, Autriche-Hongrie, Allemagne, ainsi que Russie, ce sont autant de pays qu'il sillonne et où il va pouvoir enquêter. En effet, il

« veut fonder la science sociale. Aussi choisit-il de passer par une longue phase d'accumulation d'observations au cours de laquelle il précise son objet – les familles ouvrières –, définit sa méthode – la monographie – et élabore ses techniques : l'établissement du budget familial d'une part, la collecte d'informations auprès des autorités sociales de l'autre. Multipliant, en outre, les voyages, il élargit et systématise son étude comparative<sup>3</sup> ».

Par ailleurs, ces « voyages sont justifiés par son cours à l'École qu'il actualise et la préparation de son ouvrage de science sociale *Les ouvriers européens*, notamment pour la réalisation de monographies<sup>4</sup> de familles ouvrières »<sup>5</sup> qu'il considère comme révélatrices

1. Axe 2 : *Du moral au social : pratiques et théories de l'enquête. Autour des archives du mouvement leplaysien.*, Site web du CRHIX, URL : <https://www.pantheonsorbonne.fr/unites-de-recherche/crhix/about-us/research-activities/area2/> (visité le 15/06/2020).

2. Antoine Savoye, « LE PLAY FRÉDÉRIC - (1806-1882) », *Encyclopædia Universalis*, URL : <http://www.universalis-edu.com/encyclopedie/frederic-le-play/> (visité le 12/06/2020).

3. *Idem*.

4. « La monographie selon Le Play est donc la matrice de diverses techniques qui sont aujourd'hui employées en sociologie, en ethnographie, en psychologie sociale, en histoire, en géographie humaine, etc. » *Ibid.*

5. Antoine Savoye, « Frédéric Le Play en quelques dates », dans *Frédéric Le Play : Parcours, Au-*

de l'« état social ». À la suite de ces voyages, il publie souvent des mémoires scientifiques où l'on peut suivre l'élaboration de sa pensée.

Cette pensée, il la développe dans nombre de ses ouvrages : avec la *Réforme sociale en France* en 1864, « Le Play propose une analyse globale de la société française [...] depuis l'organisation de la vie privée jusqu'aux mécanismes administratifs et de gouvernement » proposant des réformes « étayées sur ses enquêtes et celles effectuées par ses collaborateurs<sup>6</sup> ». Dans *Les Ouvriers européens*, réédité à la fin des années 1870, Le Play, d'après ses observations, souligne les éléments qui garantissent la stabilité d'une société, qui passe par deux fondements invariables, « le respect du Décalogue et le règne de l'autorité paternelle »<sup>7</sup>.

De 1857 à 1870, il tient un rôle politique de conseiller d'État, sénateur et réformateur social. Il œuvre notamment à la réforme du droit successoral et trouve un appui auprès des catholiques libéraux. Les expositions universelles sont aussi pour lui l'occasion de faire valoir ses idées<sup>8</sup>. Par ailleurs, la création des unions de la paix sociale (UPS), sorte de groupements savants et militants, lui permet de faire connaître sa pensée et élargir son influence.

Sa correspondance est aussi un moyen pour lui de faire part à ses amis polytechniciens du fruit de ses recherches pour en arriver à la Réforme sociale. Il passe de nombreuses heures par jour à s'épuiser dans une correspondance abondante afin de faire germer sa pensée dans les esprits des savants de sa connaissance.

Par la création de *La Réforme sociale*, en 1881, il se donne une revue pour développer sa pensée et la perpétuer après sa mort qui survient l'année suivante.

Les moyens de diffusion de sa pensée sont donc nombreux, et l'un d'eux, sa correspondance, est donc l'objet de notre projet.

## 2.1.2 Redécouvrir l'un des fondateurs de la sociologie

### 2.1.2.1 Explorer la méthode scientifique de Le Play, éclairer les conditions d'élaboration des enquêtes

Comme nous l'avons dit plus haut, Le Play est un précurseur des méthodes d'enquêtes sociologiques de terrain.

« Ses principes d'observation directe de la réalité et de recherche comparative, sa méthode monographique, ses techniques de quantification par le budget qu'il a appliqués à l'étude systématique des familles ouvrières font pourtant de lui

---

dience, *Héritage*, dir. Fabien Cardoni, Paris, 2013, p. 279-289

6. *Ibid.*

7. Antoine Savoye, « LE PLAY FRÉDÉRIC - (1806-1882) », *Encyclopædia Universalis*, URL : <http://www.universalis-edu.com/encyclopedie/frederic-le-play/>

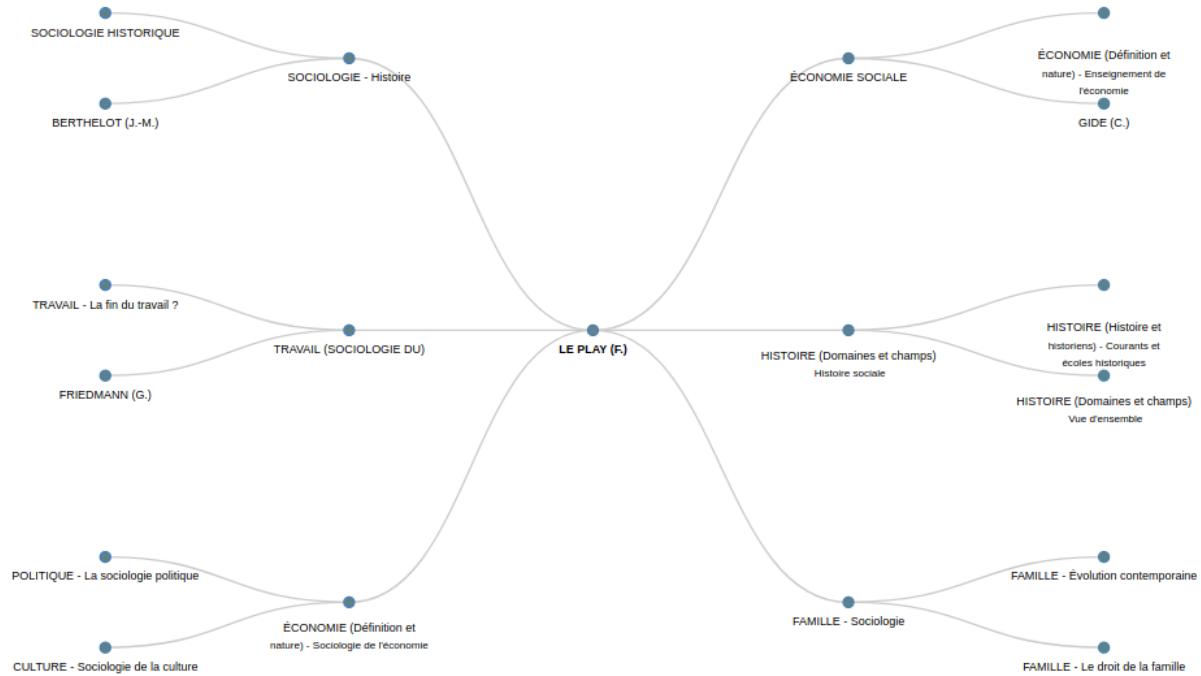
8. En 1855, il est nommé commissaire général à l'occasion de l'exposition universelle de Paris. *Idem.*

le premier théoricien de la sociologie de terrain. De plus, sa double activité de haut fonctionnaire et de sociologue éclaire l'histoire de la sociologie<sup>9</sup> ».

Notre projet scientifique consiste donc 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 gomment 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 » leplay-sienne 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.

### 2.1.2.2 Frédéric Le Play, au croisement de nombreux champs intellectuels

FIGURE 2.1 – Carte mentale de Le Play, *Encyclopædia Universalis*



Par ailleurs, Frédéric Le Play se trouve au croisement de nombreux champs intellectuels, à nous de voir comment nous pourrons le faire valoir dans notre projet. Il s'agit pour nous de relever un défi pour faire connaître un sociologue méconnu.

9. *Ibidem.*

Lors de la présentation des sources dans le troisième chapitre, nous détaillerons l'importance des sources qui sont en notre possession ou en cours d'acquisition. Avant cela, il importe de présenter le deuxième projet sur lequel nous avons travaillé, au sein du Labex OBVIL.

## 2.2 ELICOM

A travers ELICOM, le Labex OBVIL

« propose une recherche inédite sur la question de la valeur littéraire, envisagée au travers de l'étude de grands corpus numérisés qui prennent en compte, non seulement les textes eux-mêmes mais les circonstances et les modalités de leur publication et de leur réception<sup>10</sup> ».

Si les sources en elles-mêmes nous seront exposées plus en détail dans le chapitre suivant, présentons dès à présent les grands axes du projet<sup>11</sup>.

Retenu par l'appel à projet Émergence<sup>12</sup> qui entend consolider l'excellence au cœur des disciplines, accompagner les évolutions interdisciplinaires et ouvrir l'université sur la société, le projet ELICOM se révèle être ambitieux et innovant, car il est plus qu'une simple édition numérique de correspondance.

### 2.2.1 Pour une recherche collective et multidisciplinaire

Tout d'abord, le présent projet semble s'inscrire dans un mouvement de recherche collective, national ou international. Mais cette inscription en fait bien apparaître la singularité : elle sort du cadre strictement littéraire pour s'ouvrir à des corpus de philosophie, de biologie, de mathématiques. Cette multidisciplinarité répond à la volonté de s'appuyer sur les collaborations institutionnelles, à l'intérieur de Sorbonne Université avec sa Bibliothèque riche en archives scientifiques, avec la Bibliothèque Nationale de France (BNF) dont Gallica offre d'immenses ressources de correspondances éditées en format image, voire en mode texte. Le labex OBVIL poursuit ainsi sa politique patrimoniale de recherche tout en accentuant sa réflexion sur les services qu'il peut offrir à la communauté dans laquelle il s'insère et à la communauté des chercheurs.

---

10. Laboratoire d'excellence OBVIL, *Site web de l'enseignement supérieur et de la recherche*, URL : [https://cache.media.enseignementsup-recherche.gouv.fr/file/Fiches\\_Labex\\_2/63/8/OBVIL\\_207638.pdf](https://cache.media.enseignementsup-recherche.gouv.fr/file/Fiches_Labex_2/63/8/OBVIL_207638.pdf) (visité le 04/09/2020).

11. Cette présentation du projet est largement inspirée des documents qui ont été présentés afin d'obtenir les financements.

12. Voir : Appel à projets Emergence , *Site web de Sorbonne Universités*, [https://candidature.sorbonne-universites.fr/index.php?option=com\\_emundus&view=programme&id=111&Itemid=1521&lang=fr](https://candidature.sorbonne-universites.fr/index.php?option=com_emundus&view=programme&id=111&Itemid=1521&lang=fr) (visité le 07/09/2020)

### 2.2.2 Un outil de réflexion et de recherche

Cela signifie aussi que le caractère innovant du programme ne réside pas dans les correspondances elles-mêmes, même si le labex OBVIL apportera son savoir-faire et son expertise aux spécialistes d'un auteur qui éditent ou souhaitent éditer, au format numérique, une correspondance. Il réside avant tout dans la volonté forte de constituer autour d'un objet théorique – qu'est-ce que correspondre par lettres ? - une communauté de spécialistes venus de disciplines différentes pour penser les besoins éditoriaux en fonction des possibilités offertes par le numérique. La création d'une plateforme d'édition et de lecture répondra aux besoins multidisciplinaires tout en devenant un outil de réflexion et de recherche pour les communautés de chercheur.

L'ambition ultime sera de tester les résultats du présent programme sur des correspondances émanant non plus d'une élite savante, mais de communautés plus ordinaires, familiales, ou socio-professionnelles. Le projet ambitionne ce retour vers la vie ordinaire, sociale, professionnelle, citoyenne.

### 2.2.3 Trois modules

Ces corpus, une fois édités - ce qui n'est qu'un préalable au projet pris en charge par le Labex OBVIL - forment un premier échantillon multidisciplinaire qui est exploité par les trois modules de la plateforme ELICOM : l'enrichissement, la fouille et l'exploitation.

#### 2.2.3.1 Enrichissement

Ce premier module offre à chaque utilisateur la possibilité d'enrichir les textes édités à l'aide de balises TEI<sup>13</sup>, choisies à partir de ses propres questionnements, et en s'appuyant sur le manuel d'encodage pour correspondances mis au point par le consortium CAHIER<sup>14</sup>. Il lui permet également d'annoter chaque lettre en collaborant à la mise en place d'un appareil critique comme des notes savantes et des références bibliographiques, et de compléter les métadonnées. Cet enrichissement participatif permet à terme d'adapter les corpus aux perspectives de recherche de chacun, et de mettre un ensemble de données et de métadonnées à disposition de la communauté scientifique.

#### 2.2.3.2 Fouille

Ce second module ouvre ensuite à la fouille des textes et de leurs métadonnées, par un ensemble d'outils de *text mining*. Il prévoit également un travail de modélisation sémantique, réalisé au cours du séminaire qui accompagne le développement de la plateforme, et qui s'appuie à la fois sur une construction systématique de ressources linguistiques

---

13. Nous reviendrons sur tous ces termes dans la suite du mémoire.

14. Richard Walter (dir.), *L'édition numérique de correspondances – guide méthodologique*, URL : <https://cahier.hypotheses.org/guide-correspondance> (visité le 17/06/2020).

à caractère polémique et sur des corpus de correspondances annotés manuellement par des experts. L'enjeu est de pouvoir exploiter ces données par des outils du Traitement automatique des langues (TAL) dans le but d'effectuer :

- une analyse sémantique des controverses (identifier les postures et les positionnements de débats, de polémiques, etc.)
- une analyse des enchaînements temporels (différenciant récit de vie et discours)
- une analyse thématique (autour des événements publics et privés)
- une analyse des rituels de la correspondance (formules de politesse, styles selon le destinataire)
- un repérage des entités nommées<sup>15</sup>, etc.

L'ensemble de ces fonctionnalités doit permettre une exploration transversale des correspondances par le biais de requêtes croisées (par nature de la lettre, par thématique, par lieu, par âge ou sexe des correspondants, etc.), de manière à conduire l'utilisateur vers certains événements propres à la correspondance épistolaire.

### 2.2.3.3 Exploitation

Ce troisième module propose différents types de visualisation permettant une représentation et une analyse des données et des métadonnées des correspondances : des nuages de collocations, des réseaux (entités nommées et métadonnées), des cartographies par géolocalisation, des visualisations de relevés statistiques et des visualisations croisées par période et par fréquence de correspondances, par lieu et par auteur, etc.

Le projet étant très vaste, nous en sommes encore à la phase d'encodage en XML-TEI. C'est donc cette partie qui nous a été confiée lors de notre stage. Gérer un corpus aussi large est un véritable défi, notamment dans le choix des balises. C'est ce défi que nous avons dû relever. Mais avant cela, il faut penser l'édition. Ce sera l'objet de notre seconde partie. En attendant, il sera bon de considérer quelles sont les sources exploitées pour avoir meilleure vue d'ensemble du projet.

---

15. « On appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. » Voir : Maud Ehrmann, *Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*, Thèse de doctorat, Université Paris Diderot, 2008, URL : <https://hal.archives-ouvertes.fr/tel-01639190> (visité le 07/09/2020).



# Chapitre 3

## Les sources des projets : des correspondances du XIX<sup>e</sup> siècle aux formes variées

Les deux projets auxquels nous avons participé n'ont pas les mêmes caractéristiques tout en ayant de nombreux points communs, comme nous l'avons déjà souligné.

Pour le CRHXIX, l'objectif est de mettre à disposition une correspondance riche mais oubliée et peu connue et reconnue. Pour OBVIL au contraire, l'objectif est de (re)mettre en valeur une correspondance qui a déjà été publiée, à laquelle on s'est déjà intéressée, qui est déjà en ligne, mais dont on voudrait faciliter l'exploitation, le croisement des sources, l'interopérabilité.

Pour le projet d'édition numérique de la correspondance de Frédéric Le Play, nous travaillons sur un homme autour duquel se greffe tout un réseau. Pour ELICOM, c'est un réseau en soi que nous mettons en valeur.

Par ailleurs, pour le projet Le Play, c'est une approche plutôt classique d'édition. Pour ELICOM, c'est une démarche innovante.

Enfin, et c'est ce qui fait toute la différence des moyens employés, pour le projet Le Play, nous avons affaire à des manuscrits originaux numérisés ou à numériser, alors que pour ELICOM, nous travaillons sur des éditions imprimées qui ont été numérisées et sont disponibles sur Gallica.

Il est donc temps de présenter ces différents corpus du XIX<sup>e</sup> siècle.

### 3.1 La mise en valeur de manuscrits

Tout d'abord, le CRHXIX est en pleine phase d'acquisition de manuscrits de Frédéric Le Play sous une forme numérisée. .

La correspondance de Le Play est relativement dispersée. Elle se trouve principale-

ment en Europe, plus particulièrement en France, mais certains manuscrits sont conservés en Suisse et en Italie par exemple.

Il s'agit donc de contacter les différents lieux de conservation de ces manuscrits - bibliothèques, archives privées et publiques etc. - demander leur numérisation en vue de leur transcription et encodage pour la future mise en ligne du manuscrit et de sa transcription.

Ce travail a été déjà bien entamé. Un premier inventaire provisoire, de plus de deux-mille lettres, a déjà été réalisé en 2005 par Stéphane Baciocchi et Antoine Savoye - spécialiste reconnu de Frédéric Le Play - paru dans la revue des *Études sociales*<sup>1</sup>. Nous nous y sommes référée dans notre travail de prise en main du projet pour avoir une vue d'ensemble des sources. Nous nous y reportons donc également pour cette présentation des sources.

### 3.1.1 Trois fonds familiaux principaux

Les archives de Frédéric Le Play ont été localisées dans trois fonds principaux : au château de Ligoure (Haute-Vienne), à la Bibliothèque de l'Institut de France (BIF) à Paris, ainsi qu'à la Société d'économie et de science sociales (SESS) à Paris.

#### 3.1.1.1 Château de Ligoure (Haute-Vienne)

En 1856, Frédéric Le Play achète le domaine de Ligoure, y établissant une famille-souche<sup>2</sup> à travers son fils, Albert Le Play. Frédéric Le Play y séjourne pendant l'été et retourne dans sa résidence parisienne durant l'hiver.

D'après l'inventaire réalisé par Antoine Savoye et Stéphane Baciocchi, on constate que

« de nombreuses lettres de Frédéric à son fils (1865-1881), notamment celles dans lesquelles il est question de la gestion du domaine agricole, y sont encore aujourd'hui conservées et propriété de l'arrière petite-fille d'Albert, Madame Thomas-Mouzon. Elles voisinent, dans les tiroirs et placards de la bibliothèque, avec celles que F. Le Play adressa à sa belle-fille, son « héritière-associée » (1867-1880), et quelques autres lettres des années 1866-1876. En attendant un inventaire plus précis de ce fonds, on peut d'ores et déjà penser que toutes les correspondances restées à Ligoure sont une partie de celles qui y ont été

1. Antoine Savoye Stéphane Baciocchi, « La correspondance de Le Play, une source pour l'histoire des sciences sociales », *Les Études sociales*-n 142-144 (2005), p. 231-247, URL : <https://gallica.bnf.fr/ark:/12148/bpt6k9767323c/f1n284.texteBrut> (visité le 07/05/2020)

2. Principe selon lequel « l'un des fils demeure avec sa femme et ses enfants dans le foyer paternel en attendant la succession », développé par Frédéric Le Play, expliqué ici par Emmanuel Todd, « Les familles dans l'Histoire », *Herodote.net*, URL : [https://www.herodote.net/Les\\_familles\\_dans\\_1\\_Histoire-article-1287.php](https://www.herodote.net/Les_familles_dans_1_Histoire-article-1287.php) (visité le 08/09/2020).

adressées après 1864<sup>3.</sup> »

### 3.1.1.2 Bibliothèque de l’Institut de France (Paris)

Le fonds conservé à la BIF lui a été confié par Jean Albert Le Play, arrière petit-fils de Frédéric Le Play, en 1946. Il comporte un millier de lettres datées des années 1832-1882 et des liasses de notes manuscrites et de papiers divers.

Certaines sont adressées à Victor Legrand, directeur général des ponts et chaussées et des mines, au cours ou au retour de ses missions à travers l’Europe (1836-1846). Il constitue le seul fonds public qu’il nous reste de Le Play et laisse à entendre que un bon nombre de papiers ont disparu. Le gros de la correspondance conservée est en lien, d’une part, avec les éditions successives et la diffusion de ses ouvrages la *Réforme sociale en France*, *l’Organisation du travail*, *l’Organisation de la famille* et, d’autre part, avec la mise en place des UPS. Cet ensemble est daté des années 1860 et 1870.

Comme le soulignent Antoine Savoye et Stéphane Baciocchi,

« le fonds public, ouvert depuis 1946 aux chercheurs, est celui où ont puisé tous les travaux de première main sur Le Play, sans toujours s’interroger explicitement sur les effets de sources, en l’espèce particulièrement discontinues et composites. On pourrait ici pointer l’inadéquation des récits biographiques linéaires fondés sur un tel matériau<sup>4.</sup> »

D'où l'intérêt de notre travail qui permettra de mettre en ligne non seulement ce fonds déjà ouvert, mais aussi nombre de lettres encore inconnues du grand public et qui permettront de faire des recherches plus cohérentes sur Le Play et donneront l'occasion de voir apparaître de nouvelles études plus complètes.

### 3.1.1.3 Société d’économie et de science sociales (Paris)

Jean Albert Le Play conservait encore certaines lettres de son descendant. Celles-ci ont été confiées en 1986 à la SESS. On y trouve plus de 200 lettres que Frédéric Le Play adressa à sa mère Rosalie Le Play-Auxilion (1833) et à sa femme Augustine Le Play-Fouache (1837-1877).

Ces trois fonds familiaux constituent 1167 lettres adressées à 49 correspondants.

## 3.1.2 Des fonds dispersés à travers l’Europe

A côté de ces trois massifs, la correspondance est particulièrement éclatée dans de nombreuses bibliothèques et centres d’archives, tant en France (Archives nationales,

---

3. *Ibid.*

4. *Ibid.*

Bibliothèque nationale de France, bibliothèque Thiers, etc.) qu'à l'étranger (Angleterre, Belgique, Italie, Russie, etc.).

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.)

Enfin, on constate malheureusement que la correspondance autour de Le Play est lacunaire. Ainsi, aucune correspondance avec Albert de Saint-Léger et Adolphe Focillon, pourtant acteurs essentiels des premiers développements de la science sociale, n'a été retrouvée. Nous n'avons pas non plus trouvé trace d'Alexis Delaire, alors qu'il a eu un grand rôle à jouer dans l'organisation des UPS.

Par ailleurs, les correspondants étrangers sont rares, et aucune correspondance russe n'a été retrouvée, on pense surtout à Anatole de Demidoff, « commanditaire et employeur de Le Play durant une vingtaine d'années et qui lui a ouvert les portes de la Russie, découverte essentielle<sup>5</sup>. »

### 3.1.3 Une correspondance au service de l'Histoire

Le corpus, bien qu'incomplet, est largement prometteur. On y discerne déjà des thèmes principaux<sup>6</sup>, qu'il est important d'avoir en mémoire pour les futurs choix éditoriaux :

- La genèse des Ouvriers européens éclairée par la correspondance avec Jean-Baptiste Dumas et Augustin Cochin.
- L'organisation interne de la Société d'économie sociale (sur laquelle on ne possède pas d'archives pour le XIX<sup>e</sup> siècle) révélée par la correspondance avec Louis de Kergorlay.
- La volonté d'intervention publique de Le Play qui se manifeste à travers ses relations avec les catholiques libéraux (correspondance avec A. Cochin, Mgr Dupanloup et particulièrement Charles de Ribbe).
- La fondation expérimentale d'une « famille-souche » telle qu'elle apparaît dans la correspondance avec son fils, Albert.
- La conception du rôle de l'Église et la place de la religion dans la société (correspondance avec le père Hyacinthe). On y voit les importantes questions qu'y se posent autour du Concile Vatican I.
- La praxis réformatrice dont, lettres après lettres, Le Play précise sa conception à des interlocuteurs comme Emmanuel de Curzon et l'Écossais, David Urquhart.

5. *Ibid.*

6. Ces thèmes ont été relevés toujours par l'inventaire déjà nommé de 2006 que nous paraphrasons.

Nous pouvons conclure, toujours avec Messieurs Antoine Savoye et Stéphane Baciocchi, que nous avons là « de quoi alimenter un vaste programme de travail, tant sur l'émergence des sciences sociales que sur les évolutions de la pensée sociale dans la France du Second Empire et de la III<sup>e</sup> République commençante<sup>7</sup>. »

### 3.1.4 Nature des sources et première prise en main du projet

L'ensemble de cette correspondance compte 2091 lettres échangées entre Frédéric Le Play et 94 correspondants entre 1837 et 1882. Il s'agit donc de ne pas se perdre dans cette correspondance hétéroclite, à la fois active et passive.

Les premiers jours du stage ont donc été consacrés à un recensement des lettres qui nous ont été confiées<sup>8</sup>. En effet, le seul document de référence pour comprendre les

FIGURE 3.1 – Fonds numérisés, extrait de l'inventaire

Inventaire	Fonds (à défaut, nom du c)	Lieu de conservation	Cote	Correspondants	Nombre	Numérisati	Numéri
SES 2006	Fonds Dumas	Académie des sciences, Paris	Carton 29	Frédéric Le Play	23	Amateur	docx
SES 2006	Académie des sciences	Académie des sciences, Paris	Non précisé	Frédéric Le Play	1	Non	
SES 2006	Fonds Emmanuel d'Alzon	Archives assomptionnistes, Rome		Frédéric Le Play,	12	Non	
SES 2006	Fonds Montalembert	Archives départementales de Côte-d'Or	Non précisé	Frédéric Le Play	Inconnu	Non	
SES 2006	Fonds Despine	Archives départementales de Haute-Savoie	II J 437	Frédéric Le Play	26	Professionn	JPG (7)
SES 2006	Papiers du P. Félix	Archives jésuites, Vanves	JFE 80	Frédéric Le Play, P	6	Non	
SES 2006	Fonds Napoléon	Archives nationales, Pierrefitte-sur-Seine	400 AP 34	Frédéric Le Play	9	Non	
SES 2006	Sous-série F17	Archives nationales, Pierrefitte-sur-Seine	F17 4242	Joseph Boulatignie	1	Non	
SES 2006	Sous-série F17	Archives nationales, Pierrefitte-sur-Seine	F17 4142	Frédéric Le Play	2	Non	
SES 2006	Sous-série F17	Archives nationales, Pierrefitte-sur-Seine	F17 2897	Frédéric Le Play		Non	
SES 2006	Fonds Neftzger	Archives nationales, Pierrefitte-sur-Seine	113 AP	Frédéric Le Play	2	Non	
SES 2006	Fonds Emile Ollivier	Archives nationales, Pierrefitte-sur-Seine	542 AP 14	Frédéric Le Play	4	Non	
SES 2006	Fonds Persigny	Archives nationales, Pierrefitte-sur-Seine	44 AP 2, doss.	Victor de Persigny	2	Non	
SES 2006	Papiers Rouher	Archives nationales, Pierrefitte-sur-Seine	45 AP	Frédéric Le Play	4	Non	
SES 2006	Papiers Jules Simon	Archives nationales, Pierrefitte-sur-Seine	87 AP 4	Frédéric Le Play	2	Non	
Notes MBL	Sous-série F14	Archives nationales, Pierrefitte-sur-Seine	F14 2731/2, F	Frédéric Le Play	7	Amateur	JPG/PDF
SES 2006	Archives Pocquet du Haut-J	Archives Pocquet du Haut-Jussé	Non précisé	Frédéric Le Play	3	Non	
SES 2006	Fonds Lanjuinais	Archives privées B. de La Rochefoucauld	Non précisé	Frédéric Le Play	26	Non	
SES 2006	Fonds Armand de Melun	Archives privées de la vicomtesse de Mare	Non précisé	Frédéric Le Play	Inconnu	Non	
SES 2006	Archives privées Jean Rey	Archives privées de Mme Jean Reynaud (1)	Inconnu	Frédéric Le Play	10	Non	
SES 2006	Urquhart Papers	Balliol College, Oxford	IL 1, IM 1, IN 1	Frédéric Le Play	44		
SES 2006	Papiers Peruzzi	Biblioteca nazionale centrale, Florence	Cassetta XVIII	Frédéric Le Play	16	Oui (profess	JPG
SES 2006	Fonds Enfantin	Bibliothèque de l'Arsenal, Paris	7756 (93-L/8)	Frédéric Le Play	1		
SES 2006	Papiers Kérgorlay	Bibliothèque de l'Arsenal, Paris	Ms 14110, 141	Frédéric Le Play	8	Amateur	JPG
LE PLAY_BIB	Fabricant de châles parisie	Bibliothèque de l'Hôtel de Ville	Ms. 3261 f. 1 à	Fabricant de châle	1	Amateur	JPG
LE PLAY_BIB	Hausmann	Bibliothèque de l'Hôtel de Ville	Ms. 2194 f. 19	Hausmann	1	Amateur	JPG
LE PLAY_BIB	Général Mollard	Bibliothèque de l'Hôtel de Ville	Ms. 1105 f. 6	Général Mollard	1	Amateur	JPG
SES 2006	Fonds Le Play	Bibliothèque de l'Institut de France, Paris	Ms. 6062	Multiples		Professionn	JPG
LE PLAY_BIB	Divers	Bibliothèque historique de l'Hôtel de Ville	Ms. 3081 f. 13	Frédéric Le Play	3	Amateur	JPG
LE PLAY_BIB	Général Fleury	Bibliothèque historique de l'Hôtel de Ville	Ms. 3057 f. 39	Général Fleury	1	Amateur	JPG
SES 2006	Papiers Tourville	Bibliothèque nationale de France	Naf 17164 ff.	Frédéric Le Play	21	Professionn	PDF
SES 2006	Papiers Dupanloup	Bibliothèque nationale de France	Naf 24695 ff.	Frédéric Le Play	16	Professionn	PDF
SES 2006	Bibliothèque nationale de F	Bibliothèque nationale de France	Naf 22862, f. 1	Frédéric Le Play	2	Non	
	Bibliothèque nationale de F	Bibliothèque nationale de France	Naf 22862, f.	Frédéric Le Play	1	Non	
SES 2006	Jean-Baptiste Landriot	Bibliothèque nationale de France	Naf 11911, f. 4	Frédéric Le Play	1	Professionn	PDF
SES 2006	Frédéric de Mercey	Bibliothèque nationale de France	Naf, 228562, f.	Frédéric Le Play	3	Professionn	PDF

archives était l'inventaire de 2005. Nous ne disposons pas d'inventaire ou de description précise des fonds que nous avons pu récolter. Nous avons donc établi un tableau excel pour discerner quelles correspondances nous devions traiter en priorité, selon la qualité

7. *Ibid.*

8. Le stage étant en télétravail, nous avons dû procéder à un versement par WeTransfer, service de transfert de fichiers en ligne.

des numérisations que nous avions en notre possession. Nous nous sommes concentrée particulièrement sur les manuscrits écrits de la main de Frédéric Le Play afin de pouvoir commencer notre travail sur Transkribus, point que nous développerons davantage dans la deuxième partie.

Le CRHXIX est encore dans une phase d'acquisition des sources. Il en possède déjà des photocopies, matériau peu rentable pour notre future mise en ligne. Il détient également des numérisations, qui sont dans l'ensemble de qualité. Enfin, il possède des photos de médiocre qualité. Il sera nécessaire d'en demander par la suite les numérisations aux institutions en question.

Par la suite, nous avons eu accès à un fichier récapitulatif des transcriptions qui nous a été fort utile, et au cours de notre stage, un membre de l'équipe a également réalisé un fichier excel dont l'objectif était de faire un point sur la qualité des numérisations. Nous en trouvons ici<sup>9</sup> un aperçu qui nous permet de voir l'ampleur des fonds.

La prise en main du projet est donc déterminée par la nature des sources en notre possession. Ceci vaut aussi bien pour le CRHXIX que pour le Labex OBVIL.

## 3.2 De l'édition papier à l'édition numérique

Pour ELICOM, les sources du projet sont bien différentes. En effet, bien que ce soient également des lettres du XIX<sup>e</sup> siècle, nous travaillons cette fois-ci non plus sur la source directement ou sur sa numérisation, mais sur des imprimés qui ont été numérisées. Pour cela, Gallica est une mine inépuisable ou presque.

### 3.2.1 Gallica, une mine de savoirs

Gallica est la bibliothèque numérique de la BNF. Elle regroupe plus de six millions de documents, livres au format Epub, journaux, revues, images, enregistrements sonores, cartes, manuscrits et vidéos.

Parmi ces multiples documents, on trouve de nombreux ouvrages numérisés, dont d'anciennes éditions de correspondances du XIX<sup>e</sup> siècle. Ce sont elles qui ont été choisies dans le cadre du projet ELICOM pour être traitées afin d'être plus accessibles aux chercheurs et plus aisément fouillées.

Pour notre projet, un cahier des charges avait déjà été dressé il y a deux ans, mettant en exergue les sources intéressantes repérées sur le catalogue de la BNF, et redirigeant vers les sources présentes sur Gallica. Plusieurs noms ressortent de cette analyse.

---

9. Fig. 3.1

### 3.2.1.1 Pierre-Joseph Proudhon

Proudhon (1809-1865) reflète à lui seul la volonté de multi-disciplinarité d'ELICOM. Polémiste, journaliste, économiste, philosophe, politique et sociologue français, sa correspondance est riche de quatorze volumes<sup>10</sup>. Précurseur de l'anarchisme, il est le seul théoricien révolutionnaire du XIX<sup>e</sup> siècle à être issu du milieu ouvrier.

Durant notre stage, nous avons pu réaliser l'extraction du premier volume, soit près de cent lettres. Les correspondants sont relativement nombreux. On y trouve notamment le journaliste et premier disciple de Fourier, Just Muiron (1787-1881), le philologue Paul Ackermann (1812-1846) et l'orientaliste, sinologue et poète Guillaume Pauthier (1801-1873). Nous y reviendrons plus en détail dans la troisième partie de ce mémoire.

### 3.2.1.2 Louis Pasteur

Avec Louis Pasteur (1822-1895), dont nous avons retenu quatre volumes de correspondance<sup>11</sup>, nous sommes face à un corpus plus scientifique, qui souligne encore une fois la volonté trans-disciplinaire d'ELICOM.

Néanmoins, nous devons préciser que nous n'avons pas eu à traiter ce correspondant durant notre stage.

### 3.2.1.3 George Sand

Nous n'avons pas non plus traité la correspondance de George Sand (1804-1876), riche de six volumes<sup>12</sup>, correspondance cette fois d'une femme de lettres, depuis ses huit ans jusqu'à sa mort à l'âge de soixante-douze ans.

### 3.2.1.4 Alphonse de Lamartine

La correspondance de Lamartine (1790-1869), qui comporte six volumes<sup>13</sup>, est celle sur laquelle nous nous sommes le plus attardée, puisque c'est par elle que nous avons commencé notre stage et que nous nous sommes familiarisée avec les procédures à suivre.

Nous nous sommes attelée à l'extraction du premier volume, comportant près de cent lettres, où Lamartine correspond avec ses meilleurs amis, Prosper Guichard de Biennassis et Aymon de Virieu. Le deuxième volume comporte un nombre plus important de

10. *Correspondance de P.-J. Proudhon*, BNF Catalogue général, URL : <https://catalogue.bnf.fr/ark:/12148/cb35438908j> (visité le 08/09/2020).

11. *Correspondance de Pasteur, 1840-1895*, BNF Catalogue général, URL : <https://catalogue.bnf.fr/ark:/12148/cb32510650x> (visité le 08/09/2020).

12. *Correspondance : 1812-1876 / George Sand*, BNF Catalogue général, URL : <https://catalogue.bnf.fr/ark:/12148/cb31293789z> (visité le 19/05/2020).

13. *Correspondance de Lamartine*, BNF Catalogue général, URL : <https://catalogue.bnf.fr/ark:/12148/cb30725428p> (visité le 19/05/2020).

correspondants, parmi lesquels figurent encore Aymon de Virieu, et aussi Laurent de Jus sieu, Fortuné de Vaugelas, Éléonore de Canonge, la marquise de Raigecourt, le baron de Vignet, le comte de Saint-Mauris, le duc de Rohan, l'abbé Dumont et Eugène de Genoude.

### 3.2.1.5 Félicité de Lamennais

L'Abbé Félicité de Lamennais (1782-1854) présente une correspondance active forte de deux volumes<sup>14</sup>, adressée pour ce qui est du premier volume que nous avons traité, pour la plupart à des personnes de son rang, telles que Mademoiselle Cornulier de Lucinière, Monsieur le Comte de Senfft et Madame la Comtesse de Senfft, ou encore au Marquis de Coriolis<sup>15</sup>. Un volume représente en réalité trois tomes, et donc près de deux cent lettres d'une longueur assez importante en général. Le premier volume regroupe la correspondance de 1818 à 1823. Le deuxième couvre les années 1826-1827. Le troisième est entièrement consacré à l'année 1828.

### 3.2.2 Un traitement adapté à la nature des sources

Comme nous l'avons souligné plus haut, la nature des sources détermine en quelque sorte leur traitement ; même si nous avons encore une certaine liberté dans le choix des outils, ceux-ci doivent cependant s'adapter aux caractéristiques du corpus.

Ici, nous avons affaire à un corpus assez vaste et hétérogène. Les choix d'éditeurs se sont avérés être différents. Pour chaque correspondant, le cahier des charges met en avant un degré de difficulté selon les marqueurs présents ou non et permettant d'extraire les lettres avec plus ou moins de difficultés. La qualité de l'océrisation joue également un rôle et demande une relecture plus ou moins précise. Nous y reviendront plus spécifiquement dans les troisième et quatrième parties de notre mémoire.

En attendant, il convient de souligner qu'avant d'extraire des données et surtout de les traiter, il faut avoir une idée assez précise de ce que l'on veut en faire. En effet, il s'agit là, que ce soit pour ELICOM ou l'édition numérique de la correspondance de Frédéric Le Play, d'une édition numérique.

C'est d'une part une *édition* : une édition ne se fait pas à la légère. Elle s'accompagne de choix éditoriaux sentis et pesés.

C'est une édition *numérique* : le numérique apporte son enrichissement mais aussi ses contraintes. Une édition numérique ne se pense pas de la même façon qu'une édition papier.

C'est une édition numérique de *correspondance* : elle doit donc prendre en compte les caractéristiques du genre épistolaire, notamment avec les rituels épistolaires.

14. Correspondance de Lamennais, BNF Catalogue général, URL : <https://catalogue.bnf.fr/ark:/12148/cb30728369q> (visité le 08/09/2020).

15. Les correspondants sont nommés avec leurs titres dans cette correspondance.

Il est donc opportun de s'arrêter à toutes ces questions. Ce sera l'objet de notre deuxième partie.



## **Deuxième partie**

**Penser l'édition numérique de  
correspondance**



# Chapitre 4

## Bilan scientifique

### 4.1 Réflexions autour de l'édition numérique de correspondance. Une communauté scientifique grandissante

#### 4.1.1 Des communautés de réflexion, des projets et des publications...

Nombreux sont les communautés, réseaux, groupes de travail qui se retrouvent autour de réflexions sur l'édition numérique de correspondance. Celle-ci a en effet le vent en poupe depuis quelques années. Des sites se créent chaque année, nous en donnerons quelques exemples par la suite<sup>1</sup>.

En attendant, il est intéressant pour nous de voir quelles sont les personnes ou communautés qui œuvrent autour de l'édition numérique de correspondance.

##### 4.1.1.1 *TEI : Correspondence SIG*

Le *Special Interest Group* (SIG) ou pôle d'intérêt commun TEI sur la correspondance, sous la direction de Monsieur Stefan Dumont, de l'Académie des sciences de Berlin-Brandebourg, et de Madame Sabine Seifert, de l'Université de Potsdam, cherche à rassembler des universitaires intéressés par la création d'éditions scientifiques numériques de correspondance. Son but est donc de développer un ensemble de balises propres à différentes formes de correspondance en XML-TEI, ainsi que de créer des tutoriels et des modèles de bonnes pratiques<sup>2</sup>.

Par ailleurs, le *Correspondence SIG* tient une page Wiki intitulée *SIG :Correspondence*

---

1. Voir 4.2

2. *TEI : Correspondence SIG*, Site web de la TEI, URL : <https://tei-c.org/Activities/SIG/Correspondence/> (visité le 09/09/2020).

dence et mise à jour régulièrement<sup>3</sup>, et il envoie ses réflexions à une liste de diffusion, ouverte à tous ceux qui désirent s'y inscrire.

Depuis 2008, le SIG organise des journées annuelles autour de la correspondance. Celle de 2013 qui s'est tenue à Rome<sup>4</sup> a été particulièrement fructueuse, puisqu'elle a vu la création d'un groupe de travail intitulé « correspDesc » et composé de Madame Sabine Seifert, Messieurs Marcel Illetschko et Peter Stadler, dont l'objectif a été de créer et faire approuver un nouvel élément XML-TEI nommé <correspDesc>, décrivant l'action de la correspondance, notamment l'expéditeur et le destinataire<sup>5</sup>. Celui-ci est une référence aujourd'hui dans l'édition numérique de correspondance et nous a été précieux pour notre travail.

Afin de mieux appuyer leurs réflexions autour de l'édition numérique de correspondance, des publications sortent pour soutenir les chercheurs et ingénieurs de recherche dans leurs travaux d'édition. Ainsi, Stefan Dumont, Susanne Haaf et Sabine Seifert ont publié en 2018 un manuel intitulé *Encoding Correspondence. A Manual for Encoding Letters and Postcards in TEI-XML and DTABf*<sup>6</sup>. Le SIG est donc riche en initiatives pour guider ceux qui veulent se lancer dans l'édition numérique de correspondance. Il continue à être régulièrement alimenté et complété au fil des améliorations et découvertes techniques.

#### 4.1.1.2 Autres initiatives techniques du *Correspondence SIG* : le format CMIF et le service <CorrespSearch>

Le format<sup>7</sup> CMIF (*Correspondence Metadata Interchange Format*) est basé sur le module <correspDesc>. Son objectif est de pouvoir fournir les métadonnées les plus importantes permettant de partager des corpus de lettres quel que soit leur format<sup>8</sup>. Ce format CMIF permet de créer des fichiers XML rassemblant des éléments <correspDesc>, chacun représentant une lettre éditée. Chaque élément <correspDesc>, utilisé de manière

---

3. Voir *SIG :Correspondence, Page Wiki*, URL : <https://wiki.tei-c.org/index.php/SIG:Correspondence> (visité le 09/09/2020).

4. Voir le compte-rendu, *Site web de la TEI*, <https://tei-c.org/activities/sig/correspondence/tei-sig-on-correspondence-minutes-rome-oct-3-2013/> (visité le 09/09/2020).

5. *correspDesc*, *Site web de la TEI*, URL : <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-correspDesc.html> (visité le 09/09/2020).

6. *Encoding Correspondence. A Manual for Encoding Letters and Postcards in TEI-XML and DTABf*, *Site web Encoding Correspondence*, URL : <https://encoding-correspondence.bbaw.de/v1/> (visité le 05/05/2020).

7. Un format de données est la façon dont est représenté (codé) un type de données, sous forme d'une suite de bits. Voir « Format de données », *Wikipédia*, URL : [https://fr.wikipedia.org/wiki/Format\\_de\\_donnees](https://fr.wikipedia.org/wiki/Format_de_donnees) (visité le 10/09/2020).

8. Tout ce passage s'appuie largement sur *L'Édition numérique de correspondance ; Guide méthodologique*, *Site web du consortium CAHIER*, URL : [https://cahier.hypotheses.org/files/2018/03/Correspondance\\_CAHIER.pdf](https://cahier.hypotheses.org/files/2018/03/Correspondance_CAHIER.pdf) (visité le 17/06/2020).

très restrictive, se borne aux informations suivantes :

- Nom de l'expéditeur et ID contrôlée par l'autorité
- Nom du destinataire et ID contrôlée par l'autorité
- Date d'écriture et de réception
- Lieu d'écriture et de réception (nom et ID contrôlée par l'autorité)
- Numéro de la lettre dans l'édition savante
- URL de la lettre éditée

Les identifiants ou ID proviennent des fichiers d'autorité pour les personnes et les lieux, *GeoNames*<sup>9</sup> pour les lieux, VIAF (*Virtual International Authority File*) ou *data.bnf*<sup>10</sup> pour les identités<sup>11</sup>.

Le service <CorrespSearch> est une interface de programmation applicative (API, *Application Programming Interface*) destinée à être encapsulée dans une page Web, permettant de retrouver un ensemble de métadonnées identifiant des projets de correspondance (il ne s'agit pas d'un moteur de recherche avec une interface complète). Pour cela il faut qu'un projet d'édition de correspondance expose ses métadonnées afin de les rendre publiques à d'autres projets. Cet échange de données est basé sur la transmission de fichiers XML au format CMIF. Le schéma ci-dessous<sup>12</sup> permet de mieux en comprendre le fonctionnement. Une page web a par ailleurs été créée autour du CorrespSearch<sup>13</sup>

Nous reviendrons en quatrième partie sur l'importance de ces pratiques, standards, formats et leur mise en pratique dans nos projets.

#### 4.1.1.3 Le consortium CAHIER

Parmi les groupes de travail se retrouvant autour de l'édition numérique de correspondance, nommons particulièrement le consortium CAHIER (Corpus d'Auteurs pour les Humanités, Informatisation, Édition, Recherche). Celui-ci, constitué en fédération en 2011 dans le cadre de l'infrastructure « CORPUS » (désormais intégrée à la TGIR (Très

9. *The GeoNames geographical database*, URL : <https://www.geonames.org/> (visité le 10/09/2020).

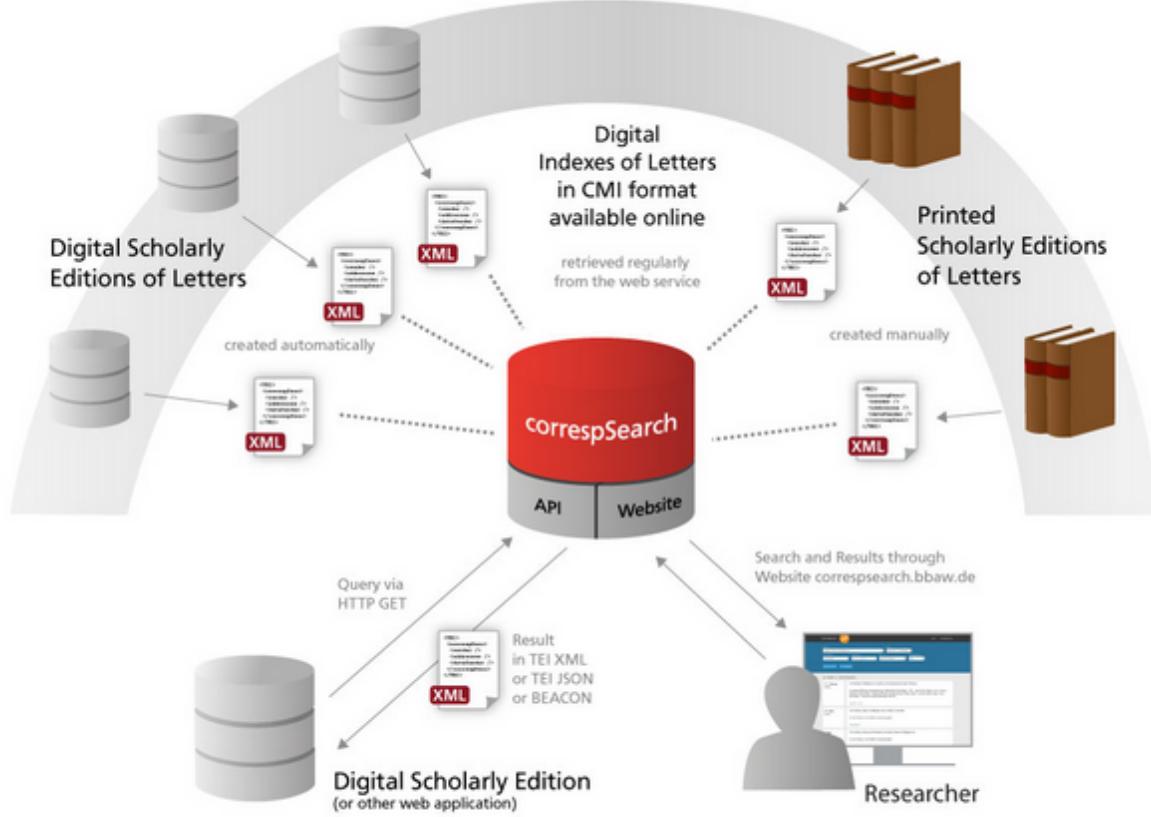
10. *Des fiches de référence sur les auteurs, les œuvres et les thèmes*, *data.bnf*, URL : <https://data.bnf.fr/> (visité le 10/09/2020).

11. Pour plus d'informations sur le CMIF, voir : *Perspectives of the further development of the Correspondence Metadata Interchange Format (CMIF)*, Site web digiversity, URL : <https://digiversity.net/2015/perspectives-of-the-further-development-of-the-correspondence-metadata-interchange-format-cmif/> (visité le 09/09/2020).

12. Fig. 4.1

13. *About our web service. The idea behind correspSearch*, Site web du CorrespSearch, URL : <https://correspsearch.net/index.xql?id=about&l=en> (visité le 09/09/2020). La présente explication s'appuie largement sur cet article.

FIGURE 4.1 – Fonctionnement du *correspSearch* - Site web du *correspSearch* en 2020



The graphic shows how *correspSearch* works.

grandes infrastructures de recherche) Huma-Num), CAHIER est un consortium interdisciplinaire de projets numériques, en accès libre, menés principalement dans les domaines des «corpus d'auteurs», qu'ils relèvent de la littérature, de la philosophie ou d'une thématique liée à une école ou à une pratique.

Comme elle le souligne elle-même<sup>14</sup>, la « fédération des différents projets existants ou projetés en France leur donne l'opportunité et les ressources » entre autres pour :

- augmenter l'acquisition de données de qualité (image et texte) tout en tenant compte des limites de taille
- proposer des normes de transcription suivant des objectifs éditoriaux clairement énoncés
- permettre leur indexation
- tester les différents modèles d'affichage du mode texte et du mode image et opérer des choix pertinents en fonction des publics

14. Accueil, Consortium CAHIER, URL : <https://cahier.hypotheses.org/le-consortium> (visité le 17/06/2020).

- offrir des métadonnées compatibles avec les standards de catalogage et d'archivage, d'identification et de protection des données ; moissonnage par Europeana, Isidore, Gallica [...]
- offrir les moyens d'évoluer vers le web sémantique, la visualisation, les entrepôts de données, les modes de représentation d'ensembles documentaires, et vers l'annotation collaborative.

Le consortium CAHIER, répondant aux objectifs qu'il s'est fixé, a publié en 2018, grâce au groupe de travail EVENT (Evaluation et Valorisation des Editions Numériques de Textes) un ouvrage intitulé *Les publications numériques de corpus d'auteurs – Guide de travail, grille d'analyse et recommandations*<sup>15</sup>. La même année, le groupe de travail « Correspondance » a publié un guide méthodologique rassemblant des recommandations pour l'édition numérique de correspondance, intitulé *L'Édition numérique de correspondance ; Guide méthodologique*<sup>16</sup> et qui nous a été très précieux pour nos deux projets.

Par ailleurs, depuis plusieurs années, la correspondance épistolaire conduit à de nombreuses réflexions philologiques menées par d'autres groupes de recherche comme l'AIRE (Association Interdisciplinaire de Recherches sur l'Épistolaire)<sup>17</sup>, fondée en 1981, mais à portée plus générale. Toutefois, elle est consciente que l'« édition électronique paraît particulièrement adaptée à ce texte discontinu, toujours inachevé et se prêtant davantage à la recherche qu'à la lecture suivie »<sup>18</sup> et a organisé dans ce sens, avec le Centre d'Étude et de Recherche Éditer/Interpréter de l'Université de Rouen (CÉRÉDI), un colloque dès 2007.

NOMBREUSES SONT donc les recherches dans ce sens.

#### 4.1.2 ...au service de normes et standards adaptés à la correspondance

Comme nous l'avons vu, tous ces groupes de travail font avancer la réflexion sur l'édition numérique de correspondance, et ceci avec des normes et standards propres. Revenons sur leur définition.

15. *Les publications numériques de corpus d'auteurs*, Site web du consortium CAHIER, URL : <https://cahier.hypotheses.org/guides-juridiques/les-publications-numeriques-de-corpus-dauteurs> (visité le 09/09/2020).

16. *L'Édition numérique de correspondance ; Guide méthodologique*, Site web du consortium CAHIER, URL : [https://cahier.hypotheses.org/files/2018/03/Correspondance\\_CAHIER.pdf](https://cahier.hypotheses.org/files/2018/03/Correspondance_CAHIER.pdf) (visité le 17/06/2020).

17. Voir *Association Interdisciplinaire de Recherches sur l'Epistolaire*, URL : <http://www.epistolaire.org/> (visité le 1/09/2020).

18. Colloque international « Éditer les correspondances », Site web de l'AIRE, URL : <http://www.epistolaire.org/evenements/editer-les-correspondances/> (visité le 10/09/2020)

#### 4.1.2.1 Normes et standards...

En anglais, un seul terme est employé, celui de *standard*. En français, on use d'une distinction.

Un **standard** est un ensemble de recommandations développées et préconisées par un groupe représentatif d'utilisateurs<sup>19</sup>. C'est donc un format élaboré par un petit nombre d'acteurs et adopté par des consortiums, des forums, c'est-à-dire des organisations non officielles.

En revanche, une **norme** est un document de référence élaboré par un organe de normalisation reconnu comme l'Organisation internationale de normalisation (ISO *International Organization for Standardization*) et l'AFNOR (Association française de normalisation) en France.

Ainsi, le W3C (*World Wide Web Consortium*) a en charge la normalisation de l'ensemble des protocoles d'Internet avec les standards de base comme HTTP, HTML et XML, ou encore des standards autour de l'interopérabilité et des services Web comme SOAP, des standards concernant les documents et le multimédia comme HTML, XML, CSS, et des standards liés à la sémantique et à la description de ressources comme XML Schema et RDF pour n'en nommer que quelques-uns<sup>20</sup>.

#### 4.1.2.2 ...pour l'édition numérique de correspondance

En parlant des communautés scientifiques, consortiums et groupes de travail qui les pensent, nous avons donc déjà évoqué certains de ces standards ou normes fort utiles pour les éditions numériques de correspondance en général et donc pour nos projets en particulier.

Ainsi, le DC (*Dublin Core*) est un standard de métadonnées consensuel établi par des professionnels provenant de diverses disciplines telles que la bibliothéconomie, l'informatique et le balisage de textes.

#### 4.1.2.3 XML

Avec l'encodage des textes, nous sommes particulièrement concernés par XML (*Extensible Markup Language*) qui provient de l'ancien standard SGML (*Standard Generalized Markup Language*). XML est « un format de données pur, très simple et documenté, conçu pour la description des documents textuels »<sup>21</sup>, ne possédant pas de jeu de balises prédéfini, et respectant les recommandations du W3C.

---

19. Voir *Indexation de ressources*, Site web du ministère de l'éducation nationale éduscol, URL : <https://eduscol.education.fr/numerique/dossier/archives/metadata/normes-et-standards> (visité le 10/09/2020) ou encore : <https://slideplayer.fr/slide/13821861/> (visité le 10/09/2020).

20. *Idem*.

21. Voir Ariane Pinche, « Séance 1 », *Cours M2 TNAH XML*, URL : [https://github.com/ArianePinche/coursTNAH\\_XML-TEI/blob/master/seance01/InitiationXML.md](https://github.com/ArianePinche/coursTNAH_XML-TEI/blob/master/seance01/InitiationXML.md) (visité le 09/10/2020).

Il a l'avantage de faciliter d'une part la lisibilité par les machines et par l'œil humain, d'autre part l'échange de données, et enfin la migration vers d'autres plates-formes, d'autres logiciels, et formats.

Nous reviendrons dans la quatrième partie sur l'intérêt d'avoir un document structuré, et sur la structure générale du document XML. Ce qui nous intéresse ici est surtout d'avoir une vue d'ensemble des efforts de réflexions faits jusqu'à ce jour autour de l'édition numérique de correspondance et de voir quels normes et standards on utilise dans ce domaine.

#### 4.1.2.4 TEI

Or, la TEI s'avère être particulièrement intéressante pour nos projets. Les avantages du XML TEI sont de s'intéresser au sens du texte plutôt qu'à son apparence, d'être indépendant de tout environnement logiciel particulier, et d'avoir été conçu par la communauté scientifique, qui est aussi en charge de son développement continu. Nous avons déjà parlé du TEI consortium plus haut. Voyons ce qu'est véritablement la TEI.

TEI est « un set de balises prédéfini et documenté dans les TEI guidelines<sup>22</sup> qui permet de procéder à une description “scientifique” et “sémantique” d'un texte »<sup>23</sup>. Il s'agit d'un format de codage de documents dit « structuré » : il a besoin d'un langage, XML, pour aider à la saisie d'un texte en lui donnant une structure compatible à la fois avec les exigences des différentes communautés qui l'utilisent et avec les possibilités des outils de consultation<sup>24</sup>.

TEI (All) n'est pas un schéma à proprement parler, mais plutôt un *framework*<sup>25</sup>, utile à la conception de son propre schéma. Il est fortement déconseillé d'utiliser un schéma englobant l'intégralité de la TEI. La conception d'un modèle adapté à ses données et son projet est extrêmement importante, et c'est ce qui va nous intéresser par la suite : comment construire un schéma propre à la correspondance en général (ce sur quoi les communautés scientifiques se penchent) et ceci fait, comment l'adapter à chacun de nos projets en particulier ?

Par ailleurs, d'autres langages s'avèrent être utiles pour générer des fichiers XML, comme le langage de programmation *Python* dont nous nous sommes servis dans le projet ELICOM, nous y reviendrons.

---

22. Voir *TEI guidelines*, Site web de la TEI, URL : <https://tei-c.org/guidelines/> (visité le 10/09/2019)

23. Ariane Pinche, « Séance 3 », *Cours M2 TNAH XML*, URL : [https://github.com/ArianePinche/coursTNAH\\_XML-TEI/blob/master/seance03/TEI.md](https://github.com/ArianePinche/coursTNAH_XML-TEI/blob/master/seance03/TEI.md) (visité le 10/09/2020).

24. Définition donnée dans le *Manuel d'encodage TEI Renaissance et temps modernes*, URL : [http://www.bvh.univ-tours.fr/XML-TEI/ManuelWeb/Manuel\\_TEI\\_BVH.html](http://www.bvh.univ-tours.fr/XML-TEI/ManuelWeb/Manuel_TEI_BVH.html) (visité le 11/09/2020).

25. Un framework est un ensemble cohérent de composants logiciels structurels, qui sert à créer les fondations ainsi que les grandes lignes de tout ou d'une partie d'un logiciel (architecture). Voir à ce sujet « *Framework* », *Wikipedia*, URL : <https://fr.wikipedia.org/wiki/Framework> (visité le 10/09/2020).

### 4.1.3 Des outils au service de l'édition numérique

Lors de nos stages, nous nous sommes appuyés sur nombre d'outils pour mener à bien nos projets.

Plusieurs d'entre eux étaient en libre accès. Parmi les éditeurs de texte, retenons particulièrement *Sublime Text*. *Oxygen XML Editor* nous a rendu aussi d'immenses services, notamment par son exigence et sa rigueur, veillant à la bonne indentation et la *TEI conformance* de nos fichiers XML. Elle est distribuée pour tous les systèmes d'exploitation. Elle est sous licence propriétaire.

Pour la formation des expressions régulières ou REGEX (*regular expression*), nous avons eu recours à un service en ligne *regular expressions 101*<sup>26</sup>. Quant à la transcription, nous avons eu recours à Transkribus, nous y reviendrons également.

En bref, le choix est vaste, les outils nombreux.

Passons du théorique à la pratique et penchons-nous désormais sur les premières réalisations d'édition numérique de correspondance. En effet, il ne s'agit pas pour nous de tout inventer à partir de rien, mais au contraire de nous inspirer de travaux de référence déjà réalisés.

## 4.2 Avec de nombreuses réalisations

Les éditions numériques de correspondances, en cours ou déjà achevées, sont nombreuses. Dressons-en une liste non-exhaustive. On peut citer notamment<sup>27</sup> :

- L'édition numérique de la correspondance de D'Alembert sous la houlette de l'Institut de France, avec plus de deux-mille lettres<sup>28</sup>.
- La plateforme EMAN (Édition de Manuscrits et d'Archives Numériques<sup>29</sup>) diffuse des corpus en respectant les standards de l'édition numérique et de l'interopérabilité. Ainsi, elle abrite le projet CORREZ (Édition des lettres internationales adressées à Émile Zola) de l'Institut des textes et manuscrits modernes (ITEM) - unité mixte de recherche entre l'ENS (École Normale Supérieure) et le CNRS (Centre national de la recherche scientifique). Plus de deux-mille quatre-cent lettres sont déjà en ligne<sup>30</sup>.
- Plus de onze-mille lettres de Juliette Drouet à Victor Hugo sont déjà en ligne dans un site dédié<sup>31</sup>.

26. *regularexpressions101*, URL : <https://regex101.com/> (visité le 10/02/2020).

27. Voir Annexes A.1

28. Voir *Accueil*, D'Alembert en toutes lettres, URL : <http://dalembert.academie-sciences.fr/Correspondance/> (visité le 10/09/2020).

29. *Accueil*, EMAN, URL : <http://eman-archives.org/EMAN/> (visité le 11/09/2020).

30. *CORREZ - Édition des lettres internationales adressées à Émile Zola*, EMAN, URL : <http://eman-archives.org/CorrespondanceZola/> (visité le 11/09/2020)

31. *Accueil*, Juliette Drouet, Lettres à Victor Hugo, URL : <http://www.juliettedrouet.org> (visité

- La correspondance inédite du géomètre Gaspard Monge (1746-1818) est déjà disponible en ligne<sup>32</sup>.
- Le Labex Obvil a déjà édité la correspondance de Jean Paulhan, disponible sur son site<sup>33</sup>.
- HUMA-NUM a édité la correspondance de l'éditeur et libraire Marc-Michel Rey<sup>34</sup>.

On pourrait allonger la liste d'une quarantaine de noms d'édition numériques de correspondance, mais ce n'est pas notre propos. Quoi qu'il en soit, il est intéressant ici de voir combien ces éditions sont de plus en plus nombreuses et présentes sur le *cloud*.

FIGURE 4.2 – Lettre d'Adolf von Buch à Louis de Beausobre (Magdebourg, 15 janvier 1761), *Lettres et textes : Le Berlin intellectuel des années 1800*

The screenshot shows a digital edition of a historical letter. On the left is a scan of the handwritten letter in German script. On the right is the XML transcription with annotations. The top right has buttons for 'Image normale' and 'Fermer la 2ème colonne'. Below that are links for 'Télécharger tout le fichier XML' and 'POWERED BY TEI'. The main area is titled 'PAGE ACTUELLE' and contains the XML code for the transcription. At the bottom, there are links for 'Scan', 'Version dipl.', 'Version de lecture', 'Métadonnées', 'Entités', and 'XML'.

Pour nous, trois éditions numériques nous ont particulièrement inspirées, pour l'architecture du site et les choix d'encodage, à savoir :

- L'édition électronique des lettres de Flaubert, dirigée par Yvan Leclerc et Danielle Girard, avec la collaboration d'une trentaine de chercheurs et mise en ligne en 2017. Celle-ci présente les manuscrits et leurs transcriptions, ce qui est également notre

le 02/07/2020)

32. Voir *La correspondance inédite du géomètre Gaspard Monge (1746-1818)*, EMAN, URL : <http://eman-archives.org/monge/> (visité le 10/09/2020)

33. Voir *Correspondance de Jean Paulhan*, Site web OBVIL, URL : <http://obvil.sorbonne-universite.site/corpus/paulhan/> (visité le 10/09/2020)

34. *Marc Michel Rey*, HUMA-NUM, URL : <http://rey.huma-num.fr/presentation> (visité le 10/09/2020)

intention, et leur site comporte une organisation intéressante que nous aimerais partiellement reproduire<sup>35</sup>.

- L'édition numérique de la correspondance de Proust. « 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 Consortium “Proust21” » : ainsi le projet est-il présenté sur leur site *Corr-Proust*<sup>36</sup>.
- L'édition numérique des correspondances berlinoises au XIX<sup>e</sup> siècle, intitulée *Lettres et textes : Le Berlin intellectuel des années 1800*<sup>37</sup> nous a particulièrement inspirée pour l'encodage, d'autant plus qu'elle a été réalisée par des spécialistes tels que Madame Sabine Seifert. Ci-dessus se trouve une capture d'écran du site avec d'un côté le manuscrit, de l'autre l'encodage en XML-TEI<sup>38</sup>.

Ainsi, nombreuses sont déjà les éditions numériques de correspondance.

Après avoir donc vu quelles communautés scientifiques guidaient les choix et les standards autour de l'édition numérique de correspondance, ainsi que les premières réalisations d'éditions, penchons-nous sur les problématiques et spécificités qu'elles supposent.

---

35. *Correspondance*, Site web du Centre Flaubert, URL : <https://flaubert.univ-rouen.fr/correspondance/edition/> (visité le 17/06/2020)

36. *Le Projet, Corr-Proust*, URL : <http://proust.elan-numerique.fr/presentation/project> (visité le 08/06/202).

37. *Lettres et textes : Le Berlin intellectuel des années 1800*, URL : <https://www.berliner-intellektuelle.eu/?fr> (visité le 19/05/202).

38. Le manuscrit a été volontairement zoomé pour la capture d'écran.

# Chapitre 5

## Problématiques et spécificités de l'édition numérique de correspondance

L'édition numérique de correspondance questionne. Nous l'avons vu à travers tous les groupes de recherche qui travaillent dessus. Considérons plus spécifiquement les problématiques et spécificités qu'elle entraîne : c'est une édition, et numérique, et de correspondance, des termes à prendre en compte et qui sont tous à peser.

### 5.1 Édition et numérique

#### 5.1.1 Trois niveaux d'édition

Tout d'abord, une édition est quelque chose de pensé. On entend par édition (papier ou numérique) le fait de reproduire et diffuser une œuvre, ici intellectuelle.

Or, le numérique est par excellence un moyen de diffusion, et il permet un accès plus large au savoir.

On distingue trois types ou niveaux d'édition dans le numérique, d'après le guide de travail intitulé *Les publications numériques de corpus d'auteurs*<sup>1</sup> :

Tout d'abord, il y a les « archives éditorialisées » qui proposent des collections d'images et, ou de textes, pour permettre la consultation de ressources rares et ce de façon rapide, avec une correction minimale des fautes d'océrisation. C'est le niveau d'édition le moins élaboré. Nous sommes allée plus loin pour nos deux projets.

Le deuxième type ou niveau de publication est l'édition de lecture ou « *reading edition* ». Le texte a été bien relu et il est documenté de choix éditoriaux et autres accompagnements. Nos projets se rapprochent davantage de ce type d'édition ainsi que du troisième qu'est l'édition enrichie.

---

1. Ioana Galleron, Marie-Luce Demonet, Cécile Meynard, Idmhand Fatiha, Elena Pierazzo, et al., *Les publications numériques de corpus d'auteurs - Guide de travail, grille d'analyse et recommandations*, 2018, URL : <https://halshs.archives-ouvertes.fr/halshs-01932519/document> (visité le 05/05/202).

Ce dernier type d'édition est le plus poussé. Le texte est très enrichi, avec de nombreuses informations documentaires et contextuelles.

Il est à noter que les frontières entre chacune de ces trois types d'édition restent un peu floues. Pour nos projets, nous nous situons donc entre les deuxième et troisième types d'édition.

### 5.1.2 Cinq dimensions à prendre en compte

Par ailleurs, les éditions numériques ont cinq dimensions fondamentales. Elles ont certains points communs avec l'édition papier, et d'autres plus spécifiques<sup>2</sup>.

#### 5.1.2.1 Le texte

Tout d'abord, tout comme l'édition papier, l'édition numérique dispose d'un texte : pour nos deux projets, ce sont des correspondances du XIX<sup>e</sup> siècle. La publication de ces textes doit respecter certaines règles, à savoir que le texte publié doit être complet, et les choix éditoriaux explicités et appliqués de façon stable sur l'ensemble du texte. Il peut être aussi intéressant de comparer la nouvelle version publiée avec les précédentes.

Tout d'abord, l'édition des textes nécessite une numérisation soigneuse et de qualité. Pour ELICOM, nous ne sommes pas concerné par cela, mais pour l'édition numérique de la correspondance de Frédéric Le Play, nous sommes encore en cours de numérisation<sup>3</sup>. En effet, « la lisibilité des images est essentielle », et il est nécessaire d'avoir une bonne résolution de DPI (*Digit per Inch* ou point par pouce) ainsi qu'une « juste évaluation des besoins de stockage et d'infrastructure matérielle pour la diffusion/communication de celles-ci »<sup>4</sup>. Pour aucun des deux projets, nous ne nous sommes penchée sur la question du stockage. Cela reste un point à éclaircir particulièrement pour le projet du CRHXIX.

Enfin, la publication en mode texte nécessite d'expliquer les choix scientifiques de normalisation et modernisation. Pour ELICOM, normalisation et modernisation sont notre politique, sachant que les chercheurs n'auront pas accès au manuscrit donc il y aura une perte de données (assez minime toutefois car cela reste de la correspondance du XIX<sup>e</sup> siècle, donc assez moderne). Pour le CRHXIX, il reste encore quelques points d'interrogation, mais de toutes façons, les chercheurs auront accès à l'image et à sa transcription, donc les pertes de données seront moindres notamment pour les chercheurs intéressés par l'histoire de la langue. Cependant, précisons à nouveau que le projet Le Play est surtout destiné aux chercheurs historiens et sociologues.

---

2. Voir en annexe (A.2) à ce sujet la grille d'évaluation des publications numériques de corpus d'auteur, tirée du guide déjà cité (note 1).

3. Nous reviendrons sur cette question de la numérisation lors de la partie III sur l'acquisition des données.

4. *Idem.* p. 7.

### 5.1.2.2 Les métadonnées et l'annotation

De plus, la question des métadonnées et de l'annotation doit retenir notre attention. On entend par métadonnées « l'ensemble structuré d'informations permettant de décrire la ressource, de la classer, de l'organiser et de caractériser des données ou du contenu »<sup>5</sup>.

On distingue plusieurs types de métadonnées. Celles-ci peuvent être **descriptives**, permettant d'identifier la source. Elles peuvent être aussi **administratives**, apportant des informations sur les droits d'accès et d'usage notamment. Il existe aussi des métadonnées **structurelles**, décrivant la structure des sources (indication des vers par des balises <1>, des paragraphes par les <p>, et plus spécifiquement pour la correspondance, indication des destinataires, expéditeurs, lieux de rédaction et dates, et en fonction de la question de recherche sous-tendant l'édition ; il peut y avoir un encodage permettant d'accéder à différentes versions du texte, une normalisée et une non-normalisée par exemple<sup>6</sup>, et un encodage des entités nommées, à savoir les noms de lieux, de personnes d'institutions, d'événements, de dates etc. à des fins d'analyse de réseaux ou autres, ce qui a été fait pour nos projets, et enfin des métadonnées **techniques**, indiquant les outils utilisés pour la production des données.

Par ailleurs, certaines métadonnées sont particulièrement intéressantes, ce sont les métadonnées d'**enrichissement**, comportant des annotations permettant d'analyser et d'interpréter la source.

Or, toutes ces métadonnées doivent respecter les normes et standards internationaux dont nous avons parlé plus haut. Pour nous, qui avons dans chacun de nos projets, réalisé une édition en XML-TEI, ces métadonnées se trouvent dans l'en-tête de chaque fichier XML-TEI, appelé <`teiHeader`>.

Par ailleurs, la présentation de ces métadonnées est normalisée. Ainsi, les dates s'écrivent sous le format AAAA-MM-JJ, les noms de lieux ainsi : PAYS, ville, les noms de personnes : NOM, Prénom<sup>7</sup>.

### 5.1.2.3 La description du projet scientifique

Comme le souligne le guide déjà cité,

« un projet scientifique d'édition numérique est défini par la qualité de sa documentation, ce qui signifie que la description du projet est fondamentale.

[...] **Une édition qui n'expose pas sa question de recherche et ne déclare pas ses critères de numérisation et de gestion des sources, n'est pas une entreprise scientifique**<sup>8</sup> ».

---

5. *Ibid.*, p. 7

6. Il est question d'un encodage de ce type pour le CRHXIX.

7. Malheureusement, des relectures attentives seront à faire pour certains de nos fichiers, car certaines normalisations laissent à désirer, notamment les noms écrits en minuscules au lieu de majuscules. En général, les dates ont été bien normalisées.

8. *Ibid.*, p. 10.

La documentation comporte donc la description des enjeux scientifiques du projet, ce que nous avons plus ou moins décrit dans la première partie de ce mémoire, avec la présentation de l'équipe et des responsabilités de chacun, la composition du corpus et la localisation des sources numérisées, l'explicitation des critères qui ont accompagné le choix des sources si nécessaire, les critères de transcriptions et le traitement des erreurs présentes dans la source, de la ponctuation, et les choix d'encodage (dernier point sur lequel nous reviendrons lorsque nous parlerons de l'ODD).

#### **5.1.2.4 L'interface de consultation**

L'interface de consultation de l'édition numérique doit être conçue en tenant compte d'une part de son accessibilité, d'autre part de sa réutilisabilité.

Tout d'abord, l'édition électronique, qui utilise le Web pour communiquer le texte édité, doit être le plus possible accessible, c'est à dire qu'elle doit pouvoir être ouvert au plus grand nombre possible d'utilisateurs. Pour cela, il convient d'avoir recours à l'utilisation de standards ouverts, à un jeu de caractère UTF-8, et de respecter les normes d'accessibilité proposées par le W3C.

Il convient de donner accès au code source du document. Pour cela, le recours aux plateformes de partage, de type « git », facilite la mise à disposition de ces codes et sources. Pour l'instant, la question se pose encore pour le CRHXIX, le projet n'étant qu'entamé. Pour ELICOM, nous utilisons déjà Github<sup>9</sup> pour notre projet.

Une double version de transcription, l'une normalisée, l'autre non, est un plus pour l'accessibilité. Par ailleurs, l'accès libre et donc non payant aux sources facilite l'accessibilité, on peut penser notamment à la licence ouverte de type Creative Commons.

Enfin, pour ce qui est de la réutilisabilité, il est bon de donner la possibilité d'explorer les contenus, de présenter des informations structurées aux lecteurs et de présenter le texte selon différents points de vue et perspectives.

Par ailleurs, la question du stockage des données et de leur maintenance se pose.

#### **5.1.2.5 La gestion des données**

Dans le numérique, les techniques et technologies évoluent à grande vitesse. Or, il est important que l'édition diffusée en ligne soit consultable à long terme, malgré l'instabilité du numérique.

Pour cela, il est « nécessaire de mettre en place un plan de conservation, qui prenne en compte tout particulièrement l'exigence de citabilité de l'édition, sans laquelle celle-

---

9. Nous reviendrons sur l'usage de Github dans la partie IV.

ci ne peut pas jouer pleinement son rôle dans le milieu académique<sup>10</sup>. » La citabilité relève de deux conditions, d'une part, un format de nommage stable de la ressource sur le Web, d'autre part, de l'indication d'une modalité de citation conforme aux normes bibliographiques.

Pour ce qui est de la pérennité des données et de leur gestion, il est bon de préparer un Plan de Gestion des données *Data Management Plan* (DMP) « document évolutif qui aide et explicite de quelles façons les données utilisées et générées par le projet seront utilisées<sup>11</sup>. »

NOMBREUSES SONT donc les dimensions à prendre en compte pour une édition numérique en général. Or, il s'agit pour nous d'édition numérique de *correspondance*. Celle-ci, comme nous l'avons déjà maintes fois souligné, a ses caractéristiques.

## 5.2 Correspondance et numérique

### 5.2.1 Importance de l'épistolaire dans la recherche

Tout d'abord, il convient de souligner à nouveau combien l'épistolaire est un atout pour la recherche. En effet, les lettres révèlent souvent l'intime de l'auteur. Elles sont une occasion pour lui de rappeler son amitié, donner des nouvelles, s'épancher sur sa vie familiale et privée, mais aussi d'exprimer ses opinions politiques, et être parfois objet de conflits<sup>12</sup>.

### 5.2.2 Spécificités de la correspondance

La lettre est un objet singulier qui a ses spécificités propres. La lettre n'est jamais seule et se trouve dans un réseau de lettres, l'expéditeur écrit à tout un réseau de destinataires, et reçoit lui-même des lettres. Comment donc rendre toute cette richesse ? Nous l'avions déjà souligné dans l'introduction, la correspondance est un genre protéiforme, réticulaire et elliptique. Il s'agit donc de prendre en compte toutes ces caractéristiques pour nos éditions.

Par ailleurs, quels objets considère-t-on comme publiable dans une édition numérique de correspondance ? Doit-on tenir compte des objets qui accompagnent la lettre ? Une fleur jointe, l'enveloppe qui l'entoure, le petit mot annexe qui y est joint ?

---

10. *Ibid.*, p. 11.

11. *Ibid.*, p. 12.

Nous n'avons pas eu à traiter ces points dans nos projets, notamment pour cause de manque de temps.

12. Voir Élisabeth Gavoille et François Guillaumont (dir.), *Conflits et polémiques dans l'épistolaire*, Tours, Presses universitaires François-Rabelais, 2015, URL : <https://books.openedition.org/pufr/10853> (visité le 02/07/2020)

Pour ELICOM, la question ne se pose pas vraiment, puisque nous nous contentons de reprendre des éditions papier où les choix ont déjà été faits, et où les objets joints et les enveloppes n'ont pas été retenus. Par ailleurs, il est important de se rappeler qu'ELICOM n'a pas la même finalité qu'une simple édition numérique de correspondance et qu'elle veut arriver à la fouille des textes et leur enrichissement, ce qui nécessite peut-être moins une focalisation sur les objets entourant le texte. On pourrait certes objecter que ces objets donnent son sens au texte, et c'est vrai, rien n'est anodin, néanmoins, nous sommes contraints à faire des choix réalistes pour pouvoir mener à terme notre projet, et ce choix se traduit dans le renoncement à tout dire et tout décrire. On est obligé de se limiter au texte.

Pour l'édition numérique de la correspondance de Le Play, dans les numérisations que nous avons reçues, nous n'avons pas encore trouvé trace d'objets insolites. Nous partons plutôt sur une édition qui se concentre sur la lettre en elle-même étant donné que les sources sont presque exclusivement des lettres sans leur enveloppe. Est-ce un choix qui a été fait en amont par la personne qui a numérisé ? C'est possible. Quoi qu'il en soit, nous n'avons pas eu à traiter d'enveloppes ni d'objets joints, mais si le cas se présentait, nous serions plutôt pour leur publication : autant profiter des avantages du numérique pour donner une vision la plus complète possible de la lettre. C'est le choix qui a été fait pour l'édition numérique de la correspondance de Proust que nous avons évoquée plus haut, et c'est vraiment un plus. Par ailleurs, la correspondance de Frédéric Le Play est parfois enrichie d'un article de journal commenté dans la lettre et annoté de sa main. Certes, un article de journal est un document *a priori* non épistolaire, mais pour nous, cela fait partie des papiers qui rentrent dans l'édition numérique de correspondance, d'autant que pour Le Play, ce sont des éléments importants pour cerner sa pensée et son évolution. Une carte postale ou carte de visite, un télégramme (nous n'en avons pas encore trouvé trace) font également partie de ce que nous pourrions éditer.

Ainsi, les objectifs éditoriaux et scientifiques permettent pour nos deux projets « de déterminer les critères de délimitation du corpus et poser les bases d'une stratégie éditoriale à long terme »<sup>13</sup>.

Une question se pose aussi pour l'édition de correspondance. Que traiter ? La correspondance active, la correspondance passive ? Pour l'édition numérique de la correspondance de Frédéric Le Play, nous partons sur une édition à la fois active et passive, contrairement à ELICOM qui traite surtout de la correspondance active. Certaines éditions papier disponibles sur Gallica glissent un peu de correspondance passive, mais c'est un phénomène très marginal. C'est pour nous l'occasion de remarquer encore une fois combien les sources et les données de départ orientent nos choix éditoriaux. OBVIL se

---

13. Richard Walter (dir.), *L'édition numérique de correspondances – guide méthodologique*, 2018, p.9, URL : <https://cahier.hypotheses.org/guide-correspondance> (visité le 17/06/2020)

base sur des sources déjà éditées, alors que le CRHXIX choisit davantage, n'étant pas limité par une précédente édition, ce qu'il veut ou non mettre en valeur.

Or, pour une édition numérique de correspondance, le traitement de la correspondance passive s'avère être un des grands avantages, d'autant que souvent, certaines lettres manquent, et que cette « gestion de l'*incertitude* » est un des grands défis de l'édition de correspondance. Dans une édition papier, nous n'avons pas le choix. Il faut savoir s'arrêter de chercher et publier. Puis, quelques années après, on retrouve une lettre dans un grenier, une autre lettre perdue dans un autre fonds et qui avait mal été classée par l'archiviste, ou tout autre chose : que faire de cette lettre, comment la mettre en valeur ? Pour les personnes ayant une correspondance innombrable, il est malaisé de publier intégralement sans manquer leurs lettres. C'est le cas de la correspondance de Jacques Maritain (1882-1973) conservée à la Bibliothèque nationale universitaire (BNU) de Strasbourg. Des éditions de sa correspondance avec Max Jacob, Jean Cocteau, Julien Green et tant d'autres correspondants sont régulièrement publiées. Et lorsqu'une nouvelle lettre est trouvée dans ce fonds sans fonds, le Cercle d'études Jacques et Raïssa Maritain se charge de publier la trouvaille dans les *Cahiers Jacques Maritain*. Ainsi, dans le numéro 67 de ces carnets se trouve un passage dédié à la « Correspondance Journet-Maritain : Nouvelles lettres retrouvées ».

L'avantage de l'édition numérique de correspondance est donc de permettre des ajouts progressifs à cette édition qu'on sait n'être jamais définitive et toujours susceptible d'être agrandie, et ces ajouts se font directement dans l'édition et non sur un cahier ou une publication extérieure. Cela permet d'avoir une meilleure vue d'ensemble, ou tout au moins une vue la plus complète possible de la correspondance.

Ainsi, la correspondance a ses spécificités par rapport à l'édition d'une autre œuvre comme un roman ou un ouvrage. Elle n'a pas été écrite pour être publiée et l'éditeur se trouve donc à devoir faire des choix et gérer nombre d'incertitudes.

Parmi ces choix se trouvent les choix de transcriptions.

### 5.2.3 Des choix éditoriaux à faire en amont

Des choix sont à faire en amont afin de mieux diriger notamment l'équipe de transcription. Comme nous l'avons déjà souligné, pour ELICOM, la question ne se pose pas autant que pour l'édition numérique de la correspondance de Frédéric Le Play dont nous allons davantage parler dans cette partie.

Certains principes de transcription nous ont paru plus évident que d'autres.

### 5.2.3.1 Principes de transcription

Une équipe de transcription a été mise en place bien avant notre arrivée afin de procéder à une première transcription des lettres de Frédéric Le Play. Certains principes avaient déjà été établis<sup>14</sup>.

Le premier principe retenu a été celui de la fidélité à la lettre originale.

Une des difficultés de l'écriture de Le Play est l'emploi fréquent de majuscules pour les noms communs. Pour résoudre ce problème, il a été convenu de respecter l'usage des majuscules par Le Play lorsqu'il semble avoir un sens précis, selon ses usages. Par exemple, Réforme plutôt que réforme ; ou les Autorités Sociales. Cependant, on retiendra l'usage actuel lorsque les majuscules n'ont pas lieu d'être : par exemple, « le concours de nos amis », et non pas « le Concours de nos Amis ». Les noms de mois ne prennent pas de majuscule : janvier, février, etc.

L'accentuation doit être modernisée pour correspondre aux usages actuels et pour la commodité de la lecture.

Il a également été convenu de respecter les abréviations utilisées par Le Play, ainsi que l'orthographe, avec indication du [sic.] pour enlever toute ambiguïté. De même pour les noms propres de personne ou de lieu mal orthographiés, il a été décidé de le transcrire tel quel et d'indiquer en note le nom correct.

Si un titre d'ouvrage est donné de manière approximative, on le transcrit tel quel et on indique en note le titre correct, avec la date d'édition à laquelle Le Play fait référence si elle est connue (sinon on indique la 1<sup>ère</sup> édition). Cette recherche se fait à partir du catalogue général de la Bibliothèque nationale de France, disponible en ligne.

Pour ce qui est de la ponctuation, elle peut paraître fantaisiste à nos yeux et gêner la compréhension. Elle peut donc être en partie corrigée pour la commodité de la lecture, tout en conservant les point-virgules et points d'exclamation originaux.

Par ailleurs, il est important de conserver la structure générale de la lettre, avec les alinéas, les changements de pages (indiqués par un double slash (//) dans la transcription).

De même, on conserve la présentation décidée par Le Play, telle qu'un mot souligné, barré, les hésitations et développements de la pensée étant intéressants.

Le Play place fréquemment des étoiles (\*) dans ses lettres, complétant sa pensée dans la marge. Il avait été convenu de reproduire les étoiles mais de placer la transcription en bas de page ou en fin de lettres. Pour notre part, nous serions plus intéressée pour, certes, reproduire les étoiles, mais placer la transcription comme Le Play l'a fait, dans la marge. Il conviendra de voir si ce souhait n'est pas trop difficile à mettre en œuvre. Nous y reviendrons lorsque nous évoquerons le balisage.

Parfois, on se heurte à la difficulté de lecture : certains mots nous semblent être illisibles. Si tel est le cas, et que le problème n'est pas résolu, malgré l'aide apportée par

---

14. Nous reproduisons presque mot à mot dans ce qui suit, les principes de transcriptions établis par Monsieur Matthieu Brejon de Lavergnée.

un autre transcripteur, on indique entre crochets combien de mots n'ont pas été transcrits.

Bien-sûr, si on reconstitue par déduction des informations manquantes pour les lieux et dates d'écriture (en fonction du contexte de la lettre), on indique ces déductions entre crochets.

Il a été établi de transcrire l'intégralité de la lettre. Or, une lettre comprend notamment :

- Le lieu de rédaction et la date
- Une « adresse » ou formule de politesse débutant la lettre : Cher ami, Éminent collègue, Monseigneur etc.
- Une formule de politesse finale
- Une signature : en général, F. Le Play pour Frédéric Le Play. Si elle n'y est pas, on indiquera qu'elle est manquante.
- Éventuellement un post-scriptum, ou des notes se trouvant après la signature.

Toutes ces caractéristiques font partie du rituel épistolaire et devront être soulignées dans l'encodage en XML-TEI, et ceci aussi bien pour l'édition numérique de la correspondance de Le Play que pour ELICOM.

Chaque transcription devra renseigner les nom et adresse à laquelle la lettre a été envoyée, si cela est mentionné, ainsi que les cachets postaux et le fonds d'où elle provient, avec son nom, sa cote, le folio, ou si c'est une copie, le type de support (photocopie, image numérisée) et le lieu où la copie est conservée.

### 5.2.3.2 Hésitations sur certains choix de transcription

Lors de la reprise de certaines transcriptions, nous avons eu quelques hésitations. En effet, Le Play ayant l'habitude d'écrire des noms communs au cours d'une phrase en mettant une majuscule au début du mot, et aussi de ne pas mettre les accents sur les « e », ou encore d'écrire des abréviations avec des chiffres pour les mois, comme 7bre pour septembre - ce qui n'est pas forcément aisément pour quelqu'un qui n'est pas habitué à ces pratiques - nous nous sommes demandés si nous ne devions pas exploiter à fond les possibilités de la TEI et faire ainsi deux versions du texte, une qui respecte l'original, une qui est normalisée. Ceci multiplie le nombre de balises à mettre, mais au sein d'un seul et unique fichier XML qui, certes, comprendra donc plus de balises, mais pourra être affiché de deux façons différentes.

La question reste en suspens étant donné que nous avons manqué de temps pour faire des tests et évaluer la durée supplémentaire de travail d'encodage que cela impliquerait. Néanmoins, cela risque d'être un surcroît de travail, probablement pas assez rentable, étant donné que, si nous avons déjà accès à une copie numérisée du manuscrit, un expert pourra se référer directement à la source s'il a un doute et donc il n'y aura pas de vraie perte de données.

Par ailleurs, un point nous fait pencher pour une seule version de transcription partiellement normalisée : on constate déjà que la qualité des transcriptions est très variable : certaines personnes n'ont pas respecté intégralement les consignes ou n'ont pas su lire le manuscrit, n'ayant pas l'œil assez entraîné (l'écriture de Le Play est parfois difficile à lire). Les transcriptions ne sont donc pas uniformisées sur la question des accents et des majuscules, même sur Transkribus<sup>15</sup> où il nous a paru important de coller au texte sans modification pour un meilleur apprentissage. La relecture sera donc un moment-clé et qui prendra beaucoup de temps et devra être extrêmement attentive à l'original tout en suivant les principes de transcription établis plus haut.

### 5.2.3.3 Indication des métadonnées

Afin de procéder par ordre et ne perdre aucune information, chaque transcripteur décrit précisément la lettre qu'il transcrit en remplissant un tableau de la façon suivante, avec dates, lieu de rédaction et de conservation, cote, nombre de pages<sup>16</sup> :

FIGURE 5.1 – Tableau récapitulatif

N° lettre	Auteu r	Corresponda nt	Lieu	Date / anné e	Date/ mois	Date / jour	Lieu de conservati on	Cote	Original / copie	Nb de pages	Transcrit par
1	Le Play	Jules Baroche	Paris	1859	février	2	Paris, Bibliothèq ue Thiers	Ms Thiers 997, folios 556-558	original		Jeanne Dupont, 2019.
	2										
	etc.										

### 5.2.3.4 Annotation des transcriptions

Le transcripteur est également chargé de l'annotation, élément clef pour la bonne compréhension et la mise en contexte de la correspondance. Ces notes se font par appel de bas de page et portent sur :

- Tous les noms propres cités : avec prénom et nom, dates de naissance et mort, biographie succincte (deux ou trois lignes maximum). On peut préciser tel élément de contexte permettant de comprendre la relation entre Le Play et son correspondant à ce moment-là.
- Tous les noms d'ouvrages cités, avec le titre exact, l'édition citée ou à défaut, la première édition.

15. La question de Transkribus sera davantage évoquée dans la troisième partie.

16. Tableau proposé par Monsieur Matthieu Brejon de Lavergnée. Nous reprenons encore ses directives.

- Un nom de lieu fautif dans la lettre doit être précisé en note. S'il est fait allusion à un lieu peu connu (village), ou situé à l'étranger, il faut donner la précision nécessaire en note.
- Chaque transcripteur, selon son appréciation, annotera tel événement cité, telle référence obscure, trouvant l'équilibre entre trop et trop peu. On peut également indiquer des références bibliographiques, renvoyer à un article.
- Une note initiale rappelant l'objet de l'échange peut s'avérer utile.

#### 5.2.3.5 Index

Une grande question qui se pose pour l'édition est celle de l'indexation.

Trois index ont été envisagés :

- Un index des noms de personne cités, par ordre alphabétique, et sous la forme suivante :  
*Cochin, Augustin (1823-1872)*
- Un index des lieux cités, sous la forme suivante :  
*Brescia (Italie)*  
*Ligoure (Haute-Vienne)*  
*Paris*
- Un index des titres (ouvrages ou revues) cités, sous la forme suivante :  
*L'Organisation du travail selon la coutume des ateliers et la loi du Décalogue..., Tours, Mame, 1870.*

À ces index, nous avons choisi d'ajouter, en accord avec le chef de projet :

- Un index des événements
- Un index des noms d'organisation
- Un index des termes sociologiques leplaysiens

Tous ces principes de transcription, d'annotation et d'indexation établis par Monsieur Matthieu Brejon de Lavergnée nous ont été très précieux. C'est à eux que nous nous sommes référée pour les adapter ensuite aux possibilités du numérique. Nous y reviendrons donc lorsque nous parlerons de l'acquisition et du traitement des données en indiquant comment transposer en numérique les principes de la correspondance et les choix éditoriaux, selon les outils qui sont à notre disposition.

Maintenant que nous avons vu ce qu'impliquait une édition numérique de correspondance, comment concevoir un site adapté à ces exigences ? Nous tenterons de répondre à cette question dans le prochain chapitre.



# Chapitre 6

## Concevoir un site adapté aux exigences de l'édition

### 6.1 Encoder, oui, mais pourquoi ?

La conception du site n'a pas été la tâche qui nous a le plus occupée durant nos stages, néanmoins, nous nous sommes un peu penchée sur la question.

Pour ce qui est du projet ELICOM, nous n'avons pas été sollicitée pour cette question fort complexe, d'autant que la plateforme envisagée est assez innovante et pleine de défis.

En revanche, pour le projet d'édition numérique de la correspondance de Frédéric Le Play, au CRHXIX, nous avons été amenée à imaginer le futur site : en effet, un encodage répond à des attentes<sup>1</sup>. Il est donc nécessaire de savoir ce que l'on veut encoder, mettre en valeur. En effet, on n'encode pas pour encoder mais dans un but précis. Ce chapitre nous aidera donc à avoir une meilleure vue d'ensemble des attentes du site, puisque nous nous penchons ici principalement sur le projet Le Play.

Tout d'abord, la mise en contexte par le biais des récits utilisateurs ou *users stories* est un atout précieux. Ceci nous permettra d'élaborer au mieux l'architecture du site, sans oublier le rôle du référencement naturel et la question de la licence. Ce sont autant de points qui seront développés dans le cahier des charges si l'on envisage une externalisation du projet de développement.

### 6.2 Les récits 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

---

1. Le principe de l'encodage est davantage expliqué dans la quatrième partie (10.1.1)

de pressentir 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.

En effet, « un récit utilisateur est une phrase simple dans le langage de tous les jours permettant de définir avec suffisamment de précision le contenu d'une fonctionnalité à développer<sup>2</sup> ».

Le récit utilisateur comprend trois éléments principaux :

**En tant que <qui>, je veux <quoi> afin de <pourquoi>.**

Le <qui> est le *persona*, le sujet fictif imaginé. Le <quoi> est le comportement ou fonctionnalité attendue. Le <pourquoi> indique l'intérêt de la fonctionnalité et justifie son développement.

Inexpérimentée que nous étions dans la pratique des récits utilisateurs, nous avons peiné à créer des récits utilisateurs entièrement satisfaisants et dignes d'être placés dans le corps de ce mémoire. Néanmoins, nous les avons mis dans les livrables, sous forme rédigée dans l'ébauche de cahier des charges et sous forme de tableau dans un fichier excel. Malheureusement, nous n'avons pas pu aller au bout de ce travail, par manque de temps. Cela fait partie d'un des points à retravailler par l'un des membres de l'équipe du CRHXIX pour pouvoir mener à terme le projet.

### 6.3 Élaboration de l'architecture du site

Pour concevoir un site adapté aux exigences de l'édition, outre les récits utilisateurs, il s'agit aussi d'imaginer une architecture du site.

Nous avons également fait une ébauche de cette architecture, la suite sera prise par un membre du CRHXIX plus expérimenté dans ce domaine. Pour concevoir un site, il est bon d'imaginer le site le plus simple possible, et le plus facile d'utilisation, avec le moins de clics à faire.

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<sup>3</sup>.

Pour cela, nous avons pris plusieurs sources d'inspirations, dont deux principales que nous avons déjà évoquées plus haut<sup>4</sup> :

- Le site de l'édition numérique de la correspondance de Gustave Flaubert<sup>5</sup>.

---

2. « Récit utilisateur », *Wikipedia*, Page Version ID 6474167, 2020, URL : [https://fr.wikipedia.org/wiki/Recit\\_utilisateur](https://fr.wikipedia.org/wiki/Recit_utilisateur) (visité le 18/06/2020).

3. La réalisation du site sera probablement confiée à un prestataire.

4. Voir la fin du chapitre 4.

5. *Correspondance*, Site web du Centre Flaubert, URL : <https://flaubert.univ-rouen.fr/>

— Celui de la correspondance de Marcel Proust<sup>6</sup>.

Tout d'abord, nous nous sommes inspirés du site d'édition électronique de la correspondance de Gustave Flaubert<sup>7</sup>, réalisé 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), soutenu par Huma-Num, cette « très grande infrastructure visant à faciliter le tournant numérique de la recherche en sciences humaines et sociales<sup>8</sup> » et son Consortium Cahier<sup>9</sup>. Sur la page d'accueil de cette édition se trouve le nombre de lettres consultables avec le récapitulatif des lettres écrites par Flaubert et celles qu'il a reçues<sup>10</sup>. 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 voit 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 notre part, nous pensons retenir la recherche par correspondant, les lieux de rédaction, éventuellement les lieux de conservation.

Quant au deuxième site retenu, c'est une édition numérique de la correspondance de Marcel Proust<sup>11</sup>. Pour chaque lettre au hasard on a la vue numérisée (si elle existe), la transcription, le texte, les notes et les informations<sup>12</sup>.

Nous pouvons donc partir de ces deux exemples pour voir ce qu'il y a à adapter par rapport à nos besoins. Le point commun à ces deux sites et qui est bien l'objectif de notre édition numérique est le vis-à-vis entre le fac-similé et sa transcription.

---

correspondance/edition/ (visité le 17/06/2020).

6. *Le Projet, Corr-Proust*, URL : <http://proust.elan-numerique.fr/presentation/project> (visité le 08/06/2020).

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

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

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

10. Voir Annexe A.4

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

12. Voir annexe A.6

Nous avons donc imaginé le futur site, de façon assez sommaire, avec ses pages et ses rubriques. Le détail est disponible dans les livrables<sup>13</sup>.

Il a donc été envisagé d'avoir une page d'accueil<sup>14</sup> donnant sur sept rubriques :

- **À propos de Le Play** : c'est une page donnant un minimum d'informations pour les personnes voulant se renseigner sommairement sur Le Play, avec une description de sa vie et de son œuvre d'une part, puis une présentation de ses correspondants, et enfin une carte de ses voyages et une visualisation de sa correspondance.
- **Correspondance de Le Play** : cette rubrique présente la correspondance par ordre chronologique (à la fois par période et par année), par correspondant, par lieu de rédaction et par lieu de conservation. En cliquant, on accède à une page présentant la lettre recherchée.
- **Index** : ce terme renferme les six index envisagés<sup>15</sup> : tout d'abord, l'index des personnes ou personnages, contemporains de Le Play, historiques ou de fiction, puis l'index des lieux, celui des ouvrages cités, des organisations, des événements évoqués et enfin, particularité de notre correspondance, un index des termes leplaysiens et sociologiques<sup>16</sup>.
- **Recherche** : il sera possible de faire une recherche avancée, par année, éventuellement par plage temporelle comme cela a été fait pour le site *Corr-Proust*, par expéditeur, destinataire, mot, lieu de conservation et lieu de rédaction.
- **Guides** : on aura accès ici à un glossaire, à un recensement des abréviations, et à un guide utilisateur expliquant comment fonctionne le site.
- **Actualités** : cette page sera à alimenter régulièrement, selon les lettres nouvellement éditées, les modifications, corrections, et ajouts. On pourra profiter de cet onglet pour faire la promotion des événements autour de Frédéric Le Play. Elle aurait l'avantage de permettre un meilleur référencement naturel mais nous reviendrons là-dessus. Il est aussi envisageable de faire une « Foire aux questions » à cet endroit.
- **À propos de l'édition**. Ici, on pourra voir une présentation du projet, de l'équipe, ainsi que celle de la politique éditoriale, point capital pour garantir la scientificité de l'édition. On présentera aussi les partenaires et soutiens qui appuient également la crédibilité du projet.

Enfin, on pourrait envisager d'ajouter un bouton d'« appel à manuscrits » pour ceux qui possèdent des manuscrits auxquels nous n'avons pas eu accès. En effet, certains ma-

13. Voir aussi les annexes, figure A.9

14. Il est bon de se référer au document joint dans les livrables intitulé `architecture_site_LP.JPG`

15. Nous n'avons pas eu le temps de mettre à jour le fichier `architecture_site_LP.JPG`, mais depuis, la réflexion a continué son cours et d'autres index ont été envisagés lors de l'encodage. En effet, si certaines étapes sont à suivre, elles ne sont pas imperméables, et elles s'enrichissent mutuellement. L'encodage a favorisé une réflexion plus profonde sur l'indexation, permettant de compléter ainsi l'architecture du site.

16. Nous y reviendrons lors de la description des choix d'encodage en dernière partie.

nuscrits de Le Play sont en vente sur internet, cela suppose que certains acquéreurs privés doivent en posséder. L'intérêt serait donc, encore une fois, de mettre à la disposition de tous une numérisation et transcription de manuscrits qui dorment dans des lieux oubliés.

On pourrait aussi ajouter un bouton « Nous contacter » permettant de joindre l'équipe du CRHXIX, pour poser des questions ou proposer des corrections, si l'un des utilisateurs remarque des coquilles qui nous auraient échappées.

Ainsi, l'architecture du site est encore en questionnement, mais elle a déjà été envisagée dans sa globalité lors de notre stage.

## 6.4 Le référencement naturel

Une question que nous nous sommes aussi posée est celle du référencement naturel, autrement dit de l'Optimisation pour les moteurs de recherche ou en anglais *Search Engine Optimization* (SEO). Sur ce sujet, les réflexions ont été à peine ébauchées. Néanmoins, c'est une piste de réflexion que nous ouvrons au CRHXIX : comment faire pour optimiser le positionnement d'une page ou de notre site dans la page des résultats de recherche d'un moteur de recherche, sachant qu'il s'agit de se trouver dans la première page des résultats du moteur de recherche ou *Search engine result page* (SERP), puisque la plupart du temps, les utilisateurs ne vont pas au-delà :ils cliquent majoritairement sur les cinq premiers résultats de la SERP. Il faut donc savoir se distinguer parmi les milliers de sites qui existent.

Le SEO est une chose sur laquelle il faut veiller chaque année car les critères varient. Lors donc de la création du site, il faut miser sur une structure du site avec un plan logique et précis<sup>17</sup>. L'adresse du site ou URL (*Uniform Resource Locator*) doit être court et précis, avec éventuellement des mots-clés. La navigation du site doit être intuitive, et le design doit être *mobile friendly*, autrement dit, il doit être visible sur les appareils à petit écran. La vitesse du site doit être optimale, au risque de perdre des utilisateurs pressés. Les titres, les en-têtes et la balise *meta description* doivent être bien pensés ; chaque page doit être composée d'un contenu unique dans son intitulé, avec des mots-clés relatifs à son contenu. Il faut également optimiser les images dans leur description. Les images trop grandes risquent de réduire la vitesse de téléchargement du site, ce sera donc un défi à relever pour notre site. En effet, on remarque sur certains sites d'édition numérique de correspondance, les images sont parfois un peu longues à charger. Un *design* attrayant est un plus et dans notre cas, cela nous permettra peut-être d'avoir un public plus large que les simples chercheurs qui savent pourquoi ils ont fait cette requête et resteront sur le site même s'il n'est pas attractif, intéressés qu'ils sont par son contenu intellectuel. Enfin, la

---

17. Voir *SEO et Webdesign : 9 quick wins à intégrer*, Le Marketing aux petits oignons, URL : <https://www.emarketing-aux-petits-oignons.com/seo-et-webdesign> (visité le 18/09/2020).

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 référencée<sup>18</sup>. » Il faudra donc prendre en compte tous ces principes pour garantir le bon classement de notre site sur le Web et gagner l'une des premières places sur la SERP afin d'avoir une audience conséquente.

## 6.5 La licence ou le cadre juridique

Un publication va de pair avec le choix d'une licence. Il est nécessaire d'associer une licence à nos jeux de données. La licence encadre et sécurise les réutilisations des données permises par le titulaire des droits d'auteur<sup>19</sup>. Leur choix est vaste, comme le souligne la description ci-dessous :

FIGURE 6.1 – Les licences



Creative commons

Partage, création, adaptation autorisés en fonction du type de licence choisie

Les licences de référence pour les administrations depuis décret de la Loi « pour une République numérique » dite Loi Lemaire



Open Database Licence

Partage, création, adaptation autorisés  
Si mention paternité, partage à l'identique



Licence Ouverte Etalab

Partage, création, adaptation autorisés  
Si mention paternité

Une particularité de la loi française dont les contours ne sont pas définis « positivement »



Domaine public

ELICOM part sur une licence CC BY-NC-ND 3.0 FR : c'est une licence Creative commons, qui autorise à partager, copier, distribuer et communiquer le matériel par tous moyens et sous tous formats. L'offrant ne peut retirer les autorisations concédées par la licence tant que les termes de cette licence sont appliqués<sup>20</sup>. Quatre conditions entrent en ligne de compte : on doit mentionner d'une part d'où viennent les informations, de plus, on n'est pas autorisé à faire un usage commercial de l'œuvre ni de modifications, ni

18. *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).

19. Gauthier Poupeau, *Open Data, Big data, Data Mining*, Module data, M2 TNAH ENC, Cours 2, 21 octobre 2019. La figure 6.1 est tirée de ses slides.

20. *Creative commons*, CC BY-NC-ND 3.0 FR, URL : <https://creativecommons.org/licenses/by-nc-nd/3.0/fr/> (visité le 18/09/2020).

de restrictions complémentaires. L'utilisateur n'est pas dans l'obligation de « respecter la licence pour les éléments ou matériel appartenant au domaine public ou dans le cas où l'utilisation [...] est couverte par une exception »<sup>21</sup>.

## 6.6 Le cahier des charges

Finalement, la clef pour concevoir un site adapté aux exigences de l'édition se trouve dans le cahier des charges. Celui-ci permet d'exprimer les attentes de l'édition et garantit une mise en pratique planifiée et réaliste.

Pour ce qui est des projets numériques, on constate que les deux tiers rencontrent de gros problèmes et un tiers se solde par un échec. Cette proportion est stable en France et dans les autres pays du monde, et aussi bien dans le public que le privé<sup>22</sup>.

Une des raisons se trouve dans le flou qui règne autour des projets, l'irréalisme, les imprécisions, l'inadéquation entre les objectifs et les moyens. Or, le cahier des charges est ainsi nommé car, à sa lecture, on peut estimer les charges. Il est un pont entre celui qui a le besoin et celui qui le résout.

Il présente à la fois le contexte du projet, ses objectifs, son périmètre, les besoins en terme de fonctionnalités, les ressources disponibles et les contraintes pour la réalisation du projet, le budget et les délais qui permettent au prestataire d'évaluer la durée de travail et de s'organiser<sup>23</sup>.

Le Labex OBVIL avait déjà établi un cahier des charges du projet il y a deux ans. Celui-ci nous a bien aidée pour définir les priorités et choisir les correspondances à traiter, de la plus facile à la plus difficile.

Pour le CRHXIX, le cahier des charges reste encore à faire. Nous voulions y résumer l'ensemble de nos attentes pour le futur site : comment voulons-nous que le site soit configuré ? Quelles fonctionnalités désirons-nous ? Nous y avons travaillé mais le résultat est encore entre un rapport et une ébauche de cahier des charges<sup>24</sup>. En effet, pour établir un cahier des charges, il faut compter un certain nombre de jours de travail, et nous n'avons travaillé qu'une petite trentaine de jours au CRHXIX. Nous avons donc dû passer assez rapidement à la suite, c'est-à-dire à la partie d'acquisition et de traitement des données. Néanmoins, les réflexions que nous avons été amenée à faire sur le cahier des charges nous ont bien éclairée pour la suite, en témoigne ce présent chapitre sur la conception d'un site adapté à l'édition numérique de correspondance.

---

21. *Idem*.

22. Voir Jean-Louis Foucard, Module de formation « Manager un projet Numérique » Master « Archives – Technologies numériques appliquées à l'histoire », École Nationale des Chartes, mars 2020.

23. *Exemple cahier des charges*, CahiersDesCharges.com, URL : <https://cahiersdescharges.com/exemple-cahier-des-charges-pdf/> (visité le 12/06/2020).

24. Voir dans les livrables.

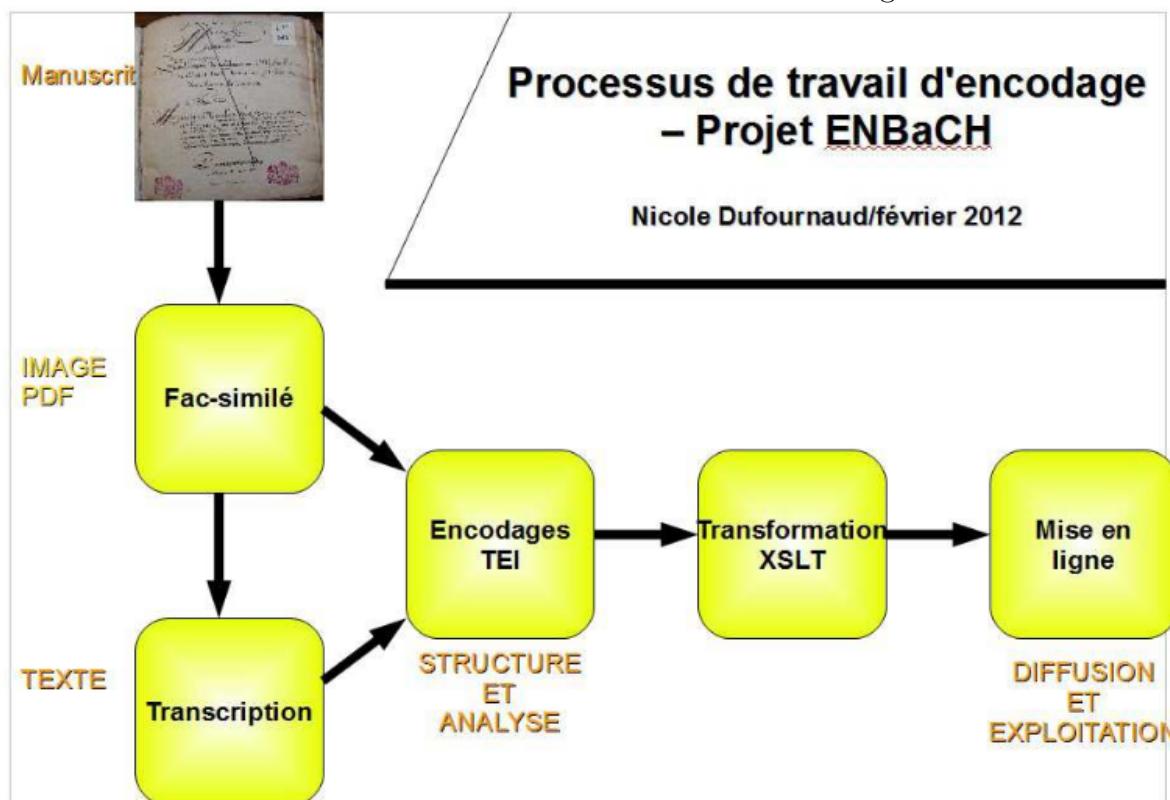
Une fois l'édition pensée, il s'agit de mettre en pratique. Ceci se fait en deux temps. Il s'agit tout d'abord d'acquérir les données (partie III) puis de les traiter (partie IV). Or, l'apprentissage machine joue un grand rôle aujourd'hui dans l'acquisition des données. C'est ce point que nous aborderons dans la partie suivante.

## Troisième partie

L'apprentissage machine au cœur de  
l'acquisition des données



FIGURE 6.2 – Processus de travail d'encodage



La suite de ce mémoire va être consacrée à l'acquisition et au traitement des données. Nous avons trouvé ici un schéma qui éclaire en partie le processus qui amène à la mise en ligne des données. Même si celui-ci diffère selon les projets, on observe de grandes constantes.

Tout d'abord, il y a la donnée brute : le manuscrit. C'est le cas pour le projet Le Play. Pour ELICOM, nous avons à faire à une édition papier imprimée. Après le manuscrit suit l'image numérisée, dont on se sert pour faire les transcriptions - c'est la partie acquisition des données - et à partir de là, l'encodage en TEI. On peut avoir recours à une transformation XSLT (*Extensible Stylesheet Language Transformations*), on parlera rapidement de cette éventualité. Et enfin, la mise en ligne, objectif de nos projets pour leur diffusion et exploitation. Puisse ce schéma éclairer les propos qui suivront.



# Chapitre 7

## L'apprentissage machine

### 7.1 Petite histoire de l'apprentissage machine

Pour mieux comprendre ce qui va suivre, il importe de revenir sur ce qu'est l'apprentissage machine afin de mieux saisir en quoi il intéresse nos projets.

Le terme *Machine Learning* (ML), traduit en français par « apprentissage machine » ou encore « apprentissage automatique », est apparu pour la première fois dans la bouche d'Arthur Samuel, en 1959, un pionnier américain dans le domaine des jeux vidéo et de l'intelligence artificielle. Il évoque la possibilité qu'a la machine d'apprendre sans être vraiment programmée<sup>1</sup>.

En effet, la machine peut être programmée pour apprendre de son expérience. On verra très nettement par la suite qu'elle fait, pour ainsi dire, des progrès. Elle apprend une écriture, à travers des « *features*<sup>2</sup> », et plus on lui donne de lettres à apprendre, plus elle progresse, jusqu'à pouvoir faire la transcription elle-même en prédisant le caractère à reconnaître par rapport à ce qu'elle a appris, avec un taux de réussite plus ou moins important<sup>3</sup>.

### 7.2 Apprentissage machine et intelligence artificielle

Comme nous l'avons rapidement souligné, l'apprentissage machine est une composante de l'intelligence artificielle (IA). Celle-ci est la « recherche de moyens susceptibles de doter les systèmes informatiques de capacités intellectuelles comparables à celles des êtres

---

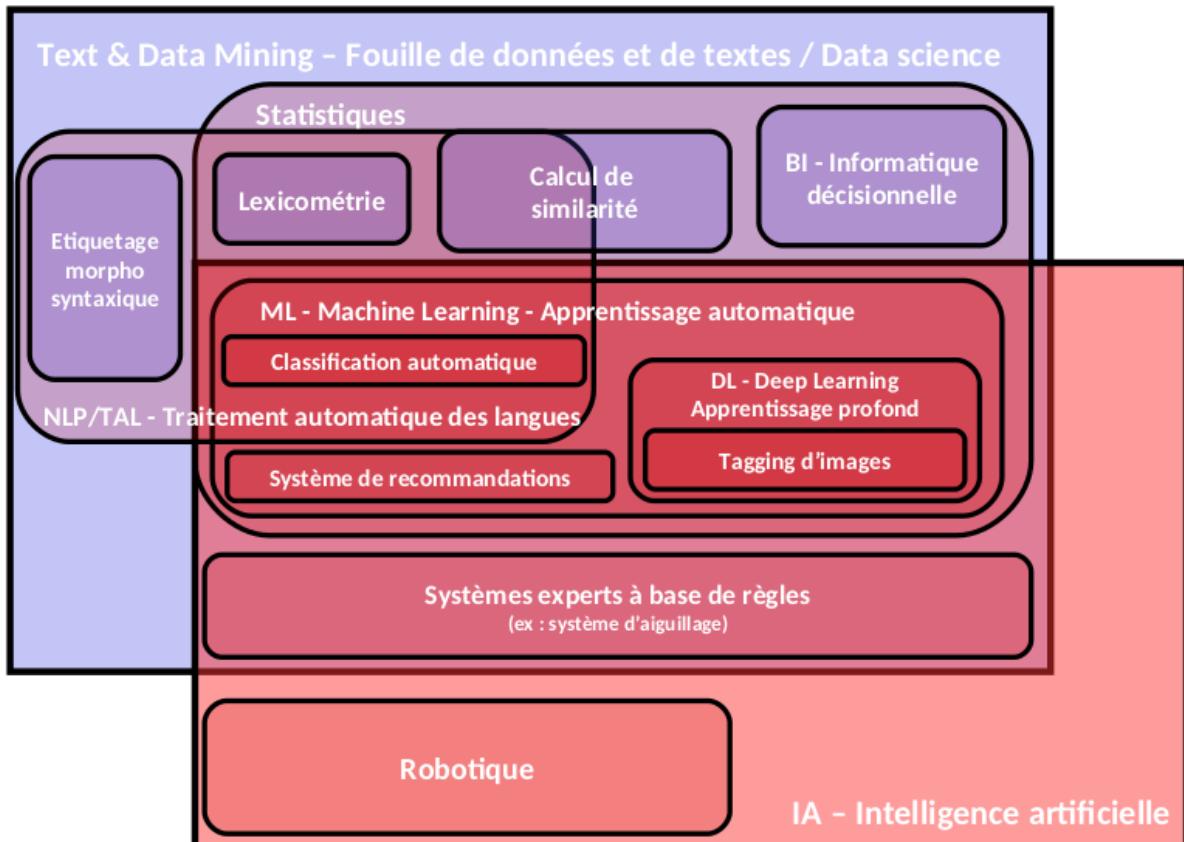
1. *An introduction to Machine Learning*, Site web GeeksforGeeks, URL : <https://www.geeksforgeeks.org/introduction-machine-learning/> (visité le 20/09/2020).

2. Une des techniques de reconnaissance est la « classification par caractéristiques (*features*) : une forme à reconnaître est représentée par un vecteur de valeurs numériques - appelées *features* en anglais - calculées à partir de cette forme ». Voir *Reconnaissance optique de caractères*, Wikipedia, URL : [https://fr.wikipedia.org/wiki/Reconnaissance\\_optique\\_de\\_caracteres](https://fr.wikipedia.org/wiki/Reconnaissance_optique_de_caracteres) (visité le 30/09/2020).

3. L'apprentissage machine a plusieurs divisions comme l'apprentissage supervisé ou non supervisé, ou encore l'apprentissage par renforcement et l'apprentissage semi-supervisé. Mais nous ne rentrons pas ici dans les détails de l'algorithme.

humains »<sup>4</sup>. Elle se traduit par « l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence »<sup>5</sup>. C'est une capacité d'un programme informatique à fonctionner comme un cerveau humain. Elle utilise, avec les réseaux de neurones la façon de fonctionner de l'intelligence humaine. Ainsi, on voit que le fait qu'une machine puisse apprendre et faire en quelque sorte des progrès fait partie de l'IA. Le schéma ci-dessous<sup>6</sup> permet de comprendre davantage la place du *Machine Learning* dans l'IA. On voit que l'apprentissage machine et l'apprentissage profond ou *deep learning* sont des sous-ensemble de l'IA. Cependant, ils ne sont pas l'IA mais un moyen de parvenir à l'IA un jour.

FIGURE 7.1 – *Machine learning* et IA  
Quelques exemples dans les différents domaines



Ainsi, pour distinguer *machine learning* et *deep learning*, on peut dire que le *machine learning* est un « système visant à accomplir des tâches à partir de caractéristiques et attributs communs (patterns) “appris” dans un ensemble de données d'exemples »<sup>7</sup> : la

4. *La Recherche*, janv. 1979, n.96, vol. 10, p. 61

5. *Intelligence artificielle*, Wikipédia, URL : [https://fr.wikipedia.org/wiki/Intelligence\\_artificielle](https://fr.wikipedia.org/wiki/Intelligence_artificielle) (visité le 20/09/2020).

6. Gauthier Poupeau, *Open Data, Big data, Data Mining*, Module data, M2 TNAH ENC, Cours 2, 21 octobre 2019

7. *Idem*, p. 27

machine apprend les données et les applique de la bonne façon ; tandis que le *deep learning* est un sous-ensemble du *machine learning*, une « technique d'apprentissage cherchant à reproduire le mécanisme des réseaux de neurones du cerveau humain »<sup>8</sup>. Il est plus poussé, plus abouti et plus autonome<sup>9</sup>.

Ainsi, nous verrons que les technologies que nous utiliserons de *machine learning* ne sont pas encore abouties et nécessiteront un traitement conséquent.

## 7.3 L'apprentissage machine dans nos deux projets

Lorsqu'on a une quantité de lettres à éditer de façon numérique, comment utiliser au mieux les capacités de la machine pour qu'elle nous aide dans l'acquisition des données ? C'est tout le rôle de l'apprentissage machine qui permet une reconnaissance automatique des caractères et permet la transcription des textes. Or, nous sommes ici face à deux types de textes : pour ELICOM, nous avons des éditions papier numérisées et disponibles sur Gallica, qui ont été imprimées. Ce sont des caractères d'imprimerie. Pour l'édition numérique de la correspondance de Frédéric Le Play, ce sont des manuscrits écrits au moins de moitié si ce n'est majoritairement de la main de Le Play. Ce sont donc des caractères d'écriture manuscrite. À chaque projet correspond donc une technologie différente. Pour les caractères d'imprimerie, on parlera d'*Optical Character Recognition* ou Reconnaissance optique de caractères (OCR), pour la reconnaissance de l'écriture manuscrite, on parlera d'*Handwritten Text Recognition* (HTR).

Ainsi, l'OCR

« désigne le processus informatique qui vise à transposer des éléments textuels présents sur une image numérique vers un fichier de texte de manière automatique. Il s'agit, en d'autres termes, de faire réaliser par l'ordinateur la tâche de copie du texte<sup>10</sup>. »

L'OCR de Gallica est un service qui permet la reconnaissance de texte dans une image<sup>11</sup>. Ainsi, le contenu des documents numérisés est extrait et transformé en fichier texte. Gallica propose le téléchargement au format .txt du résultat de l'OCR. Néanmoins,

---

8. *Ibidem*. p. 28

9. Voir *Artificial intelligence vs Machine Learning vs Deep Learning*, Site web Geeksforgeeks, URL : <https://www.geeksforgeeks.org/artificial-intelligence-vs-machine-learning-vs-deep-learning/?ref=rp> (visité le 20/09/2020).

10. Alix Chagué, *Constituer un corpus pour la fouille de texte - de la transcription des documents d'archives à l'annotation : exploration d'une méthodologie par l'ANR Time Us*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Vincent Jolivet et Éric de la Clergerie, École nationale des chartes, 2018, p.41.

11. *Mode texte et OCR*, Site web BNF.Gallica, URL : <https://gallica.bnf.fr/edit/und/consulter-les-documents> (visité le 21/09/2020).

il ne s'agit pas de texte brut mais d'un texte semi-structuré via des balises *Hypertext Markup Language* (HTML), langage de balisage conçu pour représenter les pages web. Le texte étant ainsi structuré par les balises, on peut ensuite le transformer plus facilement en vue de notre édition. Nous y reviendrons. Le service de Gallica a donc été très précieux pour notre projet d'ELICOM.

Quant à l'édition numérique de la correspondance de Frédéric Le Play, nous nous sommes servis de l'HTR via Transkribus, qui permet de transcrire cette fois l'écriture manuscrite, grâce à l'entraînement d'un modèle par main.

Ainsi, l'apprentissage machine est au cœur de l'acquisition des données pour nos deux projets. Il est temps de voir cela plus en détail, avec l'OCR et l'HTR.

# Chapitre 8

## L'OCR de Gallica

### 8.1 Un service à ne pas négliger

#### 8.1.1 Un OCR plutôt fiable

Pour le projet ELICOM du Labex OBVIL, nous avons beaucoup exploité l'OCR de Gallica. En effet, les textes que nous avons extraits sont issus d'un traitement OCR brut, sans relecture, que nous relirons sommairement par la suite. La qualité (taux OCR) dépend de l'état de la source, de la langue, mais aussi de la campagne de numérisation<sup>1</sup>. Certains livres numérisés avaient par exemple des marques de crayon de papier, ce qui est parfois mal interprété par l'OCR. En effet,

« Même si les techniques d'OCR sont en progrès constant, la qualité de reconnaissance dépend malgré tout d'un grand nombre de facteurs liés tant au document original qu'à la numérisation elle-même. Ainsi les documents patrimoniaux de Gallica présentent un certain nombre de défis pour l'OCR : dégradation du papier ou de l'encre, polices de caractères ou orthographes anciennes, etc. De plus, les anciens modes de numérisation (en noir et blanc, d'après microfilm) ont un impact négatif sur les performances<sup>2</sup>. »

Certains textes ont cependant un taux avancé d'OCR de 100%. Cependant, un taux de 100% ne signifie pas un sans fautes. De toutes façons, une relecture précise et attentive sera toujours nécessaire même sur un texte dont le taux est déclaré de 100%. En effet,

« Ces estimations donnent généralement un bon aperçu de la qualité globale d'un document, mais elles ne doivent pas être confondues avec le taux qualité réel, qui ne peut être connu (sauf à corriger le texte d'un document et comparer

---

1. C'est la même chose que pour le projet de la Très grande bibliothèque (TGB) mené aussi par OBVIL, voir : *TGB (BnF – OBVIL)*, Site web provisoire TGB, OBVIL et BNF, URL : <http://obvil.lip6.fr/tgb/> (visité le 08/09/2020).

2. *Mode texte et OCR*, Site web BNF.Gallica, URL : <https://gallica.bnf.fr/edit/und/consulter-les-documents> (visité le 21/09/2020).

cette référence avec le texte OCR, ce qui est impossible dans un contexte de numérisation de masse). De plus, ces indicateurs ne sont pas toujours calculés à partir de la totalité du document ; il se peut par exemple que des zones illisibles ou trop complexes soient exclues du calcul et que la qualité perçue par le lecteur soit ainsi très nettement inférieure à la qualité annoncée<sup>3</sup> »

L'avantage que représentent les imprimés du XIX<sup>e</sup> siècle est leur relative modernité : ils sont à la fois suffisamment anciens pour être libres de droits, contrairement aux imprimés des XX<sup>e</sup> et XXI<sup>e</sup> siècle, et ils ont aussi l'avantage d'être relativement modernes dans leur typographie, donc mieux reconnus automatiquement que les imprimés plus anciens qui comportent souvent des « s » longs ou autres caractéristiques qui sont autant d'obstacles pour l'OCR et nuisent à sa qualité.

Comme nous l'avions souligné lors de la présentation des sources<sup>4</sup>, nous avons choisi les correspondances à traiter en priorité d'une part en fonction des marqueurs du texte qui garantissent une extraction plus aisée, d'autre part en fonction de la qualité de l'océrisation. Or, la correspondance d'Alphonse de Lamartine semblait réunir ces deux avantages : pour ce qui est de l'extraction, elle était classée « relativement facile grâce aux chiffres romains qui délimitent les lettres » dans le cahier des charges, et l'OCR de Gallica indiquait un taux de réussite estimé à 100%. Nous avons donc choisi de commencer par là notre travail d'extraction et d'acquisition des données. Puis nous nous sommes concentrée sur la correspondance de Félicité de Lamennais pour finir ensuite sur celle de Pierre-Joseph Proudhon.

Une question s'est posée assez rapidement : devons-nous systématiquement utiliser le service OCR de Gallica ou est-il opportun de passer par un autre moyen pour extraire le texte, comme par exemple Transkribus ou un équivalent. Nous avons donc demandé conseil à des personnes expertes en la matière. À cela il nous a été répondu que l'OCR de Gallica est un service à ne pas négliger, et comme vu plus haut, les taux de réussite étant relativement bons, il nous a paru plus intéressant de l'exploiter au maximum, d'autant que nous ne sommes pas intéressés par une correspondance entre l'image et le texte, étant donné que nous ne mettrons en ligne que le texte extrait et non l'image de la première édition papier.

Revenons donc sur cette phase d'acquisition et de pré-traitement des données. Celle-ci s'est faite en plusieurs étapes.

### 8.1.2 Extraction de l'OCR en HTML

Pour acquérir les données, il s'agit de les extraire. Cela a occupé la première phase de notre travail. Pour chacun des auteurs sélectionnés, nous sommes allés sur le catalogue

---

3. *Idem..*

4. Voir 3.2

général de la BNF qui nous a redirigé vers la page de Gallica permettant de télécharger la correspondance sous trois formats : soit PDF (*Portable Document Format*), soit JPEG (*Joint Photographic Experts Group*), ou encore TXT (fichier texte). Nous avons choisi cette dernière option. En téléchargeant l'OCR en fichier texte brut. En effet, cela permet d'une part de s'assurer de la qualité de l'océrisation : au début de chaque fichier TXT extrait se trouve un récapitulatif sur le document et ses métadonnées essentielles, et une phrase générée automatiquement elle aussi donnant des précisions sur l'OCR. Ainsi, pour la correspondance de George Sand, on peut lire que « Le texte affiché peut comporter un certain nombre d'erreurs. En effet, le mode texte de ce document a été généré de façon automatique par un programme de reconnaissance optique de caractères (OCR). Le taux de reconnaissance estimé pour ce document est de 96 %. » L'OCR est ici brute : elle n'a pu être relue. Ces erreurs sont repérées par des caractères gris pâles au lieu des caractères noirs.

FIGURE 8.1 – L'OCR de George Sand sous format TXT

1  
A MADAME MAURI CE DUPIN' QUI ALLAIT QUtTTEtt NOHANT'  
)8i2.  
Que j'ai de regret de ne pouvoir te dire adieu T~ vois combien j'ai de chagrin de te quitter. Adieu pense à moi. et sois sûre que je ne t'oublierai point. Ta fille.  
Tu mettras la réponse derrière le portrait du vieux Dupin  
1. Mademoiselle Aurore Dupin avait alors huit ans. 2. Propriété de madame Dupin de Francueil, puis de George Sand, près la Châtre {[ndre].  
3. Portrait au pastel de M Dupin de Francueil, qui se trcuvOdans le salon de Nohant.  
J'ai reçu votre envoi, mon petit Caron, et je vous remercie de votre extrême obligeance. Toutes mes commissions sont faites le mieux du monde, et vous êtes gentil comme le père Latreille.  
II  
A LA MÊME, A PARIS  
Nohant.24Mv)'tefMt5  
Oh! oui, chère maman, je t'embrasse; je t'attends, je te désire et je meurs d'impatience de te voirie! Mon Dieu, comme tu es

Ainsi, pour le premier volume de la correspondance de George Sand<sup>5</sup>, on constate ici des mots grisés : ce sont toutes les mots soupçonnés d'avoir été mal lu par la reconnaissance automatique. La machine indique donc qu'elle estime s'être trompée sur ces parties. Et effectivement, on peut lire qu'il est écrit « )8i2. » au lieu de « 1812 », ou encore « T » au lieu de « Tu ». Cependant, on voit aussi que certaines parties en noir, et donc présumées justes, sont en réalité fausses : ainsi il est écrit « {[ndre) » au lieu de « (Indre) » : la parenthèse ouvrante a été remplacée par une accolade ouvrante, et le « I » majuscule a été remplacé par un crochet ouvrant. On a donc ici une illustration des limites de l'OCR.

Mais que faire avec ce texte brut ? Il ne nous intéresse pas particulièrement. Il s'agit

5. Correspondance : 1812-1876. George Sand, BNF.Gallica, URL : <https://gallica.bnf.fr/ark:/12148/bpt6k2065433/f1n385.texteBrut> (visité le 19/05/2020).

donc d'extraire le code source sous le format HTML.

HTML est le langage informatique de base d'Internet, utilisé pour la mise en forme des pages Web<sup>6</sup>. Il repose sur un système de balises, d'où son nom qui signifie « langage de balisage d'hypertexte ». Comme pour XML, les balises vont toujours par deux, une ouvrante (pour un paragraphe, ce sera <p>), une fermante (</p>). La balise ouvrante peut avoir un attribut pour qualifier l'élément.

En extrayant l'OCR sous le format HTML, l'avantage est d'avoir un texte déjà structuré. On peut le constater ci-dessous avec l'exemple de l'OCR de Lamartine au format HTML, avec l'éditeur XML Oxygen XML Editor. Chaque partie de texte signifiant - comme la ligne de la date, le salut, le paragraphe - est encadré par une balise <p> indiquant que c'est un paragraphe ou du moins une division particulière. Par ailleurs, les métadonnées sont comprises dans une balise head, elle-même embrassant des balises <meta> contenant des attributs comme @name ou @content. Le corps du texte quant à lui est contenu dans une balise <body>. Les balises <span> indiquent les erreurs de l'OCR.

FIGURE 8.2 – L'OCR de Lamartine en HTML sous Oxygen XML Editor

```

1  <!DOCTYPE html>
2  <html xmlns="http://www.w3.org/1999/xhtml">
3  <head>
4      <title>Correspondance de Lamartine. I. 1807-1812 / publiée par Mme Valentine de Lamartine...
5          | Gallica</title>
6      <meta name="title"
7          content="Correspondance de Lamartine. I. 1807-1812 / publiée par Mme Valentine de Lamartine... | Gallica" />
8      <meta name="description"
9          content="Correspondance de Lamartine. I. 1807-1812 / publiée par Mme Valentine de Lamartine... -- 1873-1875 -- livre" />
10     <meta name="DC.type" content="book" />
11     <meta name="DC.description" content="Contient une table des matières" />
12     <meta name="DC.description" content="Avec mode texte" />
13     <meta name="DC.creator" content="Lamartine, Alphonse de (1790-1869). Auteur du texte" />
14     <meta name="DC.title"
15         content="Correspondance de Lamartine. I. 1807-1812 / publiée par Mme Valentine de Lamartine..." />
16     <meta name="DC.date" content="1873-1875" />
17     <meta name="DC.rights" content="domaine public" />
18     <meta name="DC.rights" content="public domain" />
19     <meta name="DC.identifier" content="oai:bnf.fr:gallica/ark:/12148/bpt6k5805303r" />
20     <meta name="p:domain_verify" content="12c0fefa160572d58c3754d6158b2d99" />
21     <script type="text/javascript" src="/ruxitagentjs_ICA2SVfjqrU_10191200518082328.js" data-dtconfig="app=3c476fd10179998|cuc=7c
22     </head>
23 <body>
24     <p>I</p>
25     <p>>A monsieur Prosper Guichard de Bienassis </p>
26     <p>>A Bienassis, par Crémieu (Isère).</p>
27     <p>Milly, 24 septembre 1807. </p>
28     <p>>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 et j'y réponds ce
29

```

Le texte est donc bel et bien structuré et cela nous permettra par la suite de le transformer plus facilement en XML. En attendant, l'OCR extrait en HTML est loin d'être parfait. L'illustration présente (figure 8.2) montre l'HTML tel qu'il est après correction. En effet, toute une phase de pré-traitement s'est avérée être nécessaire avant de passer à l'étape de transformation en XML que nous développerons dans la quatrième partie de

6. Depuis 2014, on en est à la version HTML5. Voir : *HTML (HyperText Markup Language)* : définition, traduction, Site web JDN, URL : <https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203255-html-hypertext-markup-langage-definition-traduction/> (visité le 21/09/2020).

notre mémoire.

## 8.2 Un pré-traitement qui suscite des questionnements

### 8.2.1 Rendre l'HTML valide et bien indenté

Notre premier souci lors du téléchargement de l'OCR en HTML a été de vérifier la validité et la bonne indentation de notre fichier HTML. Or, les différents fichiers extraits de Gallica présentaient tous des erreurs dans les balises : les balises `<meta>` n'étaient pas fermées et il était donc impossible d'indenter le texte correctement. De même pour les balises `<br>` (élément de saut de ligne) et `<hr>` (pour *horizontal rule*, règle horizontale servant de séparation, elles indiquent ici les changements de page, on en trouve donc 391 occurrences pour Lamartine). Nous avons donc commencé par fermer toutes ces balises<sup>7</sup>. Nous avons également choisi de supprimer les ensembles de balises `</p><hr/><p>` (40 occurrences dans Lamartine) et `</p><p>` (58 occurrences dans Lamartine) quand elles divisaient des phrases, étant donc inopportunnes et rompant par là l'unité. De même, les titres courants, figurant en en-tête sont des informations superflues. Fort heureusement, elles ont été supprimées automatiquement par l'OCR pour la correspondance de Lamennais (en effet, on lisait « CORRESPONDANCE » à la page de gauche, « DE LAMENNAIS » à la page de droite), ne polluant donc pas l'HTML, contrairement à Lamartine où l'OCR de Gallica ne les avait pas supprimés.

Ensuite, même si cela n'était pas absolument nécessaire, nous avons supprimé tout ce qui concernait les rappels de la demande, les tables des matières et préfaces. Nous aurions aussi pu le faire en amont, lors du téléchargement. Ainsi, l'HTML de Lamennais comporte 12 204 lignes au départ, 8680 après nettoyage (soit le double de l'HTML de Lamartine après nettoyage).

Nous avons donc désormais un HTML valide et bien indenté. Cependant, des fautes subsistent.

### 8.2.2 Quelques fautes de l'OCR, visibles dans l'HTML

Nous avons vu plus haut que certaines fautes étaient grisées dans l'OCR en format TXT. Ceci se retrouve donc logiquement dans l'OCR en format HTML, avec les balises `<span>`. Or, on remarque que celles-ci sont parfois utilisées à bon escient, tandis que d'autres fois, elles sont absentes ou superflues.

Les balises `<span>` indiquant les doutes sur l'OCR sont absentes dans la correspondance de Lamartine. En revanche, on en trouve 510 ouvrantes et fermantes dans le premier volume de correspondance de Lamennais. C'est pour les erreurs quasiment avérées, mais

---

7. Avec la nouvelle version d'Oxygen XML Editor, cela peut se faire de façon automatique.

beaucoup d'autres erreurs se sont glissées dans le texte, sans être signalées par une balise. Une relecture sera donc nécessaire. Ainsi, à la ligne 7805 du fichier HTML de Lamennais, on voit une faute mentionnée dans une balise `<span>`, avec un attribut `@style` pour qu'il apparaisse en gris, comme on peut le constater dans la figure ci-dessous.

FIGURE 8.3 – Exemple d'une balise `<span>` dans l'HTML du premier volume de Lamennais, l. 7805

7803  
7804  
7805  
7806  
7807

et vous aviez le projet, si ce retard continuait, de vous renarer à Arona. Ainsi nous allons encore nous éloigner davantage ; car, de mon côté, je repars, le 25, pour la Bretagne. La vié</span> de ce pays-ci me fatigue extrêmement, et d'ailleurs il est nécessaire de parler dans les circonstances présentes; or à Paris cela me serait impossible. Il y a un commencement de résistance dans...

Cependant, certaines fautes ne sont pas signalées par l'OCR, comme on peut le constater vingt lignes plus loin dans le même fichier.

FIGURE 8.4 – Exemple d'une balise `<span>` manquante dans l'HTML du premier volume de Lamennais, l. 7825

7824  
7825

qu'ils devaient à Dieu. Leur conscience leur a répondu qu'il valait mieux obéir à Dieu qu'ux hommes... Ils ne résistent point; ils ne profitent pas tumultueusement des...

Ici, le « a » est remplacé par une parenthèse ouvrante et un point.

Enfin, parfois, les balises `<span>` sont présentes alors qu'aucune faute n'est à signaler, comme par exemple à la ligne 1264 où le guillemet est juste alors que signalé comme douteux, comme on peut le constater ci-dessous.

FIGURE 8.5 – Exemple d'une balise `<span>` superflue dans l'HTML du premier volume de Lamennais, l. 1264

1263  
1264  
1265

là : « Regardez bien, vous autres, afin que je ne perde pas mon temps, et que vous puissiez légalement déposer du fait! <span style="color:Slategrey;">>></span></p><p>Si, après cela, il avait le malheur d'être privé de la potence, il n'y aurait pas au...

### 8.2.3 L'apport des expressions régulières dans le nettoyage de l'HTML

L'idéal aurait été de toucher le moins possible à l'HTML et de ne faire que des traitements applicables à tous les autres volumes d'un même correspondant.

Cependant, nous avons procédé tout de même à un nettoyage sommaire de l'HTML. Pour cela, les expressions régulières nous ont été un outil précieux. On entend par « expressions régulières » ou « regex », de l'anglais *regular expression*, une chaîne de caractères, qui décrit, selon une syntaxe précise, un ensemble de chaînes de caractères possibles. En l'utilisant dans un éditeur de texte, ou sinon dans Python, on peut ainsi « matcher »,

c'est-à-dire sélectionner certaines suites de chaînes de caractères récurrentes et les sélectionner toutes ensemble, puis soit les remplacer par une autre chaîne de caractère, soit les supprimer définitivement. C'est donc très précieux pour faire des modifications dans les textes, et ici pour pré-traiter notre HTML.

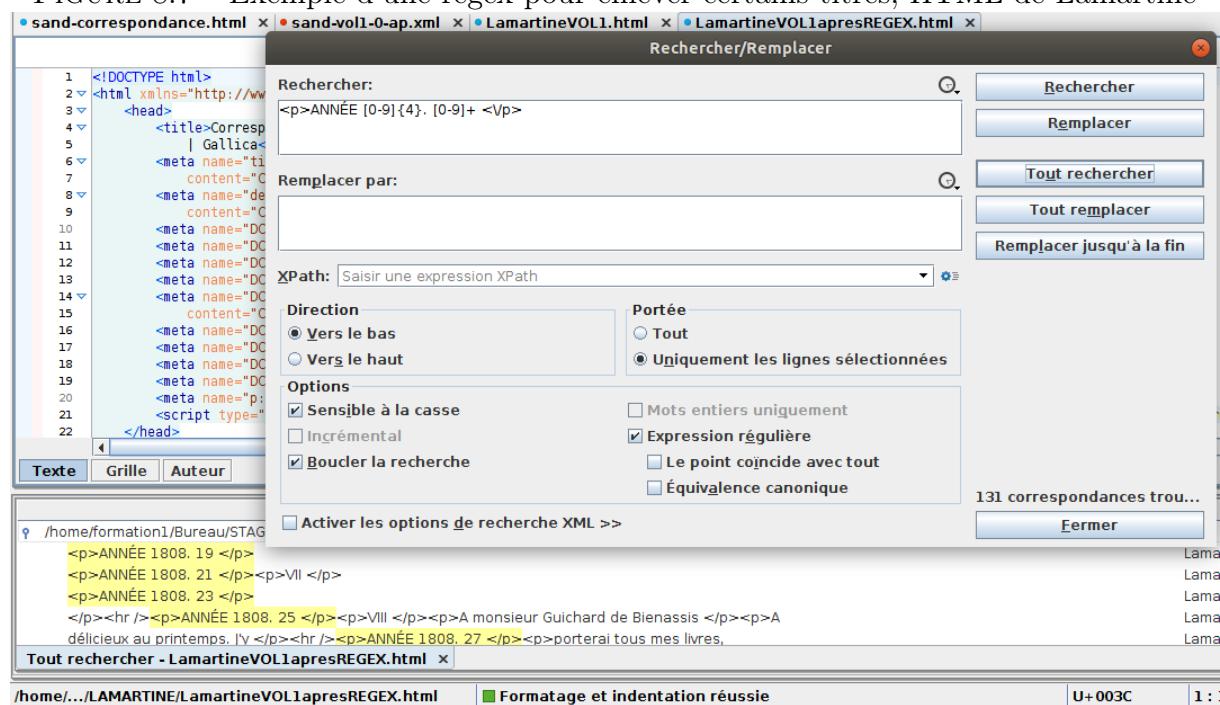
Comme nous l'avons déjà évoqué plus haut, certains titres dans Lamartine polluaient le texte, comme on peut le constater dans la figure suivante : tous les éléments inutiles y sont surlignés en jaune.

FIGURE 8.6 – Exemple des titres polluant le texte, HTML de Lamartine

```
</p><p>regarder comme le plus tendre et le plus fidèle de </p><p>tes amis.
</p><p>ALPH. DE LAMARTINE. </p><h1 /><p>30 CORRESPONDANCE DE
LAMARTINE </p><p>X
</p><p>A monsieur Guichard de Bienassis </p><p>A Belley. , </p><p>Mâcon, 8 juillet
1808.
</p><p>Rassure toi, je ne suis ni tué, ni noyé, ni surtout dégoûté d'une correspondance
qui fait tout mon bonheur, et je me dépêche, comme lu le vois, [...]
répondrai </p><hr /><p>ANNÉE 1808. 31 </p><p>un de ces jours aux deux autres lettres,
aujourd'hui c'est ton tour... Ne me critique pas tant, mon cher ami, sur ma versatilité,
sur l'inconstance de mes goûts, sur mon
```

Avec l'éditeur de texte Oxygen XML Editor, nous pouvons user des regex. C'est donc par ce biais que nous avons supprimé ces titres.

FIGURE 8.7 – Exemple d'une regex pour enlever certains titres, HTML de Lamartine



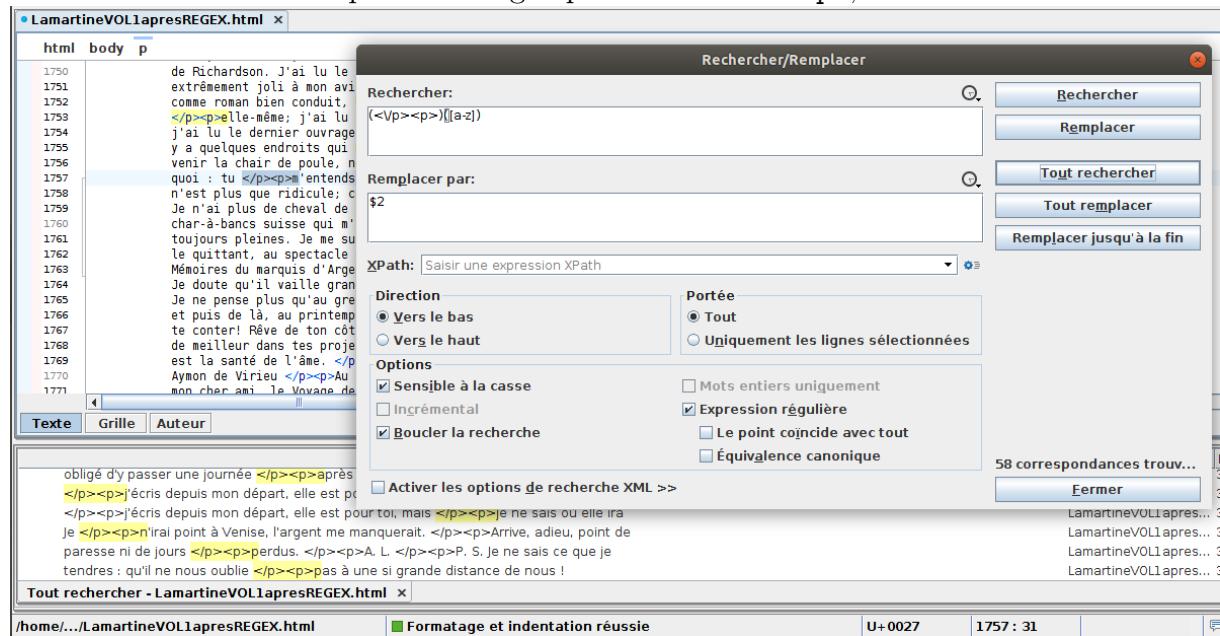
La capture d'écran ci-dessus montre comment cela se présente : on indique dans la

fenêtre la regex, et en cliquant sur « tout rechercher », 131 matches sont renseignés. Il suffit ensuite d'indiquer « tout remplacer » et l'HTML est donc nettoyé des 131 titres superflus.

Malheureusement, nous ne pouvons trouver des regex applicables à tous les corpus étant donné qu'ils se présentent différemment. Ainsi, pour la correspondance de Proudhon, nous supprimons ces titres avec une regex qui diffère : <p>[0-9]{1,3} CORRESPONDANCE <p>. Elle fait 108 matches.

De même, pour supprimer les </p><p> évoqués plus haut et coupant les phrases en deux, rompant ainsi l'unité, nous avons également utilisé les regex, cette fois-ci en indiquant que nous ne voulions supprimer que les <p> englobant des minuscules, car minuscules signifient phrases en cours.

FIGURE 8.8 – Exemple d'une regex pour enlever des <p>, HTML de Lamartine



Ici, nous matchons donc les deux balises <p> et nous mettons des parenthèses. Nous matchons celles qui se trouvent avant des minuscules (donc en milieu de phrase). Pour la substitution, nous écrivons \$2 comme cela les <p> sont supprimés et le \$2 qui fait référence à la deuxième parenthèse garde le contenu et n'enlève pas la première lettre du mot en minuscules.

Ainsi, les expressions régulières permettent un gain de temps considérable dans le nettoyage de l'HTML.

#### 8.2.4 Quelle granularité dans la correction ?

Cependant, celui-ci suscite des questionnements : dans quel mesure faut-il corriger les fautes ? Quelle correction faire ? Une correction technique ou orthographique ?

### 8.2.4.1 Une correction orthographique ?

Venant d'un parcours plus littéraire, il nous a été en effet très difficile de fermer les yeux sur certaines fautes, quitte à ce qu'elles soient corrigées par la suite. En effet, il nous a paru parfois plus judicieux de corriger en amont, c'est-à-dire dans l'HTML, certaines fautes orthographiques que nous avions remarquées, avant que celles-ci ne soient dispersées dans divers fichiers XML-TEI (en effet, la suite de la procédure est de diviser l'HTML en de multiples fichiers XML-TEI, un fichier par lettre) : corriger les fautes en amont permettrait donc de pouvoir matcher les fautes dans un seul fichier HTML. Ainsi, nous nous sommes permis de faire quelques corrections orthographiques, dont nous donnons quelques exemples.

Tout d'abord, un des problèmes de l'OCR est qu'il « recolle » un mot qui a été divisé par un tiret, lorsqu'il est 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 de l'HTML de Lamartine où l'expression « écris-moi » a été assemblée en un mot « écrismoi<sup>8</sup> ». C'est donc une fausse interprétation de l'OCR qui ne fait pas la distinction entre un mot « recollé » avec raison ou non. Tombant dessus par hasard, nous avons donc corrigé ce genre d'erreurs. De même, l'OCR confond parfois les « l » et les « t ». Ainsi, quand nous sommes tombée sur ce genre d'erreurs, nous avons remplacé « Celte » par « Cette »<sup>9</sup> ou « lu » par « tu »<sup>10</sup>.

Ainsi, lorsque cela s'y prêtait, nous nous sommes permis de corriger certaines fautes qui s'imposaient à nous. Néanmoins, ce n'était pas notre travail de les rechercher. Ce n'était pas notre objectif. Nous devions plutôt traquer les fautes techniques.

### 8.2.4.2 Avant tout, une correction technique

En effet, notre priorité était plutôt de corriger les balises comportant des problèmes, comme nous l'avons souligné plus haut (balises mal fermées, balises superflues, balises divisant des phrases en deux paragraphes).

Parmi les fautes qui nous ont retenue figurait le manque de certaines balises pouvant poser problème par la suite pour l'extraction avec Python. En effet, cette extraction se base sur certains marqueurs. Si ceux-ci sont mélangés de temps en temps au reste du texte, les lettres ne sont pas bien extraites. Ainsi, pour ce qui est de la correspondance de Lamartine, il manquait des balises <p> pour séparer la signature des chiffres romains désignant une nouvelle lettre : <p>ALPH. DE LAM. LXII</p>. Nous les avons donc ajoutées avec une regex, pour les soixante-cinq occurrences : <p>ALPH. DE LAM.</p><p>LXII</p>

En effet, il s'agit avant tout d'assurer une bonne extraction avec Python du fichier

8. 5 matches.

9. 17 matches

10. 70 matches, en faisant attention à ne pas corriger faussement certains « lu » qui sont justes (11 matches)

HTML pour qu'il soit transformé en fichiers XML. Pour cela, la dernière phase de pré-traitement est de repérer les marqueurs du texte en vue de l'extraction.

### 8.2.5 Premiers repérages des marqueurs

Les marqueurs avaient déjà été signalés dans le cahier de charges : à chaque édition imprimée correspondent des caractéristiques différentes. Ainsi, pour la correspondance de Lamartine, on remarque que les lettres sont séparées les unes des autres par un chiffre romain. Celui-ci manquait parfois, nous nous sommes donc permis de l'ajouter manuellement, comme pour la deuxième lettre de Lamartine. Pour Lamennais, les lettres sont délimitées par des chiffres arabes cette fois. Ceux-ci posent problème car ils ne sont pas bien reconnus par l'OCR. Par conséquent, ils ne peuvent pas être extraits facilement avec des regex dans Python puisqu'ils comportent des erreurs : le chiffre 5 est souvent pris pour un « a », le 5 et le 3 sont souvent confondus, et le 8 est pris pour un « S ». Au lieu de 15, on lit 1. ’1, au lieu de 19, on lit I!>; , à la place de 44, on trouve t4, et ftj pour 63, tH pour 64, pour n'en citer que quelques-uns. Nous avons donc dû retravailler ces marqueurs pour assurer une bonne extraction des lettres. Enfin, les lettres de Proudhon n'avaient pas de délimiteur. Nous nous sommes donc servie de sa signature « P.-J. PROUDHON. » pour extraire les lettres.

FIGURE 8.9 – Cahier des charges ELICOM, repérage des marqueurs de Lamennais

#### **Lamennais (Cas relativement facile grâce aux chiffres arabes qui délimitent les lettres)**

<https://catalogue.bnf.fr/ark:/12148/cb30728369q>

##### MARQUEURS DE LA LETTRE

Début des lettres : Chiffre arabe suivi d'un point et d'un tiret + adresse au destinataire en petites capitales sur le mode « A.... » (sur la même ligne)

+ lieu et date en minuscule au dessous

Fin de la lettre : les lettres sont rarement signées.

- Correspondance active
- Très longue préface dans le volume 1 intitulée « Notes et souvenirs » à enlever (jusqu'à la page 140)
- Distinction entre les lettres grâce à des chiffres arabes
- Alinéas marquent des paragraphes
- Le destinataires est indiqué en majuscule. Lettres à la suite adressées aux mêmes destinataires s'ouvrent sur « AU MÊME » ou « À LA MÊME » (voir p. 147)
- En principe date et lieux à droite sur le modèle (« Lieu, jj mois AAAA. »). Dans certains cas Lamennais mentionne également l'heure (p. 147)
- Structure de la lettre : les formules de politesse de début et de fin sont en grande majorité intégrées au corps de la lettre.

Ainsi, l'OCR de Gallica s'est avéré être assez performant dans l'ensemble et nous a rendu de précieux services pour l'extraction de la correspondance en HTML. Bien-sûr, il reste quelques fautes à corriger, aussi bien au niveau de la structure de l'HTML qu'au niveau purement orthographique. Par ailleurs, nous n'avons pas encore évoqué la question de la mise en page qui n'est pas traitée par l'OCR, ce qui nous pénalise également. Nous y reviendrons lorsque nous parlerons des fichiers XML en quatrième partie. Néanmoins, le bilan est très positif pour l'OCR de Gallica. L'apprentissage machine a donc bien été au cœur de l'acquisition des données pour notre projet d'ELICOM avec le Labex OBVIL. Il en est de même pour notre projet d'édition numérique de correspondance de Frédéric Le Play, même s'il se présente cette fois-ci sous la forme de l'HTR avec Transkribus.



# Chapitre 9

## L’HTR de Transkribus

### 9.1 Quelle procédure pour l’acquisition des données ?

#### 9.1.1 Rappels et point sur le corpus

Il s’agit désormais d’acquérir les données pour notre projet d’édition numérique de la correspondance de Frédéric Le Play avec le CRHXIX. Avant tout, faisons un point sur le corpus en notre possession. Nous pensons pour l’instant éditer les 2091 lettres échangées entre Frédéric Le Play et 94 correspondants entre 1837 et 1882.

En vue de leur mise en ligne, plusieurs étapes sont à suivre. Tout d’abord, 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 ?

#### 9.1.2 S’assurer des numérisations des manuscrits

Comme souligné plus haut, un travail de numérisation a déjà été engagé, et un récapitulatif a été fait sur les fonds numérisés<sup>1</sup>. Il est en effet capital d’avoir une vue claire là-dessus car elles constituent la matière première, la base de notre travail.

Plusieurs questions se posent sur la qualité de ces numérisations. En effet, comme nous l’avions déjà souligné dans notre deuxième partie<sup>2</sup>,

« Dans le cas de la publication de textes en fac-similés [...], la lisibilité des images est essentielle, ce qui suppose à la fois une attention aux formats d’acquisition (qualité de l’image exprimée en dpi), et une juste évaluation des besoins de stockage et d’infrastructure matérielle pour la diffusion/communication de celles-ci<sup>3</sup> »

---

1. Voir fig. 3.1

2. Voir 5.1.2

3. Ioana Galleron, Marie-Luce Demonet, Cécile Meynard, Idmhand Fatiha, Elena Pierazzo, et al., *Les publications numériques de corpus d'auteurs - Guide de travail, grille d'analyse et recommandations*, 2018,

Or, si nombre de numérisations sont satisfaisantes car elles ont été commandées à des services d'archives ou bibliothèques, pour d'autres manuscrits, nous ne possédons que des photographies prises par des particuliers et qui sont floues ou d'une qualité médiocre, tant pour la précision que pour la prise de vue. Il s'agira donc de les faire numériser par des professionnels.

Par ailleurs, nous nous posons encore la question du format de l'image. Serait-il bon d'unifier les formats ? Nous avons tantôt des fac-similés en JPEG, tantôt des PDF ou encore TIFF (*Tagged Image File Format*), plus rarement des PGN (*Portable Network Graphics*). Il serait bon d'unifier les formats mais nous ne pensons pas que cela soit indispensable<sup>4</sup>. L'important surtout est d'obtenir un fichier par page. C'est là que nous avons rencontré des problèmes avec certains PDF trop lourds, notamment ceux de la BNF et de la BIF, car ils comprenaient un trop grand nombre de pages : il faudrait donc diviser les fichiers pour qu'il y ait un seul fichier par page numérisée<sup>5</sup>, que l'on pourrait ensuite appeler dans le fichier XML-TEI comprenant la transcription correspondante.

En vue de l'édition numérique de correspondance, il serait bon de 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. Dans ce but, la numérisation en couleur est préférable. Certes, nous possédons, notamment pour les fonds Peruzzi et Loyson, des numérisations en noir et blanc, qui sont d'ailleurs de très bonnes qualité. On pourrait à long terme songer à les remplacer par des numérisations en couleur, mais cela ne nous semble pas du tout prioritaire. Le site de Flaubert que nous avons déjà évoqué ne met pas toujours en ligne des fac-similés en regard, et certains sont en couleur, d'autres en noir et blanc. Certes, ce n'est pas parfait mais cela n'enlève rien à la qualité du site.

Par ailleurs, il faudra être attentif au nommage de chaque fichier, procédant toujours de la même manière, comme par exemple `expediteur_destinataire_fonds_lettre1a.jpg`, en abrégeant les noms selon ce qui a été fixé. Ainsi, on pourrait écrire `lp_ribbe_arbaud_l1b.jpg` ce qui signifie : lettre de Frédéric Le Play à Charles de Ribbe, musée Arbaud, lettre 1, page 2 (pour le b)<sup>6</sup>.

Ainsi, avant tout, il s'agit de s'assurer de la bonne qualité des fac-similés et de leur bon nommage. Ceci fait, on peut commencer à transcrire, puisque nous ne voulons pas seulement éditer les fac-similés. Nous voulons aussi avoir leur transcription en regard. Comment y parvenir pour ce corpus si important ? N'y aurait-il pas possibilité d'automatiser tout cela pour aller plus vite ?

---

URL : <https://halshs.archives-ouvertes.fr/halshs-01932519/document> (visité le 05/05/2020).

4. Le retour en arrière aurait un coût probablement trop important alors que non indispensable

5. Par page numérisée, et non par lettre qui comprend parfois plusieurs pages

6. Cette proposition reste encore un peu longue. On pourrait trouver quelque chose d'aussi précis mais de plus court.

### 9.1.3 Transcriptions manuelles ou automatisées ?

Nombre de manuscrits<sup>7</sup> ont déjà été transcrits manuellement, comme souligné plus haut, que cela soit par des étudiants, des doctorants, des stagiaires. Or, la transcription s'avère être parfois un exercice difficile : les hommes du XIX<sup>e</sup> siècle disposent de certains codes d'écriture ou d'abréviations<sup>8</sup> qui ne nous sont pas familiers. Par ailleurs, nombre de mots sont difficiles à lire, du fait des différentes écritures. Je pense par exemple à celle de Jules Baroche<sup>9</sup>. 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. En une heure, on peut donc transcrire douze pages, sans compter la relecture, ce qui nécessite deux-cents heures de transcription pour deux-mille 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 machine. C'est ainsi que nous avons fait le choix de nous tourner vers Transkribus.

## 9.2 Transkribus, un outil de transcription

### 9.2.1 Transkribus, un pari

Transkribus est un outil utilisé par de nombreux projets, mais il n'est pas forcément approprié à tous les corpus. Une transcription « manuelle » nécessite, certes, beaucoup de temps et une relecture attentive, mais entraîner une machine telle que Transkribus nécessite également beaucoup de temps, et ceci avec l'incertitude du résultat. S'il y a plus de 90 % de taux de réussite, la transcription pourra se faire sans trop de difficultés, mais la relecture restera longue et nécessaire. S'il y a 98 % de taux de réussite, c'est à nous de voir si nous acceptons ce taux d'erreurs ou si nous préférons relire pour un rendu plus optimal, mais également plus coûteux en temps, sachant que nous aurons de toutes façons les fac-similés en regard.

Par ailleurs, comme le soulignent les tutoriels mis en ligne pour initier à Transkribus, il faut « du temps pour explorer Transkribus et se familiariser avec son fonctionnement<sup>10</sup> ». C'est donc un réel investissement au début, avec l'incertitude du résultat et la possibilité d'un échec.

---

7. Quelques centaines probablement, nous n'avons pas le chiffre exact.

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

9. Voir Annexe B.1 issue des premières pages du PDF SIM MS 6062 (p.7). Jules Baroche par exemple, lorsque deux « s » se suivent, allonge le premier. Ce sont les « s » longs que nous évoquions plus haut

10. Régis Schlagdenhauffen, *Comment utiliser Transkribus en 10 étapes (voire moins)*, Site web de l'EHESS, URL : <http://regis-schlagdenhauffen.eu/wp-content/uploads/2018/01/Comment-utiliser-Transkribus-%E2%80%93-en-10-%C3%A9tapes-ou-moins.pdf> (visité le 22/05/2020.)

Pour notre part, les résultats ont été dès le début encourageants. Sur 16 lettres, nous avons obtenu des scores de 84 % d’erreur sur les caractères en entraînement. Nous avons donc continué cette aventure virtuelle, encouragés que nous étions par ces premiers retours. Avant de présenter plus en détail notre travail sur Transkribus, il importe de présenter plus en détail cet outil.

## 9.2.2 Un outil pensé par le READ

### 9.2.2.1 Le projet

L’EADH, association européenne pour les humanités numériques (en anglais *European association for digital humanities*), fondée en 1973<sup>11</sup>, rassemble en son sein de nombreux projets pour faire avancer les humanités numériques<sup>12</sup>. Parmi eux, le projet READ<sup>13</sup>, *Recognition and Enrichment of Archival Documents* tient une place toute particulière. Comme son nom l’indique, il se consacre à la reconnaissance et à l’enrichissement des documents d’archives, visant à rendre les documents d’archives plus accessibles grâce à l’utilisation de technologies de pointe. L’objectif principal de READ est de fournir une plate-forme de services<sup>14</sup> pour la reconnaissance, la transcription et la recherche automatisées de documents historiques. Entraîner les ordinateurs à lire du texte manuscrit de cette manière promet de révolutionner l’accès aux archives écrites. C’est dans cette optique que l’outil de transcription Transkribus a été conçu.

### 9.2.2.2 L’outil

Transkribus est un logiciel de transcription collaborative. Pour l’utiliser, il suffit de s’inscrire et de le télécharger. Puis, on peut importer les manuscrits en n’importe quelle langue, les transcrire manuellement, seul ou en équipe, en liant le texte à l’image grâce à la segmentation des images en régions de textes, lignes et mots réalisée à l’aide d’outils d’analyse de disposition. Puis avec ces transcriptions, on peut entraîner un modèle qui permettra à Transkribus de reconnaître par la suite l’écriture en question et d’effectuer lui-même les transcriptions.

## 9.2.3 Transkribus et l’apprentissage machine

Transkribus s’utilise aussi bien pour l’OCR que pour l’HTR, mais ceux-ci diffèrent. En effet,

---

11. Sous le nom de *Association for Literary and Linguistic Computing* (ALLC), voir *About*, Site web de l’EADH, URL : <https://eadh.org/about>, (visité le 23/09/2020).

12. Le projet *correspSearch* évoqué dans la deuxième partie en fait partie.

13. Projet lancé en 2015

14. Voir *Accueil*, Site web de Transkribus, URL : <http://transkribus.eu> (visité le 06/03/2020).

« La reconnaissance du texte manuscrit est un champ à part entière au sein des systèmes d'OCR. On parle d'ailleurs de *Handwritten text recognition* (HTR) pour désigner le traitement des documents manuscrits, preuve qu'ils nécessitent des technologies spécifiques<sup>15</sup>. »

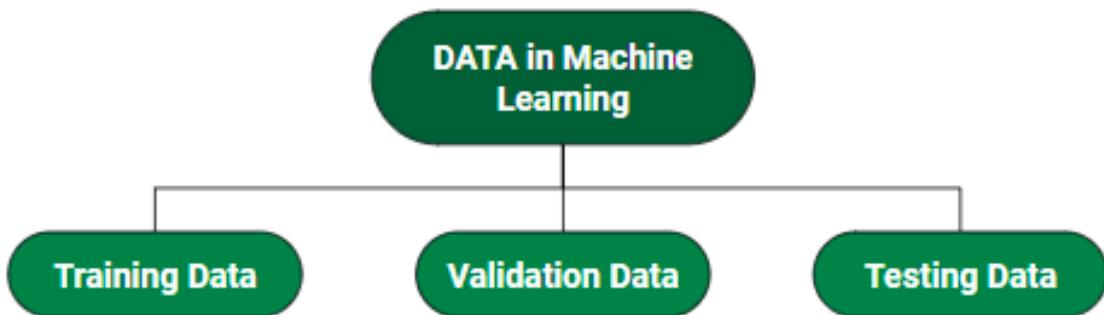
L'HTR n'est pas comme l'OCR, où, comme on a pu le remarquer avec l'OCR de Gallica, on appuie sur le bouton et le document est reconnu automatiquement. Qui dit HTR dit entraînement adapté à l'écriture d'une main particulière, sachant que cette main peut varier d'écriture ce qui rend l'entraînement d'autant plus fastidieux. Ainsi, Transkribus se doit d'entraîner des modèles<sup>16</sup>. Pour cela, il est nécessaire de transcrire préalablement 100 pages, soit 15 000 à 20 000 caractères.

En réalisant ce travail, nous participons de près ou de loin à l'objectif du READ. En effet, le but à long terme est d'entraîner le plus de styles d'écriture différents, de manière à ce que Transkribus soit en mesure de traiter la plupart des documents manuscrits sans entraînement préalable. Plus les utilisateurs travailleront avec Transkribus pour leur transcription, plus vite cet objectif ambitieux sera atteint<sup>17</sup>.

#### 9.2.4 Point sur la terminologie

Avant de voir comment nous avons procédé avec Transkribus, il est utile de faire un point sur la terminologie en usage dans l'apprentissage machine<sup>18</sup>. Par ailleurs, l'anglais étant la norme en informatique, comment rendre les termes en français ?

FIGURE 9.1 – Les données dans l'apprentissage machine



15. in Alix Chagué, *ibidem*. p.42.

16. *Handwritten Text Recognition Workflow*, Wiki de Transkribus, URL : [https://transkribus.eu/wiki/index.php/Handwritten\\_Text\\_Recognition\\_Workflow](https://transkribus.eu/wiki/index.php/Handwritten_Text_Recognition_Workflow) (visité le 23/09/2020).

17. *Questions and Answers*, Wiki de Transkribus, URL : [https://transkribus.eu/wiki/index.php/Questions\\_and\\_Answers](https://transkribus.eu/wiki/index.php/Questions_and_Answers) (visité le 23/09/2020)..

18. Voir *ML / Introduction to Data in Machine Learning*, Site web GeeksforGeeks, URL : <https://www.geeksforgeeks.org/ml-introduction-data-machine-learning/> (visité le 02/06/2020).. La figure 9.1 est tirée de cet article.

Les données dans l'apprentissage machine se divisent en trois catégories, comme nous pouvons le voir sur la figure ci-dessus<sup>19</sup> :

- Les données d'entraînement ou *Training Data* : ce sont les données que nous utilisons pour entraîner le modèle et qu'il apprend.

Dans Transkribus, on appellera ces données *Training Set*<sup>20</sup> ou set d'entraînement<sup>21</sup>.

- Les données de validation ou *Validation Data* : elles sont utilisées pour évaluer le modèle d'après les données d'entraînement.
- Les données de test ou *Testing Data* : elles permettent d'évaluer la qualité du modèle une fois qu'il est entraîné. Le modèle prédit le taux d'erreur à chaque fois avec les données de test. D'une fois à l'autre, nous pouvons ainsi voir la progression du modèle par l'expérience. En effet, plus nous rentrons de données d'entraînement, plus le modèle acquiert de l'expérience, moins il se trompe, plus le taux de réussite est important et donc le taux d'erreur moindre.

Pour Transkribus, pendant le processus de formation, quelques pages sont mises de côté à titre de test. Elles ne sont pas utilisées pour l'entraînement du modèle HTR+. Elles sont plutôt utilisées pour tester les performances de notre modèle<sup>22</sup>.

Ainsi, la plate-forme Transkribus permet aux utilisateurs d'entraîner un modèle HTR+ de reconnaissance automatique de documents. Le modèle doit être entraîné pour reconnaître un style d'écriture particulier. Cela se fait en lui "montrant" les images et les transcriptions exactes correspondantes<sup>23</sup>.

Voyons plus en détail comment procéder à l'acquisition des données pour notre projet avec l'outil de transcription Transkribus.

19. Pour nous, les données en général sont les manuscrits. Mais en soi, les données peuvent être du texte, des sons et images etc. : « *It can be any unprocessed fact, value, text, sound or picture that is not being interpreted and analyzed* » *In ibidem*.

20. Voir *How to transcribe. Train a model*, Wiki de Transkribus, URL : [https://transkribus.eu/wiki/images/3/34/HowToTranscribe\\_Train\\_A\\_Model.pdf](https://transkribus.eu/wiki/images/3/34/HowToTranscribe_Train_A_Model.pdf) (visité le 23/09/2020), traduit en français : *Entraînement d'un modèle dans Transkribus*, Wiki de Transkribus, URL : [https://transkribus.eu/wiki/images/8/84/Comment\\_entra%C3%A9ner\\_un\\_Mod%C3%A8le\\_dans\\_Transkribus.pdf](https://transkribus.eu/wiki/images/8/84/Comment_entra%C3%A9ner_un_Mod%C3%A8le_dans_Transkribus.pdf) (visité le 22/05/2020).

21. *Entraînement d'un modèle dans Transkribus*, Wiki de Transkribus, URL : [https://transkribus.eu/wiki/images/8/84/Comment\\_entra%C3%A9ner\\_un\\_Mod%C3%A8le\\_dans\\_Transkribus.pdf](https://transkribus.eu/wiki/images/8/84/Comment_entra%C3%A9ner_un_Mod%C3%A8le_dans_Transkribus.pdf) (visité le 22/05/2020).

22. *Idem.*, p.7

23. *Ibid.*, p.3

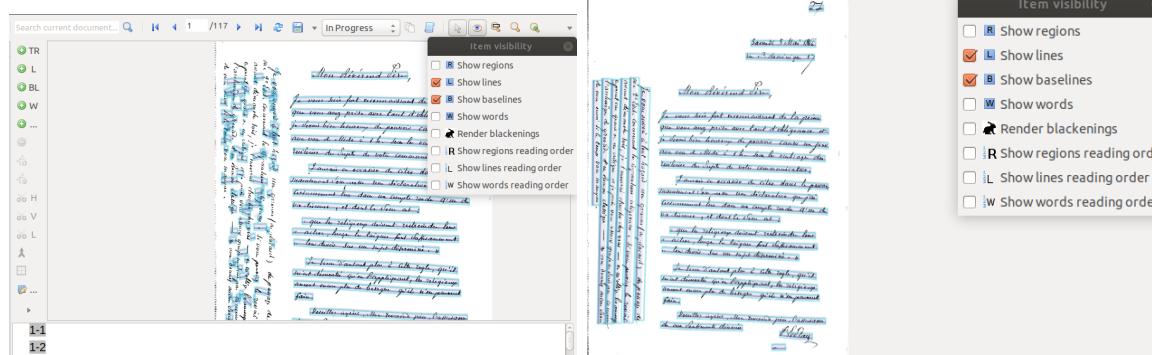
## 9.3 Chargement des données d'entraînement et premier traitement

### 9.3.1 Procédure pour le chargement des données d'entraînement

Pour charger les données d'entraînement, il y a toute une procédure à suivre. Il s'agit d'indiquer à la machine à quoi correspond tel caractère pour l'aider à apprendre l'écriture de Frédéric Le Play afin de pouvoir par la suite procéder elle-même aux transcriptions.

Après avoir importé les données, on divise le texte en *Text Region* (TR) ou régions de texte : on lui indique où se trouve l'écriture. Puis on la lance pour que, dans ce cadre donné, elle tente de reconnaître elle-même où sont les lignes et surtout les *base line* (BL), c'est-à-dire le trait à la base d'une ligne. En général, la Transkribus fait bien le travail, mais parfois, dès cette étape, on est en bute à certains dysfonctionnements lorsque la page comprend deux sens d'écriture.

FIGURE 9.2 – Résolution du problème de TR, deux sens d'écriture, capture d'écran de Transkribus



Transkribus se trompe alors totalement et reconnaît à la fois une multitude de lignes et de TR<sup>24</sup> : on peut gagner du temps en supprimant par TR (cela supprime du même coup les lignes que la TR a en son sein) mais quand il y a de multiples TR, on est obligé de tout supprimer à la main ce qui est une première perte de temps<sup>25</sup>.

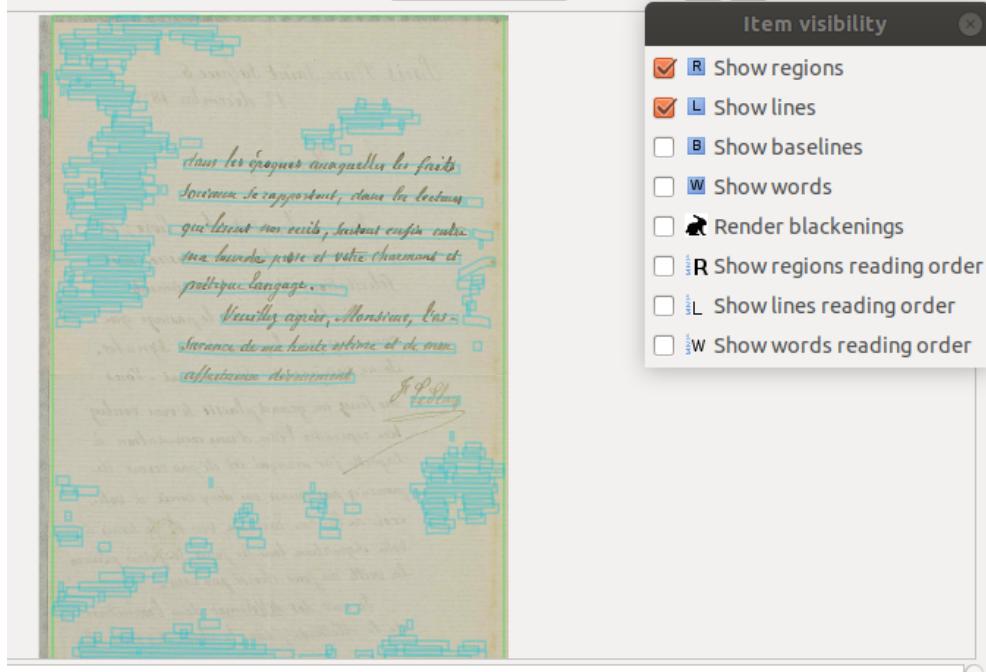
De même, quand Le Play utilise un papier légèrement strié pour écrire, celui-ci est confondu avec les caractères et considéré comme des lignes d'écriture<sup>26</sup>. Il signale donc des lignes qui n'en sont pas. Ici, le mieux est de supprimer la TR et de mettre les BL manuellement, de gauche à droite et non l'inverse, sinon, la ligne apparaît en-dessous de la BL.

24. Voir fig. 9.2

25. Voir la version agrandie de ces deux images dans les annexes B.2

26. Voir fig. 9.3

FIGURE 9.3 – Problème de TR à cause du papier, capture d'écran de Transkribus



Puis on peut procéder à la copie des transcriptions. Pour cela, on procède par copier/coller si les transcriptions ont déjà été réalisées, ce qui est notre cas. Néanmoins, cette phase nécessite une attention particulière. En effet, la mise en page et certaines graphies, lorsque l'on copie/colle, ne sont pas prises en compte. Il s'agit donc de les renseigner au fur et à mesure dans Transkribus.

Transkribus permet ainsi d'indiquer les exposants, les mots qui ne sont pas clairs ou illisibles (*unclear*), les mots barrés (*strikethrough*), les mots soulignés (*underlined*) etc.<sup>27</sup>..

Or, cette phase de chargement des données d'entraînement suscite d'autres questionnements sur les transcriptions.

### 9.3.2 Des transcriptions qui suscitent des questionnements

#### 9.3.2.1 Des transcriptions de qualité variable

Avant tout, nous avons commencé par importer les manuscrits dans Transkribus. Pour cela, il faut savoir lesquels nous désirons traiter en priorité. Notre choix s'est arrêté sur l'écriture qui ressort majoritairement du corpus en notre possession, à savoir celle de Frédéric Le Play. Ceci posé, nous avons commencé par importer dans Transkribus les manuscrits, puis les transcriptions qui avaient déjà été réalisées pour le CRHXIX. Nous avons dressé un tableau pour classer les correspondances à traiter en premier, en commençant par celles qui étaient de meilleure qualité quant à la numérisation et à la transcription, pour finir avec celles qui étaient de moindre qualité. En effet, certaines

27. Voir annexe B.3, fig. B.6, les styles de transcription en jaune.

transcriptions ont déjà été relues ce qui est une garantie de meilleure qualité.

D'autres transcriptions faites par des étudiants n'ont pas encore été relues, par manque de temps, comportant des erreurs souvent dues à un manque compréhensible de familiarité avec les documents du XIX<sup>e</sup> siècle. Par ailleurs, certaines corrections notamment dans les accentuations ont été faites dans les transcriptions. Comment gérer ces différences par rapport aux originaux, alors que nous voulons justement charger des données pour apprendre à la machine à reconnaître les caractères ?

### 9.3.2.2 Difficultés de transcriptions

En une dizaine de jours, nous avons été amenée à rentrer 20 000 caractères dans le serveur de Transkribus afin d'entraîner le modèle pour l'écriture de Frédéric Le Play.

De notre côté, nous n'avons pas fait de transcriptions à proprement parler, mais nous avons dû copier/coller les transcriptions réalisées par les étudiants et stagiaires pour le CRHXIX. Or, nous avons eu quelques doutes lors de la manipulation de ces transcriptions : celles-ci n'étant pas toujours exemptes de fautes, nous nous sommes demandée dans quelle mesure nous devions les corriger.

Plusieurs questions émergent.

Tout d'abord, d'un point de vue purement technique, comment la machine peut-elle apprendre des caractères que nous-mêmes peinons à distinguer ? Ainsi, le « z » de Le Play est parfois écrasé et ressemble fort à un « r » : cela a nécessité de notre part une certaine attention pour vérifier l'orthographe parfois incertaine des transcriptions d'une part, et d'autre part nous nous sommes demandée si la relecture des transcriptions opérées par Transkribus ne demanderait pas beaucoup de travail étant donné que les caractères sont souvent mal formés et donc difficiles à lire, et par l'œil humain, et par la machine, c'est du moins ce que nous craignons<sup>28</sup>. On constate ci-dessous que le « z » de « connaissez » pourrait être pris pour un « r ».

Par ailleurs, nous avons dû être attentives à certaines mauvaises interprétations, comme par exemple, on lit dans la transcription : « Que de bien à faire par le bon exemple ! car les sermons de vertu déposées [sic] pendant le moyen-âge ont percuté ça et là chez les vieillards. »

Alors qu'il faudrait lire ce nous semble : « Que de bien à faire par le bon exemple ! car les semences de vertu déposées pendant le moyen âge ont persisté ça et là chez les vieillards. »

Néanmoins, ces premières transcriptions même non relues sont d'une aide extrêmement précieuse : le fait d'avoir déjà une grille de lecture aide à se mettre dans l'esprit du texte et à mieux saisir les possibles erreurs. C'est plus facile que si on aborde les

28. Il est possible que par la suite, Transkribus sache mieux lire que l'homme, mais vu l'avancement du modèle de Le Play aujourd'hui, malgré les 20 000 caractères rentrés pour l'entraînement, ce n'est pas le cas.

FIGURE 9.4 – F. Le Play au R. P. Hyacinthe Loyson, 1866, capture d'écran de Transkribus

*roupera un peu après l'avec les deux am  
connaitre et qui ont comme nous le s  
de dévoiler la vérité. A partis du d  
terdu tout travail, et je ne connais pa  
able délassement que celui qui consiste  
le vrai avec un homme tel que vous.*

grosse besogne du jour.

si la Règle le permettait serait de

le soir après avec les deux amis que vous connaissez et que vous

FIGURE 9.5 – F. Le Play au R. P. Hyacinthe Loyson, 1866, capture d'écran de Transkribus

*que de bien à faire, parle bon exemple ! car les  
semeurs de vertu depuis plusieurs le moyen âge ont permis  
ça et là chez les vieillards*

textes avec un esprit totalement neuf. En outre, on peut bien dire qu'une des vertus de Transkribus est de nous permettre de réviser certaines transcriptions<sup>29</sup>

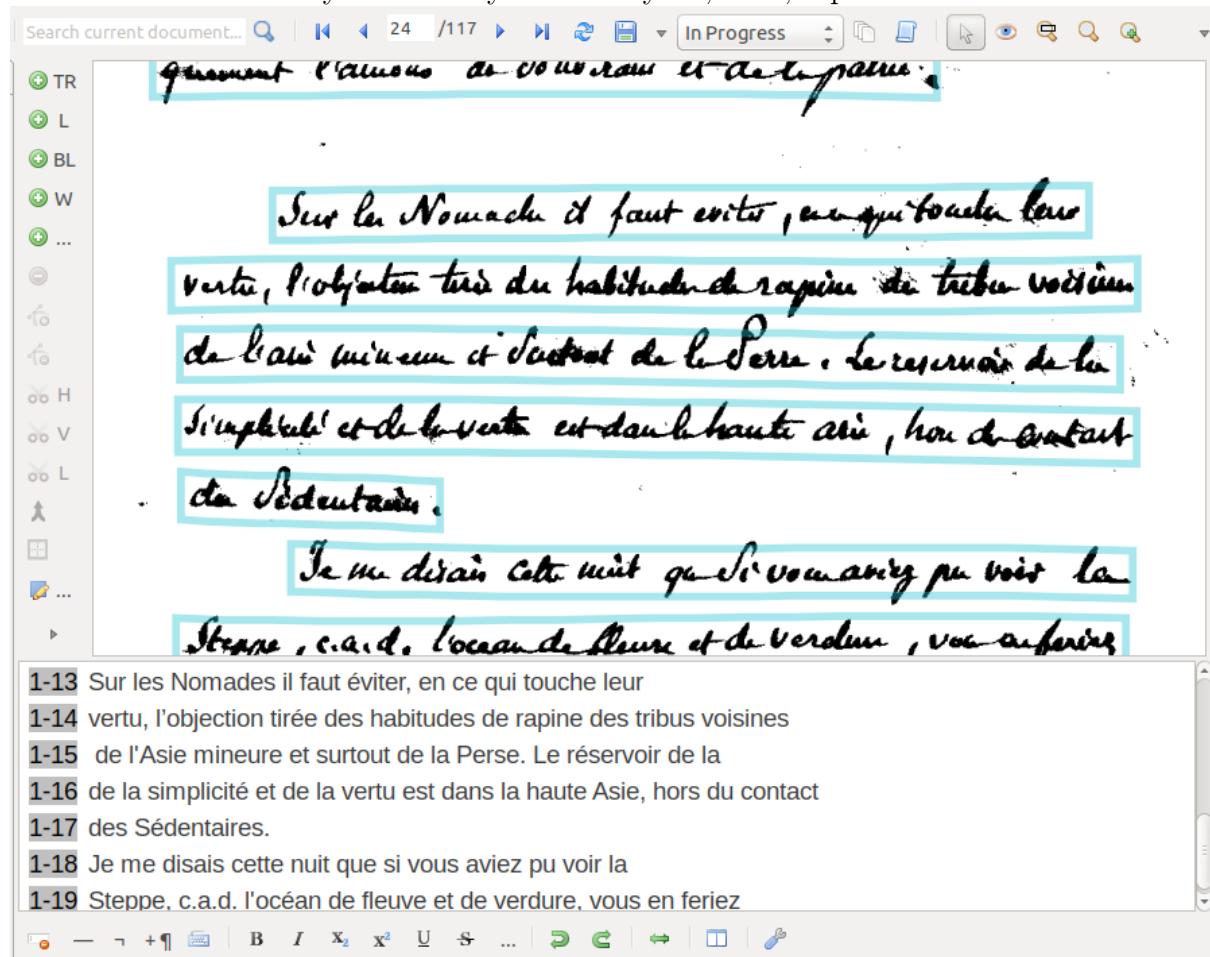
On remarque ici par ailleurs la cédille manquante au mot « ça », et l'accent circonflexe manquant sur le « a » de « âge » dans le manuscrit original. Par ailleurs, l'espace entre la préposition « par » et l'article défini « le » est si restreint qu'on croirait lire au premier regard « parle ». En trois lignes, on voit donc les complexités de transcriptions posées par l'écriture de Le Play.

Cette complexité se retrouve également face à certains mots difficilement lisibles et signalés comme illisibles par les premiers transcripteurs. Nous-même avons parfois peiné à saisir le sens de certains mots. Avec l'aide de l'équipe, nous avons pu venir à bout de certaines difficultés, comme ci-dessous<sup>30</sup> où la graphie doublée des bavures de l'encre nous ont fait buter sur certains mots.

29. À ce sujet, voir dans les livrables du CRHXIX le dossier 3-transcriptions et l'explication dans les annexes.

30. Fig. 9.6

FIGURE 9.6 – F. Le Play au R. P. Hyacinthe Loyson, 1867, capture d'écran de Transkribus



### 9.3.2.3 Transcriptions et principes de transcription

Une des plus grandes difficultés pour nous a été aussi de faire la part des choses entre les majuscules justifiées ou non. Dans les principes de transcription évoqués plus haut<sup>31</sup>, il avait été convenu de respecter l'usage des majuscules par Le Play lorsqu'il semble avoir un sens précis, selon ses usages. Par exemple, Réforme plutôt que réforme ; ou les Autorités Sociales et de retenir l'usage actuel lorsque les majuscules n'ont pas lieu d'être : par exemple, « le concours de nos amis », et non pas « le Concours de nos Amis ». Cependant, certains cas particuliers nous ont laissée (peut-être à tort) perplexe. En effet, on remarque parfois une logique dans les majuscules chez Le Play. Faut-il la conserver ? Ici, Le Play écrit « Vous remarquerez que dans mon Nouveau Plan, les anciens Chapitres<sup>32</sup> deviennent des Livres ; et que les anciens Paragraphes deviennent des Chapitres. »

En réalité, Le Play met des majuscules quand il veut insister sur un mot, qu'il souligne en plus de mettre une majuscule. Une fois que c'est établi, il continue sans ma-

31. Voir 5.2.3.1

32. Nous soulignons ici au lieu de mettre en italiques car nous reproduisons ce qui apparaîtra sur le site. Les mots soulignés et barrés apparaîtront ainsi.

FIGURE 9.7 – F. Le Play au R. P. Hyacinthe Loyson, 1866, capture d'écran du manuscrit

*et dans le Livre Premier.*

*Vous remarquerez que dans mon Nouveau Plan, les anciens chapitres deviennent des Livres ; et que les anciens paragraphes deviennent des chapitres.*

*Les 15 anciens paragraphes des deux premières divisions sont devenus 15 chapitres.*

*Le premier temps de mon Séjour ici a été consacré à*

jusques : « Les 15 anciens paragraphes des deux premières divisions sont devenus 15 chapitres ». Dans ce cas-là, pour notre part, nous serions pour conserver l'esprit de l'auteur de ces lettres en conservant la forme, et donc laisser les majuscules. Quant aux mots « Livre Premier » et « Nouveau Plan », ils répondent également à la logique d'insistance de Le Play donc nous serions pour garder les majuscules. Néanmoins, nous enlèverions la majuscule à « Séjour » et ajouterions un accent aigu sur le « e » car c'est un autre cas de figure. La frontière est donc parfois difficile à voir entre la conservation ou non des majuscules.

#### 9.3.2.4 Des transcriptions qui ont pour but d'entraîner un modèle

Une remarque s'impose : nous rentrons ici des données d'entraînement en vue de la création d'un modèle qui permettra de transcrire automatiquement les lettres de Le Play. Si nous souhaitons que le modèle soit bien entraîné, il faut rentrer les données de façon brute, sans tri (majuscule ou non ? accent ou non ?). Ces questions ne devront se poser qu'au moment des relectures de transcription, après le travail de transcription sur Transkribus. Pour l'instant, nous devons respecter l'écriture de Le Play telle qu'elle apparaît sur le manuscrit, et non enlever des majuscules, ajouter des accents.

En revanche, c'est le moment où il s'agit de déchiffrer les mots classés illisibles par les premiers transcripteurs et corriger les fautes de transcription, ce que nous avons déjà évoqué plus haut.

Ici malheureusement, nous avons dû peu à peu lâcher du lest et baisser nos exigences.

Comment pouvions-nous en si peu de temps coller entièrement au texte et corriger les premières transcriptions ? Rentrer 20 000 caractères n'est pas une mince affaire et représente une centaine de pages environ : à la fin, nous avons été contrainte de nous contenter à faire des copiés/collés et corriger simplement les fautes qui sautaient aux yeux.

Une fois les caractères rentrés, il s'agit de procéder à l'entraînement du modèle, au fur et à mesure : on peut constater ainsi la progression de l'apprentissage machine.

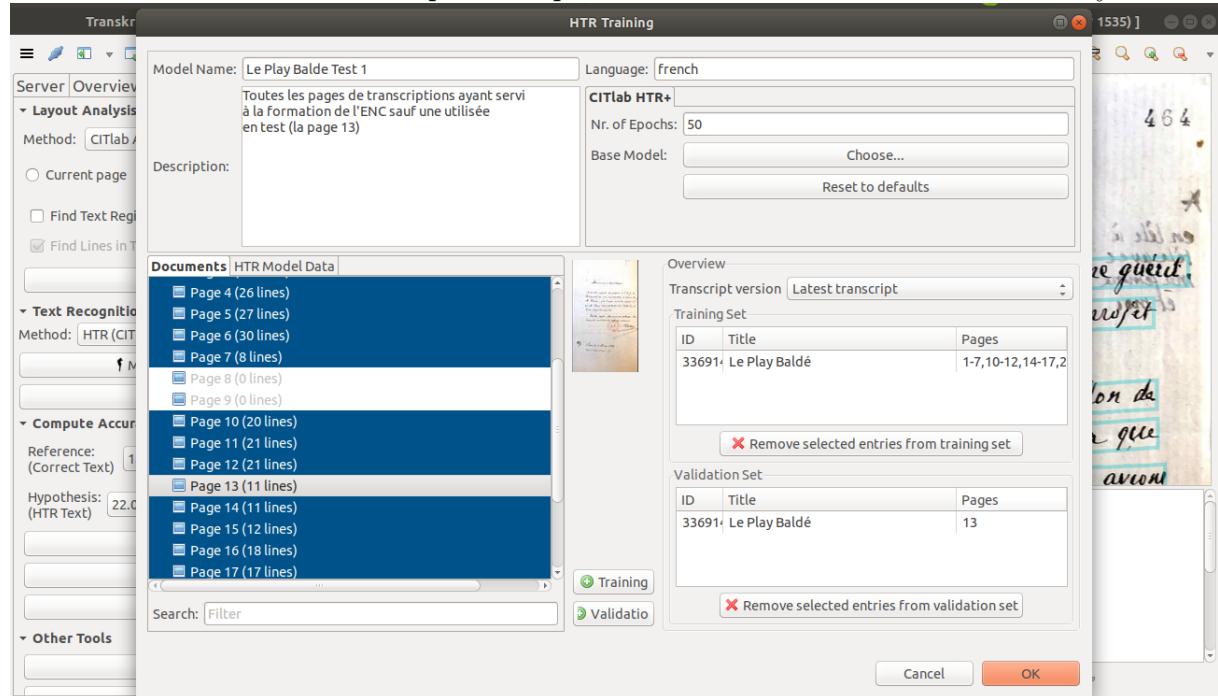
## 9.4 Entraînement d'un modèle. Quels résultats ?

### 9.4.1 Mise en place de l'entraînement

Peu à peu, on commence à entraîner le modèle. Plus on a de données d'entraînement, plus le modèle se perfectionne.

On sélectionne les pages de transcription et on les met dans le *Training Set*, puis on met de côté 10 % des données dans le *Validation Set* qui compare ce que nous avons transcrits avec ce que la machine reconnaît après avoir été entraînée<sup>33</sup>. Il est conseillé de

FIGURE 9.8 – Mise en place du premier entraînement du modèle Le Play



prendre à chaque fois le même set de validation pour se faire une meilleure idée des pro-

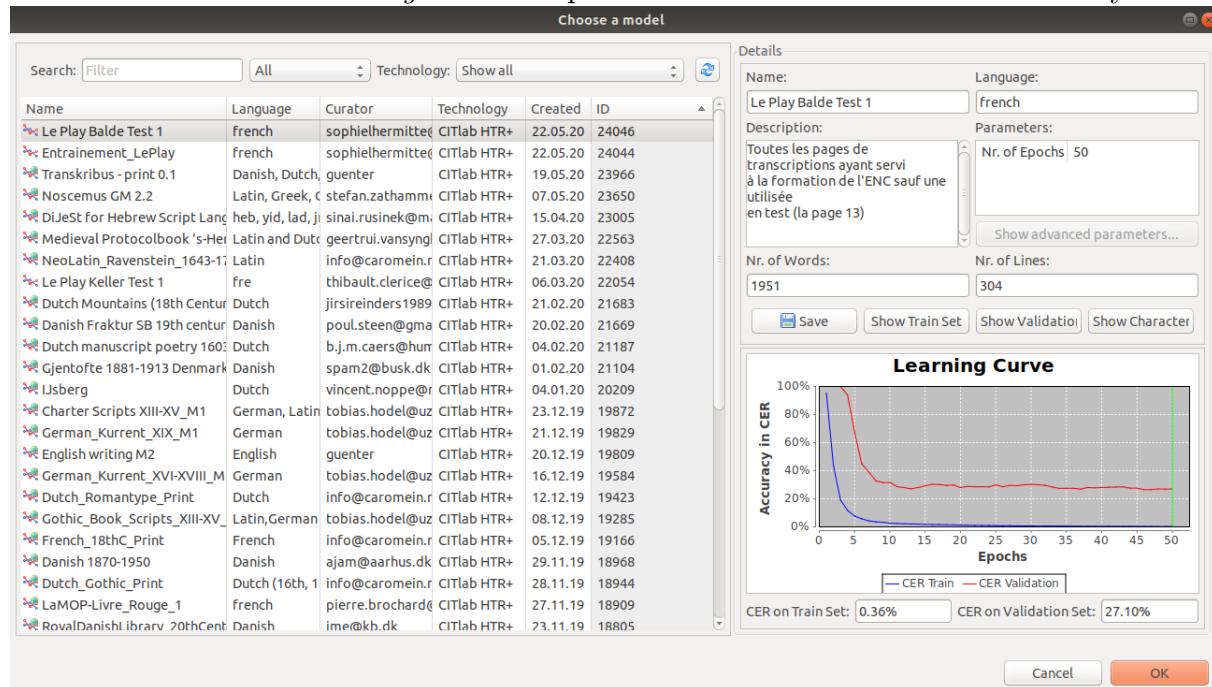
33. Voir la figure 9.8 où on peut observer la sélection entre *Training Set* et *Validation Set*, juste avant de cliquer sur « OK » pour lancer l'entraînement du modèle. Ici, nous n'avions mis qu'une page pour le set de validation ce qui était normal car nous n'avions pas encore beaucoup de données et une page représentait donc environ 10 %.

grès réalisés par l'apprentissage de la machine. De ce point de vue, notre entraînement du modèle a péché car nous n'avons pas mis assez de données de côté pour le set de validation. Nous pensions qu'il ne fallait mettre qu'une page alors qu'il fallait mettre 10 %. Nous avons donc rectifié pour les derniers entraînements, mais cela fausse donc un peu la courbe de progression.

#### 9.4.2 Quelle progression du modèle, quels résultats ?

Celle-ci est annoncée à la fin de chaque nouvel entraînement. On parle alors de *Learning Curve* ou courbe d'apprentissage. On voit donc ci-dessous<sup>34</sup> le graphique de la courbe d'apprentissage du premier entraînement du modèle Le Play. À gauche figurent tous les modèles disponibles, et tout en haut, en grisé, la ligne qui sélectionne le modèle qui nous intéresse et qui apparaît dans la colonne de droite.

FIGURE 9.9 – *Learning Curve* du premier entraînement du modèle Le Play



Le taux d'erreurs de reconnaissance de caractères ou *Character Error* (CER) définit l'efficacité du modèle. Cela représente le taux (en %) de caractères qui n'ont pas été correctement transcrits par l'HTR. L'idéal serait d'atteindre un CER inférieur à 5 %, même si en soi un taux CER inférieur à 10 % est déjà un taux satisfaisant.

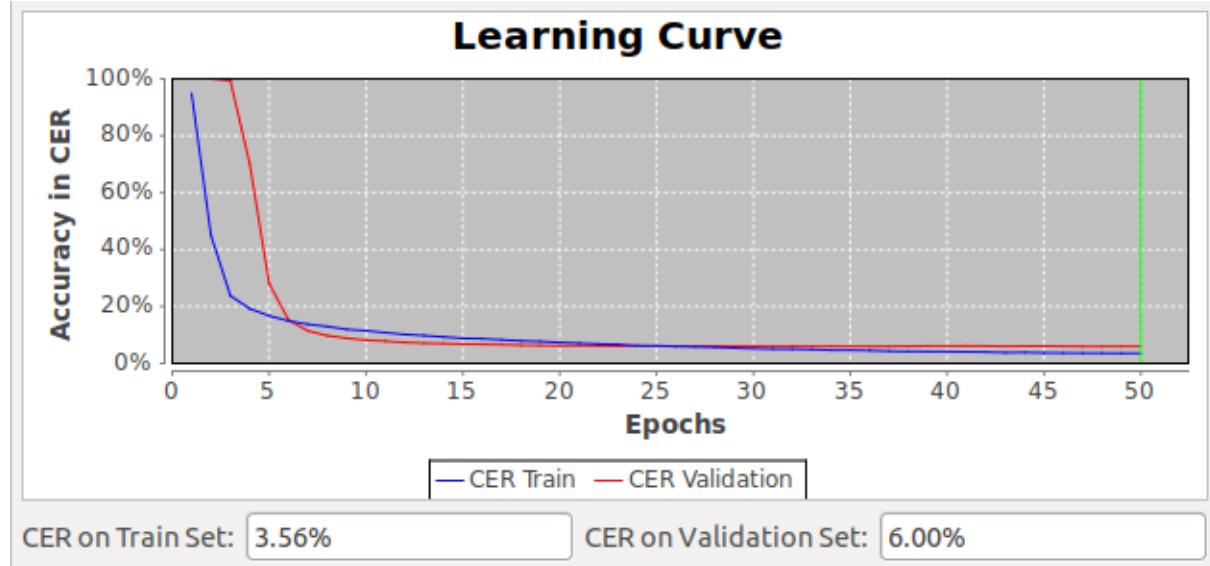
Voyons plus en précision le graphique de la courbe d'apprentissage qui illustre la précision du dernier modèle entraîné<sup>35</sup>, avec des données d'entraînement riches de 3146

34. Cf. Fig. 9.9

35. Cf. Fig. 9.10

lignes soit 23 729 mots.

FIGURE 9.10 – *Learning Curve* du dernier entraînement du modèle Le Play



Comme précisé dans le tutoriel fourni par Transkribus<sup>36</sup>, l'axe Y définit l'*Accuracy in CER* soit la précision en CER qui est affichée en pourcentage sur l'axe des ordonnées. La courbe commence toujours à 100 % et descend au fur et à mesure que l'entraînement progresse et que le modèle s'améliore.

L'axe X est défini comme *Epochs* soit époques. Pendant le processus de formation, Transkribus effectue une évaluation après chaque époque. Lorsqu'on entraîne un modèle, on spécifie le nombre d'époques dans lesquelles le *Training set* doit être divisé. Plus il y a d'époques, plus la formation dure longtemps. Ici nous en avons choisi 50.

Deux lignes apparaissent sur le graphique : une rouge qui indique l'état d'avancement des évaluations dans le jeu de tests et une bleue qui indique la progression de l'entraînement. « Le programme s'entraîne d'abord dans le *Training Set*, puis se teste à l'aide des pages du *Validation Set*.<sup>37</sup> »

« Au-dessous du graphique se trouvent deux pourcentages qui se réfèrent aux taux d'erreurs de l'ensemble d'apprentissage et de l'ensemble de test<sup>38</sup> ». Notre modèle a un taux d'erreur (CER) de 3,56 % pour le *Training Set* et de 6 % pour le *Validation Set* sachant que nous avions rentré, comme le signale la figure 9.11, 23 729 mots pour le *Train Set* et 1949 mots pour le *Validation Set* (ce qui est insuffisant car cela représente un peu moins de 10 %. Il manque environ 400 mots).

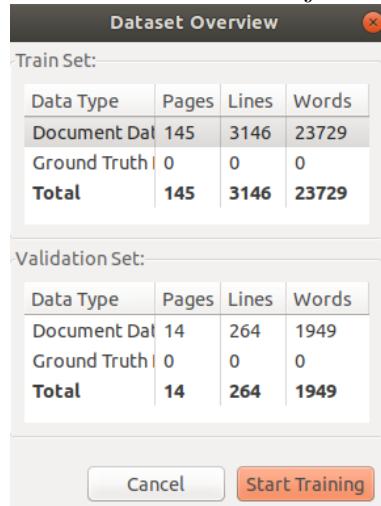
La valeur de l'ensemble de validation est significative car elle montre comment le

36. Entraînement d'un modèle dans Transkribus, Wiki de Transkribus, URL : [https://transkribus.eu/wiki/images/8/84/Comment\\_entra%C3%A9ner\\_un\\_Mod%C3%A8le\\_dans\\_Transkribus.pdf](https://transkribus.eu/wiki/images/8/84/Comment_entra%C3%A9ner_un_Mod%C3%A8le_dans_Transkribus.pdf) (visité le 22/05/2020).

37. Idem. Tout ce qui est écrit ci-dessus s'appuie largement sur ce tutoriel.

38. Ibidem.

FIGURE 9.11 – Détails sur le jeu de données



modèle se comporte sur les pages où il n'a pas été formé. Un taux de 6 % est plutôt bon.

Si l'on récapitule la progression de l'entraînement du modèle, elle s'est avérée être riche de promesses.

- Ainsi, en partant à 1849 mots, on obtenait :

CER ON TRAIN SET : 16,31 %

CER ON VALIDATION SET : 26,25 %

- Puis, avec 1 951mots :

CER ON TRAIN SET : 0,36 %

CER ON VALIDATION SET : 27,10 %

- Avec 1 977 mots :

CER ON TRAIN SET : 0,29 %

CER ON VALIDATION SET : 12,57 %

- Avec 4 207 mots :

CER ON TRAIN SET : 0,86 %

CER ON VALIDATION SET : 11,55 %

- Avec 15 928 mots :

CER ON TRAIN SET : 2,58 %

CER ON VALIDATION SET : 6,40 %

- Avec 18 531 mots :

CER ON TRAIN SET : 2,91 %

CER ON VALIDATION SET : 7,83 % (même set de validation que le précédent)

- Avec 23 729 mots :

CER ON TRAIN SET : 3,52 %

CER ON VALIDATION SET : 6,09 % (même set de validation que le précédent)

Cependant, certains de ces chiffres sont biaisés car nous n'avons pas toujours utilisé le même set de validation, excepté pour les trois derniers entraînements du modèle.

Par ailleurs, certains chiffres peuvent laisser dubitatifs car ils sont fluctuants. Cela est probablement dû à la part d'aléatoire du *machine learning* en entraînement et peut-être aussi à cause de certaines images d'une qualité insuffisante.

Nous étions plutôt confiants quant aux chiffres indiquant la progression du modèle. En effet, si on traduit le CER en taux de réussite, nous sommes près d'un taux de réussite de 95 % ce qui est très positif.

Néanmoins, 5 % de CER signifie tout de même une faute tous les 20 caractères, soit, pour une moyenne de 4 lettres par mot, 1 faute tous les 5 mots au moins.

Ceci étant établi, il est temps de voir si la réalité que cachent ces chiffres sera aussi positive qu'on l'espère : il s'agit désormais d'appliquer le dernier modèle pour que la machine opère elle-même les transcriptions. C'est l'aboutissement de notre travail.

## 9.5 Application du modèle

Ayant trouvé la progression de l'entraînement du modèle plutôt bonne quant aux chiffres, nous étions assez confiante pour l'appliquer sur des manuscrits non transcrits et voir le résultat du travail d'une petite dizaine de jours.

Nous avons donc, à cette fin, mis de côté des manuscrits venant de fonds divers et variés représentant huit pages.

Pour appliquer un modèle et donc effectuer une transcription sur un manuscrit, il faut auparavant importer les manuscrits en question sur Transkribus. Puis, il faut se rendre dans l'onglet *Tools* où se trouvent les outils, aller dans la partie *Text recognition* et cliquer sur « Run ». Une fenêtre s'ouvre alors et l'on peut choisir les modalités de transcription et sélectionner le modèle dont on se sert pour transcrire : à chaque écriture correspond en effet un modèle adapté. Pour notre part, ayant entraîné un modèle pour l'écriture de Le Play uniquement, en toute logique, nous n'avons chargé que des manuscrits ne comportant que l'écriture de Frédéric Le Play.

Ceci fait, nous sommes allée regarder quels étaient les résultats des transcriptions automatiques. Or, pour la première page, il n'y a pas un mot sans faute, à part le « Monsieur » et la signature de Le Play. Nous avons pensé que cela était peut-être dû à la mauvaise qualité de la numérisation d'une part, et d'autre part peut-être du fait que nous n'avions entraîné pour le modèle aucune lettre de ce fonds<sup>39</sup>.

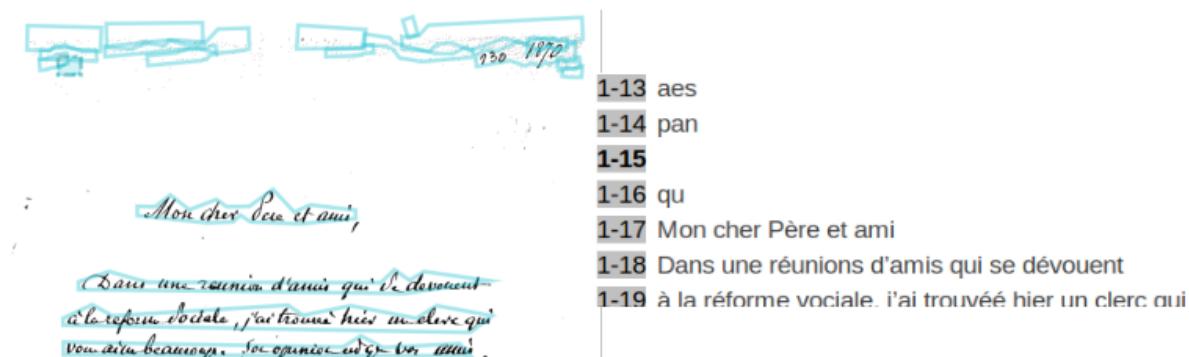
---

39. La première page transcrise automatiquement provient des archives départementales (AD) de Haute-Savoie, c'est une lettre de Le Play à Joseph Despine.

Une page est tirée du fonds de Frédéric de Mercey à la BNF. La qualité de la numérisation est donc plutôt bonne. Cependant, même si le résultat est meilleur que la précédente lettre, il reste encore beaucoup de fautes, et nous nous sommes demandée si le fait de les corriger ne prendrait pas plus de temps que de tout transcrire en partant de rien. Là encore, nous n'avions pas transcrit de lettres de ce fonds pour l'entraînement. Par ailleurs, l'écriture de Le Play varie selon son âge, et nous n'avons peut-être pas fourni de données d'entraînement comprenant toutes les périodes de la vie de Le Play. En effet, ici, l'écriture est plus fine, plus penchée, plus jeune, ce qui a peut-être posé problème pour la reconnaissance.

Une page est tirée de la bibliothèque publique universitaire de Genève : c'est une lettre de Frédéric Le Play au père Hyacinthe Loyson. Nous avons entraîné près de 90 pages de ce fonds. Nous avons trouvé malgré cela bon nombre de fautes. D'autre part, la numérisation comportant des nuances de gris assez foncées, la machine l'a pris pour de l'écriture : cela demandera beaucoup de temps de corriger ces fautes car la machine matche à chaque fois une ligne qui n'en est pas une et croit y reconnaître de l'écriture. Il faut donc tout effacer et tout corriger.

FIGURE 9.12 – Lettre de F. Le Play au R. P. Loyson, 1870



Deux pages sont tirées du fonds conservé au Château de Ligoure, ce sont des lettres de Le Play à son fils Albert. La transcription est assez propre, mais subsistent encore bon nombre de fautes. Le Play forme parfois mal ses lettres, ce qui ne facilite pas la tâche, et pour l'œil humain, et pour la machine, comme nous l'avions déjà fait remarquer plus haut, ce qui d'ailleurs est aussi visible sur la figure 9.12. Par ailleurs, même si les mots sont reconnus, les accents sont souvent absents. (Par exemple, il est écrit senat pour sénat). Par ailleurs, Le Play a tendance à gommer les différences entre les majuscules et les minuscules. Souvent, ses « s » minuscules semblent être des « s » majuscules. En revanche, le « A » de « Albert » pourrait être une minuscule, d'où de nombreuses majuscules prises pour des minuscules.

Pour une lettre de Le Play à Alfred Tylor, le résultat est plutôt satisfaisant. La

machine bute face à certains obstacles : les ratures.

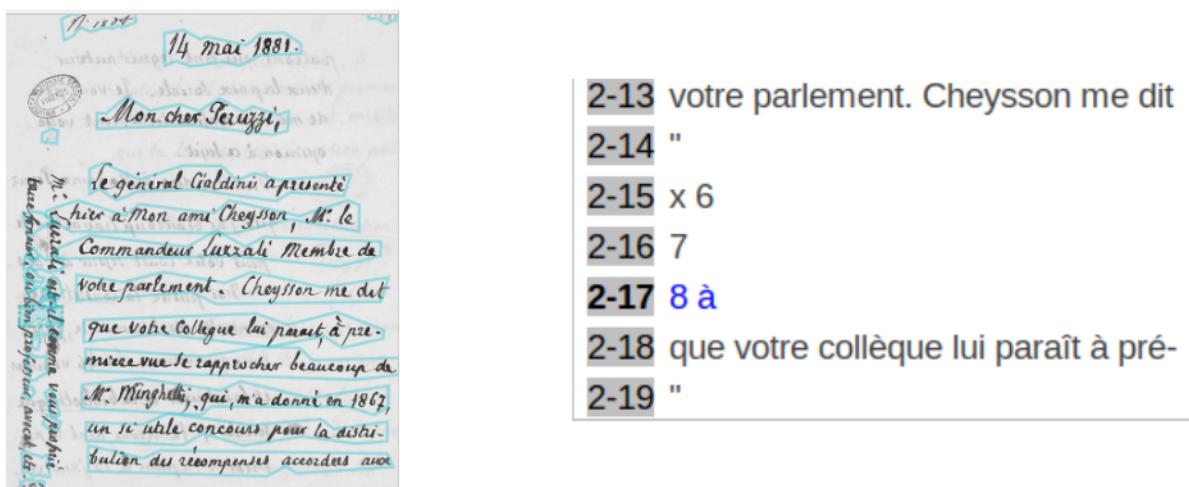
FIGURE 9.13 – Lettre de F. Le Play à Keele



Cependant, le résultat est plutôt bon dans l'ensemble, et la bonne qualité de la numérisation joue probablement un rôle dans ce sens.

Pour une lettre de Le Play à Peruzzi<sup>40</sup>, le résultat est plus que médiocre. Une des raisons de cet échec est la mauvaise délimitation des lignes : il y a deux sens d'écriture et la machine ne sait le reconnaître. Corriger ce genre d'erreurs prend plus de temps que de transcrire directement sans passer par la machine. Dans ce cas, il vaut mieux supprimer les lignes en question et transcrire soit-même.

FIGURE 9.14 – Lettre de F. Le Play à Peruzzi, 1881



Au terme de l'analyse de ces premières transcriptions réalisées avec notre modèle, nous avons été un peu surprise car nous pensions que les pourcentages annoncés promettaient moins de fautes.

Une question se pose : faut-il continuer à entraîner le modèle pour arriver à un résultat plus satisfaisant ou faut-il abandonner l'expérience ?

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

40. Biblioteca Nazionale Centrale (Florence)

d'avoir toutes les numérisations en main) 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 pourrait prendre autant de temps que de bien transcrire directement et à la main.

Quoi qu'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, ayant le manuscrit visible sur la même page. On y trouve donc nombre d'avantages, même si l'on ne pousse pas la machine jusqu'au bout de ses capacités. Transkribus est un excellent outil de transcription collaborative qui permet d'avoir l'image et sa transcription en vis-à-vis ce qui est très confortable pour transcrire, même manuellement.

Par ailleurs 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 sont également des lettres passives. Certaines sont nombreuses, comme celles par exemple de Napoléon-Joseph Bonaparte<sup>41</sup>, mais d'autres sont rares, comme celles de Louis-Joseph Buffet<sup>42</sup>.

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.

## 9.6 Rester maître de ses données

### 9.6.1 Exportation en vue de l'édition

Une fois que les transcriptions sont réalisées, il s'agit de les exporter. En effet, si elles sont toutes conservées dans le serveur de Transkribus, cela ne signifie pas que nous les avons dans nos serveurs propres. Il s'agit donc d'exporter les données pour les conserver au CRHXIX et rester maître des données : nous ne savons pas comment Transkribus va évoluer, si l'outil deviendra payant ou ne sera plus maintenu, il est donc important d'assurer ses données.

Il faudrait penser également à la question du stockage des données, mais nous n'avons pas eu à la traiter durant notre stage, cela se fera par la suite.

---

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

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

Une question surtout se pose : comment exporter les données ? Plusieurs formats sont disponibles : « le package complet comprenant les fichiers image d'origine, les fichiers XML et les métadonnées ajoutés au document, un PDF (avec le texte transcrit inclus), doc et TEI. Tous les fichiers sont stockés dans un dossier<sup>43</sup> ».

En vue de notre édition numérique de correspondance, nous avons exploré les différents formats d'exportation possibles. Les formats TXT et TEI s'avèrent être particulièrement intéressants. Lequel des deux privilégier ? Pour cela, il s'agit de considérer de plus près la question des *tags* dans Transkribus.

### 9.6.2 La question des *tags*

Dans Transkribus, on peut déjà caractériser certains mots grâce à des *tags*. Nous avons pensé dans un premier temps utiliser les *tags*<sup>44</sup> de Transkribus, afin d'annoter en direct nos données.

De nombreux *tags* sont disponibles dans Transkribus<sup>45</sup>. Nous nous sommes donc livrée à cette expérience et avons constaté que l'annotation des *tags* dans Transkribus prenait beaucoup de temps. L'exemple de la figure 9.15 ci-dessous illustre les 21 *tags* que nous avons mis pour une seule page de manuscrit.

En l'occurrence, sur cette page, nous avons mentionné la date (*date*), les abréviations (*abbrev*), les exposants (*textStyle*), les noms de lieu (*place*), les fautes dans l'original (*sic*), les noms d'organisation (*organization*), les noms d'ouvrage (*work*), les noms de personne (*person*).

L'avantage de renseigner en amont ces informations permet, lors de l'export, de les retrouver directement dans nos fichiers en XML-TEI. Cependant, nous n'avons pas opté pour ce choix pour plusieurs raisons :

- Nous avons été mise en garde sur l'export en TEI qui n'est pas encore tout à fait au point dans Transkribus
- Nous avons pris connaissance trop tard du guide d'utilisation des *tags* dans Transkribus et de la possibilité de personnaliser ses propres *tags*<sup>46</sup>.

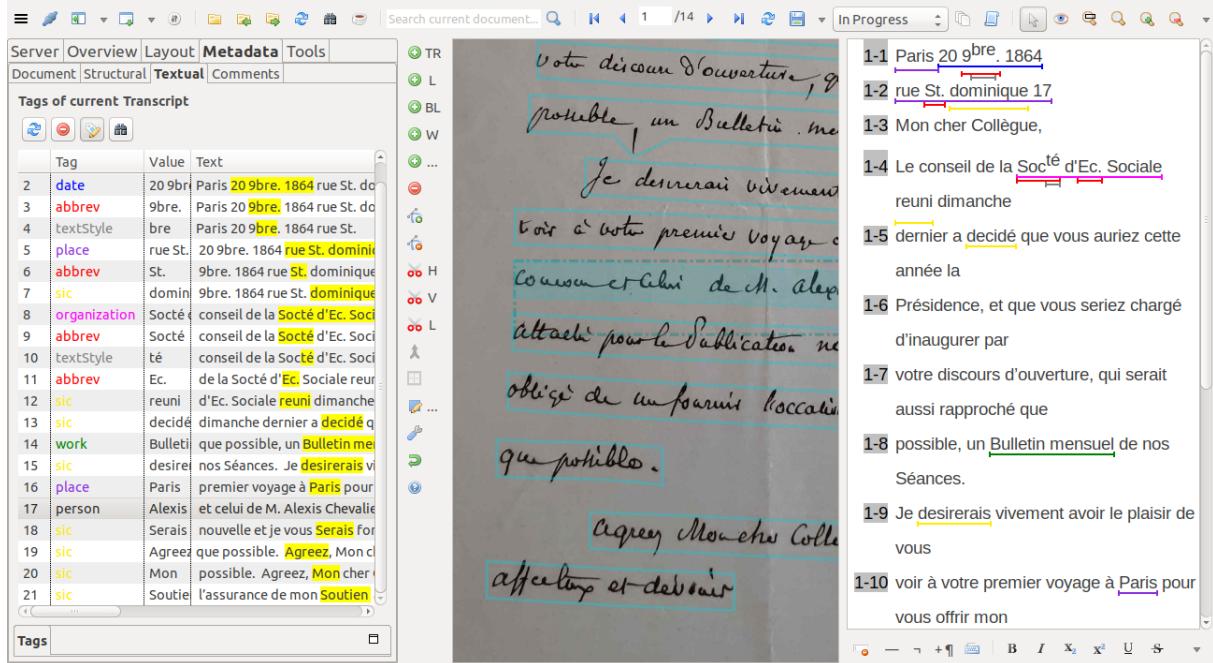
43. Régis Schlagdenhauffen, *Comment utiliser Transkribus en 10 étapes (voire moins)*, Site web de l'EHESS, URL : <http://regis-schlagdenhauffen.eu/wp-content/uploads/2018/01/Comment-utiliser-Transkribus-%E2%80%93-en-10-%C3%A9tapes-ou-moins.pdf> (visité le 22/05/2020.).

44. « Un tag (ou étiquette, marqueur, libellé) est un mot-clé (signifiant) ou terme associé ou assigné à de l'information [...], qui décrit une caractéristique de l'objet et permet un regroupement facile des informations contenant les mêmes mots-clés ». Voir *Tag (métadonnée)*, Wikipédia, URL : [https://fr.wikipedia.org/wiki/Tag\\_\(m%C3%A9tadonn%C3%A9e\)](https://fr.wikipedia.org/wiki/Tag_(m%C3%A9tadonn%C3%A9e)) (visité le 26/09/2020).

45. Voir les annexes, B.3 : Schématisation du modèle d'information de Transkribus.

46. *How to enrich transcribed documents with mark-up*, [https://transkribus.eu/wiki/images/e/e8/How\\_to\\_enrich\\_transcribed\\_documents\\_with\\_mark-up.pdf](https://transkribus.eu/wiki/images/e/e8/How_to_enrich_transcribed_documents_with_mark-up.pdf) (visité le 26/09/2020).

FIGURE 9.15 – Pour une page, 21 tags



- Nous aurions de toutes façons manqué de temps pour penser les *tags* dans Transkribus à long terme
- L'export des *tags* déjà existant s'est avéré être non conforme pour la TEI et nous avons dû corriger nombre de noms de balises. Nous aurions pu passer par une feuille de transformation (XSLT : *eXtensible Stylesheet Language Transformations*) permettant de passer du fichier TEI exporté à un fichier TEI comportant les noms de balises que nous voulons mettre, mais ceci nécessite une bonne maîtrise de la technologie et nous pensons qu'il est préférable d'apprendre d'abord à l'équipe du CRHXIX à maîtriser XML au lieu de passer directement à la maîtrise d'XSLT.
- Enfin et surtout, nous avons préféré penser le balisage directement dans XML-TEI.

Quoi qu'il en soit, nous n'excluons pas pour le CRHXIX d'utiliser l'interface graphique de Transkribus pour les *tags*. Il suffira à la personne qui prendra la suite de la partie numérique du projet de penser la chose. Nous avons de notre côté pensé surtout à l'enchaînement des balises.

L'export en TEI est tout de même intéressant pour ce qui est de la structure du texte (une balise délimite les lignes ce qui est intéressant). On peut également opter pour un export en TXT étant donné que la délimitation des lignes est également respectée.

Nous voyons que la phase d'acquisition des données s'est avérée riche en questionnements. L'apprentissage machine a eu une large part dans cette partie de nos stages, aussi

bien pour l'OCR de Gallica que pour l'HTR de Transkribus. Que ce soit pour l'OCR ou l'HTR, nous n'en sommes pas encore à du 100 % quant au résultat, néanmoins, ces technologies nous ont été d'un grand secours. La part de relecture reste tout de même importante.

Une fois les données acquises, il s'agit de les traiter en vue de leur valorisation et donc de leur mise en ligne. Qu'en est-il du traitement des données pour l'édition numérique de correspondance ? C'est l'objet de notre dernière partie.



## Quatrième partie

Traiter les données pour la réalisation  
des projets



# Chapitre 10

## Des standards et technologies au service de l'édition numérique

Les données brutes, bien qu'intéressantes, ne nous permettent pas d'aller bien loin. Une fois que nous avons les transcriptions des correspondances à éditer, il s'agit donc de les traiter pour en assurer la mise en ligne, ce qui est tout l'objet de nos deux projets. Or, que ce soit pour le projet du CRHXIX d'édition numérique de la correspondance de Frédéric Le Play ou pour celui du Labex OBVIL, ELICOM, nous avons eu recours à XML, plus spécifiquement XML-TEI. Ce langage s'avère être en effet particulièrement adapté.

### 10.1 XML, un langage particulièrement approprié

Comme nous l'avions déjà souligné dans notre deuxième partie, lors de nos réflexions sur l'édition numérique de correspondance<sup>1</sup>, XML est un langage structuré, à la fois lisible par l'œil humain et par la machine.

#### 10.1.1 XML : présentation générale

##### 10.1.1.1 Pourquoi encoder ?

Le balisage sémantique permet d'expliciter certains aspects du texte<sup>2</sup>. L'encodage est très utile pour l'édition car il permet de montrer la structure du texte, ce qu'il est, et non seulement son aspect, sa présentation. Il permet également de séparer le texte et le contenu de l'édition, et donc il permet différents modes d'affichage (notes, transcription normalisée ou non etc.).

---

1. Voir 4.1.2.3 XML

2. Voir aussi Elena Pierazzo, *Why do we encode?* URL : [https://www.youtube.com/watch?v=R0ncI\\_rr1z4&list=PL77mHK9JuenN9NXeXQbVcU0Rz7HZk-9Pv&index=2](https://www.youtube.com/watch?v=R0ncI_rr1z4&list=PL77mHK9JuenN9NXeXQbVcU0Rz7HZk-9Pv&index=2) (visité le 04/05/2020). Ce paragraphe s'en inspire.

Le balisage (*markup*) et l'encodage (*encoding*) n'ont pas été créés avec l'ordinateur, c'est un concept né avec l'imprimerie : en effet, les rédacteurs et éditeurs indiquaient par des symboles comment ils voulaient que le texte apparaisse, quelle taille de caractère, en gras ou en italiques par exemple. De même, la ponctuation et la mise en page sont aussi une sorte de balisage car elles permettent de comprendre comment diviser le texte en segments. Finalement, le balisage, c'est tout ce que nous utilisons pour rendre le texte plus lisible : l'espacement des lettres, les majuscules, le gras, l'italique, la ponctuation. Par ailleurs, encoder est un moyen de veiller à la pérennité des données, à leur conservation sur le long terme.

#### 10.1.1.2 Comment encoder ?

Le langage XML est particulièrement approprié pour nos éditions numériques de correspondance. Il a été créé par le W3C en 1996, avec une contribution importante de la communauté des Humanités Numériques et en particulier de la TEI, comme nous l'avons déjà souligné plus haut.

« Les données sont incluses dans le document XML sous forme de chaînes de caractères délimitées par un balisage les décrivant. L'unité de base qui comprend données et balisage est appelée élément »<sup>3</sup>.

Un élément est tout ce qui peut être étiqueté, tout ce qui se décrit, en quelque sorte, tout ce qui est important et devrait être annoté dans le texte.

L'élément a une balise d'ouverture qui se présente avec un chevron (<), puis le nom du *tag*, le contenu de l'élément puis la balise de fermeture qui a une barre oblique (/) :

```
<nomElement>chaineCaracteres</nomElement>
```

Certains éléments peuvent être vides, dans ce cas on met une barre oblique à la fin :

```
<nomElement/>
```

Les éléments suivent un ordre particulier : soit ils peuvent se suivre les uns après les autres, soit ils s'imbriquent, et dans ce cas, les éléments *enfants* héritent des propriétés des éléments *parents*. Jamais ils ne doivent se chevaucher.

Il faut aussi un élément qui contienne le fichier entier, on l'appelle « élément racine ».

Les éléments peuvent avoir des attributs, qui servent à préciser quelque chose de l'élément. L'attribut a une syntaxe particulière à respecter : après le nom de l'élément, il y a un espace blanc obligatoire, le nom de l'attribut suivi par un signe égal (=) et la valeur de l'attribut est entre guillemets. On peut avoir autant d'attributs qu'on veut dans l'élément mais le même attribut ne peut être utilisé qu'une seule fois<sup>4</sup>.

Un encodage qui respecte ces principes est dit bien formé.

---

3. Voir Ariane Pinche, *Séance 1*, Cours M2 TNAH XML, URL : [https://github.com/ArianePinche/coursTNAH\\_XML-TEI/blob/master/seance01/InitiationXML.md](https://github.com/ArianePinche/coursTNAH_XML-TEI/blob/master/seance01/InitiationXML.md) (visité le 09/10/2020).

4. Pour aller plus loin, voir Ariane Pinche, *ibidem* et les séances suivantes

Par ailleurs, on peut distinguer trois types de balisages<sup>5</sup> :

- Le balisage physique, formel et typographique encode les structures logique et formelle du texte
- Le balisage sémantique autorise un encodage d'un premier niveau d'analyse comme les noms, les toponymes, les citations, les discours et les dates
- Le balisage analytique complète l'apparat savant de l'encodage.

### 10.1.2 Les métadonnées

Tout d'abord, pour bien gérer les données, il est important de les décrire au fur et à mesure. Ainsi, « préalablement à tout encodage, il est nécessaire de donner des informations descriptives et déclaratives sur le texte en vue de sa diffusion, d'un échange de données ou encore de son exploitation : ces composantes – ou métadonnées – forment l'en-tête TEI nommé <teiHeader><sup>6</sup> »

On entend par métadonnées

« les données servant à représenter ou à décrire d'autres données, ici les deux textes édités. Elles contiennent des informations sur la source d'un document, sa nature, son contenu, sa localisation, son histoire, son statut juridique, etc. Elles remplissent un rôle d'indexation qui facilite l'accès au contenu et améliore la recherche. Elles doivent être normalisées grâce à des standards comme Dublin Core, EAD (Encoded Archival Description) par exemple afin de permettre l'interopérabilité et l'échange des données<sup>7</sup>. »

Les métadonnées répondent donc à des standards. Or, la correspondance a ses standards propres : nous avons parlé plus haut du <correspDesc> pensé par le *TEI SIG Correspondence*. Pour nos deux projets, nous avons suivi ces directives.

Par ailleurs, le projet ELICOM nous a éclairé dans le choix des métadonnées pour notre projet du CRHXIX, et nous retrouvons à peu près les mêmes informations, même si nous avons un peu plus détaillé le <teiHeader> pour le CRHXIX, avec notamment l'ajout de balises pour suivre les différentes mises à jour des fichiers XML-TEI.

Les métadonnées se trouvent donc dans le <teiHeader> et se subdivisent en un certain nombre de balises.

---

5. Nicole Dufournaud, Valérie Gratsac Legendre. *Manuel d'encodage XML-TEI - édition numérique de manuscrits baroques/ : Recommandations pour une application TEI*, 2012, Site web HAL, URL : <https://hal.archives-ouvertes.fr/hal-00718043/document> (visité le 28/07/2020).

6. Nicole Dufournaud, Valérie Gratsac Legendre, in *Ibidem.*, p.7.

7. *Ibidem*.

### 10.1.2.1 La description bibliographique du document avec le `<fileDesc>`

Tout d'abord, le `<fileDesc>` contient la description bibliographique du document. Pour ELICOM, il est réduit à deux éléments :

- le `<titleStmt>`<sup>8</sup> : il regroupe les informations sur le titre d'une œuvre et les personnes ou institutions responsables de son contenu intellectuel. C'est ici que sont nommés les principaux responsables du projet pour le CRHXIX. On indique les autres membres de l'équipe et ceux qui ont participé de près ou de loin au projet. Pour chaque personne, on utilise la balise `<respStmt>` qui englobe la balise `<resp>`, indiquant la responsabilité qu'a eue la personne en question, la balise `<name>` qui englobe le prénom puis le nom, et une balise `<note>` facultative, si jamais l'on a des remarques à faire. La balise `<note>` peut aussi comprendre un attribut `@resp` indiquant qui a rédigé la note.
- le `<publicationStmt>` : il informe sur la façon dont le projet est distribué ou publié. C'est donc ici qu'est renseignée la licence. Pour l'instant, nous n'avons pas encore déterminé pour le CRHXIX laquelle nous choisissons.

Pour le CRHXIX, étant donné que nous travaillons sur des fac-similés, nous y avons en plus ajouté le `<sourceDesc>`, élément capital pour l'édition numérique de correspondance. En effet, le `<sourceDesc>`, qui est une description de la source, contient le `<msDesc>` qui englobe les balises de description du manuscrit, à savoir la balise `<msIdentifier>` qui elle-même contient la balise `<country>`. Celle-ci indique le pays d'origine et peut comporter un attribut `@key` qui est facultatif. La balise `<settlement>` indique le lieu (ville ou autre), puis on renseigne l' `<institution>` qui abrite le manuscrit. La balise `<repository>` contient le nom d'un dépôt dans lequel des manuscrits sont entreposés, et qui peut faire partie d'une institution, et l'`<idno>` (identifiant) donne un numéro normalisé ou non qui peut être utilisé pour identifier une référence bibliographique.

Pour l'instant, nous ne voyons pas trop l'utilité de la balise `<msContents>` pour le projet du CRHXIX. Elle renseigne le contenu du manuscrit, décrivant le contenu intellectuel d'un manuscrit ou d'une partie d'un manuscrit, soit en une série de paragraphes `<p>`, soit sous la forme d'une série d'éléments structurés `<msItem>` concernant les items du manuscrit. Néanmoins, je laisse la possibilité de l'utiliser, mais on pourrait la supprimer par la suite.

En revanche, la balise `<physDesc>` pourra s'avérer utile. Elle contient la balise `<handDesc>` qui mentionne dans le `<handNote>` quelle main a écrit le manuscrit. Or, si Le Play écrit de sa main la plupart du temps, ce n'est pas toujours le cas, notamment quand il est malade. Pour les autres correspondants, le temps nous a manqué pour nous pencher

---

8. Pour le CRHXIX, nous avons ajouté quelques éléments à l'intérieur du `<titleStmt>` et du `<publicationStmt>`, le détail est indiqué dans l'ODD disponible dans les livrables, voir 1.4.2.1.1. Le `titleStmt`.

sur la question, néanmoins, il sera toujours intéressant pour le lecteur non spécialiste et qui ne connaît donc pas les écritures, d'être informé du rédacteur du manuscrit.

### 10.1.2.2 La description détaillée des aspects non bibliographiques avec le `<profileDesc>`

Après la description bibliographique du document dans le `<fileDesc>`, on trouve une description détaillée des aspects non bibliographiques dans le `<profileDesc>`.

Un des éléments clés qui nous intéresse ici, et qui a été mentionné dans nos deux projets, est le `<correspDesc>`. Il comprend les informations de description de l'action liée à la correspondance. Celui-ci comprend principalement la balise `<correspAction>`. Celle-ci se décline en deux temps, selon ses attributs.

- Tout d'abord, une balise `<correspAction>` avec un attribut `type` comprenant la valeur `sent` renseigne qui a envoyé la lettre. La balise `<persName>` comprend donc le prénom puis le nom de l'expéditeur (la plupart du temps il s'agit de Le Play). Elle comprend un attribut `@key` qui décrit l'expéditeur de façon normalisée et un attribut `@ref` qui renvoie vers sa fiche data.bnf. La balise `<settlement>` renseigne le lieu de rédaction de la lettre, et la balise `<date>` indique la date à laquelle la lettre a été écrite. L'attribut `@when` permet de la normaliser au format AAAA-MM-JJ.
- Puis on retrouve une deuxième balise `<correspAction>` avec un attribut `@type` comprenant cette fois la valeur `received` renseignant à qui la lettre a été envoyée. La balise `<persName>` comprend donc le prénom puis le nom du destinataire. Elle comprend également un attribut `@key` qui décrit l'expéditeur de façon normalisée et un attribut `@ref` qui renvoie vers sa fiche data.bnf. Il y a possibilité d'ajouter une balise `<settlement>` renseignant le lieu de réception de la lettre, autrement dit le lieu où se trouve le destinataire, et la balise `<date>` indiquant la date de réception. L'attribut `@when` permet de la normaliser au format AAAA-MM-JJ. Cependant, il est bien clair que ces deux dernières informations ne seront quasiment jamais renseignées<sup>9</sup>.

Ces choix du `<correspAction>` qui avaient été faits pour ELICOM nous ont donc inspiré pour le projet du CRHXIX.

Toujours dans le `<correspDesc>` l'on renseigne le `<correspContext>`, à savoir le contexte dans lequel se situe la lettre, quelle lettre la précède, quelle lettre la suit. On renseigne ces deux informations dans une balise `<ref>` avec un attribut `@type` de valeur `previous` pour la lettre précédente, de valeur `next` pour la suivante, et on le fait suivre d'un attribut `@target` qui fait pointer vers le fichier XML en question.

---

9. Tout au moins c'est l'opinion que l'on peut avoir aujourd'hui, peut-être variera-t-elle dans le temps.

Dans le `<profileDesc>` se trouve également le `<settingDesc>` et le `<particDesc>`. Nous y reviendrons lorsque nous parlerons des index.

### 10.1.2.3 L'historique du fichier avec le `<revisionDesc>`

Le dernier élément du `<teiHeader>` est le `<revisionDesc>` qui résume l'historique des révisions pour un fichier.

Chaque changement est mentionné sous la balise `<change>`. Autrement dit, à chaque modification importante du fichier XML, on renseigne quelle modification a été faite (`<change>`), à quelle date avec l'attribut `@when`, par qui avec l'attribut `@who`.

## 10.1.3 XML et le rituel épistolaire

### 10.1.3.1 Remise en contexte

Avec XML, nous pourrions en soi encoder le monde entier, avec l'inconvénient sûr de s'y perdre. Il est donc important que le schéma de balise que nous serons amenée à choisir pour nos éditions numériques de correspondance soit le plus restreint possible.

Par ailleurs, nous avons choisi de répondre aux conseils établis par la communauté scientifique et le TEI P5<sup>10</sup>. Grâce aux *TEI guidelines*, nous sommes face à des noms d'éléments définis. Il s'agit donc pour nous de suivre les directives du *TEI : Correspondence SIG*. Il n'y a pas de place ici pour la fantaisie. Tout est bien normé. Il suffit simplement de définir quelles balises nous choisissons, et le cas échéant, dans quel ordre nous voulons qu'elles apparaissent.

Or, il faut que notre balisage rende compte des caractéristiques et de la structure propre à la correspondance. Comme nous l'avons vu plus haut, il y a ce qu'on appelle le « rituel épistolaire<sup>11</sup> ». Voici donc comment il peut se traduire lorsqu'on encode :

- Le lieu de rédaction et la date apparaîtront en XML-TEI par la ligne `<dateline>` qui comprend les balises `<date>` pour la date et `<place>` pour le lieu
- Une « adresse » ou formule de politesse débutant la lettre, se traduira en balises par `<salute>`
- Une formule de politesse finale sera également transcrive dans un `<salute>`
- Une signature : en général, F. Le Play pour Frédéric Le Play. Si elle n'y est pas, on indiquera qu'elle est manquante, ceci dans la balise `<signed>`
- Éventuellement un post-scriptum qui trouvera sa place dans une balise `<postscript>`

10. Voir la partie 2 sur penser l'édition et *TEI : P5 Guideline*, TEI guidelines, URL :<https://tei-c.org/guidelines/p5/> (visité le 26/09/2020).

11. Richard Walter (dir.), *L'édition numérique de correspondances – guide méthodologique*, URL :<https://cahier.hypotheses.org/guide-correspondance> (visité le 17/06/2020). Voir particulièrement la page 13.

Ainsi, la structure de la lettre est bien mise en évidence. L'encodage répond à nos attentes. Avant de développer le rituel épistolaire, remettons-le dans son contexte.

### 10.1.3.2 Le <body>

Cette structure de la lettre apparaît donc après les métadonnées du `<teiHeader>`, dans la balise `<text>` qui comprend en général trois éléments : le `<front>` pour tous les éléments liminaires, le `<body>` pour le corps du texte proprement dit, le `<back>` pour tous les appendices, épilogues et postfaces. Pour nos éditions numériques de correspondance, nous développerons surtout la balise `<body>`.

Celle-ci comprend une division (`<div>`) avec un attribut `@type` dont la valeur est `lettre`. Si la lettre a été rédigée en deux fois, on peut donc utiliser plusieurs `<div>`, avec un attribut `@part` dont la valeur sera `I` pour le début ou partie initiale, `F` pour la partie finale, `M` pour le milieu si nécessaire. Cela permettra également d'ajouter une `<dateline>` pour les autres parties.

La première balise ensuite à mettre dans la `<div>` - uniquement pour l'édition numérique de la correspondance de Frédéric Le Play cette fois, car cela concerne les fac-similés - est la balise `<pb>`, qui permet de relier le fichier XML en question avec le manuscrit numérisé. On utilise tout d'abord un attribut `@n` pour renseigner le numéro du manuscrit, cela va de 1 à l'infini, et on peut y ajouter en plus une lettre en minuscule quand la lettre fait plusieurs pages et donc que le même fichier XML doit être mis en lien avec plusieurs images numérisées. Ainsi, la première page de la lettre 2 est numérotée 2a, la deuxième page 2b, la troisième page 2c et ainsi de suite.

L'attribut `@facs` pour les fac-similés pointe directement vers une image ou vers une partie d'une image correspondant au contenu de l'élément. On y met comme valeur le nom du fichier image, comme indiqué sur la figure 10.1.

FIGURE 10.1 – Capture d'écran de l'ODD du CRHXIX, le `<pb>`

```
<pb n="2a" facets="ChdeRibbe02A.png"/>
```

Dès que l'on change de page de manuscrit au sein du même fichier XML donc de la même lettre, il faut donc à nouveau mettre la balise `<pb>` pour redonner toutes ces informations, entre la fin de la page précédente et le début de la nouvelle page.

### 10.1.3.3 Le début de la lettre

La balise `<opener>` souligne le rituel épistolaire. Elle abrite tous les éléments qui font débuter une lettre, à savoir l'élément `<dateline>` pour la ligne mentionnant le lieu

(<placeName> avec attribut @ref, dont la valeur comprend un #, pour pointer vers l'index) et la date (<date>) de rédaction de la lettre<sup>12</sup>.

Le corps de la lettre nécessite parfois l'emploi d'autres balises pour traiter les cas particuliers<sup>13</sup>. C'est le cas des abréviations. Afin de garder les deux versions<sup>14</sup>, la version originale et la version normalisée, on utilise une balise générale <choice> qui abrite deux balises : la première <abbr> garde l'abréviation originale de l'auteur. La deuxième <expan> donne le mot entier.

Une procédure assez similaire se fait pour la correction des fautes ou la normalisation des lettres. On utilise alors une balise <choice> qui englobe une balise <sic> qui contient le texte original, et une autre <corr> qui contient le texte corrigé.

FIGURE 10.2 – Capture d'écran de l'ODD du CRHXIX, l'<opener>

```
<opener>
<dateline>
  <placeName ref="#paris">Paris</placeName>
  <date>20 <choice>
    <!-- Développement de l'abréviation sur le mois -->
    <abbr>9<hi rend="sup">bre</hi>.</abbr>
    <expan>septembre</expan>
  </choice> 1864</date>
  <placeName ref="#stDom"> rue <choice>
    <sic>St.
    </sic>
    <corr>Saint-</corr>
  </choice>
  <choice>
    <sic>dominique</sic>
    <corr>Dominique</corr>
  </choice> 17 </placeName>
</dateline>
<salute>Mon cher <choice>
  <!--On retire les majuscules -->
  <sic>Collègue</sic>
  <corr>collègue</corr>
</choice>, </salute>
</opener>
```

A cela s'ajoute la question du style. Certaines lettres sont en italiques, soulignées etc. Pour manifester ces différents styles, on utilise la balise <hi> avec son attribut @rend qui spécifie la nature du style employé :

- "italic" pour les italiques
- "bold" si c'est du gras
- "upper" pour la mise en capitales du texte sélectionné
- "small-caps" pour la mise en petites capitales
- "sup" pour la mise en exposant

12. Il faut encore décider si on normalise aussi à cet endroit la date avec un attribut @when ou si c'est inutile étant donné que cela a déjà été fait dans le <correspDesc> du <teiHeader>.

13. Encore une fois, ce paragraphe ne concerne que le CRHXIX, étant donné que nous avons affaire à des manuscrits qui n'ont jamais été publiés. La question ne se pose pas pour ELICOM qui travaille sur des éditions imprimées.

14. La question n'a pas encore été tranchée pour le CRHXIX.

- "ul" pour souligner
- "line-through" pour barrer.

La balise `<dateline>` renseignée, il s'agit ensuite de mettre dans la balise `<salute>` la salutation qui est faite au début de la lettre. Cela fait partie du rituel épistolaire.

Ceci fait, on passe au corps de la lettre.

#### 10.1.3.4 Le corps de la lettre

Le corps de la lettre est contenu dans des balises `<p>`. Autrement dit, chaque paragraphe est contenu dans une balise `<p>`. Chaque ligne de ce paragraphe est contenu dans une balise `<l>`.

Il faut ensuite gérer les cas particuliers. Parfois, l'auteur de la lettre écrit lui-même des notes dans les marges<sup>15</sup>. Dans ce cas-là, on peut utiliser une balise `<note>`. On utilise un attribut `@resp` pour indiquer que la note a été rédigée par l'auteur de la lettre lui-même, et un attribut `@place` pour indiquer où la note se trouve dans l'original.

La valeur de l'attribut `@place` peut varier. Retenons quelques exemples :

- "above" : au-dessus de la ligne
- "below" : au-dessous de la ligne
- "bottom" : en bas de page
- "inline" : dans le corps du texte
- "top" : en haut de page

FIGURE 10.3 – Capture d'écran de l'ODD du CRHXIX, l'attribut `@place`

```
<p>
<l>Regrettant beaucoup d'avoir</l>
<l>manqué l'occasion de vous voir, samedi</l>
<l>soir, nous vous prions de nous faire l'amitié </l>
<l>de venir diner jeudi *<note resp="author" place="margin-left">*24
juin</note> à 6<hi rend="sup">h</hi>1/2 <hi rend="ul">avec MM.</hi>
</l>
<l>
<hi rend="ul">
<persName ref="#rapetti">Rapetti</persName> et <persName ref="#coquille">Coquille</persName>. </hi>
</l>
</p>
<lb/>
```

Pour le corps de la lettre, nous avons été face à plusieurs questionnements pour ELICOM. En effet, les fichiers XML qui avaient été extraits de l'HTML<sup>16</sup> comportaient des fautes d'OCR ou d'autres fautes de diverses nature. Par exemple, dans un fichier XML de la correspondance de Félicité de Lamennais, deux fautes sont visibles sur la même page<sup>17</sup> : on remarque d'une part un caractère « i » indiqué à la place d'une marque

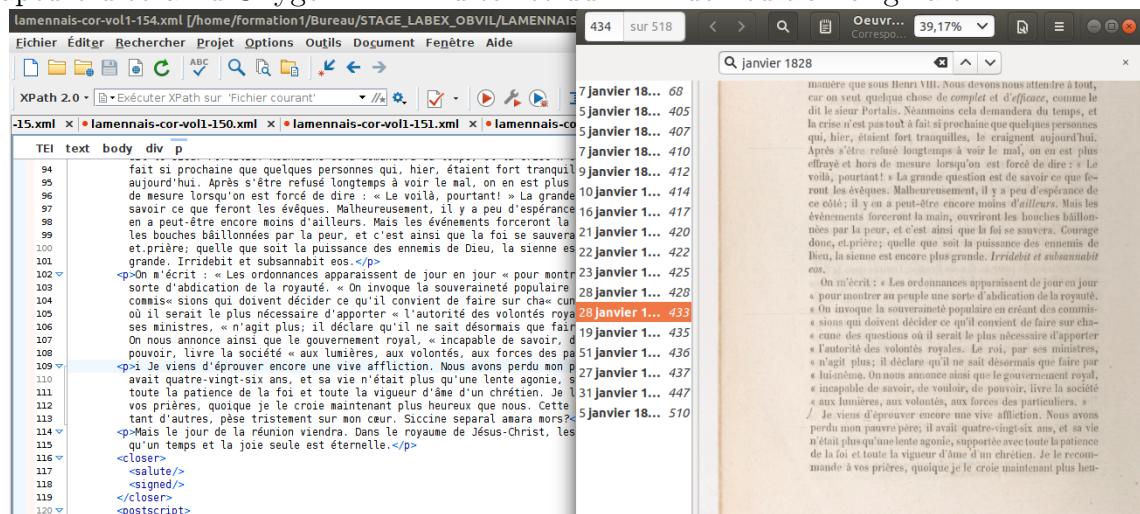
15. Tout ce qui fait référence à la mise en page dans nos dires ne concerne que le CRHXIX. Le reste concerne les deux projets, sauf avis contraire.

16. Nous reviendrons tout à l'heure sur cette extraction en parlant rapidement de Python.

17. Voir Fig. 10.4

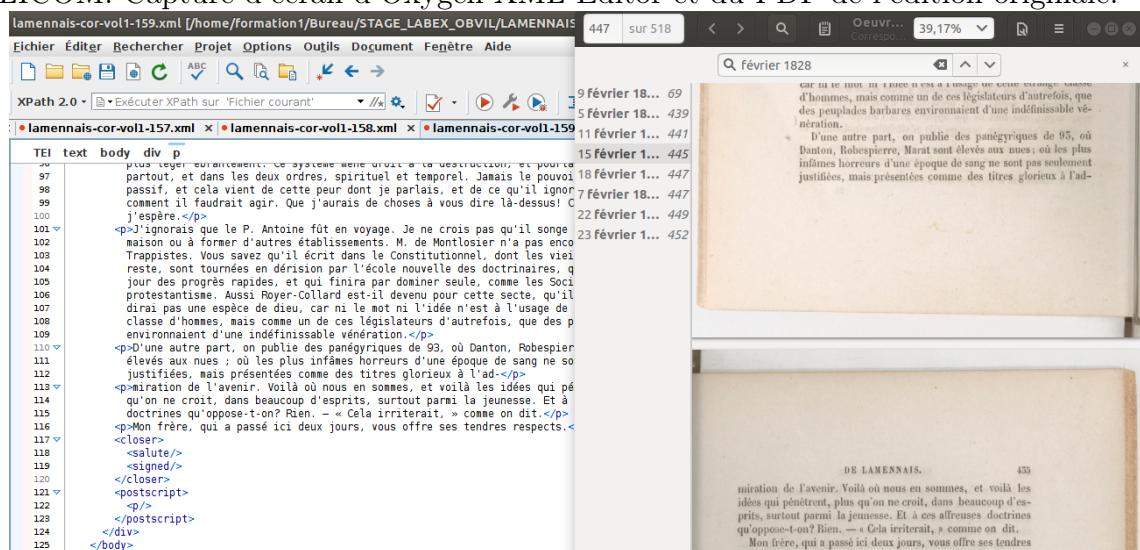
de crayon à papier. La faute d'interprétation de l'OCR est ici due à la qualité de la numérisation. Un livre usagé a été choisi pour la numérisation ce qui fausse le résultat de l'océrisation. D'autre part, on remarque un problème de guillemets à répétition. En effet, dans l'édition d'origine, les guillemets sont répétés à chaque ligne pour marquer que c'est une citation. Dans notre fichier XML, cela n'a plus de sens car ces guillemets polluent le corps du texte. Certaines modifications et corrections sont donc à apporter au fur et à mesure dans le corps du texte quand cela se présente.

FIGURE 10.4 – Des fautes dans un fichier XML (154), Félicité de Lamennais, ELICOM. Capture d'écran d'Oxygen XML Editor et du PDF de l'édition originale.



De même, on remarque de faux paragraphes qui se forment ainsi à cause d'un changement de page et qui se doublent d'un tiret séparant le mot. L'ancienne mise en page de l'édition numérique apporte donc des erreurs dans le fichier XML.

FIGURE 10.5 – Faux paragraphes dans un fichier XML (159), Félicité de Lamennais, ELICOM. Capture d'écran d'Oxygen XML Editor et du PDF de l'édition originale.



Ainsi, la mise en place et relecture du corps de la lettre a dû être relativement attentive aux problèmes d'océrisation et autres fautes.

#### 10.1.3.5 La fin de la lettre

La balise `<closer>` marque la fin d'une lettre et fait partie du rituel de correspondance.

Elle-même contient deux balises : d'une part `<salute>` qui comprend la formule de politesse concluant la lettre, d'autre part `<signed>` qui englobe la signature. Bien-sûr, la balise `<salute>` comprend des balises `<l>` encadrant chaque ligne, c'est toujours le même principe que dans le corps de la lettre. En revanche, les balises `<p>` ne sont pas admises dans la balise `<salute>`, probablement parce que les salutations ne font en général pas plus d'un petit paragraphe.

On peut mettre un attribut `@place` pour indiquer l'emplacement de la signature dans la lettre.

#### 10.1.3.6 Le post-scriptum

La balise `<postscript>` contient l'éventuel post-scriptum. Elle contient des balises `<p>` et `<l>`. C'est la dernière balise admise dans le corps du texte. Elle achève le fichier XML.

### 10.1.4 Les index

Par ailleurs, il s'agit de mettre en évidence dans le corps du texte les entités nommées déjà évoquées dans la deuxième partie de notre mémoire<sup>18</sup>. En effet, « on introduit un balisage dans un document pour l'étiqueter et l'organiser en vue d'un traitement automatisé. Si les paragraphes sont clairement marqués (balisés), alors un logiciel de mise en forme pourra les mettre en page correctement. Si les noms de lieu sont clairement marqués, un programme peut les sélectionner automatiquement pour générer un index géographique<sup>19</sup> ». Pour ELICOM, nous n'avons pas encore poussé très loin la granularité d'XML : nous avons commencé par transformer les fichiers HTML en fichiers XML sans aller encore très loin dans la description, particulièrement dans le balisage des entités nommées. En revanche, dans le cadre de l'édition numérique de Le Play, nous nous sommes bien penchée sur la question<sup>20</sup>. Comme annoncé dans la deuxième partie, nous avons choisi d'élaborer six index : index des ouvrages cités, index des événements, in-

18. 5.1.2 Cinq dimensions à prendre en compte

19. Lou Burnanrd, *Qu'est-ce que la Text Encoding Initiative ?*, Open Edition Press, 2015, URL : <https://books.openedition.org/oep/1298?lang=fr> (visité le 26/09/2020).

20. Cette partie sur les index ne concerne donc que le projet d'édition numérique de la correspondance de Frédéric Le PLay.

dex des noms de lieu, des noms de personne, des noms d'organisation, et enfin, index de vocabulaire leplaysien.

D'une part, dans le corps du texte, à chaque mot indexé, il faut donner la possibilité de pointer vers l'index. Pour les noms de lieu on utilise la balise `<placeName>`. Pour

FIGURE 10.6 – Capture d'écran d'Oxygen XML Editor, pointer vers un index

```

<l>voir à votre premier voyage à <placeName ref="#paris">Paris</placeName> pour vous
offrir mon</l>
<l>concours et celui de M. <persName ref="#alexis_chevalier">Alexis
    <!-- le tag pointe vers l'index --> Chevalier</persName> que nous nous
sommes</l>

```

les noms de personne ou personnage, on utilise la balise `<persName>`. Pour les noms d'organisation on utilise la balise `<orgName>` et pour tous les autres index, on indique la balise `<name>`. Toutes ces balises ont un attribut `@ref` qui permet de pointer vers l'index en question avec un tag (#).

D'autre part, les index en tant que tels se trouvent dans le `<teiHeader>`.

#### 10.1.4.1 L'index des ouvrages cités

Dans le `<sourceDesc>` se trouve un premier index, celui des ouvrages cités dans la correspondance. Tous se situent dans le même index, que ce soient des ouvrages écrits de la main de Le Play ou d'autres contemporains ou antérieurs. Nous avons choisi de renseigner aussi dans cet index les revues et les journaux. Pour les différencier des livres, nous avons pensé à ajouter un attribut `@type`, qui spécifie si c'est un journal ou une revue.

L'index des ouvrages est englobé dans une balise `<listBibl>`. Chaque ouvrage (livre, journal ou revue) est contenu dans une balise `<bibl>`, suivie d'une balise `<name>` pour le titre, `<author>` pour l'auteur, `<date>` normalisée avec l'attribut `@when`, et une balise `<note>` est facultative. À cet endroit, on pourra donner des informations sur l'ouvrage en question. La note pourra comporter un attribut `@resp` pour indiquer qui en est l'auteur. Cela permettra également de hiérarchiser la qualité des commentaires. Une note rédigée par un étudiant stagiaire sera considérée de moindre valeur que celle d'un spécialiste de Le Play.

La balise `<name>` doit obligatoirement avoir un attribut `@xml:id` qui permettra de pointer dans le texte vers l'index.

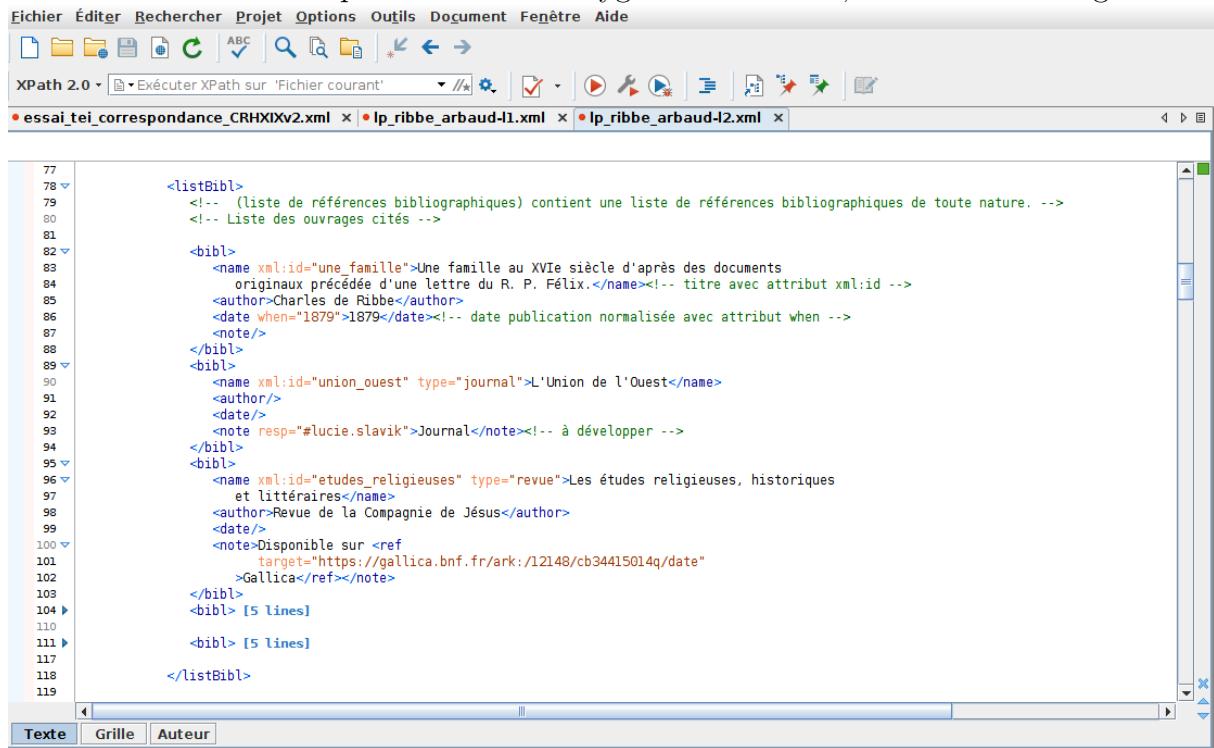
Par ailleurs, la balise `<author>` ne sera pas utile pour les journaux. Elle sera donc facultative pour les journaux mais obligatoire pour les ouvrages, distinction qui ne sera pas mentionnée dans l'ODD mais qui est néanmoins importante à avoir en tête.

La balise `<ref>` avec attribut `@target` permet de faire un lien vers un site<sup>21</sup>.

---

21. À confirmer, nous n'avons pas encore eu l'occasion de le tester.

FIGURE 10.7 – Capture d'écran d'Oxygen XML Editor, l'index des ouvrages



#### 10.1.4.2 L'index des événements

Toujours dans le <sourceDesc>, après l'index des ouvrages, on trouve l'index des événements, sous la balise <listEvent>. Il recense les événements importants qui sont mentionnés dans la correspondance. Chaque événement marquant est contenu dans une balise <event> avec un attribut @when mentionnant la date normalisée. La balise <event> englobe un <label> contenant le titre de l'événement et comportant un attribut @xml:id indispensable pour qu'on puisse faire référence à l'événement dans le corps du texte. On peut laisser un commentaire sur l'événement dans une balise <note>, avec un attribut @resp indiquant par qui elle a été rédigée.

Les autres index se trouvent dans le <profileDesc>.

#### 10.1.4.3 L'index des noms de lieu

Le `<settingDesc>` comprend l'index de noms de lieu. Les informations sur le lieu sont encadrées d'une balise `<place>` qui a un attribut `@xml:id` indispensable pour pointer ensuite du texte vers l'index. Dans le `<placeName>` est renseigné le nom du lieu, dans `<country>` le pays, dans la `<note>`, on peut écrire la description que l'on veut. On peut donc y glisser les notes rédigées par les transcripteurs, ce que nous n'avons pas pris le temps de faire durant les tests d'encodage des premières lettres en XML-TEI. Là encore, il y a possibilité d'y ajouter un attribut `@resp` pour renseigner qui a écrit la note.

#### 10.1.4.4 L'index des noms de personne

Le `<particDesc>` (toujours dans le `<profileDesc>`) comprend l'index des noms de personne et personnage mentionnés dans la correspondance. L'index est dans le `<listPerson>`.

Pour chaque personne ou personnage, il y a une balise englobante `<person>` avec un attribut `@sex` pour indiquer le genre de la personne et un attribut `@xml:id` indispensable pour lui donner un identifiant qui permettra ensuite dans le texte de pouvoir pointer vers l'index. Le prénom suivi du nom sont indiqués dans la balise `<persName>`. Une balise `<note>` permet de présenter la personne en question, mais c'est une présentation générale qui n'est pas liée à une lettre en particulier. Si l'on veut faire une note adaptée à un endroit d'une lettre en particulier, il faut la mettre dans le corps du texte. Pour la `<note>`, on peut mettre un attribut `@resp` pour indiquer qui l'a écrite. Pour son contenu, il serait intéressant de réfléchir à une façon plus ou moins normée d'écrire les notes (prénom, nom, titres, dates de naissance et de mort, rôle dans la société en général, avec la sociologie en particulier et lien avec Le Play), ou reprendre tout simplement les normes indiquées par Monsieur Matthieu Brejon de Lavergnée<sup>22</sup>.

#### 10.1.4.5 L'index des noms d'organisation

Le `<particDesc>` (toujours dans le `<profileDesc>`) comprend aussi l'index des noms d'organisation dans le `<listOrg>`. On peut si l'on veut les regrouper par type d'organisation avec un attribut `@type`, et faire donc plusieurs index (par exemple, un index d'organisations sociologiques, un index d'organisations politiques, un index d'associations etc.). Mais nous pensons plutôt ne pas faire de distinctions entre les différentes organisations et donc nous limiter à un simple index général pour simplifier au maximum l'encodage, d'autant que le profit retiré ne serait pas si grand.

Le nom de l'organisation est compris dans la balise `<org>` comprenant l'attribut `@xml:id` pour son identifiant. La balise `<orgName>` indique le nom de la société, et la balise `<note>` (plutôt que `<desc>`, initialement choisie, mais modifiée dans un souci de simplification) permet de rentrer une note générale sur l'organisation en question.

#### 10.1.4.6 L'index leplaysien

La balise `<textClass>` comprend tous les autres index. En l'occurrence, il s'agit pour nous de l'index leplaysien. Les mots à indexer sont en cours de choix, la plupart ont été sélectionnés par Messieurs Antoine Savoye, Rémy Hême de Lacotte et Matthieu Brejon de Lavergnée mais il est nécessaire d'y réfléchir encore. Par exemple, nous avons distingué « Réforme » de « Réforme morale » et « Réforme sociale ». Il serait bon de voir si l'on continue à distinguer ou si l'on met tout sous la même balise, avec le même identifiant `@xml:id`.

---

22. Voir 5.2.3 Des choix éditoriaux à faire en amont.

Quoiqu'il en soit, voici comment se présente l'index : on met tout d'abord une balise <textClass> comprenant une balise <keywords> comprenant elle-même une balise <list> avec, pour valeur de l'attribut @type, le nom de l'index dont il est question.

Chaque balise <item> comprend le nom du terme leplaysien avec un attribut @xml:id qui permettra de pointer vers le terme depuis le texte. Une balise <note> permet d'y laisser des commentaires. Il sera possible d'y ajouter un attribut @resp pour indiquer qui en est l'auteur.

Chaque <keywords> ouvrant un index, si l'on avait besoin de faire un autre index, on pourrait le mettre à cet endroit.

La réflexion autour de l'indexation a donc été riche. Nous avons d'ailleurs fait appel aux conseils de Monsieur Antoine Savoye, spécialiste de Le Play, pour nous éclairer quant au vocabulaire leplaysien à indexer.

### 10.1.5 La normalisation

Un des points importants dans XML est la normalisation. Nous l'avons déjà évoqué. Pour ELICOM, nous nous sommes surtout attachée à normaliser, outre les dates (AAAA-MM-JJ), les noms de personne dans le <correspAction>.

Nous avons repéré en amont les destinataires des différentes correspondances, puis nous avons cherché s'ils étaient présents sur *data.bnf*. En effet, sur *data.bnf*, on trouve des fiches de référence sur les auteurs, les œuvres et les thèmes, et renvoyer à ce site permet l'interopérabilité<sup>23</sup>. Cependant, tous n'étaient pas recensés dans *data.bnf* notamment

FIGURE 10.8 – La normalisation du <correspAction>, correspondance de Lamartine (lamartine-col-vol1-12.xml), capture d'écran d'Oxygen XML Editor

```

34 <profileDesc>
35   <correspDesc>
36     <correspAction type="sent">
37       <persName key="de Lamartine, Alphonse (1790-1869)">
38         ref="https://data.bnfr/fr/11910800/alphonse_de_lamartine/">Alphonse de
39         Lamartine</persName>
40       <date when="1808-09-10" resp="author">10 septembre 1808.</date>
41     </correspAction>
42     <correspAction type="received">
43       <persName key="Guichard de Bienassis, Prosper (1789-1857)">
44         ref="https://data.bnfr/en/16601127/nicolas_prosper_guichard_de_bienassis/">Prosper
45         Guichard de Bienassis</persName>
46     </correspAction>

```

parmi les correspondants de Félicité de Lamennais et de Proudhon. Lorsque c'était le cas, nous avons donc supprimé l'attribut @ref renvoyant à *data.bnf* et nous avons mis des parenthèses vides (...-...), à la place des dates de naissance et de mort, en espérant

23. Voir *Accueil*, Site web data.bnfr, URL : <https://data.bnfr/> (visité le 17/06/2020).

pouvoir les remplir un jour<sup>24</sup>.

## 10.2 L'ODD et la pérennité des données

### 10.2.1 Un schéma et une documentation pour la pérennité des données

Une fois les balises XML définies, il s'agit de créer une documentation dessus pour expliquer nos choix. Or, il existe une technologie qui permet de gérer à la fois le schéma de balises et sa documentation, c'est un document qui fait tout, autrement dit *One document does it all* (ODD).

Tout d'abord, comme le souligne Lou Burnard, « l'ODD est le langage de définition et de maintenance du système TEI. Il permet la maintenance du code et de sa documentation d'une manière intégrée, à partir d'une seule source XML. Il [...] fournit une manière efficace d'assurer la pérennité [des] données, en [...] obligeant de documenter leur usage d'une manière standardisée »<sup>25</sup>.

En bref, l'ODD comprend :

- Un schéma formel. Pour nous, nous avons choisi un schéma RELAX NG. Il contrôle l'édition, détermine quelles sont les balises disponibles, dans quels contextes, avec quels attributs, avec quelles valeurs, en respectant les contraintes et enchaînements. Nous y avons déjà réfléchi plus haut.
- Une documentation pour expliciter aux utilisateurs et développeurs les principes éditoriaux, les principes de choix de balises etc. Nous l'avons fait dans XML puis nous l'avons édité dans un format HTML.

La figure ci-dessous<sup>26</sup> permet de mieux saisir ce qu'est l'ODD<sup>27</sup>. Elle résume nos propos.

### 10.2.2 Créer l'ODD et l'associer à un document XML-TEI

Un ODD est un document TEI. Pour ce qui est du projet Le Play, nous l'avons construit avec oXygen XML Editor. Pour faciliter sa construction, nous avons utilisé

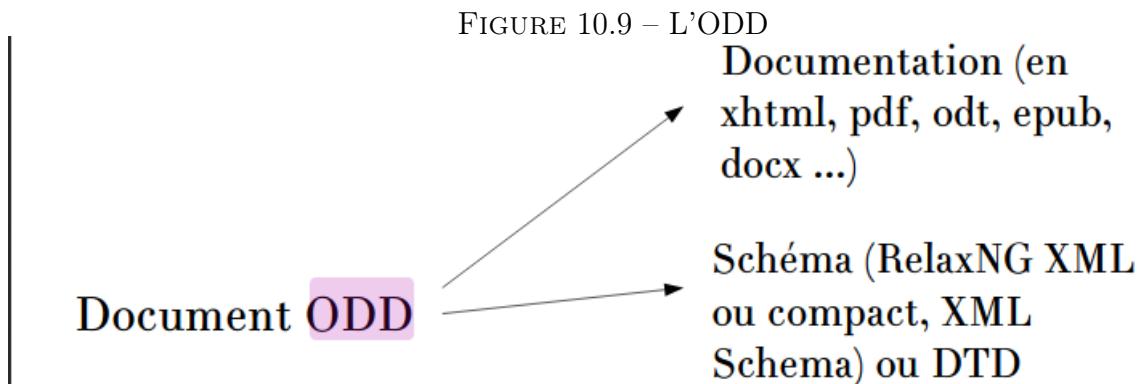
---

24. Voir dans les livrables, les dossiers de remarque des différentes correspondances du Labex OBVIL.

25. Lou Burnard, *Comment maîtriser le tigre TEI*, URL : <https://cahier.hypotheses.org/files/2018/08/ODD-diapos.pdf> (visité le 11/06/2020.).

26. Fig. 10.9

27. J.B. Camps, *ODD Structuration des données et des documents : balisage XML. Personnaliser la TEI : One Document Does it all*, M2 TNAH, ENC, 2017, p.44 URL : [https://halshs.archives-ouvertes.fr/cel-01706530/file/06\\_TEI\\_ODD\\_Camps\\_20170202.pdf](https://halshs.archives-ouvertes.fr/cel-01706530/file/06_TEI_ODD_Camps_20170202.pdf) (visité le 12/06/2020).



*ODD By Example*<sup>28</sup>.

Après avoir installé notre scénario *oddbyexample*, nous avons transformé l'ODD en RELAX NG, puis nous avons associé le schéma RELAX NG au fichier XML-TEI. Ainsi, l'association de l'ODD et du fichier XML-TEI se fait via le fichier RELAX NG.

Une fois que notre fichier TEI est lié au schéma qui le constraint, il doit lui correspondre. Dès qu'un enchaînement de balises, constraint dans le schéma, n'est pas respecté, le fichier XML-TEI le signale : cela pousse donc à une certaine rigueur. Par ailleurs, si l'on a des doutes sur la façon d'encoder, on peut toujours se reporter à l'ODD qui explique les choix. Ainsi, si la personne qui encode se rend compte d'une erreur, elle peut modifier et mettre à jour l'ODD, en indiquant que c'est elle qui a fait la modification.

FIGURE 10.10 – Extrait de la table des matières de l'ODD pour le projet Le Play

1. [Documentation pour l'encodage de la correspondance active et passive de Le Play](#)
  - 1.1. [Introduction](#)
  - 1.2. [Avertissement préliminaire](#)
  - 1.3. [Présentation du projet d'édition numérique Le Play](#)
    - 1.3.1. [Objectifs scientifiques](#)
    - 1.3.2. [Description du corpus](#)
    - 1.3.3. [Objectifs de cet encodage](#)
  - 1.4. [Structure de l'encodage](#)
    - 1.4.1. [Présentation générale](#)
    - 1.4.2. [Structure du teiHeader](#)
      - 1.4.2.1. [Le fileDesc](#)
        - 1.4.2.1.1. [Le titleStmt](#)
        - 1.4.2.1.2. [Le publicationStmt](#)
        - 1.4.2.1.3. [Le sourceDesc](#)
        - 1.4.2.1.3.1. [Le msDesc](#)
      - 1.4.2.2. [Le profileDesc](#)

### 10.2.3 L'ODD dans nos projets

Nous avons eu un rôle très différent quant à l'ODD pour nos deux projets. Pour ELICOM, l'ODD n'avait pas encore été réalisé. Nous n'avons donc été chargée que du

28. Pour plus d'informations sur le scénario *oddbyexample*, voir Ariane Pinche, Séance 11, « Personnaliser son ODD », Cours M2 TNAH, URL : [https://github.com/ArianePinche/coursTNAH\\_XML-TEI/tree/master/seance11](https://github.com/ArianePinche/coursTNAH_XML-TEI/tree/master/seance11) (visité le 18/02/2020).

travail en amont. En effet, avant de réaliser l'ODD, il faut avoir une connaissance assez large du corpus. Néanmoins, il a été parfois difficile pour nous de ne pas avoir d'ODD auquel nous pouvions nous référer pour garder une cohérence et une constance dans le choix des balises.

Pour le projet Le Play, c'est nous qui avons été chargée de la mise en place de l'ODD. Nous avons donc été au cœur de certains choix. Néanmoins, il est à noter que cet ODD est encore perfectible. De nombreux cas n'ont pas encore été traités. C'est donc un document susceptible d'être modifié. Nous pensons notamment à l'encodage des documents joints à certaines lettres ainsi que nombre d'autres cas particuliers que nous n'avons pas eu le temps de traiter durant ce stage relativement court. Il sera donc nécessaire de mettre l'ODD du CRHXIX à jour, au fil des cas particuliers rencontrés<sup>29</sup>.

L'ODD est donc un incontournable pour qui veut encoder des données en XML-TEI, de façon cohérente et pérenne.

XML a donc été le langage de base dont nous nous sommes servies pour nos éditions numériques de correspondance. Néanmoins, nous avons eu également recours à d'autres technologies.

## 10.3 Au service d'XML

### 10.3.1 XSLT

XSLT est un langage basé sur XML, permettant de styliser ou transformer des fichiers XML ou HTML. Comme nous l'avions déjà souligné dans la troisième partie<sup>30</sup> nous aurions pu passer par une feuille de transformation pour passer du fichier XML-TEI exporté de Transkribus et criblé de fautes, à un fichier XML-TEI comportant les noms de balises conformes. Cela reste une possibilité mais pour notre part, nous n'avons pas poussé plus loin la réflexion. En revanche, nous avons eu recours au langage de programmation Python.

### 10.3.2 Python

#### 10.3.2.1 Un *script* Python pour extraire les fichiers XML

Dans le cadre du projet ELICOM, Python nous a servi à extraire du résultat de l'OCR, le texte des différentes lettres pour les encoder en XML-TEI dans des fichiers séparés<sup>31</sup>.

Pour cela, nous avons installé un environnement virtuel basé sur Python 3 avec plusieurs *packages* dont **lxml** qui est un parseur pour les fichiers XML et HTML en Python,

---

29. Notre ODD pour le CRHXIX est joint à nos livrables.

30. Voir 9.6.2 La question des *tags*

31. Le *script* Python sert à la fois à distinguer les lettres les unes des autres et à passer en TEI.

utilisé ici pour construire les fichiers XML, ainsi que **beautifulsoup4** « qui permet une interaction simplifiée en Python avec les fichiers XML et HTML parsés »<sup>32</sup>. Par ailleurs, BeautifulSoup fonctionne sur les fichiers HTML mal formés, ce qui est le cas ici. Nous avons également importé le module **re** pour faire des regex dans Python.

Un *script* Python avait déjà été réalisé par un membre de l'équipe du Labex OB-VIL<sup>33</sup>. Nous avons donc eu un squelette commun, une base commune traitant les points communs aux différents corpus, mais à adapter à chacun. En effet, chaque corpus est tout de même à traiter différemment car les données n'apparaissent pas toujours de la même manière, à cause des différentes stratégies d'édition.

FIGURE 10.11 – Extrait du *script extraction-elicom.py*, capture d'écran de Sublime Text

```

1  #!/usr/bin/env python
2
3  """
4  Extraire le texte des 'xml' pour la partie Correspondances
5  Script différent pour les Works
6  """
7  |           ###Pour les libraires
8  import glob #-> pas utilisé. Module qui permet d'itérer à travers une arborescence de dossiers.
9  import re #-> regex
10 from bs4 import BeautifulSoup, NavigableString, Tag #-> pour utiliser beautifulsoup qui est le parseur html ()
11
12 from lxml import etree
13 from lxml.builder import ElementMaker
14
15 RE_ROMAN_NUMERALS = re.compile(r"^[MDCLXVI]+\$") #récuperer les chiffres romains (quand ils sont bien écrits)
16 RE_NOTES = re.compile(r"^(?:\d|[a-z])\." ) #supprimer les notes de bas de page
17 RE_AURORE = re.compile(r"AURORE") #récuperer le mot aurore
18 RE_SAND = re.compile(r"(?:\w+[A-Z]+ )?SAND")
19 RE_DEST = re.compile(r"^(?:A [A-Z][A-Z.]| )|^(:A[A-Z.]| )" ) #avoir le destinataire
20 RE_MAJ = re.compile(r"^(?:[A-Z][A-Z]+ ?)+") #trouver tout ce qui est en maj
21 RE_MEME = re.compile(r"^\u00c2\u0080\u00a3 \w+ M[A-Z\u00e9]{3,3}(,[^$]+\$)?") #à la meme : apparait très souvent
22

```

### 10.3.2.2 Les regex dans Python

Avec les regex dans python, nous nous sommes surtout attachée à matcher les chiffres romains pour Lamartine, les chiffres arabes pour Lamennais, et les signatures pour Proudhon, afin de pouvoir extraire les lettres<sup>34</sup>.

On peut constater sur la figure ci-dessus<sup>35</sup> que les lettres de George Sand (*script extraction-elicom.py*) ont été aussi extraites au moyen du repérage des chiffres romains via une regex.

Ainsi, à chaque chiffre arabe ou romain ou signature, une lettre était extraite et mise directement en XML-TEI grâce au module **lxml.builder**. La figure ci-dessous<sup>36</sup> montre

32. Voir Alix Chagué, *Ibidem*, p.85

33. Voir Fig. 10.11

34. Voir 8.2.5 Premiers repérages des marqueurs : pour Lamartine, les lettres sont toujours introduites par des chiffres romains, pour Lamennais par des chiffres arabes. Pour Proudhon, nous nous sommes servis de sa signature.

35. Fig. 10.11

36. Fig. 10.12

FIGURE 10.12 – Extrait du *script* de Proudhon, capture d'écran de Github

```

63 teifile = E.TEI (
64     E.teiHeader (
65         E.fileDesc (
66             E.titleStmt(E.title("Correspondance Pierre-Joseph Proudhon"),
67                         E.author("Pierre-Joseph Proudhon"),
68                         E.respStmt(E.resp("Encodage réalisé pour Obvil dans le cadre d'un stage M2 TNAH de l'ENC, sous la direction d'Arthur Pro",
69                                     E.persName(E.forename("Lucie"), E.surname("Slavik")))),
70                         E.editionStmt(E.edition()),
71                         E.publicationStmt(
72                             E.publisher("Obvil"),
73                             E.date(when='2020'),
74                             E.idno(),
75                             E.availability(E.licence(E.p, target="http://creativecommons.org/licenses/by-nc-nd/3.0/fr/"), status="restricted")
76                         ),
77                         E.sourceDesc(E.bibl())),
78         ),
79         E.profileDesc (
80             E.correspDesc(
81                 E.correspAction(
82                     E.persName("Pierre-Joseph Proudhon", key="Proudhon, Pierre-Joseph (1809-1865)", ref="https://data.bnfr.fr/fr/1192"),
83                     E.date(date, when="%s %date, resp="),
84                     type="sent"
85                 ),
86                 E.correspAction(
87                     E.persName(to, key="%s (.....)" %to),
88                     type="received"
89                 ),
90                 E.correspContext(
91                     E.ref(type="prev", target="foo.xml"),
92                     E.ref(type="next", target="bar.xml"),
93                 )
94             ),
95             E.creation(E.date(when='%s %date)),

```

l’arborescence demandée depuis le *script* Python de Proudhon.

Par ailleurs, nous nous sommes également servis des expressions régulières pour extraire les métadonnées et les récupérer ensuite dans le <teiHeader>. Par exemple, pour le *script* de Proudhon, nous avons écrit `RE_DEST = re.compile(r"A M[A-Z].+")` pour matcher les destinataires de Proudhon, et `RE_DATE = re.compile(r"([,]+), (.+)\$")` pour matcher la date d’écriture de la lettre, puis nous avons mis ces informations dans le <teiHeader><sup>37</sup>.

### 10.3.2.3 Les limites de l’OCR entravent l’efficacité du *script* Python

Cependant, tout n’est pas parfait du premier coup. Certes, Python nous avance beaucoup et fait gagner beaucoup de temps dans l’extraction des lettres et leur transformation en fichiers XML-TEI en vue de leur édition. Toutefois, certaines choses restent à modifier.

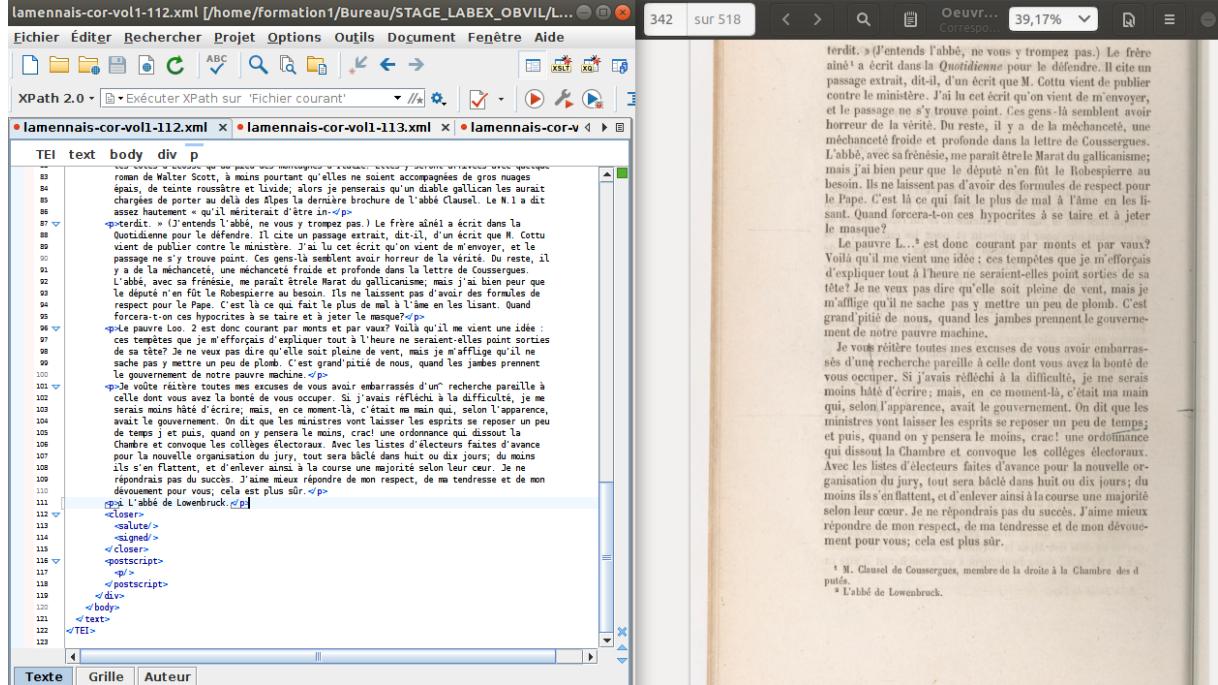
Tout d’abord, certaines modifications doivent parfois être faites en amont pour une meilleure extraction : c’est le cas lorsque des chiffres arabes ou romains ont été mal océrisés et empêchent la regex de les reconnaître : ainsi, les fautes de l’OCR rendent l’extraction difficile.

Par ailleurs, après l’application du *script*, on remarque également certaines défaillances, encore dues à la mauvaise océrisation. En effet, nous avons élaboré pour la correspondance de Lamennais une regex afin de matcher et supprimer les notes de bas de pages qui étaient disséminées au sein du texte. Des centaines de notes ont été ainsi

37. Voir les *scripts* dans leur intégralité dans les livrables.

supprimées. Cependant, de nombreuses notes n'ont pas été matchées par la regex (et ont

FIGURE 10.13 – Lettre de Lamennais, notes mal matchées à cause de l'océrisation, lamennais-cor-vol1-112.xml



donc échappé à la suppression) car l'océrisation avait été mauvaise. L'exemple de la figure 10.13 est typique : on constate que la première note a été supprimée mais que la deuxième subsiste : le « 2 » a été pris pour un « i » par l'OCR. La note n'a donc pas été matchée et il faut donc la supprimer autrement<sup>38</sup>.

Enfin, le résultat de l'OCR n'étant balisé que par des `<p>` et des `<span>`, et qu'aucune autre indication stylistique ou bien de structure n'apparaît, cela empêche un traitement plus fin des données pour la mise en page ou le signalement des vers. Ainsi, l'OCR ne tient pas compte des italiques, mais nous n'avons pas traité ce point. En revanche, pour ce qui est de la mise en page, nous nous sommes penchée sur la question : Lamartine surtout et parfois Lamennais glissaient des vers dans leurs lettres. Pour Lamartine, c'était assez fréquent. Il lui arrivait même parfois d'écrire une lettre intégralement en vers<sup>39</sup>.

Afin de les remettre en valeur, nous avons choisi d'indiquer les strophes par les `<lg>` et les vers par les `<1>`, pour améliorer les repérages et la visualisation, dans le but de poursuivre l'objectif de cette édition numérique qui veut permettre une plus grande facilité d'accès au texte pour y effectuer des recherches plein texte et sur la structure, via l'interrogation des balises et des métadonnées.

Pour effectuer ce travail, nous avons dû mettre en regard les numérisations pour voir quelles lettres contiennent des vers. Nous avons donc fait ces ajouts manuellement. On

38. Pour notre part, nous l'avons fait manuellement.

39. La lettre 38 par exemple.

FIGURE 10.14 – Mise en place des `<lg>` et des `<l>` dans la correspondance de Lamartine

	<pre> &lt;lg&gt; &lt;l&gt; Tandis que d'un léger coton&lt;/l&gt; &lt;l&gt; Mon visage frais se colore,&lt;/l&gt; &lt;l&gt; Que tout sourit à mon aurore,&lt;/l&gt; &lt;l&gt; Et que raisonner en Caton &lt;/l&gt; &lt;l&gt; Chez moi serait risible encore,&lt;/l&gt; &lt;l&gt; De mon espoir, de mes désirs&lt;/l&gt; &lt;l&gt; Je veux divertir ta paresse,&lt;/l&gt; &lt;l&gt; Et, laissant l'ingrate vieillesse&lt;/l&gt; &lt;l&gt; S'affliger sur ses souvenirs,&lt;/l&gt; &lt;l&gt; Une heure ou deux de ma jeunesse,&lt;/l&gt; &lt;l&gt; Parler au moins de mes plaisirs.&lt;/l&gt; &lt;p&gt;Sur une plus courte mesure Pour toi je vais mouler mes vers Et dans mille sentiers divers Courir à huit pieds sans césure.&lt;/p&gt; </pre>
	<pre> &lt;lg&gt; &lt;l&gt; Sur une plus courte mesure&lt;/l&gt; &lt;l&gt; Pour toi je vais mouler mes vers &lt;/l&gt; &lt;l&gt; Et dans mille sentiers divers&lt;/l&gt; &lt;l&gt; Courir à huit pieds sans césure.&lt;/l&gt; &lt;/lg&gt; </pre>

peut constater sur la figure 10.14 les modifications qui ont été faites. Sur la gauche, on voit le fichier XML qui a été extrait : les deux strophes apparaissent dans deux paragraphes (`<p>`), sans aucune distinction des vers. Sur la droite, les deux paragraphes sont devenus des strophes reconnues par les `<lg>`, et au sein des strophes, chaque vers (`<l>`) est marqué par un retour à la ligne.

XML est donc poussé au bout de ses possibilités avec l'apport d'autres technologies telles que Python, XSLT et les regex. Néanmoins, la qualité pas toujours optimale de l'OCR nécessite tout de même de nombreuses reprises à la fois techniques et orthographiques.

Toutes ces technologies sont donc au service de l'édition numérique de correspondance et elles permettent de traiter les données acquises.

# Chapitre 11

## Relever les défis du projet numérique

### 11.1 Des difficultés à surmonter

Nous avions déjà évoqué plus haut<sup>1</sup> les difficultés que rencontrent nombre de projets numériques, que cela soit dans le public ou le privé : deux tiers rencontrent de gros problèmes au cours de leur élaboration, et un tiers se solde par un échec<sup>2</sup>.

Nos deux projets n'échappent donc pas à cette constante. Pour ce qui est du projet au CRHXIX, nous pensons que le manque de subventions pourrait être un des freins au développement du projet, qui toutefois avance peu à peu.

Pour ELICOM, nous n'avons pas assez de recul pour voir les difficultés que nous pourrions rencontrer. Une fois que l'ODD sera faite, ce sera bien-sûr un grand pas en avant, tant pour éclairer l'encodage que pour la pérennité des données.

Pour prévenir les difficultés, notre rôle en tant que stagiaire a été d'être un relais et d'assurer la transmission de notre travail, d'autant que nous avons réalisé la totalité de notre travail sur nos deux projets en télétravail.

### 11.2 Prévenir le « facteur d'autobus »

Pour la réussite d'un projet, il s'agit d'éviter ce que les britanniques appellent le *bus factor*, autrement dit, le « facteur d'autobus », expression qui vient de la phrase « Combien de personnes clés dans votre équipe peuvent se faire renverser par un autobus avant que votre projet échoue ? ». Ainsi, « le “facteur d’autobus” est le nombre minimum de membres de l’équipe qui peuvent disparaître soudainement d’un projet avant que celui-ci ne s’arrête par manque de personnel compétent ou bien informé »<sup>3</sup>.

---

1. 6.6 Le cahier des charges

2. Voir Jean-Louis Foucard, Module de formation « Manager un projet Numérique » « Master Archives – Technologies numériques appliquées à l’histoire », École Nationale des Chartes, mars 2020.

3. *Facteur d'autobus*, Wikipédia, URL : [https://fr.wikipedia.org/wiki/Facteur\\_d%27autobus](https://fr.wikipedia.org/wiki/Facteur_d%27autobus) (visité le 28/09/2020).

Autrement dit, en tant que stagiaire, il faut qu'une fois que nous avons quitté le projet, les membres de l'équipe, que ce soit celle du CRHXIX ou du Labex OBVIL, aient en main tout notre travail et puisse profiter de ce que nous avons appris, compris, repris durant notre stage. Il s'agit d'assurer la transmission des savoirs et le transfert de compétences.

### 11.2.1 Les avantages de GitHub

Pour cela, GitHub est un moyen intéressant pour le travail en équipe et la transmission. GitHub est un « service web d'hébergement et de gestion de développement de logiciels, utilisant le logiciel de gestion de versions Git »<sup>4</sup>. Il permet donc le versionnage<sup>5</sup> ainsi que le travail en équipe avec le partage de code. Nous ne l'avons pas utilisé pour le projet Le Play mais pour le projet ELICOM<sup>6</sup>. Nous avons donc pu transmettre notre travail au fur et à mesure et faire des rapports au jour le jour pour une meilleure gestion de projet. À cela se sont ajoutés les appels réguliers avec le tuteur technique pour faire le point, et un rapport général en fin de stage.

Si nous n'avons pas usé de cette stratégie pour notre travail au CRHXIX, des contacts fréquents avec l'équipe, aussi bien par courriel que par téléphone ou encore par visioconférence ont eu lieu. Le transfert du travail se fera sous peu. Par ailleurs, pour le transfert de compétences, nous avons mis en place des tutoriels.

### 11.2.2 Des tutoriels et des fiches de savoir pour assurer la transmission

Ainsi, pour le CRHXIX, nous avons constitué divers tutoriels selon les technologies utilisées, afin que les membres de l'équipe puissent prendre la suite de la partie numérique.

Pour ce qui est de notre travail sur Transkribus, nous avons réalisé quatre tutoriels, pour chaque étape :

- Un premier tutoriel pour le chargement des données et leur premier entraînement<sup>7</sup>
- Un second dédié à l'entraînement du modèle<sup>8</sup>
- Un troisième consacré à l'application du modèle définitif<sup>9</sup>
- Un quatrième qui traite de l'exportation des données depuis le serveur Transkribus<sup>10</sup>

4. GitHub, Wikipédia, URL : <https://fr.wikipedia.org/wiki/GitHub> (visité le 28/09/2020).

5. On entend par versionnage la « Gestion des différentes versions d'un même document ». Voir *versionnage*, Wiktionnaire, URL : <https://fr.wiktionary.org/wiki/versionnage> (visité le 30/09/2020)

6. Voir OBVIL/Elicom, Github, URL : <https://github.com/OBVIL/elicom>

7. Voir dans les livrables *Tuto1\_preparation\_des\_donnees\_transkribus.pdf*

8. Voir dans les livrables *Tuto2\_entrainement\_du\_modele.pdf*

9. Voir dans les livrables *Tuto3\_application\_du\_modele.pdf*

10. Voir dans les livrables *Tuto4\_exportation\_transkribus.pdf*

Par ailleurs, nous avons réalisé un point sur notre travail pour dire où nous en étions et quelles conclusions nous tirions de cette première expérience<sup>11</sup>.

Pour ce qui est de notre travail sur XML-TEI et ODD, nous avons également travaillé à la transmission, pour le CRHXIX :

- Par la réalisation d'un tutoriel<sup>12</sup> pour la prise en main d'Oxygen XML Editor afin de réaliser les encodages
- Par la transmission d'un tutoriel sur l'ODD<sup>13</sup>.

Par ailleurs, la documentation de l'ODD<sup>14</sup> est le moyen par excellence pour prévenir le « facteur d'autobus ». Nos choix y sont justifiés, et nos questionnements son présentés. Celui qui prendra la suite du projet pourra donc comprendre la logique de notre travail, perfectionner l'ODD et résoudre certains problèmes ou questionnements.

Quant à ELICOM, nous avons, au fur et à mesure de notre travail, constitué des fichiers rassemblant nos remarques sur l'HTML et le XML et à l'occasion du *script* Python de chacun des auteurs. Nous y avons entre autres écrit quelques expressions régulières. Ces fichiers sont perfectibles et auraient pu être perfectionnés. Nous avons quand-même choisi de les mettre dans les livrables à titre d'exemple. À la fin de notre travail sur Lamartine, nous avons également réalisé un fichier récapitulatif sur le code Python. Celui-ci

**FIGURE 11.1 – Rapport sur le code Python de Lamartine, capture d'écran de GitHub  
Récapitulatif sur le code python *extraction-elicom\_lamartine.py***

- 96 lettres sur 97 sont matchées. La dernière a disparu, je ne me l'explique pas. Elle est pourtant bien présente dans le fichier HTML. Il faudra donc créer un fichier xml pour la lettre 97, qu'on intitulera lamartine-cor-vol1-96 (puisque la numérotation commence à 0), et que l'on construira selon l'arborescence utilisée pour les autres lettres.
- Pour ce qui est de l'arborescence, dans le `<teiHeader>` :

- J'ai ajouté dans le `<teiHeader>` les balises `<respStmt/>`, après le `<titleStmt/>` ce qui donne :

```
<respStmt>
<resp>Encodage réalisé pour Obvil dans le cadre d'un stage M2 TNAH de l'ENC, sous la direction d'Arthur Provenier
<persName>
<forname>Lucie</forname>
<surname>Slavik</surname>
</persName>
</respStmt>
```

est disponible sur GitHub<sup>15</sup>. La figure ci-dessus<sup>16</sup> donne un échantillon de ce rapport

11. Voir dans les livrables *point\_transkribus.odt*

12. Voir dans les livrables *Tuto\_XML-TEI\_CHRXIX.pdf*

13. Ce tutoriel n'a pas été réalisé par nos soins en revanche.

14. L'ODD du CRHXIX est disponible dans les livrables.

15. Voir [https://github.com/OBVIL/elicom/blob/master/extraction\\_cor\\_stage2020/cor\\_lamartine/remarques/recapitulatif\\_lamartine.md](https://github.com/OBVIL/elicom/blob/master/extraction_cor_stage2020/cor_lamartine/remarques/recapitulatif_lamartine.md)

16. Fig. 11.1

rédigé au moyen du langage de balisage Markdown.

Par ailleurs, le cahier des charges du CRHXIX ayant été un peu détourné de sa fin en se transformant en rapport, il a toutefois permis également de faire un petit bilan de la situation pour voir où nous en étions et quelles étaient les prochaines étapes à suivre.

Enfin, nous avons réalisé des fiches pour accélérer certaines normalisations. Ainsi, pour le CRHXIX, nous avons écrit sept fiches pour les normalisations des index, que ce soit pour le recensement des correspondants dans le <teiHeader> ou les six index.

De même, pour ELICOM, nous avons constitué des index de correspondants pour pouvoir normaliser plus facilement chacun des destinataires lors des corrections des fichiers XML.

### 11.2.3 Documenter son code

Enfin, un des moyens d'assurer la transmission est la documentation du code en direct. En effet, pour ce qui est des *scripts* Python pour ELICOM, nous avons essayé, du moins au début, de bien documenter notre code, à la fois pour mieux le comprendre et aussi pour mieux le faire comprendre. Cela fait partie des bonnes pratiques pour nos projets numériques. De même pour les premiers essais d'encodage en XML-TEI pour le CRHXIX, nous avons commenté chaque balise pour la justifier et l'expliquer. Nous avons été particulièrement attentive à cela dans un souci de transmission, et ceci nous a d'ailleurs aidée nous-même au moment de la rédaction de l'ODD car il n'y avait plus qu'à suivre les indications déjà présentes dans les fichiers XML commentés au fur et à mesure.

Nous avons donc tenté de relever au cours de nos stages les défis de ces deux projets d'édition numérique de correspondance sur des corpus du XIX<sup>e</sup> siècle. Cette partie de traitement des données s'est réalisée autour du langage XML, avec le souci de transmettre non seulement les réalisations de notre travail mais également les étapes de réflexion par lesquelles nous sommes passée.

# Conclusion



Nos deux stages nous ont permis d'approfondir les enjeux de l'édition numérique de correspondance par deux approches différentes, sur des corpus du XIX<sup>e</sup> siècle. Cette double expérience nous a permis d'avoir une vue plus large et de découvrir différents projets auxquels nous avons pris plaisir à participer.

Dans une première partie, nous nous sommes penchée sur les contextes des projets, les buts qu'ils se proposent, et la matière première si l'on peut dire, sur laquelle ils se basent pour atteindre les objectifs fixés. Ces éléments sont essentiels pour avoir une meilleure vue d'ensemble des projets et cerner ensuite quel rôle sera le nôtre, comment nous nous inscrivons dans ces projets, et quels moyens prendre pour y contribuer et les faire avancer . Nous avons donc analysé le contexte universitaire et intellectuel des projets : notre but est de faire avancer la recherche historique et culturelle en général. Cependant, ces deux projets, bien que travaillant tous deux à la diffusion de correspondances du XIX<sup>e</sup> siècle, diffèrent dans leur but. Certes, ils ont les points communs du genre épistolaire et du cadre dans lequel les lettres ont été produites, le XIX<sup>e</sup> siècle, mais le CRHXIX œuvre à une édition plus classique, centrée sur une personne autour de laquelle se greffe un réseau de correspondants, le but restant la recherche historique, alors que pour le Labex OBVIL, ELICOM est un moyen de pousser les possibilités du numérique en terme de fouilles de données, d'enrichissement et de visualisation, ainsi que la recherche en humanités numériques, avec un séminaire de recherche consacré au projet et aux avancées de l'édition numérique de correspondance. Bien sûr, certains outils restent les mêmes, mais l'optique est différente. Cette différence s'accentue lorsque l'on considère les sources qui servent de point de départ aux deux projets. Pour le CRHXIX, nous avons affaire à des manuscrits originaux, qui ont été numérisés ou devront l'être. Certes, nous n'avons pas accès directement à la source mais à son fac-similé. Nous sommes néanmoins assez proches de l'original. Par ailleurs, nous ne nous appuyons aucunement sur des éditions précédentes, en ce sens-là, notre travail d'édition est conséquent puisque nous ouvrons la voie. Pour ELICOM, au contraire, nous travaillons sur des sources indirectes. Nous n'avons pas accès aux manuscrits ni à des fac-similés mais à de précédentes éditions imprimées. Nous prenons de la distance avec elles dans les notes et certains choix, mais elles restent tout de même notre point de départ. Notre rapport à la source est donc différent, dans l'un et l'autre projet. Or, la source a son rôle dans le choix des moyens car on les adapte à elle. Ce sont ces considérations qui nous ont occupée pendant la première partie.

Après avoir exposé nos projets et leurs différentes facettes, nous avons souhaité prendre un peu de hauteur par rapport à l'édition numérique de correspondance aujourd'hui, et faire un état de l'art : où en sommes-nous, quels sont les moyens mis à notre disposition ? Toutes ces réflexions vont influencer notre manière d'agir et d'aborder nos projets. Nous avons donc établi un petit bilan scientifique sur les réflexions de la communauté scientifique autour de l'édition numérique de correspondance, et les divers déploiements d'outils pour favoriser son essor. Nous avons abordé les problématiques propres

à l'édition numérique de correspondance, sachant que l'aspect numérique est essentiel. Nos considérations sur l'édition numérique de correspondance se sont conclues par la pratique : nous avons donc envisagé notre futur site pour le CRHXIX, concluant ainsi notre deuxième partie.

Après la théorie, nous sommes donc passée à la pratique, tout d'abord pour ce qui est de l'acquisition des données, ce qui a fait l'objet de notre troisième partie. Or, ici, nous avons constaté que pour nos deux projets, la nouvelle technologie de l'apprentissage machine s'est avérée être au cœur de l'acquisition des données, tant pour ELICOM via l'OCR de Gallica, que pour notre projet d'édition numérique de la correspondance de Frédéric Le Play, via l'HTR de Transkribus. Nous avons particulièrement développé ce point de Transkribus, outil de transcription collaborative fort utile. Néanmoins, des interrogations subsistent quant à la rentabilité du modèle, au traitement et au mode d'importation des données. Vaudrait-il mieux passer par XSLT ? Devrait-on penser davantage au *tags* dans Transkribus et les personnaliser ? Ce sont autant de questions qui se posent encore. Quant à l'OCR de Gallica, nous avons pu voir qu'il conditionnait beaucoup le pré-traitement et traitement des données sur lequel nous nous sommes plus attardée en quatrième partie.

Après avoir considéré l'acquisition des données, nous nous sommes penchée sur leur traitement. Dans cette partie, nous avons vu combien le langage XML est utilisé dans les éditions numériques de correspondance. Celui-ci qui a, en soi, de multiples possibilités, est restreint en l'occurrence à nos besoins d'édition de correspondance. Nous avons donc expliqué nos choix de balises, leur documentation via l'ODD, ainsi que les technologies utilisées pour pousser au maximum les possibilités d'XML, à savoir XSLT - qui reste encore un moyen à utiliser éventuellement - et Python. Puis nous avons exposé les difficultés inhérentes aux projets numériques, les défis à relever, et les moyens que nous avons employés dans ce but, pour que les projets soient menés à terme. Nous n'avons été qu'un maillon de la chaîne, mais il ne doit pas manquer, au risque de briser cette chaîne.

Le contexte de télétravail a rendu l'expérience de ces stages particulièrement inédite. Nous avons dû relever le défi de travailler seule et de n'être reliée aux différentes équipes que par le net. Nous avons en quelque sorte travaillé avec des équipes qui nous ont paru un peu virtuelles, quoique toujours là pour répondre à nos questions.

Malgré tout, le bilan reste très positif. Nous avons pu mettre en pratique les connaissances reçues à l'ENC, et nous familiariser toujours plus aux diverses technologies. Nous avons particulièrement apprécié le fait de voir concrètement, dans des projets bien réels, les avantages et les possibilités du numérique.

Pour ce qui est de notre stage au Labex OBVIL, au total, 388 lettres ont été extraites, dont 264 fichiers ont été corrigés durant ces 20 jours de stage. Nous avons donc eu la satisfaction d'avoir pu participer au projet ELICOM. Dans l'ensemble, ce stage nous a bien aidée à consolider nos connaissances en XML et nous a indirectement aidée pour le projet Le Play, grâce à leur point commun d'édition de corpus épistolaires du XIX<sup>e</sup> siècle.

Le plus grand défi de ce stage s'est situé dans la rédaction du code Python que nous avons appris à mieux maîtriser, même si de grands progrès restent à faire.

Quant à notre stage au CRHXIX, il a été extrêmement enrichissant car nous avons été chargée de penser la partie numérique du projet. Durant tout le confinement, nous avons été la seule personne de l'équipe, à quelques exceptions près, à pouvoir nous consacrer totalement à ce projet. Sans les remarques du chef de projet et de membres de l'équipe, et sans l'aide des professeurs de l'École, particulièrement celle de notre tuteur, ainsi que le soutien de notre tuteur d'OBVIL, nous n'aurions pas pu aller aussi loin dans le projet. Durant une petite trentaine de jours, nous avons dû faire notre les enjeux du projet, prendre en main Transkribus qui nous était presque totalement étranger, entraîner un modèle avec plus de 20 000 mots, penser à l'exportation en XML-TEI, se faire une idée du futur site et en fonction, penser l'encodage en TEI. Cela a donc été un stage particulièrement riche par sa diversité. Il n'a pas été exempt de limites : nous n'avons pu réaliser de véritable cahier des charges ni de récits utilisateurs dignes de ce nom. Néanmoins, il était difficile de tout mener de front en si peu de temps.

Ces différentes approches de l'édition numérique de correspondance ont donc été très enrichissantes. Nous avons pu constater que l'épistolaire à l'ère du numérique se situe réellement entre des perspectives historiques, ou plus largement des perspectives de sciences humaines, d'aide à la recherche et autres, et des innovations technologiques. Nous sommes entre les sciences humaines et le numérique, nous sommes tout simplement dans les humanités numériques, avec peut-être un aspect plus « humanités » pour le projet Le Play, et plus « numérique » pour ELICOM.

Ainsi, ces projets sont l'illustration que les humanités numériques, comme le souligne le *Manifeste des digital humanities*<sup>17</sup>, « ne font pas table rase du passé, mais s'appuient, au contraire, sur l'ensemble des paradigmes, savoir-faire et connaissances propres à ces disciplines, tout en mobilisant les outils et les perspectives singulières du champ du numérique ».

---

17. *Manifeste des digital humanities*, THATCamp, Paris, 2010 URL : <https://tcp.hypotheses.org/318> (visité le 29/09/2020)



## **Annexes**



Les annexes contiennent certaines illustrations du mémoire ainsi que la description des fichiers disponibles dans les livrables sur Github à cette adresse : <https://github.com/LaureRossignol96/MemoireTNAH2020>.

Nous présentons d'abord les annexes qui illustrent le mémoire et auxquelles nous renvoyons explicitement dans le mémoire : ce sont les annexes A, B et C.

Les annexes D décrivent les livrables du CRHXIX, les annexes E ceux du Labex OBVIL.



## Annexe A

# Édition numérique de correspondance

### A.1 Exemples d'éditions déjà existantes

FIGURE A.1 – *D'Alembert en toutes lettres*. Édition numérique de la correspondance de D'Alembert

Édition numérique des Œuvres complètes de D'Alembert (1717-1783)  
Série V : correspondance générale



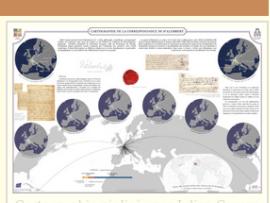
***D'Alembert en toutes lettres***

  
L'édition critique de la correspondance est l'objet de la série V des Œuvres complètes de D'Alembert, l'enfant trouvé devenu académicien, mathématicien, philosophe, coéditeur de l'Encyclopédie, secrétaire perpétuel de l'Académie française... L'Inventaire analytique qui a paru en 2009 a permis de recenser, dater, décrire et résumer les 2200 lettres envoyées ou reçues par D'Alembert, d'en distinguer les lettres ostensibles (épitres, articles publiés dans les imprimés sous le nom de « lettres ») et d'autres documents autographes, d'en éliminer les doublons et lettres factices. L'interrogation de la base de données correspondante (tenue à jour des nouvelles lettres) est proposée dans le bandeau de droite de l'interface.

Les saisies des lettres sont progressivement mises en ligne, ainsi que les manuscrits quand les institutions qui les possèdent nous en donnent l'autorisation. L'interface permet la consultation de ces saisies (lorsque la saisie est disponible, la lettre est suivie de l'icône ), leur mise en regard avec le facsimilé des manuscrits. Le tout est enrichi d'informations critiques sur les lettres et l'historique des documents présentés.

Le lien ci-dessous vous conduira par défaut aux corpus des lettres, présentées par fonds, dont les manuscrits sont actuellement consultables. Pour accéder aux lettres de la correspondance, utilisez le

» 2300 lettres connues, datées, décrites  
» 278 lettres transcrives  
» 135 lettres avec les manuscrits

  
Cartographie réalisée par Julien Caverio (cartographe du labex TransfēS de l'ENS) avec la collaboration d'I. Passeron, A. Guillaud (IMJ-PRG) et de M.-L. Massot (CAPHES)

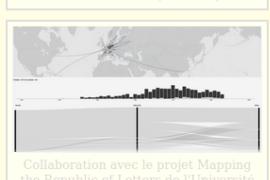
  
Collaboration avec le projet Mapping the Republic of Letters de l'Université de Stanford : « Mapping the Republic of Letters »

FIGURE A.2 – Édition numérique de la correspondance de Jean Paulhan, Labex OBVIL

 Observatoire  
de la vie littéraire

[RETOUR DANS LA BIBLIOTHÈQUE](#)

Téléchargements : tei, epub, kindle, texte brut, iramuteq, html.

Rechercher de mots

De  à

↓ N° ↑	↓ Auteur ↑	↓ Date ↑	↓ Titre ↑
1	Adamov, Arthur	1955	Arthur Adamov à Jean Paulhan. Correspondance (1950-1955)
2	Alain ; Chartier-Alain, Gabrielle ; Paulhan, Jean	1952	Alain, Gabrielle Chartier-Alain & Jean Paulhan. Correspondance (1927-1952)
3	Albert Thibaudet	1936	Albert Thibaudet à Jean Paulhan. Correspondance (1926-1936)
4	Allan, Blaise	1958	Blaise Allan à Jean Paulhan. Correspondance (1950-1958)
5	Amrouche, Jean ; Paulhan, Jean	1954	Jean Amrouche à Jean Paulhan. Correspondance (1951-1954)
6	Andreu, Pierre	1935	Jean Paulhan à Pierre Andreu. Correspondance (1931-1935)
7	André Berne-Joffroy	1958	André Berne-Joffroy à Jean Paulhan. Correspondance (1950-1958)
8	Arabia, Jean	1958	Jean Arabia à Jean Paulhan. Correspondance (1951-1958)
9	Arban, Dominique	1954	Dominique Arban à Jean Paulhan. Correspondance (1954)
10	Arland, Marcel	1936	Marcel Arland à Jean Paulhan. Correspondance (1925-1936)
11	Arland, Marcel	1958	Marcel Arland à Jean Paulhan. Correspondance (1950-1958)
12	Baldensperger, Fernand	1958	Fernand Baldensperger à Jean Paulhan. Correspondance

FIGURE A.3 – Édition numérique de la correspondance de Marc Michel Rey, HUMA-NUM

Marc  
Michel  
**REY**

recherche (2 caractères min)

[Recherche avancée ➔](#)

---

INVENTAIRE      CATALOGUES      INDEX

---

**PIÈCE N° REY17451203**

Type	Lettre
Identifiant	Rey17451203
Date de composition	1745-12-03
Certitude sur la date	haute
Date de réception	/
Expéditeur	Rey, Marc Michel
Destinataire	anonyme
Lieu d'envoi	Amsterdam
Lieu de réception	/
Adresse	/
Lieu de conservation	Haarlem, Noord Hollands archief, KNIW
Cote	inv. 32
Cote (copie)	/
Imprimé	/



The letter is written in French cursive. The signature 'M. Rey' is at the top. The text discusses sending a present to a friend in Paris and mentions a letter from a Mr. de L'Isle.

FIGURE A.4 – Accueil de l'édition numérique de la correspondance de Gustave Flaubert, Centre Flaubert

FIGURE A.5 – Lettre à Théophile Gauthier, Édition numérique de la correspondance de Gustave Flaubert, Centre Flaubert

FIGURE A.6 – Lettre Lionel Hauser, Édition numérique de la correspondance de Marcel Proust, Corr-Proust

The screenshot shows a digital edition of a letter from Marcel Proust to Lionel Hauser, dated Mardi soir [29 août 1916]. The interface includes a header with navigation links (Corr-Proust, Lettres, Présentation, Guides, Partenaires, Mentions légales, Crédits, Connexion, Inscription, FR) and a toolbar with search and zoom functions.

**Left Panel:** Shows a grid of thumbnail images of the handwritten letter. Below the thumbnails is a toolbar with icons for search, zoom, and other functions.

**Right Panel:**

- Header:** CP 03189 Marcel Proust à Lionel Hauser Mardi soir [29 août 1916]
- Section Headers:** Images, Transcription, Texte, Notes, Informations (highlighted in blue).
- Textual Content:**
  - Texte:** "Mardi soir 1
  - Notes:** "Mon cher Lionel"
  - Transcription:**

Je reçois à l'instant ta lettre  
et comme j'ai très mal aux yeux  
je me borne provisoirement à  
une réponse fort incomplète. Si  
les 3 jeunes hommes (pas si jeunes  
puisque je suis moi-même vieux  
hélas — et d'ailleurs sensiblement leur  
âgé) sont ceux que je suppose, je
- Bottom:** Logos for ILLINOIS, UGA, CEF, & ENS, and item.

## A.2 Attentes liées à chaque type de publication

FIGURE A.7 – Grille d'évaluation des publications numériques de corpus d'auteurs

Cette grille résume les attendus liés à chaque type de publication, même si des zones de superposition et de flottement entre les différents types d'édition y apparaissent clairement. Elle précise, en outre, si l'élément ou la caractéristique concernée est indispensable ou bien si son absence ne peut pas être considérée, malgré tout, comme un élément diminuant la valeur de l'édition. Deux codes sont employés, E désignant une exigence essentielle, et O une demande optionnelle.

	Type 1	Type 2	Type 3
<b>Traitement des sources</b> - Fac-similé (jpeg, pdf, tiff, etc. ; sauf restrictions liées aux droits) - Transcription (ou OCR, ou similaire), au format brut - Transcription selon des critères scientifiques établis par le projet (diplomatique, semi-diplomatique, normalisée)	E  E  O	O  E  E	O  O  E
<b>Métadonnées</b> <ul style="list-style-type: none"> <li>• format standard (normalisé)</li> <li>• métadonnées descriptives (bibliographiques, y. c. les responsabilités dans la création de l'édition)</li> <li>• métadonnées administratives (techniques, de droits, etc.)</li> <li>• métadonnées structurelles</li> <li>• enrichissements (annotation et/ou balisage)</li> </ul>	E  E  E  O  O	E  E  E  E  O	E  E  E  E  E
<b>Description du projet scientifique</b> <ul style="list-style-type: none"> <li>• enjeux scientifiques (motivation, apports, etc.)</li> <li>• présentation de l'équipe et des responsabilités de chacun</li> <li>• éventuellement, critères ayant accompagné le choix des sources (témoins)</li> <li>• présentation du corpus</li> <li>• présentation des critères de transcription</li> <li>• précisions sur le traitement des erreurs présentes dans la source (scribes, typographes, lectures antérieures, etc.)</li> <li>• choix de traitement de la ponctuation et des graphies</li> <li>• choix d'encodage</li> <li>• informations sur la genèse du texte</li> </ul>	E  E  O  O  E  O  O  O	E  E  E  O  E  O  O	E  E  E  O  E  E  O

FIGURE A.8 – Grille d'évaluation, suite

<b>Accessibilité</b>	E O	E O	E O
• version de lecture (interface minimale)	E	E	E
• possibilité de lire le texte en différentes versions	O	O	E
• possibilité de télécharger les sources			
◦ en format texte	O	O	O
◦ en format image (si libres de droits)	O	O	O
◦ en format xml	O	O	E
• possibilité de télécharger le texte en différents formats	O	O	O
<b>Plan de gestion des données</b>			
• usage de standards	E	E	E
• archivage pérenne	E	E	E
• présence d'un identifiant permanent (handle, ark, purl)	O	E	E
• programme de maintenance des données (périodicité, responsabilité)	E	E	E



## A.3 Ébauche d'une arborescence

FIGURE A.9 – Ébauche de l'accueil du futur site d'édition numérique de la correspondance de Frédéric Le Play, CRHXIX

**LOGO<sup>1</sup> DU CRHXIX**

**NOTA BENE :** Cette page peut être considérée comme le *container.html* (architecture qui se retrouve sur toutes les autres pages du site, pour les hauts et bas de page)

**Onglets haut de page<sup>2</sup>**

<b>A propos de Le Play</b>	<b>Correspondance de Le Play</b>	<b>Index</b>	<b>Recherche</b>	<b>Guides</b>	<b>Actualités</b>	<b>A propos de l'édition</b>
Sa vie, son œuvre	par ordre chronologique	des personnes	simple	abréviations	Le projet	
Ses correspondants	par correspondant	des lieux	avancée	glossaire <sup>3</sup>	L'équipe	
	par lieu de rédaction	des publications personnelles <sup>4</sup>	utilisateurs		Politique éditoriale	
	par lieu de conservation	des ouvrages cités			Partenaires et soutiens	
		leplayrien			Etc...	
		général ?				

**Corps de la page**

Définir ce qu'on y met.

De préférence une présentation du projet d'édition numérique.

Ou quelque chose de très simple, qui ne nécessite pas de scroller et permet une vue d'ensemble, comme pour l'édition de Flaubert. <https://flaubert.univ-rouen.fr/correspondance/edition/>

On peut y ajouter un bouton **d'appel à manuscrits**.

Pour l'instant, on part sur une édition numérique sans possibilité de connexion. Ce sera plus simple ainsi.

---

**Bas de page**

[Mentions légales](#)

[Crédits](#)

[Nous contacter](#)

+ **LOGOS DE SOUTIEN (Sorbonne etc.)**

<sup>1</sup> On pourrait créer un logo particulier pour le site... A voir. Ou reprendre celui du CRHXIX s'il en a un, tel quel ou adapté pour LP

<sup>2</sup> **AVERTISSEMENT : la nomination des onglets est importante pour le SEO ! Les reconSIDéRer en fonction, je n'ai pas le temps de m'y attarder.**

<sup>3</sup> Voir si l'on met ici un glossaire de termes leplaysiens que l'on distingue des index ou si on se contente des index (je serais plutôt pour un glossaire pour ma part).

<sup>4</sup> Voir si on le fusionne avec les ouvrages cités.

## Annexe B

### Transcription et Transkribus



## B.1 Les fac-similés

On peut voir ici quelques exemples des manuscrits que nous avons au CRHXIX. On peut constater les variations d'écriture de Le Play, les différentes qualités de numérisation et les nombreux fonds différents, recélant de la correspondance de Frédéric Le Play. Nous avons choisi de ne présenter que des lettres écrites par Le Play (excepté celle de Jules Baroche qui illustre nos propos du mémoire). Il faut cependant savoir qu'il existe bien mille lettres de correspondance passive, mais nous n'avons traité que la correspondance active donc c'est celle que nous avons choisi de mettre en valeur.

FIGURE B.1 – Lettre de Jules Baroche à Frédéric Le Play, BIF, Paris)

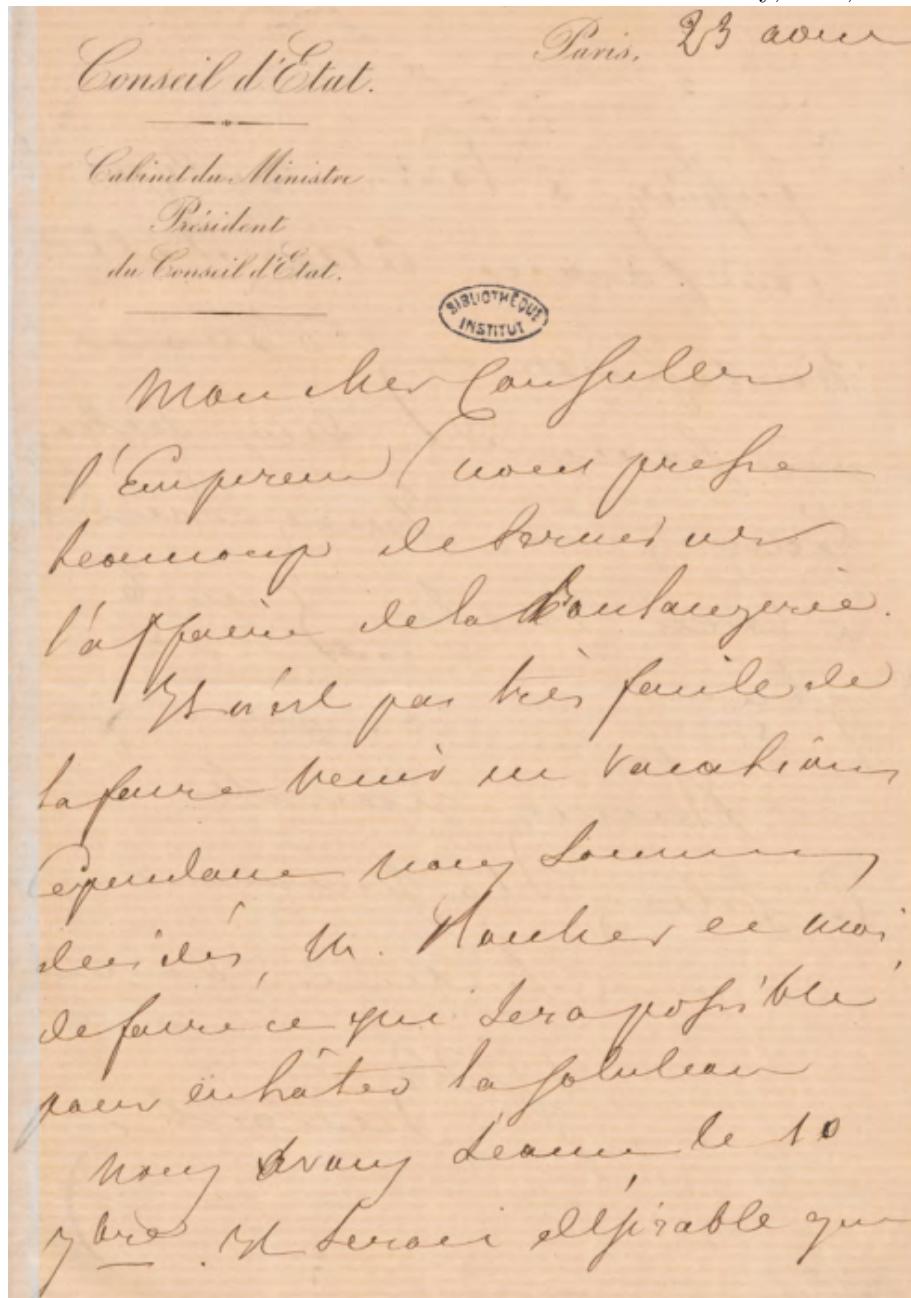


FIGURE B.2 – Lettre de Le Play à Louis de Kergorlay, 1864, Bibliothèque de l'Arsenal, Paris. Exemple typique d'un fac-similé médiocre.

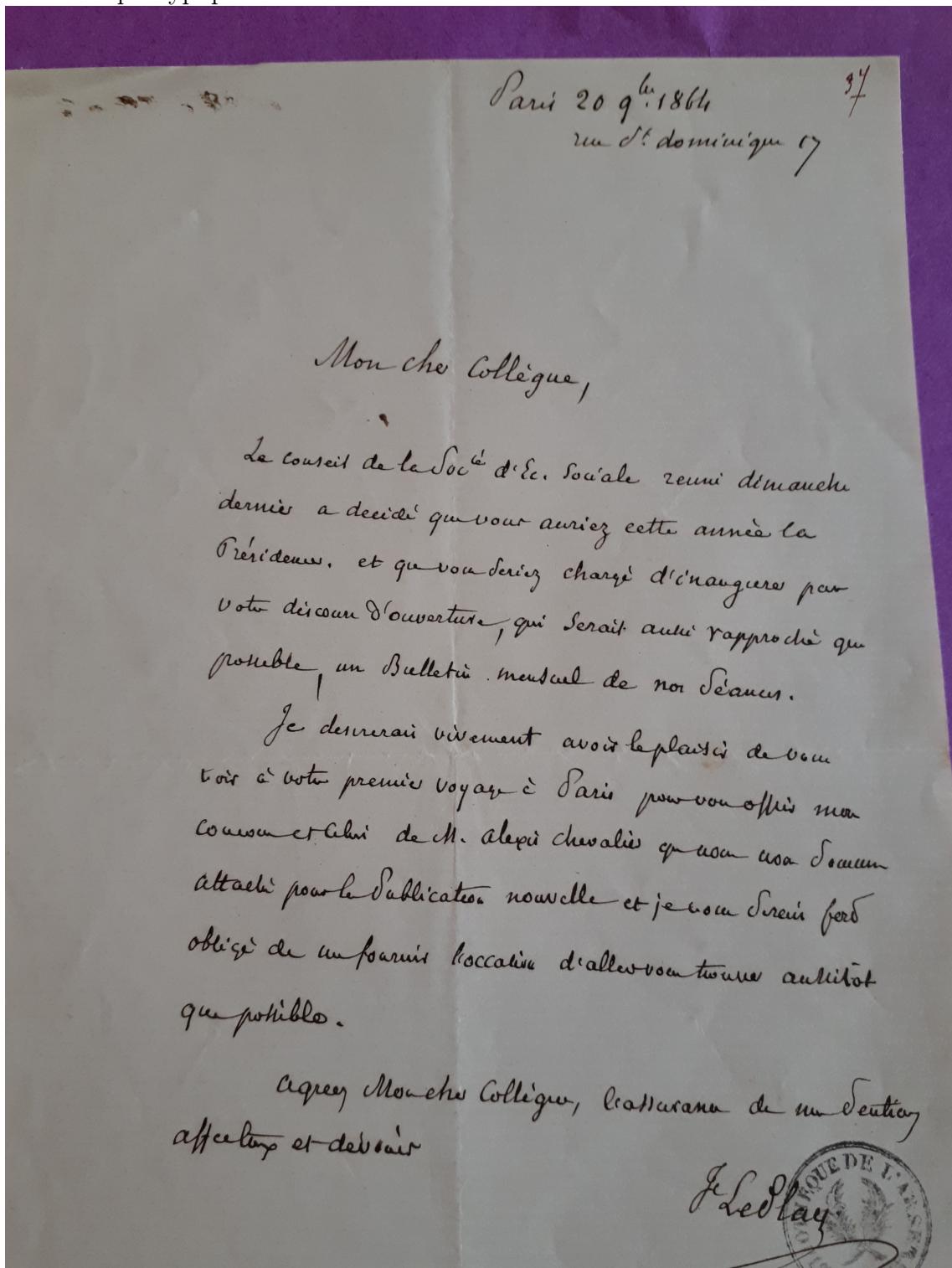


FIGURE B.3 – Lettre de Le Play à Ubaldino Peruzzi, 1857, BNC, FLorence. L'écriture de Le Play est plus penchée, liée et arrondie.

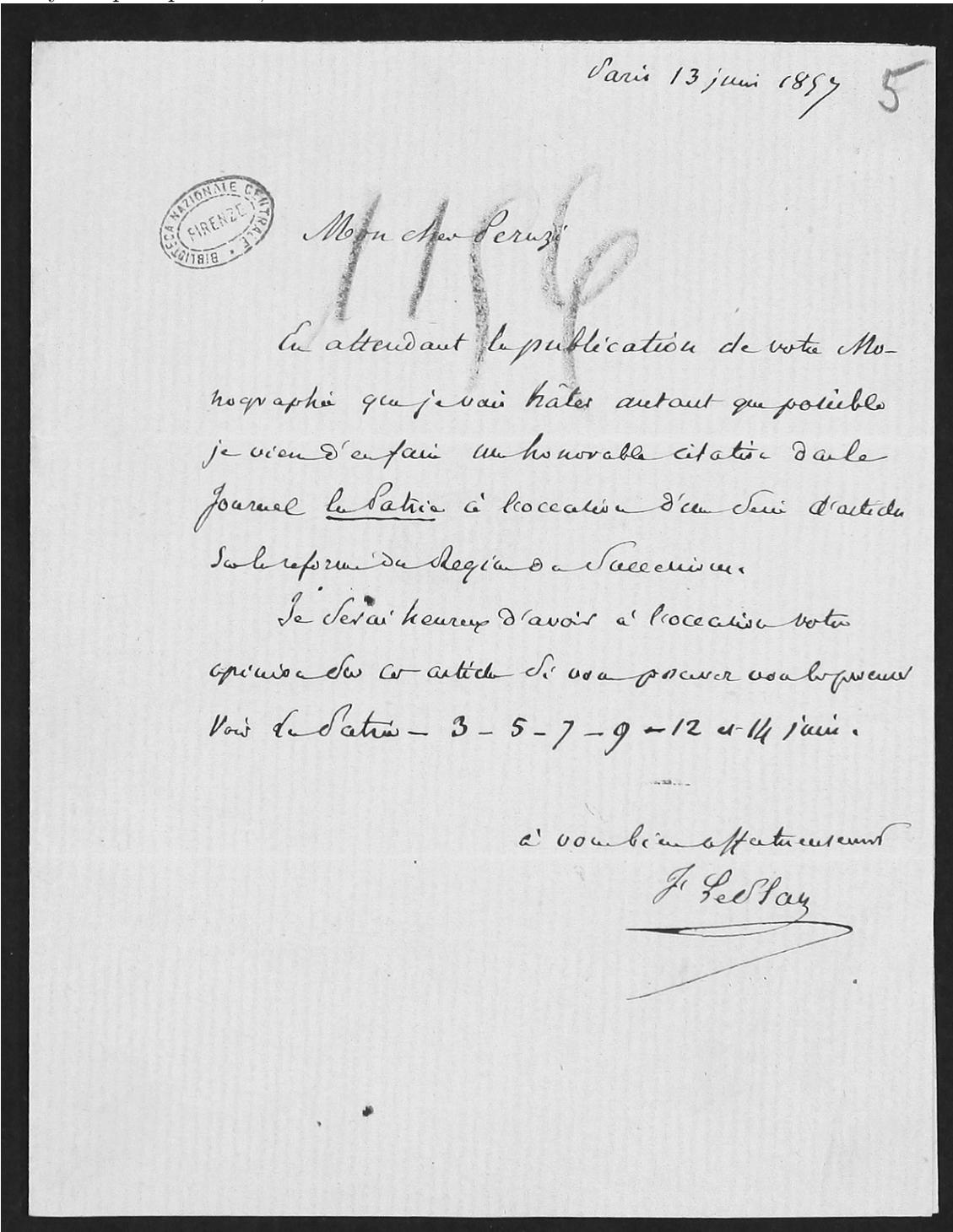


FIGURE B.4 – Lettre de Le Play à M<sup>gr</sup> Félix Dupanloup, 1873, BNF, Paris. L'écriture de Le Play est très appliquée, plus vieille également, le papier légèrement strié, la plume plus épaisse.

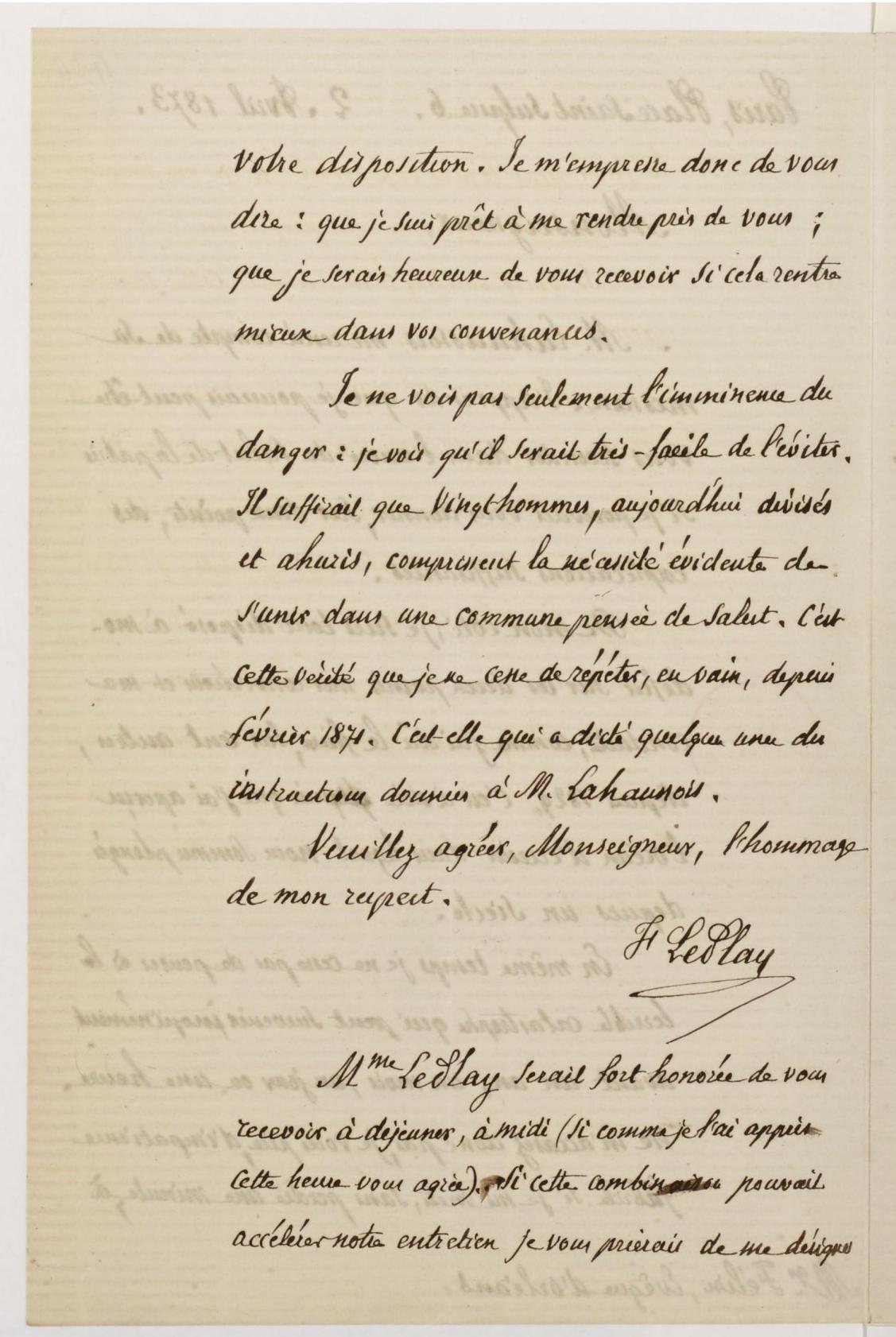


FIGURE B.5 – Lettre de Le Play à Frédéric de Mercey, 1856, BNF, Paris. L'écriture de Le Play est plus hâtive, la plume plus fine.

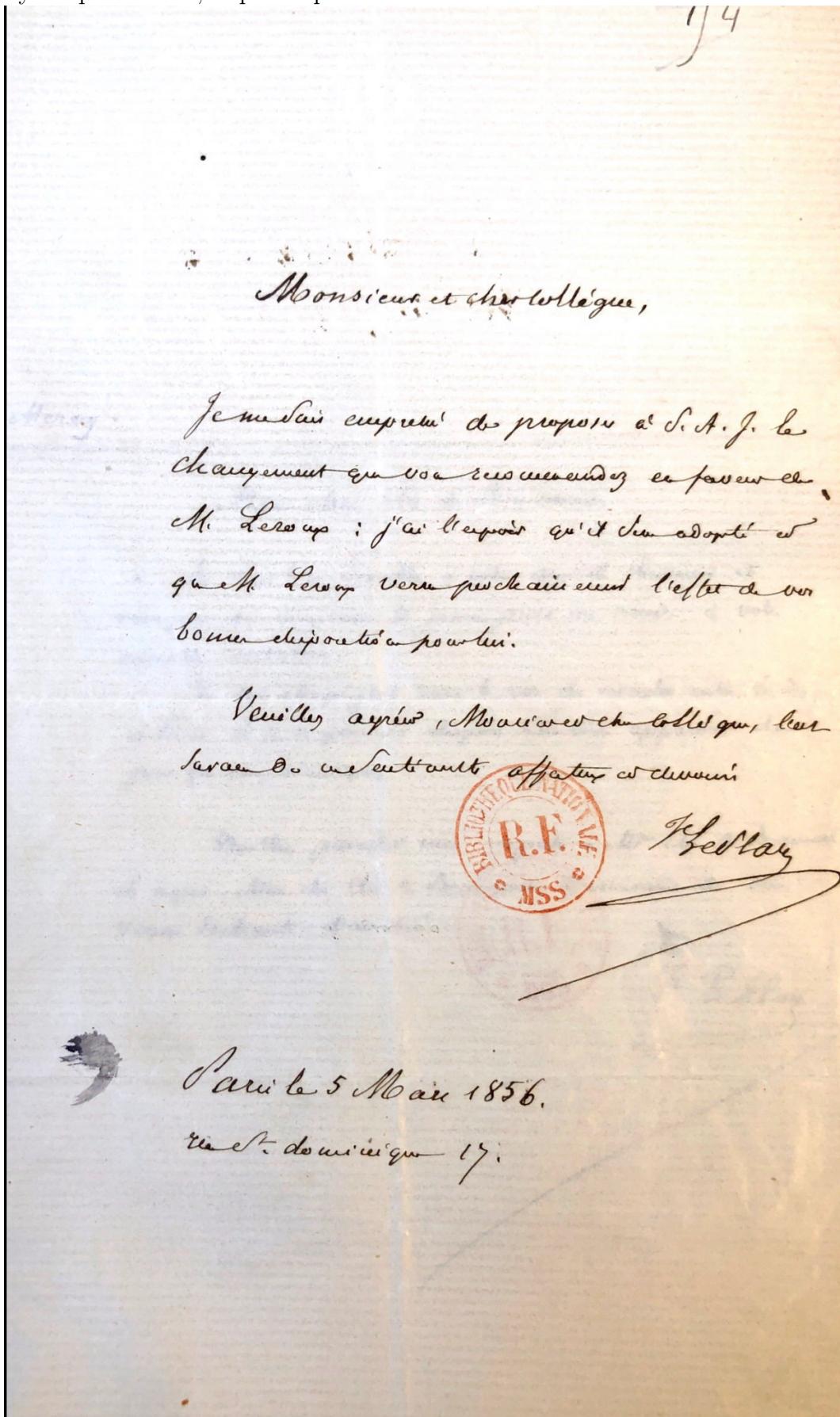


FIGURE B.6 – Lettre de Le Play son fils Albert, 1865, Château de Ligoure. Le « A » de Albert s'apparente à un « a » minuscule.

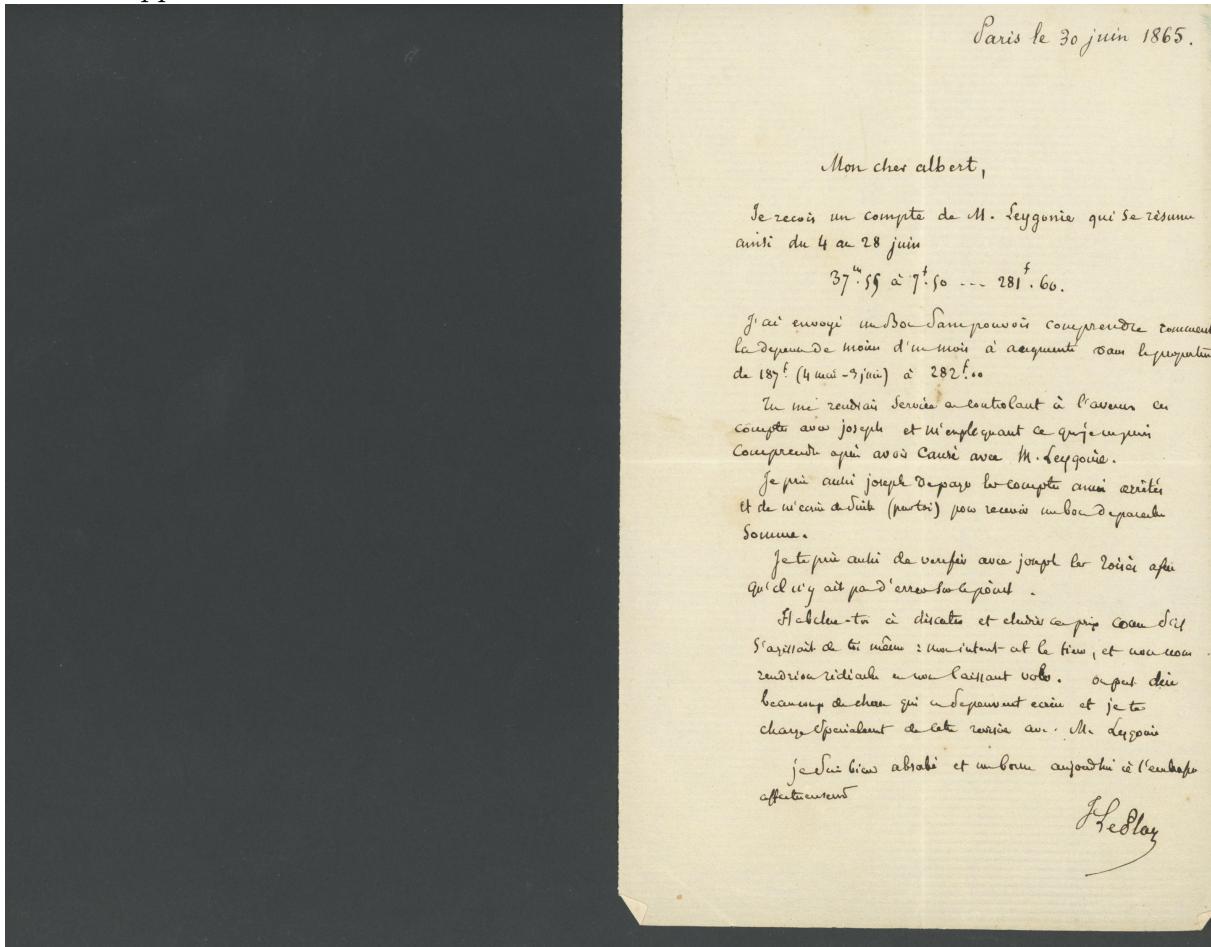


FIGURE B.7 – Lettre de Le Play Charles de Ribbe, 1869, Musée Arbaud, Aix-en-Provence.  
L'écriture est plus penchée, moins jeune.

Paris le 6 aout 1869  
Place Saint-Sulpice 6.

Mon cher ami,

J'ai été très sensible au témoignage de bon souvenir que vous m'avez donné dernièrement. J'ai appris avec plaisir que vous étiez revenu dans vos foyers, après votre exil au Nord. J'espère que vous y trouverez toute la satisfaction compatible avec les difficultés matérielles de la vie humaine.

Nous sommes d'accord depuis longtemps sur les vues de la Société actuelle. Il faut faire de se soustraire au découragement que tend à propager un si déplorable état de choses. Il faut se dire qu'en gagnant personnellement un homme à la vérité, on a peut-être jeté le base de la régénération complète de la race. Un haut fonctionnaire me disait un jour que si le concile avait mis à bas le régime social sous le coup de Louis Napoléon, la réforme serait

## B.2 Chargement des données d'entraînement

FIGURE B.8 – Problème de TR, capture d'écran de Transkribus

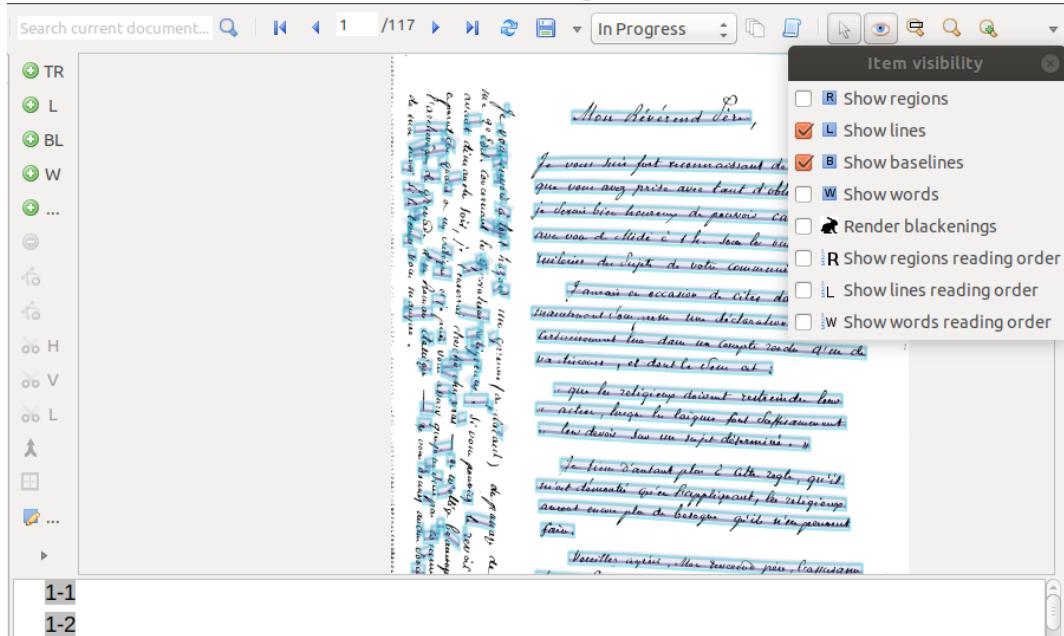
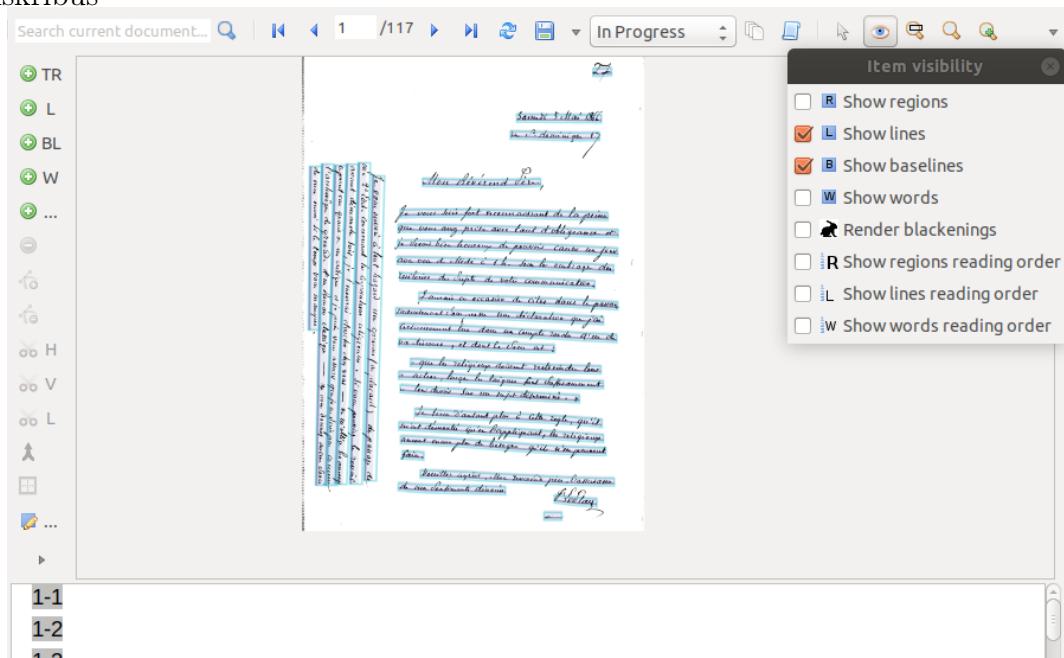


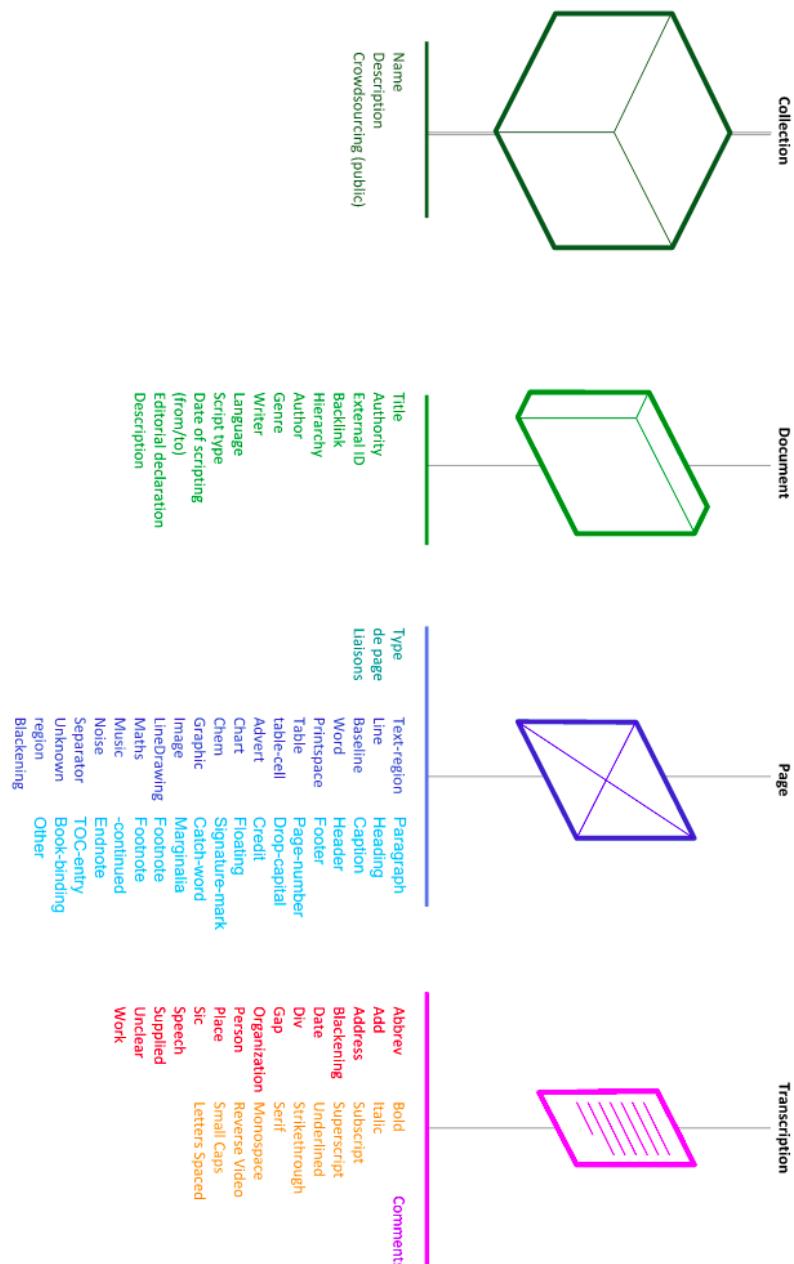
FIGURE B.9 – Résolution du problème de TR, deux sens d'écriture, capture d'écran de Transkribus



### B.3 Schématisation du modèle d'information de Transkribus

Schéma modèle Transkribus proposé par l'INHA<sup>1</sup>.

FIGURE B.10 – Le modèle de métadonnées Transkribus, INHA



1. Schéma modèle Transkribus, Site web de l'INHA, URL : <https://skylab.inha.fr/EditionsEnrichies/Documents/Schema-Modele-Transkribus.pdf> (visité le 18/06/2020)

FIGURE B.11 – La transcription diplomatique numérique via Transkribus, INHA

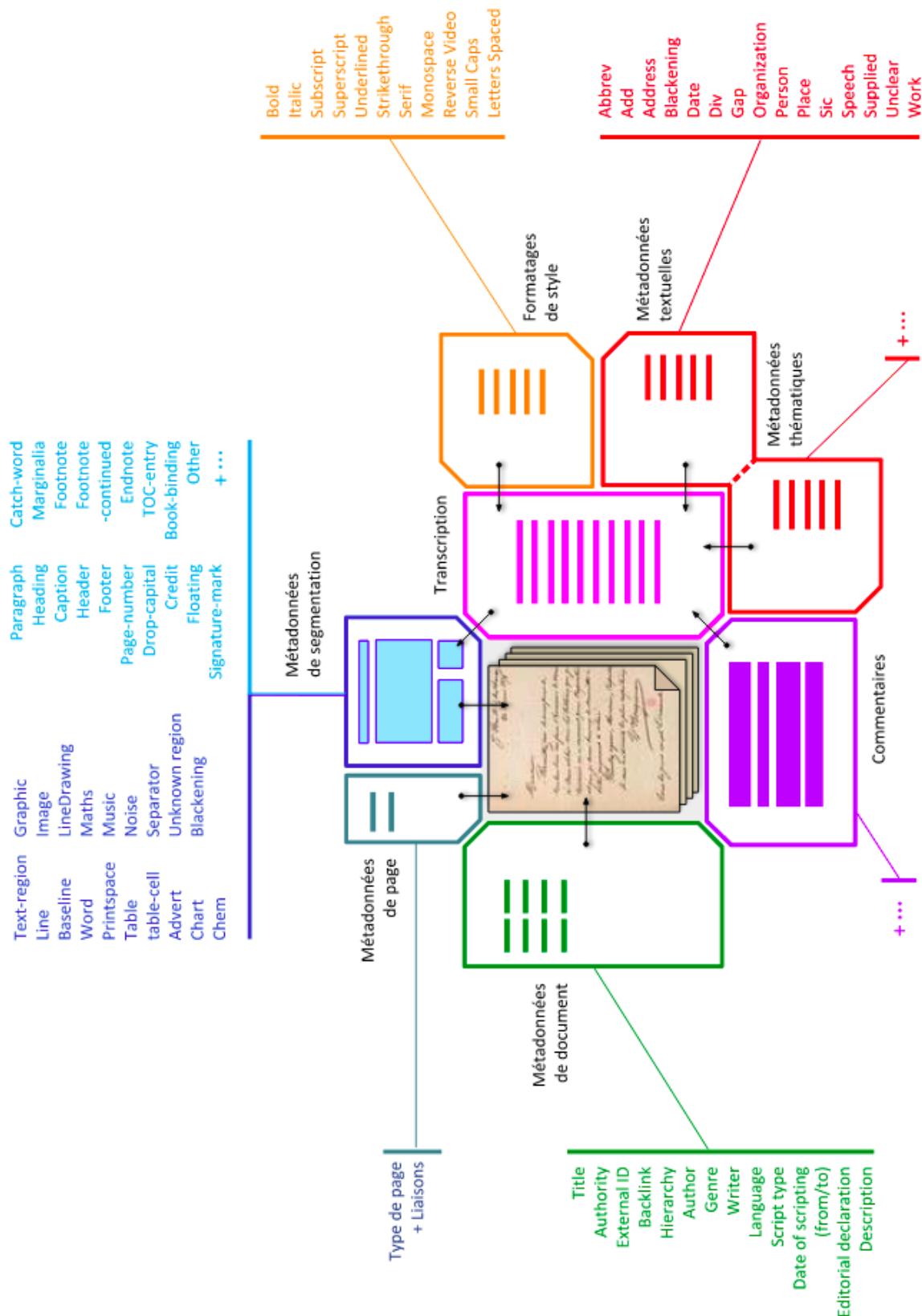


FIGURE B.12 – « Cheatsheet » métadonnées Transkribus, INHA

Champs de Document		Régions de Segmentation		Éléments de Segmentation	
<u>Title</u>	Titre de document	Text-region	Zone de texte	Paragraph	Paragraphe de texte
<u>Authority</u>	Institution	-> Line	Bloc de ligne d'écriture	Heading	Titre de paragraphe
<u>External ID</u>	Id. dans le système source	-> Baseline	Ligne de base d'écriture	Caption	Légende
<u>Backlink</u>	Lien vers info. de référence	Word	Mot	Header	Entêtes de page
<u>Hierachy</u>	Dossiers parents dans le système source	Printspace	...?	Footer	Pied de page
<u>Author</u>	Auteur du document	Table	Tableau	Page-number	Numéro de page
<u>Genre</u>	Genre du document	Table-cell	Cellule de tableau	Drop-capital	Lettrine
<u>Writer</u>	Scripteur du document	Advert	Annonce publicitaire	Credit	Crédits
<u>Language</u>	Langue du document	Chart	Diagramme	Floating	voir TEI <floatingText>
<u>Script type</u>	Type d'écriture [ handwritten   printed ]	Chem	Symbole chimique	Signature-mark	Cachet - Poinçon
<u>Date of scripting</u>	( from + to ) / Date d'écriture	Graphic	Graphisme	Catch-word	Slogan - Accroche
<u>Editorial declaration</u>	Déclaration éditoriale de la transcription	Image	Reproduction - Illustration ?	Marginalia	Note de marge
<u>Description</u>	Description du document	LineDrawing	...	Footnote	Note de bas de page
<b>Champs de Page</b>		Maths	Symbole mathématique	Footnote_continued	Note de bas de page suite
<u>Type</u>	Type de page [ Front-cover   Back-cover   Title   Table-of-contents   Index   Content   Blank   Other ]	Music	Notation musicale	Endnote	Note de fin
<u>Liaisons</u>	liaisons blocs et notes (à vérifier...)	Noise	Bruit	TOC-entity	Elément de sommaire
<b>Styles de Transcription</b>		Separator	Séparateur	Book-binding	Reliure
<u>Texte en gras</u>	<u>Reverse Video</u>	Unknown	Zone inconnue	Other	Autre
<u>Texte en italique</u>	<u>Small Caps</u>	<u>region</u>	<u>Blackening</u>	Caviardage	
<u>Texte en indice</u>	<u>Señí</u>				
<u>Texte en exposant</u>					
<u>Texte souligné</u>	<u>Monospace</u>				
<u>Strikethrough</u>	Texte barré				
<u>Letters Spaced</u>	Lettres espacées ?				
<b>Tags de Transcription</b>		<b>Tags de Transcription</b>			
<u>Bold</u>		<u>Abbrev</u>	Abbréviation → tei	<u>Person</u>	Personne identifiable → tei
<u>Italic</u>		<u>Add</u>	Insertion → tei	<u>Place</u>	Information géographique → tei
<u>Subscript</u>		<u>Address</u>	Adresse → tei	<u>Sic</u>	Sic → tei
<u>Superscript</u>		<u>Blackening</u>	Caviardage	<u>Speech</u>	Locuteur d'un texte dramatique → tei
<u>Underlined</u>		<u>Date</u>	Date → tei	<u>Supplied</u>	Texte fourni par le transcriveur → tei
		<u>Gap</u>	Transc. manquante → tei	<u>Unclear</u>	Texte illisible ou inaudible → tei
		<u>Organization</u>	Nom d'organisation → tei	<u>Work</u>	Références bibliographiques - œuvre

## Annexe C

# L'encodage en XML-TEI

### C.1 Les index

FIGURE C.1 – L'index de vocabulaire leplaysien, capture d'écran de l'ODD

```
<!-- Autres index --><textClass>
<keywords>
<!-- Index des termes leplaysiens -->
<list type="index_socio_lp">
<item>
  <name xml:id="corporation">Corporation</name>
<!-- Voir s'il faut mettre des majuscules pour les index -->
  <note/>
</item>
<item>
  <name xml:id="reforme">Réforme</name>
  <note/>
</item>
<item>
  <name xml:id="ref_morale">Réforme morale</name>
  <note/>
</item>
<item>
  <name xml:id="ref_sociale">Réforme sociale</name>
  <note/>
</item>
<item>
  <name xml:id="masses">Masses</name>
  <note/>
</item>
</list>
</keywords>
<keywords>
<!-- On peut faire d'autres index si nécessaires -->
<list>
<item>
  <name/>
  <note/>
</item>
</list>
</keywords>
</textClass>
```



## Annexe D

# Livrables du CRHXIX

Pour mieux se repérer dans les livrables, il est conseillé de consulter le fichier `arborescence_detaillee.txt` qui donne une arborescence détaillée des livrables, ou alors, pour embrasser tout dans une rapide vue d'ensemble, on consultera de préférence le fichier intitulé `arborescence.txt` qui ne fait pas le détail pièce par pièce.

### D.1 CRHXIX/rapport\_fin\_stage\_CRHXIX.pdf

Ces livrables ont été faits au cours de notre stage au CRHXIX qui recouvrait une période de 31,5 jours de travail. Dans l'ensemble, les 15 premiers jours ont été consacrés à la prise en main du projet et le travail sur Transkribus, les 15 autres ont été employés à l'établissement du cahier des charges et de récits utilisateurs, dans une forme d'ébauche, permettant ainsi d'arriver au schéma XML-TEI et à l'ODD.

Pour avoir une petite vue d'ensemble du travail réalisé au cours de ce stage, il sera bon de lire le fichier `rapport_fin_stage_CRHXIX.pdf` qui fait une synthèse des tâches accomplies.

### D.2 CRHXIX/1-inventaires

- Pour mieux comprendre ce qu'a été le travail de prise en main du projet, nous avons mis le fichier excel `Inventaire_cor_LP.xlsx` fait d'après le premier inventaire de 2005-2006 réalisé par Stéphane Baciocchi et Antoine Savoye dans *La correspondance de Le Play, une source pour l'histoire des sciences sociales*, Stéphane Baciocchi in *Les Études Sociales / Cairn.info*, 142-143-144 (II-2005-2006).
- Nous y joignons cet inventaire `Inventaire_LP_plain_text_gallica.odt` dans lequel nous avons fait mention des archives numériques en notre possession lors de ce stage.

A noter : ces documents sont encore à un stade de travail et devraient être perfectionnés. Ils ont surtout servi à la prise en main du projet à court terme.

FIGURE D.1 – Aperçu de l'inventaire de prise en mains du projet, capture d'écran, mai 2020.

Inventaire_cor_LP.xlsx - LibreOffice Calc						
	A	B	C	D	E	F
1		NUMERISATIONS	NUMERISATIONS	NUMERISATIONS	NUMERISATIONS	NUMERISATIONS
2	Dossier	AD Haute Savoie	BNF	BNF	BNF	BNF
3	Sous-dossier	BNF 1 et 3	BNF 2	BNF 4	BNF 5	BNF 6
4	Intitulé	Correspondance reçue de L	Correspondance reçue de L	Correspondance reçue de L	Correspondance reçue de L	Correspondance reçue de L
5	Destinataire	DESPINE Joseph		CHEVALIER Michel	?	QUATREFAGE DE BREAU Ar
6	Expéditeur	Le Play	Le Play/Ad.Blaise ?	Le Play	Le Play	Le Play
7	Notice biographique	Joseph Despine (1792-1859), ingénieur des mines orig		Michel Chevalier (1806-1879), ancien élève de l'école p	Armand de Quatrefage (18	Jean-Baptiste Landriot (181
8	Lieu de conservation/Archiv	Archives départementales c	BNF (Paris)	Bibliothèque de l'Arsenal (P	BNF (Paris)	BNF (Paris)
9	Cote	Fonds Despine 11 J 437	Naf 18165	Fonds Enfantin 7756 (93-L	Naf 28420	Naf 11824, ff. 141-142
10	Nombre de lettres	26	1	1	5	1
11	Nombre de pages numérisé	71 + 1 facture		1	18	4
12	Dates extrêmes	1839-1858	1865	1832	1875-1886	1876
13	Format de fichier	JPG	PDF	PDF	PDF	PDF

La couleur rouge mentionne les points obscurs et les données incertaines.

### D.3 CRHXIX/2-transcriptions

On trouve dans ce dossier un échantillon d'une transcription réalisée par la stagiaire M. F. On peut y lire en note de bas de page les autres suggestions faites à la lecture de certains mots qui ne nous semblaient pas cohérents selon le contexte. Nous avons fait cela très occasionnellement, le but de notre stage étant surtout de gérer la partie technique. On voit ici l'avantage de notre formation initiale en sciences humaines qui nous permet de comprendre le fond et d'avoir une réflexion à son sujet (et donc de pouvoir corriger certaines erreurs) alors que nous sommes en train de travailler sur la partie plus technique. Cela peut être également un frein car l'important reste aussi la gestion de la partie technique. On peut voir un exemple de ces ponctuelles corrections à la page 11 de cette transcription *LePlay\_Loyson\_MF.pdf*.

### D.4 CRHXIX/3-trankribus

Ce dossier comprend les tutoriels que nous avons réalisés pour la prise en main de Transkribus. Cela permettra à l'équipe du CRHXIX de continuer notre travail après

notre départ, afin d'éviter le « facteur d'autobus » (en anglais « *bus factor* ») provoqué par l'absence de partage d'informations et de compétences.

- Les membres de l'équipe du CRHXIX pourront donc, grâce au fichier `Tuto1_préparation_des_donnees_transkribus.pdf` apprendre à préparer les données en vue de l'entraînement d'un modèle expliqué dans le fichier `Tuto2_entraînement_du_modèle.pdf`.
- Le fichier `Tuto3_application_du_modèle.pdf` explique ensuite comment l'appliquer pour transcrire automatiquement. Ils pourront donc, s'ils estiment que c'est pertinent, outre le fait de continuer à entraîner le modèle Le Play, en créer d'autres sur les correspondants les plus importants. Et quoi qu'il en soit, ils pourront prendre en main Transkribus pour les simples transcriptions ce qui sera déjà un bon point pour le projet.
- Enfin, le fichier `Tuto4_exportation_transkribus.pdf` explique comment exporter les données du serveur Transkribus à notre ordinateur. En effet, il est important de garder la main sur nos données.

En guise de complément à ce tutoriel, se trouve dans le dossier `export_test` un test d'export des données de Transkribus vers l'ordinateur. Il est à noter que les fichiers ne sont pas encore nommés dans la forme qu'ils prendront, telle qu'on a pu la constater dans le dossier numérisation.

- On trouve ainsi dans le dossier `BNF9enJPG`, comprenant les lettres de Le Play à Monseigneur Félix Dupanloup, un export des fichiers transcrits en XML-TEI avec les numérisations correspondantes.
- De même pour les lettres de Le Play à Charles de Ribbe `LP_à_Charles_de_Ribbe1` : on trouve toujours un fichier XML-TEI regroupant l'ensemble des lettres, ici `LP_à_Charles_de_Ribbe_te1.xml`, puis dans le dossier `page` un fichier XML-TEI par page transcrise (et non par lettre transcrise). Il est intéressant de constater ici le test qui a été fait : nous avons utilisé, notamment à la page 4, les *tags* Transkribus et nous avons ensuite exporté en XML pour voir ce que cela donnerait. Or, ces tags se traduisent par des balises et attributs qui ne sont visibles que dans le fichier XML général et non par page. Par ailleurs, leur nommage est souvent invalide. Tout ceci est développé dans le mémoire.
- Le dossier `LP_à_Charles_de_Ribbe2` comprend l'export d'une sélection plus restreinte de lettres mais avec des *tags* plus poussés. Il a été fait plus récemment. Il comprend également un export en `.txt`, ce qui permet une comparaison avec l'export `.xml`.
- Dans le dossier `Set_entraînement_le_play`, l'export a été fait dans plusieurs formats. Pour les fichiers `.docx`, `.pdf`, `.xlsx`, cela n'a pas fonctionné. En revanche, les

formats `.txt` et `.xml` ne présentent pas d'anomalie. Or, c'est ce qui nous intéresse donc l'expérience s'avère concluante.

Le dossier Transkribus comprend également un rapport réalisé sur l'apport de Transkribus au projet. Ce rapport `Point_Transkribus.pdf`, rédigé à la moitié de notre stage, a été transmis au CRHXIX en juin. Il a été partiellement repris dans notre mémoire avec des réflexions mises à jour selon le recul que nous avons actuellement.

## D.5 CRHXIX/4-cahier\_des\_charges

Une partie importante de la mise en pratique du projet a été l'établissement d'un cahier des charges. C'est un bien grand mot étant donné que nous n'avons pas eu le temps de le mener à terme. C'est donc plutôt un rapport menant à la réalisation d'un futur cahier des charges qui sera vraiment réalisé par la suite.

En complément de cette ébauche de cahier des charges, on trouve un fichier excel nommé `users_stories_LP_V3.ods` dans lequel sont écrits des récits utilisateurs ou *users stories* permettant de mieux cerner les attentes des futurs visiteurs et utilisateurs de l'édition numérique et donc d'adapter notre édition à leurs besoins. Là encore, ces US gagneraient à être améliorés.

Nota : à chaque fois, on trouve le nommage V2 ou V3 à la fin des noms de fichiers. Cela signifie que nous avons fait plusieurs versions et que nous mettons la dernière (V1 = version 1, V2 = version 2, V3 = version 3 et ainsi de suite).

## D.6 CRHXIX/5-site

Il a fallu penser le site de l'édition numérique. C'est encore une ébauche de réflexion qui sera continuée par un membre de l'équipe du CRHXIX.

On trouve ici :

- Dans le dossier `pages` les pages du site telles qu'elles ont été pensées pour l'édition numérique de la correspondance de Le Play.
- Le fichier `architecture_site_LP.JPG` présente l'architecture. Les numéros qui y sont inscrits renvoient aux fichiers du dossier `page`

## D.7 CRHXIX/6-index

Une fois que l'on s'est fait une idée plus précise des attentes des utilisateurs et des besoins du site, il s'agit d'entreprendre la mise en oeuvre technique pour répondre à ces attentes. Un des points importants est la réalisation d'index. Il a été convenu pour les choix éditoriaux que nous ferions plusieurs index : un index pour les noms de personnes, un

pour les noms de lieux, un pour les grands événements, un pour les noms d'organisation, un pour les ouvrages cités, un pour les termes leplaysiens ou tout au moins sociologiques. Pour ce dernier, il a été nécessaire de mener une réflexion plus importante pour savoir en amont quels termes nous indexerions. Pour cela, nous avons sélectionné certains mots qui nous paraissaient importants dans les lettres que nous avions déjà croisées, et nous avons soumis cet index au Professeur Antoine Savoye. Le fichier nommé **Vocabulaire leplaysien - mots à ajouter.doc** contient les mots que nous lui avons soumis et qui ont été retenus. Il ne nous a pas paru intéressant de mettre les autres index dans les livrables. En revanche, nous les avons mis dans leur forme provisoire, tels qu'ils seront utiles pour l'encodage TEI dans le dossier suivant.

## D.8 CRHXIX/7-index\_tei

Ce dossier comprend les index en cours de constitution, tels qu'ils apparaîtront dans les lettres encodées en XML-TEI. À chaque index correspond un fichier recensant les noms dans leur forme finale en XML-TEI. Cela permettra à la personne qui prendra la suite, d'une part d'avoir un modèle à suivre, d'autre part de pouvoir lister chaque nom rencontré et de se contenter de faire un copier/coller lors de l'encodage suivant.

On trouve donc ici :

- l'index des noms d'événement : **index\_evenements\_tei.pdf**
- l'index des termes leplaysiens et sociologiques : **index\_leplaysien\_tei.pdf**
- l'index des noms de lieu : **index\_lieux\_tei.pdf**
- l'index des noms d'organisation : **index\_organisations\_tei.pdf**
- l'index des noms d'ouvrages : **index\_ouvrages\_tei.pdf**
- l'index des noms de personnes et personnages : **index\_personnes\_tei.pdf**

En outre, une recension des formes normalisées de correspondants à renseigner dans le **<teiHeader>** se trouve dans le fichier intitulé **index\_correspondants\_teiHeader.pdf**.

On aurait pu faire un index pour aider les transcripteurs à remplir les informations sur les origines du document, mais nous avons estimé que le fichier **xxxx\_base.xml** (voir le point 8 sur la tei) fait pour chacun des documents suffisait.

## D.9 CRHXIX/8-tei

Dans ce dossier se trouvent les premiers essais d'encodage qui ont permis de réaliser ensuite l'ODD.

- le dossier **lp\_dupanloup\_xml** comprend les premiers essais d'encodage en XML-TEI.

Ils sont nommés respectivement **lp\_dupanloup\_bnf-11.xml** et **lp\_dupanloup\_bnf-12.xml**

(`lp` pour leplay, on indique toujours l'auteur des lettres en premier ; `dupanloup` pour le destinataire, `bnf` pour le lieu de conservation du fonds, `1xxx` pour le numéro de la lettre. Ici, les lettres 1 et 2). Le fichier `lp_dupanloup_bnf_base.xml` comprend tous les éléments communs aux manuscrits écrits de Le Play à Dupanloup. Il sert de base à l'encodage de chacune des lettres.

- même principe pour le dossier `lp_ribbe_arbaud_xml`
- le fichier `essai_tei_correspondance_CRHXIXv2.xml` est un essai (très perfectible) d'encodage qui a aidé à la réflexion d'encodage en vue de l'ODD.
- le fichier `Tuto_XML-TEI_CHRXIX.pdf` est un tutoriel qui a été réalisé pour le CRHXIX par nos soins, afin d'aider l'équipe du Centre dans les futurs encodages.

## D.10 CRHXIX/9-odd

Les réflexions sur l'encodage en XML-TEI aboutissent à l'ODD (*One Document Does it all*). Ce dossier comprend :

- l'ODD en html `ODD_LEPLAY_V1.html`
- un dossier `LP_testODD` qui contient les test réalisés sur un fichier XML lié à l'ODD par un schéma relaxNG.

L'ODD est perfectible. En quinze jours travaillés, nous n'avons pas eu le temps d'avoir une vue assez complète des milliers de lettres en notre possession, mais c'est déjà un bon point de départ pour l'encodage, et nous pensons qu'il ne devra pas être beaucoup modifié. Mais il devra l'être un peu, c'est inévitable.

## Annexe E

# Livrables du Labex OBVIL

Le stage au Labex Obvil recouvre une totalité de 20 jours travaillés.

Le fichier `rapport_fin_stage_OBVIL.pdf` est le rapport qui a été écrit en fin de stage dans le but d'informer le chef de projet, Monsieur Glenn Roe, du travail accompli.

On trouvera sur Github les rapports qui ont été faits au quotidien à partir d'une certaine date, et une bonne partie des fichiers réalisés : [https://github.com/OBVIL/elicom/tree/master/extraction\\_cor\\_stage2020](https://github.com/OBVIL/elicom/tree/master/extraction_cor_stage2020). Néanmoins, il nous a paru plus clair d'en faire une sélection ici et de l'expliquer de la même manière que nous l'avons fait pour le CRHXIX, afin de garder une cohérence de moyens.

### E.1 Les fichiers XML-TEI

Le dossier intitulé `extraction_cor_stage2020` rassemble l'ensemble des livrables d'Obvil.

- `extraction_cor_stage2020/cor_lamartine` : les fichiers en lien avec la correspondance d'Alphonse de Lamartine. Total de 97 fichiers XML-TEI extraits et corrigés.
- `extraction_cor_stage2020/cor_lamennais` : ceux qui traitent de la correspondance de l'Abbé Félicité de Lamennais. Total de 204 fichiers. Certains sont nommés a, b, c car ils ont dû être divisés a posteriori. Nous avons corrigé 80 fichiers. Il faut encore corriger 124 fichiers, à savoir les fichiers 164 à 195, 2 à 9, 17 à 99.
- `extraction_cor_stage2020/cor_proudhon` : ceux concernant Pierre-Joseph Proudhon. Total de 87 fichiers extraits et corrigés.

### E.2 Organisation des dossiers

Les dossiers sont toujours organisés de la même manière au sein de chaque dossier `cor_nomDuCorrespondant`, reflet de ce travail plus systématique. Dans chacun on trouve les dossiers suivants :

### E.2.1 corpus

Il rassemble d'une part le PDF extrait sur Gallica et dont nous nous servons pour repérer la mise en page et les passages versifiés afin de bien encoder, d'autre part l'extraction de l'OCR de ce même volume, en HTML. L'HTML a été corrigé par nos soins pour en assurer la validité.

### E.2.2 *script*

Comme son nom l'indique, il renferme le *script* Python qui permet de passer de l'OCR HTML à un fichier XML par lettre. Le squelette est commun pour les auteurs, mais adapté aux variations de chacun. Il est de qualité inégale. Le temps nous a manqué pour parfaire le premier *script* réalisé (Lamartine).

### E.2.3 dump

Il contient les extractions XML corrigées (sauf pour Lamennais où seulement une partie est corrigée, comme indiqué dans le rapport).

### E.2.4 remarques

Sur Github, on peut voir les rapports écrits au jour le jour mais qui n'ont pas grand intérêt. Nous avons mis ici les fiches faites au fur et à mesure de notre travail, mais qui ont un aspect souvent de réflexion et d'ébauche, car ils ont été faits selon le cours de nos réflexions, qui sont peut-être plus avancées aujourd'hui. Certains des questionnements ont été résolus. On trouve également des index de noms normalisés pour permettre de faire par la suite des copier/coller et avancer plus vite dans la correction des fichiers XML-TEI.

# Bibliographie



# Le Labex OBVIL et le CRHXIX

*Appel à projets Emergence*, Sorbonne Université, URL : [https://candidature.sorbonne-universites.fr/index.php?option=com\\_emundus&view=programme&id=111&Itemid=1521&lang=fr](https://candidature.sorbonne-universites.fr/index.php?option=com_emundus&view=programme&id=111&Itemid=1521&lang=fr) (visité le 07/09/2020).

*CRHXIX : 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 18/06/2020).

*CRHXIX : 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 18/06/2020).

*CRHXIX : Les activités de recherches 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 18/06/2020).

*CRHXIX : Présentation du Centre*, Site web de l'Université Paris-1 Panthéon Sorbonne, URL : <https://www.pantheonsorbonne.fr/unites-de-recherche/crhxix/aboutthecenter/> (visité le 18/06/2020).

*Observatoire de la vie littéraire*, Agence nationale de la recherche, URL : <https://anr.fr/ProjetIA-11-LABX-0059> (visité le 04/09/2020).

*Présentation d'OBVIL*, Observatoire de la vie littéraire, URL : <http://obvil.sorbonne-universite.site/obvil/presentation> (visité le 04/09/2020).



# Frédéric Le Play

*CRHXIX : Axe 2 : Du moral au social : pratiques et théories de l'enquête. Autour des archives du mouvement leplaysien.* Site web de l'Université Paris-1 Panthéon Sorbonne, URL : <https://www.pantheonsorbonne.fr/unites-de-recherche/crhix/about-us/research-activities/area2/> (visité le 15/06/2020).

SAVOYE (Antoine), *LE PLAY FRÉDÉRIC (1806-1882)*, dans, URL : <http://www.universalis-edu.com.janus.bis-sorbonne.fr/encyclopedie/frederic-le-play/> (visité le 15/06/2020).

SAVOYE (Antoine), « Frédéric Le Play en quelques dates », dans *Frédéric Le Play : Parcours, Audience, Héritage*, dir. Fabien Cardoni, Paris, 2013, p. 279-289.

STÉPHANE BACIOCCHI (Antoine Savoye), « La correspondance de Le Play, une source pour l'histoire des sciences sociales », *Les Études sociales* n° 142-144 (2005), p. 231-247, URL : <https://gallica.bnf.fr/ark:/12148/bpt6k9767323c/f1n284.texteBrut> (visité le 07/05/2020).

TODD (Emmanuel), *L'origine des systèmes familiaux . Les familles dans l'Histoire*, Herodote.net, URL : [https://www.herodote.net/Les\\_familles\\_dans\\_l\\_Histoire-article-1287.php](https://www.herodote.net/Les_familles_dans_l_Histoire-article-1287.php) (visité le 08/09/2020).



# L'édition numérique de correspondance

*About EADH - The European Association for Digital Humanities*, EADH, URL : <https://eadh.org/about> (visité le 23/09/2020).

*Accueil*, EMAN, URL : <http://eman-archives.org/EMAN/> (visité le 11/09/2020).

*Accueil*, Juliette Drouet, Lettres à Victor Hugo, URL : <http://www.juliettedrouet.org/lettres/#.X3MsdYY69uQ> (visité le 02/07/2020).

*Accueil*, Lettres et textes : Le Berlin intellectuel des années 1800, URL : <https://www.berliner-intellektuelle.eu/?fr> (visité le 19/05/2020).

*Accueil*, Le blog d'Huma-Num et de ses consortiums, URL : <https://humanum.hypotheses.org/6089> (visité le 19/06/2020).

*Association Interdisciplinaire de Recherches sur l'Epistolaire*, Epistolaire.org, URL : <http://www.epistolaire.org/> (visité le 10/09/2020).

CHAPRON (Emmanuelle), *Journée d'études sur l'édition numérique de correspondances du consortium CAHIER : quelques réflexions*, Archives savantes des Lumières, URL : <https://seguier.hypotheses.org/211> (visité le 02/09/2020).

*Correspondance de Flaubert*, Université de Rouen, URL : <https://flaubert.univ-rouen.fr/correspondance/edition/> (visité le 17/06/2020).

*Correspondance Jean Paulhan*, Observatoire de la vie littéraire, URL : <http://obvil.sorbonne-universite.site/corpus/paulhan/> (visité le 10/09/2020).

*CORREZ, Édition des lettres internationales adressées à Émile Zola*, EMAN, URL : <http://eman-archives.org/CorrespondanceZola/> (visité le 11/09/2020).

*Edition de la correspondance et des archives de Marc Michel Rey*, Marc Michel Rey, URL : <http://rey.huma-num.fr/presentation> (visité le 10/09/2020).

*Édition numérique de la correspondance de D'Alembert*, D'Alembert en toutes lettres, URL : <http://dalembert.academie-sciences.fr/Correspondance/> (visité le 10/09/2020).

*L'édition numérique*, dans *Wikipédia*, Page Version ID : 173657501, 2020, URL : [https://fr.wikipedia.org/w/index.php?title=%5C%C3%89dition\\_num%C3%C3%5C%A9rique&oldid=173657501](https://fr.wikipedia.org/w/index.php?title=%5C%C3%89dition_num%C3%C3%5C%A9rique&oldid=173657501) (visité le 03/09/2020).

*La correspondance inédite du géomètre Gaspard Monge (1746-1818)*, EMAN, URL : <http://eman-archives.org/monge/> (visité le 10/09/2020).

- LA MENNAIS (Félicité de), *Correspondance*, 2 t., Didier, Paris, 1863, URL : <http://gallica.bnf.fr/ark:/12148/bpt6k9761794z> (visité le 08/09/2020).
- LAMARTINE (Alphonse de), *Correspondance de Lamartine*, 6 t., Hachette, Furne, Jouvet et Cie, Paris, 1873, URL : <http://gallica.bnf.fr/ark:/12148/bpt6k5805303r> (visité le 19/05/2020).
- Le consortium*, Consortium Cahier, URL : <https://cahier.hypotheses.org/le-consortium> (visité le 19/06/2020).
- Le Projet*, Corr-Proust, URL : <http://proust.elan-numerique.fr/presentation/project> (visité le 08/06/2020).
- PROUDHON (Pierre-Joseph), *Correspondance*, 7 t., Slatkine, Genève, 1971, URL : <http://gallica.bnf.fr/ark:/12148/bpt6k55985> (visité le 08/09/2020).
- RAGEOT (Laurence), *Projets d'édition numériques de correspondances : approches et spécificités – journée d'études du groupe Correspondance*, Consortium Cahier, URL : <https://cahier.hypotheses.org/2172> (visité le 02/09/2020).
- SAND (George), *Correspondance : 1812-1876*, 6 t., C. Lévy, Paris, 1883, URL : <http://gallica.bnf.fr/ark:/12148/bpt6k2065433> (visité le 19/05/2020).
- SINATRA (Michaël E.) et VITALI-ROSATI (Marcello), « Introduction », dans *Pratiques de l'édition numérique*, dir. Michael E. Sinatra, Montréal, 2014, p. 7-11, URL : <http://books.openedition.org/pum/308> (visité le 03/09/2020).
- Pratiques de l'édition numérique*, dir. Marcello Vitali-Rosati, Montréal, 2014 (Parcours numérique), URL : <http://books.openedition.org/pum/306> (visité le 03/09/2020).
- WALTER (Richard (dir.)), « L'édition numérique de correspondances – guide méthodologique », *Consortium Cahier* (), URL : <https://cahier.hypotheses.org/guide-correspondance> (visité le 26/06/2020).

# L'apprentissage machine, OCR et HTR

*An introduction to Machine Learning*, GeeksforGeeks, URL : <https://www.geeksforgeeks.org/introduction-machine-learning/> (visité le 20/09/2020).

*Artificial intelligence vs Machine Learning vs Deep Learning*, GeeksforGeeks, URL : <https://www.geeksforgeeks.org/artificial-intelligence-vs-machine-learning-vs-deep-learning/> (visité le 20/09/2020).

*Intelligence artificielle*, dans *Wikipédia*, Page Version ID : 174966188, 2020, URL : [https://fr.wikipedia.org/w/index.php?title=Intelligence\\_artificielle&oldid=174966188](https://fr.wikipedia.org/w/index.php?title=Intelligence_artificielle&oldid=174966188) (visité le 20/09/2020).

*ML / Introduction to Data in Machine Learning*, GeeksforGeeks, URL : <https://www.geeksforgeeks.org/ml-introduction-data-machine-learning/> (visité le 02/06/2020).

*Mode texte et OCR*, BNF. Gallica, URL : <https://gallica.bnf.fr/edit/und/consulter-les-documents> (visité le 21/09/2020).

POUPEAU (Gauthier), *Open Data, Big data, Data Mining, Module data ENC M2TNAH*, 2019.

*Reconnaissance de l'écriture manuscrite*, dans *Wikipédia*, Page Version ID : 171738712, 2020, URL : [https://fr.wikipedia.org/w/index.php?title=Reconnaissance\\_de\\_l%27%C3%A9criture\\_manuscrite&oldid=171738712](https://fr.wikipedia.org/w/index.php?title=Reconnaissance_de_l%27%C3%A9criture_manuscrite&oldid=171738712) (visité le 02/09/2020).

TUFFÉRY (Stéphane), *Data mining et statistique décisionnelle - 4ème édition*, 4e édition, Paris, 2012.



# Transkribus

- Accueil*, Transkribus, URL : <https://transkribus.eu/Transkribus/> (visité le 06/03/2020).
- CHAGUÉ (Alix), *Constituer un corpus pour la fouille de texte - de la transcription des documents d'archives à l'annotation : exploration d'une méthodologie par l'ANR Time Us*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Vincent Jolivet et Éric de la Clergerie, École nationale des chartes, 2018.
- Entraînement d'un modèle dans Transkribus*, Transkribus Wiki, URL : <https://docplayer.fr/177163907-Entrainement-d-un-modele-dans-transkribus.html> (visité le 22/05/2020).
- Handwritten Text Recognition Workflow*, Transkribus Wiki, URL : [https://transkribus.eu/wiki/index.php/Handwritten\\_Text\\_Recognition\\_Workflow](https://transkribus.eu/wiki/index.php/Handwritten_Text_Recognition_Workflow) (visité le 23/09/2020).
- How To enrich transcribed documents with mark-up*, Transkribus Wiki, URL : [https://transkribus.eu/wiki/images/e/e8/How\\_to\\_enrich\\_transcribed\\_documents\\_with\\_mark-up.pdf](https://transkribus.eu/wiki/images/e/e8/How_to_enrich_transcribed_documents_with_mark-up.pdf) (visité le 26/09/2020).
- How to transcribe. Train a model*, Transkribus Wiki, URL : [https://transkribus.eu/wiki/images/3/34/HowToTranscribe\\_Train\\_A\\_Model.pdf](https://transkribus.eu/wiki/images/3/34/HowToTranscribe_Train_A_Model.pdf) (visité le 23/09/2020).
- Questions and Answers*, Transkribus Wiki, URL : [https://transkribus.eu/wiki/index.php/Questions\\_and\\_Answers](https://transkribus.eu/wiki/index.php/Questions_and_Answers) (visité le 23/09/2020).
- Schéma modèle Transkribus*, Site web de l'INHA, URL : <https://skylab.inha.fr/EditionsEnrichies/Documents/Schema - Modele - Transkribus.pdf> (visité le 18/06/2020).
- SCHLAGDENHAUFFEN (Régis), *Comment utiliser Transkribus en 10 étapes (voire moins)*, EHESS, URL : <http://regis-schlagdenhauffen.eu/wp-content/uploads/2018/01/Comment-utiliser-Transkribus-%C2%80%C93-en-10-%C3%A9tapes-ou-moins.pdf> (visité le 22/05/2020).
- Tag (métadonnée)*, dans Wikipédia, Page Version ID : 169495366, URL : [https://fr.wikipedia.org/w/index.php?title=Tag\\_\(m%C3%A9tadonn%C3%A9e\)&oldid=169495366](https://fr.wikipedia.org/w/index.php?title=Tag_(m%C3%A9tadonn%C3%A9e)&oldid=169495366) (visité le 26/09/2020).



# L'encodage, XML-TEI et ODD

- About our web service. The idea behind correspSearch, correspSearch, URL : <https://correspsearch.net/index.xql?id=about&l=en> (visité le 09/09/2020).*
- Accueil, TGB, OBVIL et BNF, URL : <http://obvil.lip6.fr/tgb/> (visité le 08/09/2020).*
- BURNARD (Lou), « La TEI et le XML », dans *Qu'est-ce que la Text Encoding Initiative ?*, Marseille, 2015, URL : <http://books.openedition.org/oep/1298> (visité le 26/09/2020).
- *Comment maîtriser le tigre TEI*, URL : <https://cahier.hypotheses.org/files/2018/08/ODD-diapos.pdf> (visité le 11/06/2020).
- CAMPS (Jean-Baptiste), *ODD Structuration des données et des documents : balisage XML. Personnaliser la TEI : One Document Does it all*, 2017.
- DUFOURNAUD (Nicole (dir.)), *Manuel d'encodage TEI Renaissance et temps modernes*, BVH, URL : [http://www.bvh.univ-tours.fr/XML-TEI/ManuelWeb/Manuel\\_TEI\\_BVH.html](http://www.bvh.univ-tours.fr/XML-TEI/ManuelWeb/Manuel_TEI_BVH.html) (visité le 11/09/2020).
- DUFOURNAUD (Nicole) et GRATSAC LEGENDRE (Valérie), *Manuel d'encodage XML-TEI - édition numérique de manuscrits baroques*, 2012, URL : <https://hal.archives-ouvertes.fr/hal-00718043> (visité le 28/07/2020).
- EHRMANN (Maud), *Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*, Theses, Paris Diderot University, 2008, URL : <https://hal.archives-ouvertes.fr/tel-01639190> (visité le 07/09/2020).
- Encoding Correspondence. A Manual for encoding letters and postcards in TEI-XML and DTABf*, Encoding Correspondence, URL : <https://encoding-correspondence.bbaw.de/v1/> (visité le 05/05/2020).
- GALLERON (Ioana), DEMONET (Marie-Luce), MEYNARD (Cécile), FATIHA (Idmhand), PIERAZZO (Elena), WILLIAMS (Geoffrey), ROGER (Julia) et BUARD (Pierre-Yves), *Les publications numériques de corpus d'auteurs - Guide de travail, grille d'analyse et recommandations*, Huma-Num, 2018, URL : <https://halshs.archives-ouvertes.fr/halshs-01932519> (visité le 05/05/2020).
- Les publications numériques de corpus d'auteurs – Guide de travail, grille d'analyse et recommandations*, Consortium Cahier, URL : <https://cahier.hypotheses.org/guides-juridiques/les-publications-numeriques-de-corpus-dauteurs> (visité le 09/09/2020).

*Mise à disposition de la Très Grande Bibliothèque du Labex OBVIL / api.bnf.fr*, Site web de la BNF, URL : <http://api.bnf.fr/mise-disposition-de-la-tres-grande-bibliotheque-du-labex-obvil> (visité le 04/09/2020).

*Perspectives of the further development of the Correspondence Metadata Interchange Format (CMIF)*, digiversity, URL : <https://digiversity.net/2015/perspectives-of-the-further-development-of-the-correspondence-metadata-interchange-format-cmif/> (visité le 09/09/2020).

PIERAZZO (Elena), *Why Do We Encode ?*, Youtube, URL : [https://www.youtube.com/watch?v=R0ncI\\_rr1z4&list=PL77mHK9JuenN9NXeXQbVcU0Rz7HZk-9Pv&index=2](https://www.youtube.com/watch?v=R0ncI_rr1z4&list=PL77mHK9JuenN9NXeXQbVcU0Rz7HZk-9Pv&index=2) (visité le 04/05/2020).

PINCHE (Ariane), *CoursTNAH\_XML-TEI Scéance 1*, GitHub, URL : [https://github.com/ArianePinche/coursTNAH\\_XML-TEI](https://github.com/ArianePinche/coursTNAH_XML-TEI) (visité le 09/10/2019).

— *CoursTNAH\_XML-TEI, scéance 3*, GitHub, URL : [https://github.com/ArianePinche/coursTNAH\\_XML-TEI](https://github.com/ArianePinche/coursTNAH_XML-TEI) (visité le 11/09/2020).

*SIG : Correspondence*, TEI Wiki, URL : <https://wiki.tei-c.org/index.php/SIG:Correspondence> (visité le 09/09/2020).

*TEI : Correspondence SIG*, Text Encoding Initiative, URL : <https://tei-c.org/Activities/SIG/Correspondence/> (visité le 09/09/2020).

*TEI Guidelines*, Text Encoding Initiative, URL : <https://tei-c.org/guidelines/> (visité le 10/09/2020).

*TEI SIG on Correspondence – Minutes Rome, Oct 3, 2013*, Text Encoding Initiative, URL : <https://tei-c.org/activities/sig/correspondence/tei-sig-on-correspondence-minutes-rome-oct-3-2013/> (visité le 09/09/2020).

# Le cahier des charges, la conception du site et le SEO

*CC BY-NC-ND 3.0 FR*, Creative Commons, URL : <https://creativecommons.org/licenses/by-nc-nd/3.0/fr/> (visité le 18/09/2020).

*Comment optimiser l'architecture d'un site web pour le SEO ?*, La fabrique du net, URL : <https://www.lafabriquedunet.fr/seo/articles/optimiser-architecture-site-web-seo/> (visité le 25/06/2020).

*Exemple de cahier des charges*, Cahiers des charges, URL : <https://cahiersdescharges.com/exemple-cahier-des-charges-pdf/> (visité le 12/06/2020).

*Facteur d'autobus*, dans Wikipédia, Page Version ID : 174361862, URL : [https://fr.wikipedia.org/w/index.php?title=Facteur\\_d%27autobus&oldid=174361862](https://fr.wikipedia.org/w/index.php?title=Facteur_d%27autobus&oldid=174361862) (visité le 28/09/2020).

FOUCARD (Jean-Louis), *Manager un projet Numérique, Module M2 TNAH, ENC*, 2020.  
*Récit utilisateur*, dans Wikipédia, Page Version ID : 172279518, 2020, URL : [https://fr.wikipedia.org/w/index.php?title=R%C3%A9cit\\_utilisateur&oldid=172279518](https://fr.wikipedia.org/w/index.php?title=R%C3%A9cit_utilisateur&oldid=172279518) (visité le 24/08/2020).

*SEO et Webdesign : 9 quick wins à intégrer*, Emarketing aux Petits Oignons, URL : <https://www.emarketing-aux-petits-oignons.com/seo-et-webdesign> (visité le 18/09/2020).



# Les humanités numériques et le numérique en général

*Accueil*, data.bnf, URL : <https://data.bnf.fr/> (visité le 10/09/2020).

BOULÉTREAU (Viviane) et HABERT (Benoît), « Les formats », dans *Pratiques de l'édition numérique*, Montréal, 2014, p. 145-159, URL : <http://books.openedition.org/pum/329> (visité le 03/09/2020).

*Format de données*, dans *Wikipédia*, Page Version ID : 173750621, URL : [https://fr.wikipedia.org/w/index.php?title=Format\\_de\\_donn%C3%A9es&oldid=173750621](https://fr.wikipedia.org/w/index.php?title=Format_de_donn%C3%A9es&oldid=173750621) (visité le 10/09/2020).

*Framework*, dans *Wikipédia*, Page Version ID : 173493836, URL : <https://fr.wikipedia.org/w/index.php?title=Framework&oldid=173493836> (visité le 10/09/2020).

*GeoNames*, geonames.org, URL : <https://www.geonames.org/> (visité le 10/09/2020).

*GitHub*, dans *Wikipédia*, Page Version ID : 174838124, URL : <https://fr.wikipedia.org/w/index.php?title=GitHub&oldid=174838124> (visité le 28/09/2020).

*HTML (HyperText Markup Langage) : définition, traduction*, Journal du Net, URL : <https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1203255-html-hypertext-markup-langage-definition-traduction/> (visité le 21/09/2020).

*Humanités numériques*, dans *Wikipédia*, Page Version ID : 173210622, 2020, URL : [https://fr.wikipedia.org/w/index.php?title=Humanit%C3%A9s\\_num%C3%A9riques&oldid=173210622](https://fr.wikipedia.org/w/index.php?title=Humanit%C3%A9s_num%C3%A9riques&oldid=173210622) (visité le 03/09/2020).

*Manifeste des Digital humanities*, Hypotheses, URL : <https://tcp.hypotheses.org/318> (visité le 29/09/2020).

MOUNIER (Pierre), « Les Humanités numériques, gadget ou progrès ? », *Revue du Crieur*-7 (2017), p. 144-159, URL : <https://www.cairn.info/revue-du-crieur-2017-2-page-144.htm> (visité le 02/09/2020).

*Normes et standards*, éduscol, URL : <https://eduscol.education.fr/numerique/dossier/archives/metadata/normes-et-standards> (visité le 10/09/2020).

*Online regex tester and debugger : PHP, PCRE, Python, Golang and JavaScript, regular expressions101*, URL : <https://regex101.com/> (visité le 10/02/2020).

*Rich Text Format*, dans Wikipédia, Page Version ID : 174007898, 2020, URL : [https://fr.wikipedia.org/w/index.php?title=Rich\\_Text\\_Format&oldid=174007898](https://fr.wikipedia.org/w/index.php?title=Rich_Text_Format&oldid=174007898) (visité le 02/09/2020).

SINATRA (Michaël E.) et VITALI-ROSATI (Marcello), « Histoire des humanités numériques », dans *Pratiques de l'édition numérique*, Montréal, 2014, p. 49-60, URL : <http://books.openedition.org/pum/317> (visité le 03/09/2020).

*versionnage*, Wiktionnaire, URL : <https://fr.wiktionary.org/wiki/versionnage> (visité le 30/09/2020).

# *Packages Python utilisés*

*Beautiful Soup 4.4.0 : documentation, URL : <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>* (visité le 19/05/2020).

*lxml - Processing XML and HTML with Python, URL : <https://lxml.de/>* (visité le 19/05/2020).

*Regular expression operations, URL : <https://docs.python.org/3/library/re.html>* (visité le 19/05/2020).



# Logiciels, services et outils utilisés

*Github*, URL : <https://github.com> (visité le 03/06/2020).

*Oxygen XML Editor*, URL : <https://www.oxygenxml.com/> (visité le 19/05/2020).

*ShareLatex*, URL : <https://fr.sharelatex.com/> (visité le 02/09/2020).

*Sublime Text*, URL : <https://www.sublimetext.com/> (visité le 19/05/2020).

*Transkribus*, URL : <https://transkribus.eu/Transkribus/> (visité le 14/05/2020).

*Transkribus*, URL : <https://transkribus.eu/Transkribus/> (visité le 14/05/2020).



# Glossaire

- **Bibliothèque numérique** : Ensemble organisé de documents nativement numériques ou numérisés accessibles à distance par internet.
- **Format** : Manière normalisée de représenter des données ou des fichiers sous la forme d'informations binaires.
- **ID** : *IDentifier*, abréviation anglaise de identification ou identifiant, il sert à identifier un objet précis dans un ensemble d'objets.
- **Interface graphique** : Par opposition à l'interface en ligne de commande, l'interface graphique désigne la manière dont est présenté un logiciel à l'écran pour l'utilisateur, permettant l'interaction entre l'humain et la machine. Une interface graphique bien conçue est ergonomique et intuitive afin que l'utilisateur la comprenne tout de suite.
- **Module** : En Python, un fichier pouvant contenir des fonctions, des classes et des données, et pouvant être importé dans un script.
- **ODD** (*One Document Does it all*) : Langage de définition et de maintenance du système TEI. Il permet la maintenance du code et de sa documentation d'une manière intégrée, à partir d'une seule source XML. Il se compose d'un schéma formel (utilisant un langage informatique tel que DTD, RELAX NG, W3C Schema, Schematron) pour contrôler l'édition et d'une documentation explicant aux utilisateurs ou développeurs les principes éditoriaux et choix de balises.
- **Python** : Langage de programmation informatique à usage général, multi-plateforme et *open-source*.
- **Package, library** : Un ensemble de modules contenant des outils tels que des fonctions. Pour être utilisé, il doit être importé entièrement ou partiellement, par module.
- **Parser** : Processus d'analyse d'un élément textuel le rendant intelligible par la machine, sous la forme d'un encodage numérique.
- **Script** : Un script désigne un programme, entier ou extrait, chargé d'exécuter une action prédéfinie quand un utilisateur réalise une action ou qu'une page web est en cours d'affichage sur un écran. Il s'agit d'une suite de commandes simples et souvent

peu structurées qui permettent l'automatisation de certaines tâches successives dans un ordre donné.

- **Standard** : Texte de référence reconnu, documenté et élaboré par un groupe de travail spécialisé, visant à harmoniser l'activité d'un secteur donné. Pour XML, les standards prennent la forme de schémas et de règles de balisage permettant de créer des documents de structures comparables au sein d'un même standard.
- **TEI** : *Text Encoding Initiative* - Standard de description de documents textuels pour XML. Développé par le TEI Consortium.
- **Transkribus** : Plateforme de transcription automatique de textes manuscrits. Fondé sur l'intelligence artificielle, le moteur de reconnaissance de texte manuscrit (Handwritten Text Recognition ou HTR) doit être préparé avec des données d'apprentissage, obtenues par la transcription d'une centaine de pages minimum, en établissant la correspondance ligne à ligne entre l'image du texte numérisé et sa transcription.
- **User Story** : Un récit utilisateur, ou « *user story* » en anglais, est une description simple d'un besoin ou d'une attente exprimée par un utilisateur et utilisée dans le domaine du développement de logiciels et de la conception de nouveaux produits pour déterminer les fonctionnalités à développer.
- **VIAF** : *Virtual International Authority File* ou Fichier d'autorité international virtuel, est un fichier d'autorité international servant à identifier les personnes ou les collectivités contenues dans d'autres fichiers d'autorité
- **Wiki** : Application web dont le contenu peut être édité par les visiteurs, ce qui permet la création et la modification des pages de manière collaborative. Il est généralement dédié à un projet ou à une thématique précise.
- **XML** : *eXtensible Markup Language* - Un langage de balisage générique permettant de décrire des informations de manière organisée et standardisée. Le XML est une recommandation du W3C [<https://www.w3.org/XML/>].
- **XSLT** : *Extensible Stylesheet Language Transformations* - Un langage basé sur XML permettant de styliser ou transformer des fichiers XML ou HTML.

# Table des figures

2.1	Carte mentale de Le Play, <i>Encyclopædia Universalis</i> . . . . .	18
3.1	Fonds numérisés, extrait de l'inventaire . . . . .	27
4.1	Fonctionnement du <i>correspSearch - Site web du correspSearch</i> en 2020 . . .	38
4.2	Lettre d'Adolf von Buch à Louis de Beausobre (Magdebourg, 15 janvier 1761), <i>Lettres et textes : Le Berlin intellectuel des années 1800</i> . . . . .	43
5.1	Tableau récapitulatif . . . . .	54
6.1	Les licences . . . . .	62
6.2	Processus de travail d'encodage . . . . .	67
7.1	<i>Machine learning</i> et IA . . . . .	70
8.1	L'OCR de George Sand sous format TXT . . . . .	75
8.2	L'OCR de Lamartine en HTML sous Oxygen XML Editor . . . . .	76
8.3	Exemple d'une balise <span> dans l'HTML du premier volume de Lamennais, l. 7805 . . . . .	78
8.4	Exemple d'une balise <span> manquante dans l'HTML du premier volume de Lamennais, l. 7825 . . . . .	78
8.5	Exemple d'une balise <span> superflue dans l'HTML du premier volume de Lamennais, l. 1264 . . . . .	78
8.6	Exemple des titres polluant le texte, HTML de Lamartine . . . . .	79
8.7	Exemple d'une regex pour enlever certains titres, HTML de Lamartine . .	79
8.8	Exemple d'une regex pour enlever des <p>, HTML de Lamartine . . . . .	80
8.9	Cahier des charges ELICOM, repérage des marqueurs de Lamennais . . . . .	82
9.1	Les données dans l'apprentissage machine . . . . .	89
9.2	Résolution du problème de TR, deux sens d'écriture, capture d'écran de Transkribus . . . . .	91
9.3	Problème de TR à cause du papier, capture d'écran de Transkribus . . . . .	92
9.4	F. Le Play au R. P. Hyacinthe Loysen, 1866, capture d'écran de Transkribus	94

9.5	F. Le Play au R. P. Hyacinthe Loyson, 1866, capture d'écran de Transkribus	94
9.6	F. Le Play au R. P. Hyacinthe Loyson, 1867, capture d'écran de Transkribus	95
9.7	F. Le Play au R. P. Hyacinthe Loyson, 1866, capture d'écran du manuscrit	96
9.8	Mise en place du premier entraînement du modèle Le Play . . . . .	97
9.9	<i>Learning Curve</i> du premier entraînement du modèle Le Play . . . . .	98
9.10	<i>Learning Curve</i> du dernier entraînement du modèle Le Play . . . . .	99
9.11	Détails sur le jeu de données . . . . .	100
9.12	Lettre de F. Le Play au R. P. Loyson, 1870 . . . . .	102
9.13	Lettre de F. Le Play à Keele . . . . .	103
9.14	Lettre de F. Le Play à Peruzzi, 1881 . . . . .	103
9.15	Pour une page, 21 <i>tags</i> . . . . .	106
10.1	Capture d'écran de l'ODD du CRHXIX, le <pb> . . . . .	117
10.2	Capture d'écran de l'ODD du CRHXIX, l'<opener> . . . . .	118
10.3	Capture d'écran de l'ODD du CRHXIX, l'attribut @place . . . . .	119
10.4	Des fautes dans un fichier XML (154), Félicité de Lamennais, ELICOM. Capture d'écran d'Oxygen XML Editor et du PDF de l'édition originale. . . . .	120
10.5	Faux paragraphes dans un fichier XML (159), Félicité de Lamennais, ELI-COM. Capture d'écran d'Oxygen XML Editor et du PDF de l'édition originale. . . . .	120
10.6	Capture d'écran d'Oxygen XML Editor, pointer vers un index . . . . .	122
10.7	Capture d'écran d'Oxygen XML Editor, l'index des ouvrages . . . . .	123
10.8	La normalisation du <correspAction>, correspondance de Lamartine ( <i>lamartine-col-vol1-1</i> ). capture d'écran d'Oxygen XML Editor . . . . .	125
10.9	L'ODD . . . . .	127
10.10	Extrait de la table des matières de l'ODD pour le projet Le Play . . . . .	127
10.11	Extrait du <i>script extraction-elicom.py</i> , capture d'écran de Sublime Text	129
10.12	Extrait du <i>script</i> de Proudhon, capture d'écran de Github . . . . .	130
10.13	Lettre de Lamennais, notes mal matchées à cause de l'océrisation, <i>lamennais-cor-vol1-112.xml</i>	
10.14	Mise en place des <1g> et des <1> dans la correspondance de Lamartine . . . . .	132
11.1	Rapport sur le code Python de Lamartine, capture d'écran de GitHub . . . . .	135
A.1	<i>D'Alembert en toutes lettres</i> . Édition numérique de la correspondance de D'Alembert . . . . .	147
A.2	Édition numérique de la correspondance de Jean Paulhan, Labex OBVIL . . . . .	148
A.3	Édition numérique de la correspondance de Marc Michel Rey, HUMA-NUM	148
A.4	Accueil de l'édition numérique de la correspondance de Gustave Flaubert, Centre Flaubert . . . . .	149

A.5 Lettre à Théophile Gauthier, Édition numérique de la correspondance de Gustave Flaubert, Centre Flaubert . . . . .	149
A.6 Lettre Lionel Hauser, Édition numérique de la correspondance de Marcel Proust, Corr-Proust . . . . .	150
A.7 Grille d'évaluation des publications numériques de corpus d'auteurs . . . . .	151
A.8 Grille d'évaluation, suite . . . . .	152
A.9 Ébauche de l'accueil du futur site d'édition numérique de la correspondance de Frédéric Le Play, CRHXIX . . . . .	154
B.1 Lettre de Jules Baroche à Frédéric Le Play, BIF, Paris) . . . . .	156
B.2 Lettre de Le Play à Louis de Kergorlay, 1864, Bibliothèque de l'Arsenal, Paris. Exemple typique d'un fac-similé médiocre. . . . .	157
B.3 Lettre de Le Play à Ubaldino Peruzzi, 1857, BNC, Florence. L'écriture de Le Play est plus penchée, liée et arrondie. . . . .	158
B.4 Lettre de Le Play à Mgr Félix Dupanloup, 1873, BNF, Paris. L'écriture de Le Play est très appliquée, plus vieille également, le papier légèrement strié, la plume plus épaisse. . . . .	159
B.5 Lettre de Le Play à Frédéric de Mercey, 1856, BNF, Paris. L'écriture de Le Play est plus hâtive, la plume plus fine. . . . .	160
B.6 Lettre de Le Play son fils Albert, 1865, Château de Ligoure. Le « A » de Albert s'apparente à un « a » minuscule. . . . .	161
B.7 Lettre de Le Play Charles de Ribbe, 1869, Musée Arbaud, Aix-en-Provence. L'écriture est plus penchée, moins jeune. . . . .	162
B.8 Problème de TR, capture d'écran de Transkribus . . . . .	163
B.9 Résolution du problème de TR, deux sens d'écriture, capture d'écran de Transkribus . . . . .	163
B.10 Le modèle de métadonnées Transkribus, INHA . . . . .	164
B.11 La transcription diplomatique numérique via Transkribus, INHA . . . . .	165
B.12 « Cheatsheet » métadonnées Transkribus, INHA . . . . .	166
C.1 L'index de vocabulaire leplaysien, capture d'écran de l'ODD . . . . .	167
D.1 Aperçu de l'inventaire de prise en mains du projet, capture d'écran, mai 2020. . . . .	170



# Table des matières

Résumé	iii
Remerciements	v
Liste des sigles et abréviations	vii
Introduction	3
<b>I Des projets portés par des institutions culturelles</b>	<b>9</b>
<b>1 Un contexte universitaire</b>	<b>11</b>
1.1 Le Centre de Recherche et d'Histoire du XIX <sup>e</sup> siècle . . . . .	11
1.1.1 Une institution dédiée à la recherche autour du XIX <sup>e</sup> siècle . . . . .	11
1.1.2 Un Centre dynamique . . . . .	12
1.1.3 Les acteurs du projet . . . . .	12
1.1.4 Partenaires susceptibles d'être mobilisés . . . . .	12
1.1.5 Soutiens financiers . . . . .	13
1.2 Le Labex OBVIL . . . . .	13
1.2.1 Un laboratoire d'excellence pour les humanités numériques . . . . .	13
1.2.2 De nombreux partenaires . . . . .	14
1.2.3 Les acteurs du projet . . . . .	14
<b>2 Deux projets ambitieux</b>	<b>15</b>
2.1 L'édition numérique de la correspondance de Frédéric Le Play . . . . .	15
2.1.1 Au service de l'histoire des sciences sociales . . . . .	15
2.1.2 Redécouvrir l'un des fondateurs de la sociologie . . . . .	17
2.2 ELICOM . . . . .	19
2.2.1 Pour une recherche collective et multidisciplinaire . . . . .	19
2.2.2 Un outil de réflexion et de recherche . . . . .	20
2.2.3 Trois modules . . . . .	20

<b>3 Les sources des projets : des correspondances du XIX<sup>e</sup> siècle aux formes variées</b>	<b>23</b>
3.1 La mise en valeur de manuscrits . . . . .	23
3.1.1 Trois fonds familiaux principaux . . . . .	24
3.1.2 Des fonds dispersés à travers l'Europe . . . . .	25
3.1.3 Une correspondance au service de l'Histoire . . . . .	26
3.1.4 Nature des sources et première prise en main du projet . . . . .	27
3.2 De l'édition papier à l'édition numérique . . . . .	28
3.2.1 Gallica, une mine de savoirs . . . . .	28
3.2.2 Un traitement adapté à la nature des sources . . . . .	30
<b>II Penser l'édition numérique de correspondance</b>	<b>33</b>
<b>4 Bilan scientifique</b>	<b>35</b>
4.1 Réflexions autour de l'édition numérique de correspondance. Une communauté scientifique grandissante . . . . .	35
4.1.1 Des communautés de réflexion, des projets et des publications... . .	35
4.1.2 ...au service de normes et standards adaptés à la correspondance . .	39
4.1.3 Des outils au service de l'édition numérique . . . . .	42
4.2 Avec de nombreuses réalisations . . . . .	42
<b>5 Problématiques et spécificités de l'édition numérique de correspondance</b>	<b>45</b>
5.1 Édition et numérique . . . . .	45
5.1.1 Trois niveaux d'édition . . . . .	45
5.1.2 Cinq dimensions à prendre en compte . . . . .	46
5.2 Correspondance et numérique . . . . .	49
5.2.1 Importance de l'épistolaire dans la recherche . . . . .	49
5.2.2 Spécificités de la correspondance . . . . .	49
5.2.3 Des choix éditoriaux à faire en amont . . . . .	51
<b>6 Concevoir un site adapté aux exigences de l'édition</b>	<b>57</b>
6.1 Encoder, oui, mais pourquoi? . . . . .	57
6.2 Les récits utilisateurs . . . . .	57
6.3 Élaboration de l'architecture du site . . . . .	58
6.4 Le référencement naturel . . . . .	61
6.5 La licence ou le cadre juridique . . . . .	62
6.6 Le cahier des charges . . . . .	63

<b>III L'apprentissage machine au cœur de l'acquisition des données</b>	<b>65</b>
<b>7 L'apprentissage machine</b>	<b>69</b>
7.1 Petite histoire de l'apprentissage machine . . . . .	69
7.2 Apprentissage machine et intelligence artificielle . . . . .	69
7.3 L'apprentissage machine dans nos deux projets . . . . .	71
<b>8 L'OCR de Gallica</b>	<b>73</b>
8.1 Un service à ne pas négliger . . . . .	73
8.1.1 Un OCR plutôt fiable . . . . .	73
8.1.2 Extraction de l'OCR en HTML . . . . .	74
8.2 Un pré-traitement qui suscite des questionnements . . . . .	77
8.2.1 Rendre l'HTML valide et bien indenté . . . . .	77
8.2.2 Quelques fautes de l'OCR, visibles dans l'HTML . . . . .	77
8.2.3 L'apport des expressions régulières dans le nettoyage de l'HTML . .	78
8.2.4 Quelle granularité dans la correction? . . . . .	80
8.2.5 Premiers repérages des marqueurs . . . . .	82
<b>9 L'HTR de Transkribus</b>	<b>85</b>
9.1 Quelle procédure pour l'acquisition des données? . . . . .	85
9.1.1 Rappels et point sur le corpus . . . . .	85
9.1.2 S'assurer des numérisations des manuscrits . . . . .	85
9.1.3 Transcriptions manuelles ou automatisées? . . . . .	87
9.2 Transkribus, un outil de transcription . . . . .	87
9.2.1 Transkribus, un pari . . . . .	87
9.2.2 Un outil pensé par le READ . . . . .	88
9.2.3 Transkribus et l'apprentissage machine . . . . .	88
9.2.4 Point sur la terminologie . . . . .	89
9.3 Chargement des données d'entraînement et premier traitement . . . . .	91
9.3.1 Procédure pour le chargement des données d'entraînement . . . . .	91
9.3.2 Des transcriptions qui suscitent des questionnements . . . . .	92
9.4 Entraînement d'un modèle. Quels résultats? . . . . .	97
9.4.1 Mise en place de l'entraînement . . . . .	97
9.4.2 Quelle progression du modèle, quels résultats? . . . . .	98
9.5 Application du modèle . . . . .	101
9.6 Rester maître de ses données . . . . .	104
9.6.1 Exportation en vue de l'édition . . . . .	104
9.6.2 La question des <i>tags</i> . . . . .	105

<b>IV Traiter les données pour la réalisation des projets</b>	<b>109</b>
<b>10 Des standards et technologies au service de l'édition numérique</b>	<b>111</b>
10.1 XML, un langage particulièrement approprié . . . . .	111
10.1.1 XML : présentation générale . . . . .	111
10.1.2 Les métadonnées . . . . .	113
10.1.3 XML et le rituel épistolaire . . . . .	116
10.1.4 Les index . . . . .	121
10.1.5 La normalisation . . . . .	125
10.2 L'ODD et la pérennité des données . . . . .	126
10.2.1 Un schéma et une documentation pour la pérennité des données . . . . .	126
10.2.2 Créer l'ODD et l'associer à un document XML-TEI . . . . .	126
10.2.3 L'ODD dans nos projets . . . . .	127
10.3 Au service d'XML . . . . .	128
10.3.1 XSLT . . . . .	128
10.3.2 Python . . . . .	128
<b>11 Relever les défis du projet numérique</b>	<b>133</b>
11.1 Des difficultés à surmonter . . . . .	133
11.2 Prévenir le « facteur d'autobus » . . . . .	133
11.2.1 Les avantages de GitHub . . . . .	134
11.2.2 Des tutoriels et des fiches de savoir pour assurer la transmission . . . . .	134
11.2.3 Documenter son code . . . . .	136
<b>Conclusion</b>	<b>139</b>
<b>Annexes</b>	<b>145</b>
<b>A Édition numérique de correspondance</b>	<b>147</b>
A.1 Exemples d'éditions déjà existantes . . . . .	147
A.2 Attentes liées à chaque type de publication . . . . .	151
A.3 Ébauche d'une arborescence . . . . .	154
<b>B Transcription et Transkribus</b>	<b>155</b>
B.1 Les fac-similés . . . . .	156
B.2 Chargement des données d'entraînement . . . . .	163
B.3 Schématisation du modèle d'information de Transkribus . . . . .	164
<b>C L'encodage en XML-TEI</b>	<b>167</b>
C.1 Les index . . . . .	167

<i>Glossaire</i>	209
<b>D Livrables du CRHXIX</b>	<b>169</b>
D.1 CRHXIX/rapport_fin_stage_CRHXIX.pdf . . . . .	169
D.2 CRHXIX/1-inventaires . . . . .	169
D.3 CRHXIX/2-transcriptions . . . . .	170
D.4 CRHXIX/3-trankribus . . . . .	170
D.5 CRHXIX/4-cahier_des_charges . . . . .	172
D.6 CRHXIX/5-site . . . . .	172
D.7 CRHXIX/6-index . . . . .	172
D.8 CRHXIX/7-index_teи . . . . .	173
D.9 CRHXIX/8-teи . . . . .	173
D.10 CRHXIX/9-odd . . . . .	174
<b>E Livrables du Labex OBVIL</b>	<b>175</b>
E.1 Les fichiers XML-TEI . . . . .	175
E.2 Organisation des dossiers . . . . .	175
E.2.1 <i>corpus</i> . . . . .	176
E.2.2 <i>script</i> . . . . .	176
E.2.3 <i>dump</i> . . . . .	176
E.2.4 <i>remarques</i> . . . . .	176
<b>Bibliographie</b>	<b>179</b>
<b>Glossaire</b>	<b>199</b>
<b>Table des figures</b>	<b>201</b>
<b>Table des matières</b>	<b>205</b>