

ÉCOLE NATIONALE DES CHARTES

Jean-Damien Généro

Licencié ès histoire

Diplômé de Master

VALORISER LE TRAITEMENT
AUTOMATIQUE DES DONNÉES :

Le cas des Ouvriers des deux mondes

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

2020

Résumé

Résumé d'une trentaine de lignes à placer en tête du mémoire, accompagné d'une dizaine de mots-clés destinés à décrire le mémoire et des informations bibliographiques nécessaires pour le citer. Ce résumé et ces mots-clés sont destinés à compléter la notice bibliographique du mémoire dans la future bibliothèque numérique des mémoires.

N.B. : ne pas dépasser une page pour le tout.

Mots-clés : XML ; TEI ; Python ; traitement automatique des données ; transcription automatique ; édition numérique ; ALTO ; OCR ; *Kraken* ; *Transkribus* ; *GitLab* ; Frédéric Le Play ; *Les Ouvriers des deux mondes* ; enquêtes sociologiques ; monographies de familles ; *Time Us* ; Inria.

Informations bibliographiques : Jean-Damien Généro, *Valoriser le traitement automatique des données : Le cas des Ouvriers des deux mondes*, mémoire du Master « Technologies numériques appliquées à l'histoire », dir. A. Chagué et V. Jolivet, École nationale des chartes, 2020.

Liste des sigles et abréviations

Institutions

- ALMAnaCH : *Automatic Language Modelling and Analysis & Computational Humanities* (Inria Paris)
- ANR : Agence nationale de la recherche
- CMH : Centre Maurice-Halbwachs (EHESS et ENS Paris)
- CNRS : Centre National de Recherche Scientifique
- CRH : Centre de recherches historiques (EHESS)
- EA : Équipe d'accueil
- EHESS : École des Hautes Études en Sciences Sociales
- ENS : École normale supérieure
- ICT : Identités, Cultures et Territoires (Université de Paris)
- Inria : Institut national de recherche en informatique et en automatique
- LARHRA : Laboratoire de Recherche Historique Rhône-Alpe (Lyon 2)
- TGIR : Très grande infrastructure de recherche
- TELEMM : Temps, Espaces, Langages, Europe Méridionale-Méditerranée (Université d'Aix-Marseille)
- UMR : Unité mixte de recherche

Programmes de recherche

- DAHN : *Digital edition of historical manuscripts*
- MetaLEX : Métalexicographie numérique des langues historiques du droit en Europe
- READ : *Recognition and Enrichment of Archival Documents*
- Time Us : *time usage* (Rémunérations et usages du temps des femmes et des hommes en France de la fin du XVII^e siècle au début du XX^e siècle)

Informatique et nouvelles technologies

- ALTO : *Analyzed Layout and Text Object*
- API : *Application Programming Interface*
- CSV : *Comma-separated values*
- GPU : *Graphics processing unit*
- HTML : *Hypertext Markup Language*
- IETF : *Internet Engineering Task Force*
- IIIF : *International Image Interoperability Framework*
- JPEG : *Joint Photographic Experts Group*
- JP2 : *JPEG 2000*
- JSON : *JavaScript Object Notation*
- LSE-OD2M : *Logical Structure Extraction from Les Ouvriers des Deux Mondes*
- OCR : *Optical Character Recognition*
- ODD : *One Document Does it all*
- PDF : *Portable Document Format*
- Relax NG : *Regular Language for XML Next Generation*
- RFC : *Request for comments*
- RNG : *cf. relax NG*
- TEI : *Text Encoding Initiative*
- URI : *Uniform Resource Identifier*
- XML : *eXtensible Markup Language*
- XSL : *eXtensible Stylesheet Language*

Introduction

Les *Ouvriers des deux mondes* au sein de l'ANR *Time Us*

« Reconstituer les rémunérations et les budgets temps des travailleuses et des travailleurs du textile dans quatre villes industrielles françaises (Lille, Paris, Lyon, Marseille) dans une perspective européenne et de longue durée »¹ est l'objectif du programme ANR *Time Us*. Son intitulé exact, « Rémunérations et usages du temps des femmes et des hommes en France de la fin du XVII^e siècle au début du XX^e siècle »² met en évidence sa double articulation autour d'un temps long qui court depuis les premières manufactures de l'époque Moderne jusqu'à la fin de la révolution industrielle au tournant des XIX^e et XX^e siècles, et du temps individuel et quotidien qui est celui d'un ouvrier ou d'une ouvrière au sein de ces grands mouvements historiques. Les relations entre la rémunération et le temps sont au cœur des questionnements du programme. Cette attention portée à l'usage du temps se traduit dans la dénomination courante du programme, « *Time Us* », abréviation de l'anglais « *time usage* ».

Initialement financé sur une durée de trente-six mois (janvier 2017-janvier 2020), le programme ANR, coordonné par le professeur Manuela Martini de l'Université Lumière Lyon 2, est porté par une équipe pluri-institutionnelle. Celle-ci est composée de quatre unités mixtes de recherche (UMR) formées d'universitaires et de chercheurs du Centre National de la Recherche Scientifique³, d'une équipe d'accueil (EA)⁴ de l'Université de Paris⁵ et de l'équipe projet ALMAAnCH⁶ de l'Institut national de recherche en informatique et en automatique (Inria⁷ Paris). Stéphane Baciocchi, ingénieur de recherche du Centre de recherches historiques de l'EHESS, participe également au projet. Le stage qui a donné lieu au présent mémoire était placé sous la responsabilité administrative du laboratoire ICT de l'Université de Paris et a été effectué sous le tutorat scientifique et professionnel d'Inria, en la personne d'Alix Chagué, ingénierie de recherche et de développement au sein de l'équipe ALMAAnCH.

-
1. Présentation du programme sur le site de l'ANR (<https://anr.fr/Projet-ANR-16-CE26-0018>, consulté le 21 septembre 2020).
 2. Présentation du programme sur le site du LARHRA (<http://larhra.ish-lyon.cnrs.fr/anr-time-us>, consulté le 21 septembre 2020).
 3. Le Laboratoire de recherches historiques Rhône-Alpes (LARHRA, Lyon 2), le laboratoire Temps, Espaces, Langages, Europe Méridionale-Méditerranée (TELEMMe, Université d'Aix-Marseille), l'Institut de Recherches Historiques du Septentrion (Université de Lille) et le Centre Maurice-Halbwachs (CMH, EHESS et ENS Paris).
 4. Le laboratoire Identités, Cultures et Territoires (ICT) .
 5. Université instituée par le décret n° 2019-209 du 20 mars 2019 et regroupant les établissements Paris V-Descartes et Paris VII-Diderot.
 6. *Automatic Language Modelling and Analysis & Computational Humanities*.
 7. « Inria » est depuis 2011 une marque; aussi l'élision par le biais d'un « l' » liminaire n'est-elle plus requise dans son écriture courante (« Rappel : désormais *L'INRIA* s'écrit *Inria* », <https://web.archive.org/web/20111113064215/https://www.inria.fr/institut/inria-en-bref/charter-logo-inria/charter>, consulté le 21 septembre 2020).

Time Us se concentre principalement sur les femmes et plus encore sur les ouvrières du textile, industrie dans laquelle « elles sont présentes dans toutes les phases du processus productif »⁸. Le programme tend à combler le biais des genres dans l'historiographie du travail industriel en réalisant une opération de collecte et de traitement de la documentation manuscrite et imprimée relative à l'emploi et aux activités quotidiennes des femmes⁹. Ainsi, le moteur de *Time Us* est moins la production d'une réflexion scientifique autour du travail des femmes que la constitution d'un corpus documentaire sériel et prêt à être exploité par des chercheurs d'horizons multiples. La pluridisciplinarité est un aspect majeur du programme, qui souhaite une utilisation de ses données dans un maximum de champs de recherche des sciences humaines et sociales, notamment « en histoire économique et sociale, en histoire de la famille et du genre, en histoire des conflits du travail et de la culture des classes populaires »¹⁰.

La documentation est essentiellement constituée de documents manuscrits conservés par la Bibliothèque municipale de Lyon et les dépôts d'archives départementaux et municipaux lillois, parisien, lyonnais et marseillais. Très diverse, elle est issue d'organes officiels, à l'instar des conseils prud'homaux, des chambres de commerce ou encore des tribunaux de commerce, mais aussi des archives de personnes physiques ou morales de droit privé, telles que celles du tisseur lyonnais Pierre Charnier (1795-1857)¹¹, ou les registres comptables (1817-1821) et le dossier de faillite (1821-1822) de la filature parisienne Dupuis-Drouet¹².

En sus de cette documentation manuscrite, le programme *Time Us* s'appuie sur trois grands corpus d'imprimés. Le premier se compose de neuf titres de la presse ouvrière lyonnaise, entièrement numérisés sur le site *Numelyo* de la Bibliothèque municipale de Lyon¹³. L'intérêt du programme pour ce corpus porte sur les nombreux compte rendus d'audience du Conseil des prud'hommes qui s'y trouvent, ainsi que sur les reproductions ou extraits de discours, les lettres ou encore les analyses économiques et sociales relayées par ces journaux. Quatre des titres sont publiés dans la première moitié des années 1830 (*L'Écho de la fabrique*, *L'Écho des travailleurs*, *L'Indicateur* et *La Tribune prolétaire*), les cinq restants l'étant dans les années 1840 (*L'Écho des ouvriers*, *L'Écho de la Fabrique*

8. Alix Chagué, Manuela Martini, Victoria Le Fournier et Éric Villemonte de la Clergerie, *Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ?*, DH Nord 2019, 2020, p. 1.

9. *Ibid.*

10. *Ibid.*, p. 2.

11. Les papiers du canut Pierre Charnier font partie du fonds « Fernand Rude » de la Bibliothèque municipale de Lyon, nommé d'après l'historien qui les possédait avant leur versement (<https://www.bm-lyon.fr/collections-patrimoniales-et-specialisees/explorer-les-collections/article/fernand-rude>, consulté le 21 septembre 2020).

12. Archives de Paris, D 12U1, n° 375-376 (http://archives.paris.fr/arkotheque/inventaires/ead_ir_consult.php?a=4&ref=FRAD075_000727, consulté le 21 septembre 2020).

13. Le corpus se trouve à l'adresse <https://collections.bm-lyon.fr/PER003> (consulté le 21 septembre 2020).

de 1841, *L'Écho de l'industrie, L'Avenir, La Tribune Lyonnaise*)¹⁴.

Un second corpus d'imprimés est constitué de monographies collectées sur le site *Gallica* de la Bibliothèque nationale de France au format PDF. Très divers, on y trouve à la fois *L'ouvrière* de Jules Simon (1861), un *Dictionnaire général des tissus anciens et modernes* (1859-1863) ou encore un *Traité complet sur la fabrication des étoffes de soie* (1859)¹⁵. Pour le LARHRA, il s'agit d'une base complémentaire au projet, qui doit faciliter la contextualisation de la base archivistique principale, notamment en permettant de suivre l'évolution du vocabulaire afférent au textile sur la période étudiée. Inria, dans un optique de traitement automatique du langage (TAL) considère quant à lui que ces ouvrages servent à la « connaissance du domaine », c'est-à-dire que les modèles de langue en tireront le vocabulaire spécifique, les tournures syntaxiques et les associations fréquentes de mots propres au domaine du textile.

Le troisième et dernier corpus d'imprimés est composé des treize volumes de la série des *Ouvriers des deux mondes*.

*

Initiés par le sociologue Frédéric Le Play (1806-1882), les *Ouvriers des deux mondes* sont des enquêtes sociologiques conduites par les membres de la Société internationale des études pratiques d'économie sociale¹⁶ de 1857 à 1928. Répartie en trois séries comptant un total de cent vingt-six monographies¹⁷, il s'agit de la deuxième entreprise d'étude empirique de Le Play, après celle des *Ouvriers européens* dont la première édition a lieu en 1855¹⁸. Les enquêtes sont couramment désignées sous le vocable de « monographies de familles » ou plus simplement « monographies ».

Dans sa préface du numéro spécial de la revue *Les Études sociales* consacré aux monographies leplaysiennes, Alain Chenu relève la prégnance de l'assimilation de celles-ci à des « mines » dans lesquelles les chercheurs peuvent « [puiser] » à leur guise, tant les sujets abordés et les étendues géographiques traitées sont nombreux¹⁹. *Les Ouvriers des deux mondes*, dont le titre fait écho à *La revue des deux mondes* fondée en 1829, présentent en effet une succession de « trajectoires et récits de vie de familles ouvrières »²⁰ établies

14. Présentation de ce corpus sur le wiki du programme : http://timeusage.paris.inria.fr/mediawiki/index.php/Documentation_régionale_-_Presse_lyonnaise (consulté le 21 septembre 2020).

15. Liste complète sur le wiki du programme : http://timeusage.paris.inria.fr/mediawiki/index.php/Aperçu_des_états#Imprimés_divers (consulté le 21 septembre 2020).

16. Actuelle Société d'économie et de sciences sociales, désormais abrégée en SESS.

17. Anthony Lorry, « Les monographies des *Ouvriers européens* (1855, 1877-1879) et des *Ouvriers des deux mondes* (1857-1930). Inventaire et classification », *Les Études sociales, Les monographies de famille de l'École de Le Play*, 131-132 (2000), p. 93-181, p. 95.

18. *Ibid.*

19. Alain Chenu, « Préface », *Les Études sociales, Les monographies de famille de l'École de Le Play*, 131-132 (2000), p. 5-9, p. 5.

20. Stéphane Baciocchi et Jérôme David, « Trajectoires et récits de vie. Singulariser les expériences sociales », *Les Études sociales, Frédéric Le Play : anthologie et correspondance*, n° 142-143-144 (II-2005-2006), p. 193-194, URL : https://www.academia.edu/5925885/Fr%C3%A9d%C3%A9ric_Le_Play_%C3%

de part et d'autre de la Méditerranée, et parfois plus loin encore. Se succèdent ainsi une enquête consacrée à un charpentier de Paris²¹, à un métayer de la banlieue de Florence²² et à un menuisier-charpentier de Tanger²³.

Les *Ouvriers des deux mondes* présentent un double intérêt pour *Time Us*. D'une part, leur attention dirigée de manière exclusive envers les familles ouvrières en font un matériau privilégié pour le programme, d'autant que les monographies se focalisent sur le budget et son usage²⁴. Aucun individu de la cellule familiale n'est ignoré : l'ouvrier — il s'agit le plus souvent du père, qui peut être accompagné de ses frères ou de ses fils —, ses descendants, ses ascendants, sa femme évidemment, mais aussi les domestiques²⁵ et parfois les esclaves²⁶. Ainsi, si seulement quatorze enquêtes ont pour sujet des familles travaillant dans l'industrie du textile, l'ensemble reste intéressant à l'échelle du programme ANR dans la mesure où chaque enquête s'attache à établir le budget affecté aux matières textiles et à son utilisation par la famille enquêtée. L'angle de lecture adopté par *Time Us* concerne donc la totalité du texte pour quatorze monographies, et se focalise sur les paragraphes et tableaux relatifs aux individus et à leurs budgets pour les cent douze restantes.

D'autre part, la reproduction systématique d'une même structure logique dans chaque monographie permet d'envisager un traitement informatique de l'ensemble du corpus et, à travers celui-ci, la mise à disposition des textes structurés dans des fichiers XML. L'idée des « mines » leplaysiennes est en effet inexacte, et Alain Chenu dénonce le caractère restrictif de cette image : les monographies ne sont pas une succession d'enquêtes indépendantes. Le Play et la SESS ont conservé une même « grille d'observation » depuis la première (1856) jusqu'à la dernière enquête (1928), construisant *de facto* un « système dont les éléments prennent sens les uns par rapport aux autres »²⁷.

Time Us a souhaité transcrire automatiquement les volumes des *Ouvriers des deux mondes* à partir de leurs numérisations, avant d'implémenter la structure logique d'origine au sein du texte brut obtenu. Le but visé est de mettre à disposition de la communauté

A91%C3%A9ments_d'%C3%A9pist%C3%A9mologie_et_de_science_sociale, p. 193.

21. Frédéric Le Play et Adolphe Focillon, « Charpentier de Paris (Seine - France), de la Corporation des compagnons du Devoir », dans *Les Ouvriers des deux mondes*, Paris, 1857 [janvier 1858] (série 1 (1)), t. 1, chap. 1, [27]-68, URL : <https://archive.org/details/lesouvriersdesde01sociuoft/page/26>.

22. Ulbadino Péruzzi, « Métayer de la banlieue de Florence (Grand-Duché de Toscane) », dans *Les Ouvriers des deux mondes*, Paris, 1857 [janvier 1858] (série 1 (1)), chap. 5, p. 221-262, URL : <https://archive.org/details/lesouvriersdesde01sociuoft/page/220>.

23. Narcisse Cotte, « Menuisier-charpentier (Nedjar) de Tanger (Province de Tanger - Maroc) », dans *Les Ouvriers des deux mondes*, Paris, 1858 (série 1 (2)), chap. 12, p. 105-144, URL : <https://archive.org/details/lesouvriersdesde02sociuoft/page/104>.

24. Le budget est « à la fois la méthode et le résultat » des monographies : Fabien Cardoni, « Aux sources du budget domestique selon Le Play », *Les Études sociales*, 155-1 (2012), p. 11-46, DOI : 10.3917/etsoc.155.0011, p. 11

25. Ernest Delbet, « Paysans en communauté et en polygamie de Bousrah (Esky Cham), dans le pays de Haouran (Syrie - Empire Ottoman) », dans *Les Ouvriers des deux mondes*, Paris, 1858 (série 1 (2)), chap. 18, p. 363-446, URL : <https://archive.org/details/lesouvriersdesde02sociuoft/page/362>.

26. Narcisse Cotte, « Menuisier-charpentier... », *op. cit.*

27. A. Chenu, « Préface »..., p. 5.

scientifique ces textes afin qu'elle puisse y porter un regard nouveau à l'aide des outils numériques ; il s'agit de faire entrer ce « matériau exceptionnel sur les sociétés des cinq continents »²⁸ dans le champ des humanités numériques.

★

Les rapports entre le programme ANR *Time Us* et le Master « Technologies numériques appliquées à l'histoire » de l'École nationale des chartes remontent à 2018. Deux étudiantes, Alix Chagué (prom. 2018) et Victoria Le Fournier (prom. 2019) y ont effectué leur stage de fin d'étude. Alix Chagué a travaillé sur la collecte et le traitement de la documentation²⁹, tandis que Victoria Le Fournier s'est intéressée à la structuration de cette documentation en prenant pour exemple la juridiction prud'homale³⁰. Leurs mémoires correspondent à deux phases successives du projet, et la présente étude entend décrire sa phase finale de valorisation des fichiers obtenus. Si cet aspect l'inscrit dans la suite des précédents, il n'en reste pas moins fondamentalement différent.

En premier lieu, il porte une attention exclusive à un corpus du programme, les *Ouvriers des deux mondes*. Alix Chagué balayait pour sa part l'ensemble de la documentation et Victoria Le Fournier s'appuyait quant à elle sur une de ses composantes majeures et transversales. Les documents afférents aux conseils des prud'hommes se trouvent en effet à la fois dans les archives propres à ces institutions et, pour la juridiction lyonnaise, sous forme de comptes rendus d'audience publiés dans la presse. Le présent travail ne cherche pas à remplacer les *Ouvriers des deux mondes* au sein de la documentation globale du programme.

Au commencement du stage ayant donné lieu à ce mémoire en avril 2020, les *Ouvriers des deux mondes* étaient déjà entièrement mis en forme et structurés au format XML-TEI. Ainsi — c'est la deuxième différence d'avec nos prédecesseures — nous sommes intervenus dans la phase de post-traitement, c'est-à-dire que l'essentiel du stage a été consacré au contrôle qualité des fichiers, à la réparation des erreurs issues du traitement automatique et à une réflexion sur la valorisation des données ainsi produites.

Plusieurs objectifs ont été poursuivis durant ce stage, et les questionnements qu'ils portent sous-tendent le propos que nous allons tenir entre ces pages. Il faut tout d'abord rappeler que le mémoire du Master « Technologies numériques appliquées à l'histoire » est

28. Antoine Savoye, « Éditorial », *Les Études sociales, Les monographies de famille de l'École de Le Play*, 131-132 (2000), p. 3-4.

29. A. 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*, dir. Éric de la Clergerie et Vincent Jolivet, mémoire du Master « Technologies numériques appliquées à l'histoire », École nationale des chartes, 2018, URL : https://github.com/alix-tz/M2TNAH_memoire-de-stage.

30. V. Le Fournier, *Étude de la structuration automatique et de l'éditionnalisation d'un corpus hétérogène, L'exemple des sources du conseil des prud'hommes pour le textile au XIX^e siècle du projet Time Us*, dir. Éric de la Clergerie et Ariane Pinche, mémoire du Master « Technologies numériques appliquées à l'histoire », École nationale des chartes, 2019, URL : <https://github.com/Victorialf/M2TNAH-memoireDeStage>.

un exercice à mi-chemin entre le mémoire de recherche et le rapport de stage. Il ne s'agit pas d'établir un compte rendu des missions reçues et des actions menées, ni d'exposer une réflexion scientifique sur un sujet délimité. Il s'agit de mettre en perspective de manière critique l'intervention du stagiaire en la replaçant au sein des dynamiques scientifiques propres au projet et à l'équipe où le stage a été effectué.

Dans cet optique, notre intervention avait un premier objectif technique et professionnel. Il s'agissait d'analyser et de reprendre des fichiers XML-TEI afin de corriger leurs erreurs structurelles dans le but d'amorcer leur valorisation. Pour mener à bien cette opération, nous avions la possibilité de recourir à des techniques automatiques ou semi-automatiques en appliquant des scripts Python, ou bien d'effectuer les corrections en manuel. Le second objectif a résulté de cette façon de faire et interrogeait la valeur ajoutée de l'automatisation pour l'édition numérique d'une documentation imprimée dans le cadre d'un programme de recherche.

Nous tiendrons dans ce mémoire un propos en trois parties. Pour commencer, nous présenterons les différents états du corpus des *Ouvriers des deux mondes*, qui est à la fois imprimé, numérisé et structuré en fichiers informatiques. Cela sera pour nous l'occasion de décrire la structure logique mise en place par Le Play et de faire la part entre la volonté de continuité de ses successeurs dans la réutilisation de celle-ci et les innovations qu'ils ont pu mettre en place au fil du temps. Nous conclurons cette première partie par la description des opérations qui ont permis d'obtenir les fichiers des monographies, qui nous furent transmis sous la forme d'un dépôt au sein de l'espace *GitLab* d'Inria.

La seconde partie abordera les reprises que nous avons effectuées, en commençant par définir la méthode de travail suivie et notamment les outils de développements et les espaces d'échange mis en place pour la gestion du projet. Après avoir présenté une typologie des erreurs identifiées à l'issue de l'analyse des documents, nous exposerons nos interventions tant manuelles qu'automatiques en procédant du général au particulier, c'est-à-dire du niveau global de l'encodage documentaire à celui, plus fin, de l'encodage scientifique.

Nous réfléchirons enfin aux différentes manières de valoriser ces fichiers dans la troisième partie, en insistant sur trois voies possibles. La première consiste à déterminer l'opportunité de conserver un lien entre le texte de la page et son image d'origine, ainsi que la nature que celui-ci pourrait prendre (stockage local, utilisation des ressources de la plate-forme d'hébergement des numérisations, recours à un protocole IIIF³¹). La seconde concerne l'indexation des individus enquêtés et la place que pourrait tenir l'automatisation dans cette opération. La dernière porte enfin sur la possibilité de corriger les transcriptions pour que les fichiers soient exploitables non seulement au niveau des données qu'ils contiennent, mais aussi au niveau du texte en lui-même, c'est-à-dire de permettre et à la

31. Abréviation d'*International Image Interoperability Framework*, généralement prononcée « triple I F » en français.

machine et à l'humain de les utiliser³².

32. Ce mémoire est entièrement rédigé avec le langage LATEX. Il est consultable et téléchargeable sur le GitHub de l'auteur : <https://github.com/jeandamien-genero/Memoire-TNAH>.

Bibliographie

Bibliographie

Bibliographie générale

- BACIOCCHI (Stéphane) et DAVID (Jérôme), « IV. Ramifications épistémologiques. Décrire, définir, évaluer, comparer », *Les Études sociales, Frédéric Le Play : anthologie et correspondance*, n° 142-143-144 (II-2005-2006), p. 87-90, URL : https://www.academia.edu/5925885/Fr%C3%A9d%C3%A9ric_Le_Play_%C3%A9t%C3%A9ments_d%C3%A9pist%C3%A9mologie_et_de_science_sociale.
- « Trajectoires et récits de vie. Singulariser les expériences sociales », *Les Études sociales, Frédéric Le Play : anthologie et correspondance*, n° 142-143-144 (II-2005-2006), p. 193-194, URL : https://www.academia.edu/5925885/Fr%C3%A9d%C3%A9ric_Le_Play_%C3%A9t%C3%A9ments_d%C3%A9pist%C3%A9mologie_et_de_science_sociale.
- BELAÏD (Abdel), RANGONI (Yves) et FALK (Ingrid), *Représentation des données en XML pour l'analyse d'images de documents*, juil. 2007, URL : <http://lodel.irevues.inist.fr/cide/index.php?id=147>.
- BREWSTER (Kahle), « Preserving the Internet », *Scientific American*, 276 (3 — mars 1997), p. 82-83, URL : <https://www.jstor.org/stable/24993660>.
- *Archiving the Internet*, 1996, URL : http://web.archive.org/web/19971011050140/http://www.archive.org/sciam_article.html.
- BREWSTER (Kahle) et PAREJO VADILLO (Ana), « The Internet Archive : An Interview with Brewster Kahle », 19, *Interdisciplinary Studies in the Long Nineteenth Century*, 21 (2015), DOI : 10.16995/tnn.760.
- BURNARD (Lou), *Qu'est-ce que la Text Encoding Initiative ?, Comment ajouter un balisage intelligent aux ressources numériques*, Marseille, 2015, DOI : 10.4000/books.oep.1237.
- CARDONI (Fabien), « Aux sources du budget domestique selon Le Play », *Les Études sociales*, 155-1 (2012), p. 11-46, DOI : 10.3917/etsoc.155.0011.
- 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*, dir. Éric de la Clergerie et Vincent Jolivet, mémoire du Master « Techno-

- logies numériques appliquées à l'histoire », École nationale des chartes, 2018, URL : https://github.com/alix-tz/M2TNAH_memoire-de-stage.
- CHAGUÉ (Alix), *Constitution d'un corpus textuel sur les monographies de Le Play*, carnet de recherche de *Time Us*, URL : <https://timeus.hypotheses.org/626>.
- CHAGUÉ (Alix), LE FOURNER (Victoria), MARTINI (Manuela) et VILLEMONTE DE LA CLERGERIE (Éric), « Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ? », dans *Colloque DHNord 2019 "Corpus et archives numériques"*, MESHS Lille Nord de France, Lille, France, 2019, URL : <https://hal.inria.fr/hal-02448921>.
- *Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ?*, DH Nord 2019, 2020.
- CHENU (Alain), « Préface », *Les Études sociales, Les monographies de famille de l'École de Le Play*, 131-132 (2000), p. 5-9.
- CHIFFOLEAU (Floriane), *Vers un alignement de traductions et d'éditions à partir d'un lexique et à travers un corpus multilingue, Travail sur Dei Delitti e delle Pene du marquis de Beccaria*, dir. Thibault Clérice, mémoire du Master « Technologies numériques appliquées à l'histoire », École nationale des chartes, 2019, URL : <https://github.com/FloChiff/memoire-M2>.
- CLÉRICE (Thibault), « Les outils CapiTainS, l'édition numérique et l'exploitation des textes », *Médiévales*, 73 (automne 2017), p. 115-131, DOI : [10.4000/medievales.8211](https://doi.org/10.4000/medievales.8211).
- DUVAL (Frédéric), « Pour des éditions numériques critiques. L'exemple des textes français », *Médiévales*, 73 (automne 2017), p. 13-29, DOI : [10.4000/medievales.8165](https://doi.org/10.4000/medievales.8165).
- ELAGOUNI (Khaoula), GARCIA (Christophe), MAMALET (Franck) et SÉBILLOT (Pascale), « Combining Multi-Scale Character Recognition and Linguistic Knowledge for Natural Scene Text OCR », dans *10th IAPR International Workshop on Document Analysis Systems, DAS*, Gold Coast, Queensland, Australia, 2012, p. 120-124, URL : <https://hal.archives-ouvertes.fr/hal-00753908>.
- GÉNÉRO (Jean-Damien), *Les ouvriers des deux mondes : des images aux urls*, carnet de recherche de *Time Us*, URL : <https://timeus.hypotheses.org/645>.
- HAWKINS (Kevin), DALMAU (Michelle), MYLONAS (Elli) et BAUMAN (Syd), *Best Practices for TEI in Libraries, a guide for mass digitization, automated workflows, and promotion of interoperability with XML using the TEI*, Consortium TEI, 2018 (septembre), URL : <https://tei-c.org/extra/teiinlibraries/4.0.0/bpt1-driver.html>.
- JOLIVET (Vincent), « Éditions ou données ? API et (re)publications », dans *Actes royaux et princiers à l'ère du numérique (Moyen Âge — Temps moderne)*, dir. Olivier Canteaut, Olivier Guyotjeannin et Olivier Poncet, Pau, 2020, p. 59-68, URL : <https://www.nakala.fr/nakala/data/11280/eae11732>.

- KARPINSKI (Romain) et BELAID (Abdel), *Rapport Evaluation des OCR*, Research Report, LORIA - Université de Lorraine, 2016, URL : <https://hal.inria.fr/hal-01356824>.
- KIESSLING (Benjamin), *Kraken — a Universal Text Recognizer for the Humanities*, rapp. tech., Université PSL et Université de Leipzig, 2019, DOI : 10.34894/Z9G2EX.
- KIESSLING (Benjamin), THOMAS MILLER (Matthew), ROMANOV (Maxim G.) et BOWEN SAVANT (Sarah), « Important New Developments in Arabographic Optical Character Recognition (OCR) », *Al-‘Usur al-Wusta : The Journal of Middle East Medievalists*, 25 (2017), p. 1-13, DOI : 10.17613/M6TZ4R.
- LE FOURNER (Victoria), *Étude de la structuration automatique et de l'éditorialisation d'un corpus hétérogène, L'exemple des sources du conseil des prud'hommes pour le textile au XIX^e siècle du projet Time Us*, dir. Éric de la Clergerie et Ariane Pinche, mémoire du Master « Technologies numériques appliquées à l'histoire », École nationale des chartes, 2019, URL : <https://github.com/Victorialf/M2TNAH-memoireDeStage>.
- LORRY (Anthony), « Les monographies des *Ouvriers européens* (1855, 1877-1879) et des *Ouvriers des deux mondes* (1857-1930). Inventaire et classification », *Les Études sociales, Les monographies de famille de l'École de Le Play*, 131-132 (2000), p. 93-181.
- NOUVEL (Damien), ANTOINE (Jean-Yves), FRIBURGER (Nathalie) et SOULET (Arnaud), « Fouille de règles d'annotation pour la reconnaissance d'entités nommées », dir. Sophia Ananiadou, Nathalie Friburger et Rosset Sophie, *Traitemen Automatique des Langues, Entités Nommées*, 54-2 (), p. 13-41, URL : <https://www.atala.org/content/fouille-de-r%C3%A9gles-d%5C%C3%5C%A8gles-d%5C%E2%5C%80%5C%99annotation-pour-la-reconnaissance-d%5C%E2%5C%80%5C%99entit%C3%5C%C3%5C%A9s-nomm%C3%5C%C3%5C%A9es>.
- PANOS (Patrick), « Technotes : The Internet Archive, An End to the Digital Dark Age », *Journal of Social Work Education*, n° 39 (2003-2), p. 343-347, URL : www.jstor.org/stable/23044068.
- POUPEAU (Gautier), *Et le wiki devint sémantique*, Les Petites cases, URL : <https://www.lespetitescases.net/et-le-wiki-devint-semantique>.
- ROBINSON (Peter), « Towards a Theory of Digital Editions », *Variants, The Journal of the European Society for Textual Scholarship*, 10 (2013), p. 105-131.
- SAGOT (Benoît) et GÁBOR (Kata), « Détection et correction automatique d'entités nommées dans des corpus OCRisés », dans *Traitemen Automatique du Langage Naturel 2014*, Marseille, France, 2014, URL : <https://hal.inria.fr/hal-01022378>.
- SAVOYE (Antoine), « Les continuateurs de Le Play au tournant du siècle », dir. Philippe Besnard, *Revue française de sociologie, Sociologies françaises au tournant du siècle. Les concurrents du groupe durkheimien*, 22-3 (1981), DOI : 10.2307/3321155.

- SAVOYE (Antoine), « Éditorial », *Les Études sociales, Les monographies de famille de l'École de Le Play*, 131-132 (2000), p. 3-4.
- « La monographie sociologique : jalons pour son histoire (1855-1974) », *Les Études sociales, Les monographies de famille de l'École de Le Play*, 131-132 (2000), p. 11-46.

Les Ouvriers des deux mondes

- BADIER (Alexis-Félix), « Compositeur-typographe de Paris (Seine - France) », dans *Les Ouvriers des deux mondes*, Paris, 1862 (série 1 (4)), chap. 33, p. 241-282, URL : <https://archive.org/details/lesouvriersdesde04sociuoft/page/240>.
- BIGOT (Maximilien) et ESCARD (François), « Paysans corses en communauté, porchers-bergers des montagnes de Bastelica », dans *Les Ouvriers des deux mondes*, Paris, 1890 (série 2e (2)), p. 433-524, URL : <https://archive.org/details/s2lesouvriersdes02sociuoft/page/n495>.
- CORONEL (Samuël Mozes), « Précis d'une monographie d'un tisserand d'Hilversum (Hollande septentrionale - Pays-Bas) », dans *Les Ouvriers des deux mondes*, Paris, 1892 (série 2e (3)), p. 143-172, URL : <https://archive.org/details/s2lesouvriersdes03sociuoft/page/142>.
- CORONEL (Samuël-Mozes) et ALLAN (F), « Pêcheur côtier, maître de barques, de Marken (Hollande septentrionale - Pays-Bas) », dans *Les Ouvriers des deux mondes*, Paris, 1862 (série 1 (4)), chap. 37, p. 405-460, URL : <https://archive.org/details/lesouvriersdesde04sociuoft/page/404>.
- COTTE (Narcisse), « Menuisier-charpentier (Nedjar) de Tanger (Province de Tanger - Maroc) », dans *Les Ouvriers des deux mondes*, Paris, 1858 (série 1 (2)), chap. 12, p. 105-144, URL : <https://archive.org/details/lesouvriersdesde02sociuoft/page/104>.
- COTTE (Narcisse) et HARAÏRI (Soliman el), « Parfumeur de Tunis (Régence de Tunis - Afrique) du bazar appelé : El Attharin-el-kebar (les grands parfumeurs) », dans *Les Ouvriers des deux mondes*, Paris, 1861 (série 1 (3)), chap. 25, p. 285-326, URL : <https://archive.org/details/lesouvriersdesde03sociuoft/page/284>.
- DARASSE (Vincent), « Paysans en communauté et colporteurs émigrants de Tabou-Douchd-El-Baar (Grande Kabylie - Province d'Alger) », dans *Les Ouvriers des deux mondes*, Paris, 1885 (série 1 (5)), chap. 46, p. 459-502, URL : <https://archive.org/details/lesouvriersdesde05sociuoft/page/n499>.
- DELAIRE (Alexis), « Paysan-paludier du Bourg de Batz (Loire-Inférieure - France) », dans *Les Ouvriers des deux mondes*, Paris, 1887 (Fascicule 1 (01)), p. 1-56, URL : <https://archive.org/details/s2lesouvriersdes01sociuoft/page/n23>.

- DELAIRE (Edmond Augustin), « Petit fonctionnaire de Pnom-Penh (Cambodge) », dans *Les Ouvriers des deux mondes*, Paris, 1899 (série 2e (5)), p. 437-483, URL : <https://archive.org/details/2serlesouvriersde05sociuoft/page/n487>.
- « Précis d'une monographie d'un manœuvre-coolie de Pnom-Penh (Cambodge) », dans *Les Ouvriers des deux mondes*, Paris, 1899 (série 2e (5)), p. 484-500, URL : <https://archive.org/details/2serlesouvriersde05sociuoft/page/484>.
- DELBET (Ernest), « Paysans en communauté et en polygamie de Bousrah (Esky Cham), dans le pays de Haouran (Syrie - Empire Ottoman) », dans *Les Ouvriers des deux mondes*. Paris, 1858 (série 1 (2)), chap. 18, p. 363-446, URL : <https://archive.org/details/lesouvriersdesde02sociuoft/page/362>.
- DONNAT (Léon) et OUANG-TCHING-YONG, « Paysans en communauté du Ning-Po-Fou (province de Tché-Kian - Chine) », dans *Les Ouvriers des deux mondes*, Paris, 1862 (série 1 (4)), chap. 30, p. 83-158, URL : <https://archive.org/details/lesouvriersdesde04sociuoft/page/82>.
- ESCARD (François), « Précis d'une monographie du pêcheur-côtier du Finmark (Laponie - Norvège) », dans *Les Ouvriers des deux mondes*, Paris, 1892 (série 2e (3)), p. 125-142, URL : <https://archive.org/details/s2lesouvriersdes03sociuoft/page/n163>.
- GUÉRIN (Urbain), « Ouvrier cordonnier de Malakoff (Seine - France) », dans *Les Ouvriers des deux mondes*, Paris, 1885 (série 1 (5)), chap. 41, p. 145-200, URL : <https://archive.org/details/lesouvriersdesde05sociuoft/page/n165>.
- « Paysan-résinier de Lévignacq (Landes - France) », dans *Les Ouvriers des deux mondes*, Paris, 1885 (série 1 (5)), chap. 44, p. 315-386, URL : <https://archive.org/details/lesouvriersdesde05sociuoft/page/n347>.
- « Tourneur-mécanicien des usines de la Société Cockerill, de Seraing (Belgique) », dans *Les Ouvriers des deux mondes*, Paris, 1890 (série 2e (2)), p. 1-52, URL : <https://archive.org/details/s2lesouvriersdes02sociuoft/page/n32>.
- « Fileur en peigné et régleur de métier de la Manufacture du Val-des-Bois (Marne - France) », dans *Les Ouvriers des deux mondes*, Paris, (série 2e (5)), p. 73-136, URL : <https://archive.org/details/2serlesouvriersde05sociuoft/page/n98/>.
- LE PLAY (Frédéric), « Introduction. III, Appréciation des deux procédés communément employés pour observer les faits sociaux : supériorité des enquêtes directes sur les recherches scientifiques », dans *Les Ouvriers européens*, Paris, 1855, p. 9-12, URL : <https://gallica.bnf.fr/ark:/12148/bpt6k1057844n/f17>.
- « Instruction sur la méthode d'observation dite des monographies de famille, propre à l'ouvrage intitulé *Les ouvriers européens* », dans *Les Ouvriers des deux mondes*, Paris, 1862 (série 1 (4)), p. 15-32, URL : <https://archive.org/details/lesouvriersdesde04sociuoft/page/14/>.

- LE PLAY (Frédéric) et FOCILLON (Adolphe), « Charpentier de Paris (Seine - France), de la Corporation des compagnons du Devoir », dans *Les Ouvriers des deux mondes*, Paris, 1857 [janvier 1858] (série 1 (1)), t. 1, chap. 1, [27]-68, URL : <https://archive.org/details/lesouvriersdesde01sociuoft/page/26>.
- MAROUSSEM (Pierre du), « La société générale des papeteries du Limousins », dans *Les Ouvriers des deux mondes, fascicule supplémentaire A*, Paris, 1904 (série 3e (1)), p. 1-48, URL : <https://archive.org/details/lesouvriersdesde0108sociuoft/page/n9>.
- « Usine hydraulique d'éclairage et de transport de force », dans *Les Ouvriers des deux mondes, fascicule supplémentaire A*, Paris, 1908 (série 3e (3)), p. 1-32, URL : <https://archive.org/details/lesouvriersdesde0108sociuoft/page/n9>.
- PARISET (Félicien), « Bûcheron usager de l'ancien Comté de Dabo (Lorraine allemande) », dans *Les Ouvriers des deux mondes*, Paris, 1885 (série 1 (5)), chap. 45, p. 387-458.
- PASOLINI (Marie), « Précis d'une monographie d'un ouvrier agriculteur de la campagne de Ravenne (Romagne - Italie) », dans *Les Ouvriers des deux mondes*, Paris, 1899 (série 2e (5)), p. 234-252, URL : <https://archive.org/details/2serlesouvriersde05sociuoft/page/234>.
- PERETZ (Alexandre de), « Précis d'une monographie de l'armurier des manufactures impériales de Toula (Grande-Russie), Ouvrier propriétaire et chef de métier dans le système du travail sans engagements, par le général A. Peretz (de Saint-Petersbourg) [1886] », dans *Les Ouvriers des deux mondes*, Paris, 1887 (série 2e (1)), p. 113-132, URL : <https://archive.org/details/s2lesouvriersdes01sociuoft/page/n143>.
- PÉRUZZI (Ulbadino), « Métayer de la banlieue de Florence (Grand-Duché de Toscane) », dans *Les Ouvriers des deux mondes*, Paris, 1857 [janvier 1858] (série 1 (1)), chap. 5, p. 221-262, URL : <https://archive.org/details/lesouvriersdesde01sociuoft/page/220>.
- REVIER DE MAUNY (Jacques), « Serrurier-forgeron de Paris (Seine - France) », dans *Les Ouvriers des deux mondes*, Paris, 1885 (série 1 (5)), chap. 42, p. 201-259, URL : <https://archive.org/details/lesouvriersdesde05sociuoft/page/n225>.
- SIMON (Gabriel-Eugène) et ESCARD (Paul), « Précis d'une monographie d'un pêcheur-côtier, maître de barques, de l'archipel Chusan (Chine) », dans *Les Ouvriers des deux mondes*, Paris, 1904 (série 3e (1)), p. 61-87, URL : <https://archive.org/details/lesouvriersdesde0108sociuoft/page/n127>.
- « Avertissement », dans *Les Ouvriers des deux mondes*. éd. Société internationale des études pratiques d'économie sociale, Paris, 1887, p. I-VIII, URL : <https://archive.org/details/s2lesouvriersdes01sociuoft/page/n11>.

- « Avertissement », dans *Les Ouvriers des deux mondes*, éd. Société internationale des études pratiques d'économie sociale, Paris, 1875 (série 1 (5)), p. I-III, URL : <https://archive.org/details/lesouvriersdesde05sociuoft/page/n9>.
- « Avertissement », dans *Les Ouvriers des deux mondes*, éd. Société internationale des études pratiques d'économie sociale, Paris, 1887 (série 2e (1)), p. I-VIII, URL : <https://archive.org/details/s2lesouvriersdes01sociuoft/page/n11>.
- TOYTOT (Ernest de), « Gantier de Grenoble (Isère - France) », dans *Les Ouvriers des deux mondes*, Paris, 1887 (série 2e (1)), p. 465-520, URL : <https://archive.org/details/s2lesouvriersdes01sociuoft/page/n520/>.
- VALLIN (Charles), « Précis d'une monographie du pêcheur côtier, maître de barque, d'Étretat (Seine-Inférieure - France) », dans *Les Ouvriers des deux mondes*, Paris, 1890 (série 2e (2)), p. 153-172, URL : <https://archive.org/details/s2lesouvriersdes02sociuoft/page/152>.
- WILBOIS (Adolphe), « Brigadier de la Garde républicaine de Paris (Seine - France) », dans *Les Ouvriers des deux mondes*, Paris, 1885 (série 1 (5)), chap. 43, p. 261-313, URL : <https://archive.org/details/lesouvriersdesde05sociuoft/page/n289>.

Première partie

Un corpus déjà structuré

Le corpus sur lequel nous avons travaillé est un ensemble de fichiers XML-TEI. Ces derniers résultent d'une version numérisée des *Ouvriers des deux mondes* réalisée et hébergée par le site *Internet Archive*, elle-même issue du traitement des volumes physiques de l'Université de Toronto. Dans cette première partie, nous allons décrire ces différents états des *Ouvriers des deux mondes* et commenter les opérations qui ont mené à leur structuration, tout en interrogeant les raisons ayant conduit le programme *Time Us* à les privilégier par rapport à d'autres versions similaires.

Chapitre 1

Un corpus d'imprimés

1.1 Une longue publication

La publication des trois séries des *Ouvriers des deux mondes* s'échelonne sur soixante-treize années (1857-1930). Elle n'est véritablement continue que sur trois périodes, de 1857 à 1862, puis de 1885 à 1913 et enfin de 1928 à 1930. L'interruption entre 1913 et 1928 est une conséquence de la première Guerre mondiale et des difficultés budgétaires auxquelles la SESS doit faire face après celle-ci. La première interruption, entre 1862 et 1885, est présentée *a posteriori* comme un recalibrage des objectifs de la SESS, qui considère que « continuer à recueillir des faits sans essayer d'en faire sortir aucune conclusion (...) c'eût été (...) une preuve d'impuissance et de stérilité »¹. Elle se concentre dès lors sur la parution de son *Bulletin* jusqu'au début des années 1880, où « le besoin de conclusions pratiques était amplement satisfait » et « l'œuvre des monographies de familles » redevenait possible².

Une tentative de reprise avait déjà été effectuée en 1875 où « un premier fascicule composé de trois monographies » avait paru³. Deux autres paraissent en 1883 et 1884, et les trois sont finalement rassemblés en 1885 dans le cinquième volume de la première série⁴. Ce volume apparaît ainsi à bien des égards comme un moment charnière dans l'histoire de la publication des *Ouvriers des deux mondes*. En effet, les monographies suivantes des deuxième (1887-1899) et troisième série (1904-1930) sont publiées sous forme de fascicules trimestriels qui sont ensuite reliés en volumes, là où les quatre premiers tomes étaient directement parus en volumes⁵. Des éléments de paratexte sont systématiquement fournis aux relieurs, notamment une page de titre, une introduction au volume, un index,

1. « Avertissement », dans *Les Ouvriers des deux mondes*, éd. Société internationale des études pratiques d'économie sociale, Paris, 1875 (série 1 (5)), p. I-III, URL : <https://archive.org/details/lesouvriersdesde05sociuoft/page/n9>, p. I.

2. *Ibid.*, p. II.

3. *Ibid.*, p. III.

4. *Ibid.*

5. A. Lorry, « Les monographies des *Ouvriers européens* (1855, 1877-1879) et des *Ouvriers des deux mondes* (1857-1930). Inventaire et classification »..., p. 124.

des *errata* et une table des matières⁶.

Le passage des volumes reliés aux fascicules trimestriels est acté par l'éditeur dans l'*Avertissement* liminaire du premier fascicule de la deuxième série :

« La nouvelle série des *Ouvriers des deux mondes* s'ouvre avec le présent fascicule. Le grand nombre des travaux soumis à la Société d'Économie sociale assure l'avenir et la régularité de la publication. *Les monographies paraîtront désormais en fascicules trimestriels*⁷. »

Dans l'*Avertissement* général de ce volume, édité en 1887 et placé immédiatement après la page de titre et le sommaire, le changement est à nouveau annoncé, mais également justifié :

« Quand la Société d'Économie sociale, en 1882, perdit celui qui avait réglé son rôle auprès de lui et après lui⁸, elle trouva le cinquième tome des *Ouvriers des deux mondes* arrêté au premier tiers de sa publication : elle acheva de le mettre au jour. Puis, voulant laisser bien distincts des travaux accomplis sous l'œil du maître, ceux dont il lui fallait dès lors prendre seule la responsabilité, *elle commença en juillet 1885, sous le même titre, une nouvelle série de monographies de familles paraissant par fascicules trimestriels*⁹. »

Deux versions de ce même volume sont à notre disposition. L'une résulte de la numérisation de l'exemplaire déposé à la Bibliothèque nationale de France à Paris et conservé par le département Philosophie, histoire, sciences de l'homme¹⁰, la seconde se trouve à l'Université de Toronto et a été numérisée par *Internet Archive*¹¹.

Des différences de composition existent et concernent notamment l'emplacement des page liminaires des fascicules. Ainsi, dans l'exemplaire de Toronto (fig. 1.1a, 1.1b, 1.1c), les premières pages ont été laissées à leur emplacement : la première porte au recto le titre des *Ouvriers des deux mondes* et au verso l'*Avertissement* que nous avons cité, la seconde est la page de titre du fascicule avec un verso vierge. Dans l'exemplaire de la Bibliothèque nationale, la première page de la monographie n° 47 suit la fin de l'*Avertissement* général du volume : deux feuillets sont donc manquant. Ils ont été éliminés lors du processus de

6. Nous utilisons le terme *paratexte* par commodité, tout en étant au fait des problèmes qu'il pose. Il suppose notamment une unité de ces textes, ce qui n'est pas le cas : cohabitent ainsi le paratexte propre aux fascicules et celui propre aux volumes finaux.

7. Alexis Delaire, « Paysan-paludier du Bourg de Batz (Loire-Inférieure - France) », dans *Les Ouvriers des deux mondes*, Paris, 1887 (Fascicule 1 (01)), p. 1-56, URL : <https://archive.org/details/s2lesouvriersdes01sociuoft/page/n23>, « Avertissement ». Nous soulignons.

8. Frédéric Le Play est mort le 5 avril 1882.

9. « Avertissement », dans *Les Ouvriers des deux mondes*. éd. Société internationale des études pratiques d'économie sociale, Paris, 1887, p. I-VIII, URL : <https://archive.org/details/s2lesouvriersdes01sociuoft/page/n11>, spec. p. II. Nous soulignons.

10. Consultable sur *Gallica* : <https://gallica.bnf.fr/ark:/12148/bpt6k54465138> (consulté le 21 septembre 2020).

11. Consultable à cette adresse : <https://archive.org/details/s2lesouvriersdes01sociuoft> (consulté le 21 septembre 2020).

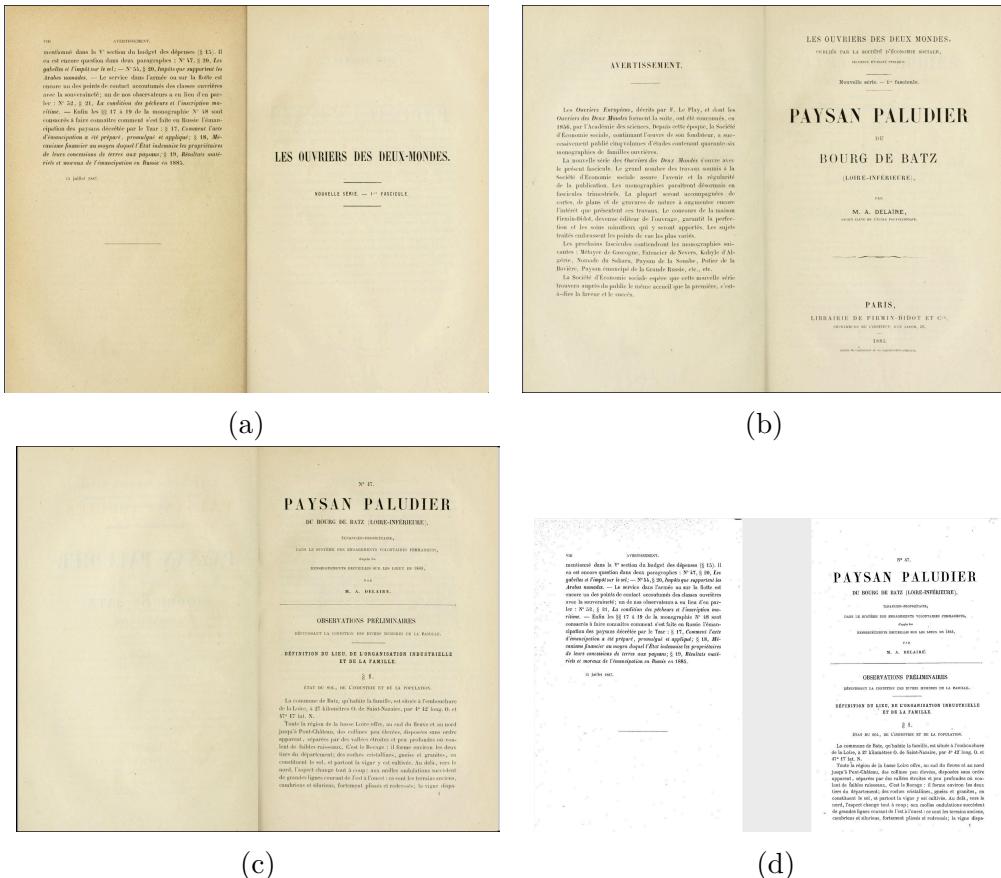


FIGURE 1.1 – Comparaison des exemplaires de Toronto et de Paris. (a) Toronto. Fin de l'*Avertissement* général du volume et début du fascicule. (b) Toronto. *Avertissement* du fascicule et page de titre de la monographie. (c) Toronto. Début de la monographie. (d) Paris. Fin de l'*Avertissement* général du volume et début de la monographie.

reliure. Les mêmes manques s'observent dans les volumes qui suivent ; on peut cependant observer que les pages propres aux fascicules ont été placées en fin de volume, après la table des matières. Il n'y a donc pas eu d'élimination systématique.

Un autre élément symptomatique de ces différences de composition se remarque dans le corpus de Toronto. La monographie n° 56 est la première du deuxième volume de la deuxième série¹², or elle se trouve également à la fin du volume précédent. La table des matières de ce dernier n'y fait pourtant pas référence. Ce doublon trouve sans doute son origine dans l'existence d'un fascicule surnuméraire que le relieur a choisi de conserver, quitte à l'implanter au mauvais endroit.

Ces exemples illustrent le fait que laisser aux acheteurs le soin de relier les fascicules pour composer le volume final revient à éloigner ce dernier de l'œuvre originelle, chaque nouvelle opération de reliure donnant naissance à un nouveau document. Les relieurs ont en effet pu choisir de conserver ou de ne pas conserver les pages propres aux fascicules,

12. Urbain Guérin, « Tourneur-mécanicien des usines de la Société Cockerill, de Seraing (Belgique) », dans *Les Ouvriers des deux mondes*, Paris, 1890 (série 2e (2)), p. 1-52, URL : <https://archive.org/details/s2lesouvriersdes02sociuoft/page/n32>.

à l'instar des feuillets de l'exemplaire de Toronto. Il n'existe donc pas un modèle ayant autorité dans la composition des *Ouvriers des deux mondes* — en dépit du fait que les volumes de la Bibliothèque nationale de France, issus du dépôt légal, ont dû être versés par l'éditeur, c'est-à-dire la Société internationale des études pratiques d'économie sociale.

Pour le projet *Time Us*, cela signifie que le corpus retenu pour le traitement informatisé et la publication finale ne peut être qu'une version de l'œuvre que représente les *Ouvriers des deux mondes*.

1.2 La structure logique des monographies

Un élément remarquable traverse la centaine de monographies des *Ouvriers des deux mondes* : une même structure logique que les éditeurs ont tenté de maintenir tout au long de la publication.

Cette structure est l'incarnation de la méthodologie leplaysienne des monographies de familles. Le terme *monographie*, « emprunté à l'histoire naturelle et à la médecine », qualifie au XIX^e siècle une « étude scientifique, minutieuse et détaillée, portant sur un objet ou un phénomène circonscrit »¹³. La démarche monographique cherche à embrasser à travers l'observation directe des phénomènes inatteignables pour la démarche statistique, à laquelle elle s'oppose¹⁴.

Avant les *Ouvriers des deux mondes*, Frédéric Le Play s'attelle aux *Ouvriers européens*, recueil paru en 1855 et réédité de 1877 à 1879. Cette série contient trente-six (1^{re} éd.) puis cinquante-sept (2^e éd.) monographies issues d'enquêtes menées entre 1833 et 1853¹⁵. Les *Ouvriers des deux mondes* se présentent comme une suite élargie à des espaces extra-européens de ce premier mouvement, dont ils reprennent la démarche, la structure et auquel ils font régulièrement référence par un système de renvoi. La structure logique des *Ouvriers des deux mondes*, dont le premier volume est publié en 1855, est donc déjà présente dans *Les Ouvriers européens*.

Elle procède d'une méthode d'observation sur laquelle Le Play revient en 1862, au moment de la parution du quatrième volume des *Ouvriers des deux mondes*. Le titre de cette méthodologie précise néanmoins qu'elle a déjà fait ses preuves pour les *Ouvriers*

13. A. Savoye, « La monographie sociologique : jalons pour son histoire (1855-1974) », *Les Études sociales, Les monographies de famille de l'École de Le Play*, 131-132 (2000), p. 11-46, p. 12.

14. Voir en particulier la charge de F. Le Play contre les statisticiens dans son introduction aux *Ouvriers européens* de 1855 : « La méthode des statisticiens n'est pas l'observation directe des faits ; c'est la compilation et l'interprétation plus ou moins plausible de faits recueillis à des points de vue fort différents, étrangers pour la plupart à l'intérêt scientifique. Malgré leur généralité apparente et leur séduisante régularité, les statistiques ont médiocrement contribué au progrès de la science sociale » : F. Le Play, « Introduction. III, Appréciation des deux procédés communément employés pour observer les faits sociaux : supériorité des enquêtes directes sur les recherches scientifiques », dans *Les Ouvriers européens*, Paris, 1855, p. 9-12, URL : <https://gallica.bnf.fr/ark:/12148/bpt6k1057844n/f17>, p. 11.

15. Inventaire complet dans A. Lorry, « Les monographies des *Ouvriers européens* (1855, 1877-1879) et des *Ouvriers des deux mondes* (1857-1930). Inventaire et classification »..., p. 106-112.

européens : *Instruction sur la méthode d'observation dite des monographies de familles, propre à l'ouvrage intitulé Les ouvriers européens*¹⁶.

L'*Instruction* est divisée en quatre parties suivies d'une cinquième faisant office de conclusion. Dans la première, Le Play circonscrit son champ d'observation à la famille ouvrière¹⁷, érigée en « clef de compréhension des organisations sociales »¹⁸. La phase d'observation est ensuite divisée en étapes¹⁹, savoir l'observation des faits, l'interrogatoire de l'ouvrier et celui des personnes tiers²⁰.

La troisième partie est celle où Le Play expose la structure logique qu'il utilise à partir de 1855 et que ses « continuateurs »²¹ reprennent jusqu'en 1930. Elle est construite en miroir aux différentes phases de l'observation sur le terrain : à l'observation des faits répond la section des *Observations préliminaires*, à l'interrogatoire de l'ouvrier celle des *Budgets* et aux contacts avec les tiers celle des *Notes*²².

Le miroir est cependant légèrement déformant, afin de satisfaire au « hiatus » propre à la démarche leplaysienne, qui différencie les « faits observés » des « faits décrits »²³. L'observation est en effet « contrainte par l'évidence des choses » là où la description, « parce qu'elle suppose de désigner précisément ce qu'on évoque, repose sur un calibrage sémantique »²⁴. Les monographies se trouvent du côté de la description des faits, leur structure logique n'est donc pas qu'un simple plan de rédaction, elle procède d'une démarche intellectuelle réglée par Le Play qui doit être appliquée pour garantir la validité de la démonstration.

Cette structure (Annexe A.2) possède trois niveaux de titre : les parties (A, B, C), les sections (I, II, etc.) et les paragraphes (§1, §2, etc.). La partie C (*Notes*) n'a pas recours au deuxième niveau et ne contient que des paragraphes.

16. F. Le Play, « Instruction sur la méthode d'observation dite des monographies de famille, propre à l'ouvrage intitulé *Les ouvriers européens* », dans *Les Ouvriers des deux mondes*, Paris, 1862 (série 1 (4)), p. 15-32, URL : <https://archive.org/details/lesouvriersdes04sociuoft/page/14/>.

17. *Ibid.*, I., « Remarques préliminaires sur l'étude des faits sociaux et sur la méthode des monographies de familles », p. 15-16.

18. A. Savoye, « La monographie sociologique : jalons pour son histoire (1855-1974) »..., p. 15.

19. F. Le Play, « Instruction sur la méthode d'observation dite des monographies de famille, propre à l'ouvrage intitulé *Les ouvriers européens* »..., II., « Règles à suivre pour procéder à l'observation des faits sociaux », p. 17-19.

20. A. Savoye, « La monographie sociologique : jalons pour son histoire (1855-1974) »..., p. 16.

21. Nous désignons par ce terme le groupe d'individus qui a participé à l'entreprise des *Ouvriers des deux mondes* après la mort de Le Play, sans entrer dans les débats sur l'existence ou non d'une « école leplaysienne ». De fait, un schisme a lieu à la fin des années 1880 : *Id.*, « Les continuateurs de Le Play au tournant du siècle », dir. Philippe Besnard, *Revue française de sociologie, Sociologies françaises au tournant du siècle. Les concurrents du groupe durkheimien*, 22-3 (1981), DOI : 10.2307/3321155.

22. F. Le Play, « Instruction sur la méthode d'observation dite des monographies de famille, propre à l'ouvrage intitulé *Les ouvriers européens* »..., III., « Précis des faits à observer — Établissement des budgets », p. 20-31.

23. S. Baciocchi et J. David, « IV. Ramifications épistémologiques. Décrire, définir, évaluer, comparer », *Les Études sociales, Frédéric Le Play : anthologie et correspondance*, n° 142-143-144 (II-2005-2006), p. 87-90, URL : https://www.academia.edu/5925885/Fr%C3%A9d%C3%A9ric_Le_Play_%C3%A9pist%C3%A9mologie_et_de_science_sociale, p. 87.

24. *Ibid.*, p. 87-88.

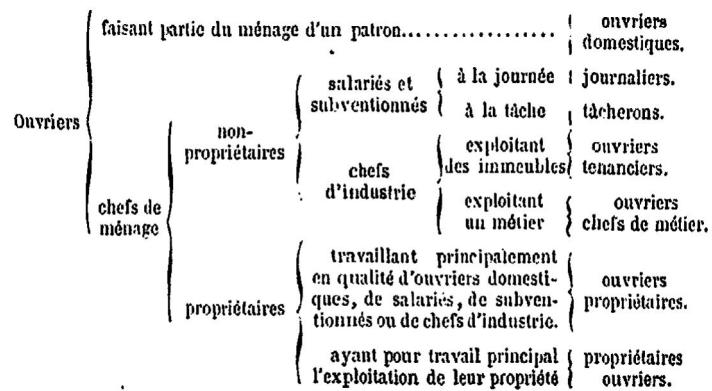


FIGURE 1.2 – « Tableau des sept situations principales que les ouvriers peuvent occuper successivement dans les quatre systèmes sociaux pour s'élever des rangs inférieurs de la hiérarchie industrielle à la condition de propriétaires ou de chefs d'industrie. », permettant de construire le titre d'une monographie de famille (capture d'écran de *Gallica*).

La première partie (A) est la page de titre. Le titre de la monographie est présenté comme le « résumé » de celle-ci, et doit pour cela contenir quatre éléments imprescriptibles : « 1° la profession de l'ouvrier, 2° la population dont il fait partie, 3° la nature des engagements qu'il contracte pour se procurer des moyens de travail, 4° la situation qu'il occupe dans l'organisation sociale caractérisée par cet engagement »²⁵. Ainsi, en dépit du fait que Le Play ne considère pas explicitement la page de titre comme une composante de sa structure, l'importance qu'il donne au titre en lui-même a conduit le projet *Time Us* à l'interpréter comme telle. Un tableau pour construire ce titre est d'ailleurs mis à la disposition des monographies (fig. 1.2²⁶).

1.3 Continuité et discontinuité dans la structure logique

En dépit de son aspect monolithique, la structure logique des monographies de familles subit plusieurs ajustements au cours de la publication. Certains, uniquement de circonstance, sont des évènements mineurs, là où d'autres sont des changements significatifs adoptés pour la suite de la publication.

La monographie n° 85 représente ainsi un tournant dans la numérotation des titres de paragraphe, sans pour autant changer leurs libellés (Annexe A.2). En effet, la numérotation cardinale est étendue au dernier paragraphe des budgets (*Comptes annexés aux budgets*) et remplace la numérotation alphabétique utilisée auparavant dans la partie *Notes*.

25. F. Le Play, « Instruction sur la méthode d'observation dite des monographies de famille, propre à l'ouvrage intitulé *Les ouvriers européens* »..., p. 20.

26. *Ibid.*, p. 21 .

Pourquoi cette uniformisation intervient-elle à partir de cette monographie, en plein milieu du cinquième volume ? Nous avons vu plus haut qu'il représentait un tournant dans la publication des *Ouvriers des deux mondes*, notamment en raison de la mort de Frédéric Le Play qui intervient avant qu'il ne soit achevé. Il est possible que le départ du fondateur de la série ait permis d'opérer des changements dans la structure qu'il avait créée, ou bien que ce soit Le Play lui-même qui ait souhaité ces changements. Quoi qu'il en soit, il y a une volonté claire de renforcer la solidité de la démonstration en rassemblant ces différents paragraphes en un seul bloc par le biais d'une numérotation cardinale continue. Le nombre moyen de paragraphes est de vingt-et-un ; la monographie la plus longue, composée de trente-cinq paragraphes, étant la n° 64, intitulée *Paysans corses en communauté, porchers-bergers des montagnes de Bastelica*²⁷.

Ce même volume est l'occasion pour la SESS d'agrémenter aux *Ouvriers des deux mondes* un type de monographie déjà introduite dans la réédition des *Ouvriers euroépens* (1877-1879), le précis²⁸. Elle le définit comme un document qui, « n'ayant pas la forme complète des travaux auxquels ce recueil est destiné [les monographies de familles], offre néanmoins assez d'intérêt pour avoir été annexé à la description d'une famille de même nationalité »²⁹. Il s'agit d'un texte court, d'une quinzaine de pages environ, qui, tout en ayant le même propos qu'une monographie, est fortement « allégé »³⁰.

Cet allègement se remarque à travers deux traits caractéristiques. Le premier est la suppression du paragraphe 16 et la réduction des paragraphes 14 et 15 consacrés aux budgets. Ils sont en effet réduits à quelques lignes, là où il s'agit de tableaux de plusieurs pages dans les monographies de familles. Le précis n° 92 bis est le seul à posséder un paragraphe 16, mais les 14 et 15 sont rassemblés en une seule section³¹. Le second trait caractéristique consiste en des notes extrêmement courtes (un à trois paragraphes, à l'exception du précis n° 66 ter qui en compte six³²).

Les libellés des titres peuvent également être modifiés. Celui de la deuxième partie

27. Maximilien Bigot et François Escard, « Paysans corses en communauté, porchers-bergers des montagnes de Bastelica », dans *Les Ouvriers des deux mondes*, Paris, 1890 (série 2e (2)), p. 433-524, URL : <https://archive.org/details/s2lesouvriersdes02sociuoft/page/n495>.

28. A. Lorry, « Les monographies des *Ouvriers européens* (1855, 1877-1879) et des *Ouvriers des deux mondes* (1857-1930). Inventaire et classification »..., p. 102.

29. « Avertissement », dans *Les Ouvriers des deux mondes*, éd. Société internationale des études pratiques d'économie sociale, Paris, 1887 (série 2e (1)), p. I-VIII, URL : <https://archive.org/details/s2lesouvriersdes01sociuoft/page/n11>, p. II.

30. A. Chenu, « Préface »..., p. 7.

31. Gabriel-Eugène Simon et Paul Escard, « Précis d'une monographie d'un pêcheur-côtier, maître de barques, de l'archipel Chusan (Chine) », dans *Les Ouvriers des deux mondes*, Paris, 1904 (série 3e (1)), p. 61-87, URL : <https://archive.org/details/lesouvriersdes0108sociuoft/page/n127>, p. 78-81.

32. Samuël Mozes Coronel, « Précis d'une monographie d'un tisserand d'Hilversum (Hollande septentrionale - Pays-Bas) », dans *Les Ouvriers des deux mondes*, Paris, 1892 (série 2e (3)), p. 143-172, URL : <https://archive.org/details/s2lesouvriersdes03sociuoft/page/142>, p. 158-172.

(*Observations préliminaires*) est absent des précis n° 58 bis³³ et 66 bis³⁴. Ceux des paragraphes sont également placés en italique au début du texte dans quatre précis (n° 58 bis, 66 bis, 85 bis³⁵, 90 bis³⁶).

Les trois premiers précis sont des cas particuliers, puisqu'à l'inverse des suivants ils ne sont pas formellement détachés de la monographie dont ils découlent (n° 41³⁷, 42³⁸ et 43³⁹) mais sont traités comme une section de la partie *Notes*. Il n'y a ni grandes parties ni titres de paragraphe, seul le niveau intermédiaire ayant été retenu : les sections I à IV dans les trois, et un court budget dans les deux derniers.

Un dernier facteur de discontinuité au sein des *Ouvriers des deux mondes* et de sa structure logique doit être signalé : il s'agit des deux monographies d'ateliers que l'on trouve dans la troisième série. La première s'intéresse à la société générale des papeteries du Limousin (1904⁴⁰) et la seconde à une usine hydraulique d'éclairage et de transport de force sur la Loire (1908⁴¹). La monographie d'atelier est théorisée par Émile Cheysson, une des grandes figures de la SESSAprès la mort de Le Play⁴², pour qui le « foyer cesse d'être le centre unique de notre activité et cède à l'atelier une partie de ses attributions primitives »⁴³.

Il s'agit d'une rupture importante avec l'objectif initial des *Ouvriers des deux mondes*, mais également d'un signe de la volonté de la collection de s'adapter à son temps. Ces

33. Charles Vallin, « Précis d'une monographie du pêcheur côtier, maître de barque, d'Étretat (Seine-Inférieure - France) », dans *Les Ouvriers des deux mondes*, Paris, 1890 (série 2e (2)), p. 153-172, URL : <https://archive.org/details/s2lesouvriersdes02sociuoft/page/152>, p. 153.

34. F. Escard, « Précis d'une monographie du pêcheur-côtier du Finmark (Laponie - Norvège) », dans *Les Ouvriers des deux mondes*, Paris, 1892 (série 2e (3)), p. 125-142, URL : <https://archive.org/details/s2lesouvriersdes03sociuoft/page/n163>, p. 125.

35. Marie Pasolini, « Précis d'une monographie d'un ouvrier agriculteur de la campagne de Ravenne (Romagne - Italie) », dans *Les Ouvriers des deux mondes*, Paris, 1899 (série 2e (5)), p. 234-252, URL : <https://archive.org/details/2serlesouvriersde05sociuoft/page/234>.

36. Edmond Augustin Delaire, « Précis d'une monographie d'un manœuvre-coolie de Pnom-Penh (Cambodge) », dans *Les Ouvriers des deux mondes*, Paris, 1899 (série 2e (5)), p. 484-500, URL : <https://archive.org/details/2serlesouvriersde05sociuoft/page/484>.

37. U. Guérin, « Ouvrier cordonnier de Malakoff (Seine - France) », dans *Les Ouvriers des deux mondes*, Paris, 1885 (série 1 (5)), chap. 41, p. 145-200, URL : <https://archive.org/details/lesouvr05sociuoft/page/n165>, p. 188-196.

38. Jacques Reviers de Mauny, « Serrurier-forgeron de Paris (Seine - France) », dans *Les Ouvriers des deux mondes*, Paris, 1885 (série 1 (5)), chap. 42, p. 201-259, URL : <https://archive.org/details/lesouvr05sociuoft/page/n225>, 246-258.

39. Adolphe Wilbois, « Brigadier de la Garde républicaine de Paris (Seine - France) », dans *Les Ouvriers des deux mondes*, Paris, 1885 (série 1 (5)), chap. 43, p. 261-313, URL : <https://archive.org/details/lesouvr05sociuoft/page/n289>, p. 300-305.

40. Pierre du Maroussem, « La société générale des papeteries du Limousins », dans *Les Ouvriers des deux mondes, fascicule supplémentaire A*, Paris, 1904 (série 3e (1)), p. 1-48, URL : <https://archive.org/details/lesouvr0108sociuoft/page/n9>.

41. Id., « Usine hydraulique d'éclairage et de transport de force », dans *Les Ouvriers des deux mondes, fascicule supplémentaire A*, Paris, 1908 (série 3e (3)), p. 1-32, URL : <https://archive.org/details/lesouvr0108sociuoft/page/n9>.

42. A. Savoye, « Les continuateurs de Le Play au tournant du siècle »..., p. 336-337.

43. Émile Cheysson, « La monographie d'atelier et les Sociétés d'économie sociale », *La Réforme sociale*, 15 mai 1887, p. 545 — cité dans Id., « La monographie sociologique : jalons pour son histoire (1855-1974) »..., p. 23.

deux textes ne peuvent néanmoins pas reprendre la structure logique initiale, uniquement applicable aux familles. Néanmoins, on trouve dans leur organisation des traits similaires : trois parties (*Organisation commerciale de l'atelier*, *Organisation du travail* et des *Appendices* faisant fonction de notes), un second niveau signalé par une numérotation en romain et des paragraphes qui peuvent être titrés ou non⁴⁴.

Ces modulations dans l'application de la structure définie par Frédéric Le Play sont importantes pour le projet *Time Us*. Son but étant de mettre à la disposition de la communauté scientifique les monographies des *Ouvriers des deux mondes* sous la forme de fichiers XML-TEI, il est nécessaire qu'il étudie cette structure et qu'il fasse la part entre ses moments de continuité et de rupture. Ceux-ci doivent en effet être pris en compte dans l'établissement du schéma d'encodage.

44. Id., « Les continuateurs de Le Play au tournant du siècle »..., p. 337.

Chapitre 2

Des numérisations multiples

Avant d'établir le schéma d'encodage des *Ouvriers des deux mondes*, il est nécessaire de disposer d'une version digitale du corpus « papier » pour réaliser une opération de transcription automatique. Or, comme nous l'avons déjà signalé, les *Ouvriers des deux mondes* ont bénéficié des programmes de numérisation des ressources des bibliothèques.

Un corpus se trouve notamment sur le site *Google Books*, mais il n'est important que d'un point de vue quantitatif et non pas qualitatif. En effet, les treize volumes ne sont pas accessibles dans leur intégralité. Cet accès restreint est regrettable, dans la mesure où *Google Books* n'a pas numérisé un seul corpus, mais plusieurs issus de bibliothèques italiennes, françaises et américaines (Annexe A.3). Disposer de tous ces exemplaires aurait permis un relevé plus fin des différences de composition. Notons que par l'intermédiaire de la *HathiTrust Digital Library*, une bibliothèque numérique lancée en 2008 et agrégeant les livres numérisés par *Google Books*, *Internet Archive* et certaines bibliothèques universitaires, il est possible de consulter à des numérisations pour lesquelles l'accès est restreint par Google.

Le projet *Time Us* s'est donc orienté vers les deux seuls corpus complets, hébergés par la Bibliothèque nationale de France et le site *Internet Archive*.

2.1 Les volumes de la Bibliothèque nationale de France

Les trois séries des *Ouvriers des deux mondes* ont été mises en ligne sur le site *Gallica* de la Bibliothèque nationale de France le 15 octobre 2007¹.

Les volumes sont téléchargeables en trois formats :

- PDF (*Portable Document Format*), qui préserve la mise en page du document telle qu'elle a été définie par son auteur ;
- TXT, qui contient uniquement du texte brut, c'est-à-dire des chaînes de caractères (ce qui n'intéresse pas le programme *Time Us*, qui souhaite obtenir les

1. Notice de la numérisation : <https://gallica.bnf.fr/ark:/12148/cb32830863r/> (consulté le 21 septembre 2020).

- images des pages de chaque volume) ;
- JPEG (*Joint Photographic Experts Group*), qui permet de compresser les images avant de les enregistrer.

Le téléchargement au format JPEG n'est possible que pour la page en cours de visualisation par l'utilisateur. Seul le téléchargement au format PDF permet d'obtenir l'ensemble des images, avec l'inconvénient qu'elles sont toutes rassemblées en un seul fichier, ce qui complique le traitement à la chaîne.

Le corpus numérisé sur *Gallica* n'est ainsi pas exploitable par le programme *Time Us*. Cela est regrettable dans la mesure où, comme nous l'avions signalé plus haut, il s'agit des exemplaires issus du dépôt légal et donc déposés par l'éditeur des *Ouvriers des deux mondes*. Il s'agissait de l'unique occasion d'avoir avec certitude la vision que Frédéric Le Play et la SESS avaient de leur travail.

2.2 Les volumes d'*Internet Archive*

Internet Archive est un organisme privé à but non lucratif fondé en 1996 par Brewster Kahle. Son objectif premier, tel qu'annoncé dans son *manifeste*, est de « collecter les données publiques sur Internet pour bâtir une bibliothèque digitale »². Cet archivage du web est accessible grâce à l'outil *Wayback Machine*, lancé en octobre 2001³ mais prévu dès le lancement d'*Internet Archive*⁴. À partir de 2000, l'organisation commence à archiver les programmes de télévision, les musiques et les films numériques ; en 2001, c'est au tour des livres de faire leur entrée sur ses serveurs⁵.

Internet Archive commence ses digitalisations en tant que participant au projet *Million Book* de l'Université Carnegie-Mellon de Pittsburgh, et se met à digitaliser par lui-même en se rendant dans les bibliothèques universitaires à partir de 2005⁶. La digitalisation est faite page par page au moyen de scanners⁷. Les livres ainsi numérisés sont accessibles et téléchargeables librement, sans aucune restriction, ce qui constitue le principal avantage du site, à l'image de *Gallica* et à l'inverse d'autres plates-formes comme

2. « *The Internet Archive is such a new organization that is collecting the public materials on the Internet to construct a digital library* » : Kahle Brewster, *Archiving the Internet*, 1996, URL : http://web.archive.org/web/19971011050140/http://www.archive.org/sciam_article.html (cité dans K. Brewster et Ana Parejo Vadillo, « The Internet Archive : An Interview with Brewster Kahle », 19, *Interdisciplinary Studies in the Long Nineteenth Century*, 21 (2015), DOI : 10.16995/ntn.760).

3. Patrick Panos, « Technotes : The Internet Archive, An End to the Digital Dark Age », *Journal of Social Work Education*, n° 39 (2003-2), p. 343-347, URL : www.jstor.org/stable/23044068, p.344.

4. *We possess the capability of supplying documents that are no longer available from the original publisher, an important function if the Web's hypertext system is to become a medium for scholarly publishing. Such a service could also prove worthwhile for business research.* : K. Brewster, « Preserving the Internet », *Scientific American*, 276 (3 — mars 1997), p. 82-83, URL : <https://www.jstor.org/stable/24993660>, p.83.

5. K. Brewster et A. Parejo Vadillo, « The Internet Archive : An Interview with Brewster Kahle »..., p. 3-4.

6. *Ibid.*, p. 4.

7. *Ibid.*

*Google Books*⁸. Par rapport à *Gallica*, *Internet Archive* offre beaucoup plus de formats pour le téléchargement de ses fichiers⁹.

Les exemplaires des *Ouvriers des deux mondes* numérisés sur *Internet Archive* sont conservés à la *John P. Robarts Research Library* de l'Université de Toronto. Leurs mises en ligne ont été réalisées en novembre 2008, mise à part le fascicule 17 bis de la série 3, ajouté en mars 2010.

2.3 Choix du corpus à numériser



FIGURE 2.1 – Exemples d'images exclues du lot à transcrire (source : Alix Chagué, <https://timeus.hypotheses.org/626>).

Le téléchargement depuis *Internet Archive* permet d'obtenir un dossier contenant le fichier image de chaque page au format JPEG (en l'occurrence JPEG 2000, abrégé en JP2) avec une qualité maximale¹⁰. C'est donc pour ce corpus et ce format qu'a opté le programme *Time Us*. Cela représente 7 190 fichiers images, pour un poids de 3,3 Go¹¹.

L'ensemble a été versé sur un espace de stockage créé grâce au service *ShareDocs* de la TGIR Huma-Num. Adopté en mai 2018 à l'initiative de l'équipe ALMAnaCH, ce service est gratuit, français et sur-mesure (l'espace de stockage est alloué à la demande)¹². Le *ShareDocs* permet également de centraliser l'ensemble des fichiers image du programme *Time Us*.

8. *Ibid.*, p. 1.

9. Liste complète : A. Chagué, *Constituer un corpus pour la fouille de texte — de la transcription des documents d'archives à l'annotation...*, p. 33.

10. Id., *Constitution d'un corpus textuel sur les monographies de Le Play*, carnet de recherche de *Time Us*, URL : <https://timeus.hypotheses.org/626>.

11. *Ibid.*

12. *Ibid.*, p. 33-34.

Choisir le corpus d'*Internet Archive* n'est pas sans conséquence, notamment en raison du caractère incomplet du corpus. Les fascicules contenant les six dernières monographies (n° 109 à 114) et les précis qui les accompagnent (n° 109 bis à 111 bis) n'ont en effet pas été numérisés. De fait, nous ne les avons trouvés qu'à une seule adresse sur Internet : il s'agit de l'exemplaire de l'Université de Princeton numérisé par *Google Books* et accessible par le truchement de la *HathiTrust Digital Library* (Annexe A.3). Il contient les monographies n° 108 à 112 et constitue donc le troisième volume de la troisième série. Nous n'avons pas trouvé d'exemplaires numérisés des deux dernières monographies, publiées à la fin des années 20 du xx^e siècle.

Ce corpus comporte également divers biais liés à l'impression, dont des titres appartenant à la structure logique qui n'ont pas été imprimés (Annexe B.1.1), sans que l'on puisse déterminer s'il s'agit d'une erreur de composition ou d'une suppression volontaire par les éditeurs.

Notons qu'en dépit de ces écueils, toutes les monographies relatives au textile et qui intéressent directement le programme *Time Us* sont présentes dans le corpus d'*Internet Archive*.

À l'issue du téléchargement, il a fallu procéder à un contrôle de la qualité des fichiers, afin de détecter et de retirer les images présentant des imperfections¹³. Ces dernières étaient principalement dues à des erreurs de cadrage ou à des vues prises alors que la main de l'opérateur ou de l'opératrice était toujours présente dans le champ de l'image (fig. 2.1).

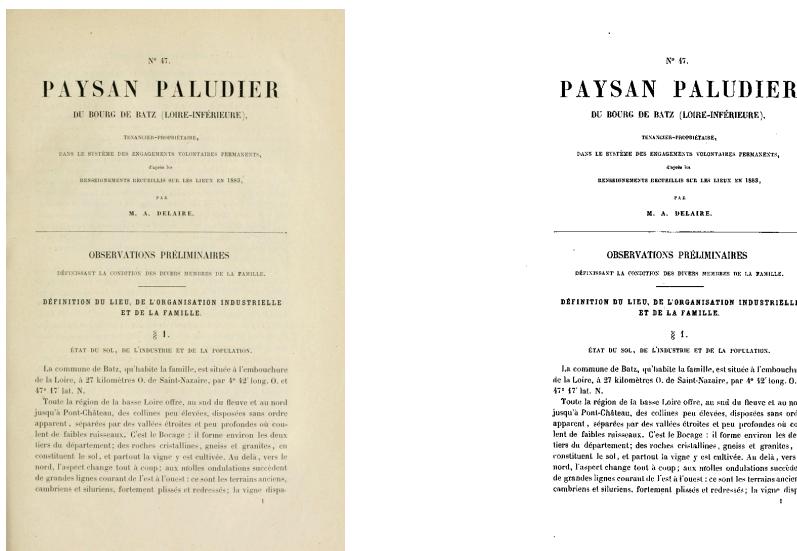


FIGURE 2.2 – Exemple d'une binarisation : s. 2 vol. 1, n° 47, p. 1.

Enfin, les images retenues ont été binarisées, c'est-à-dire passées en valeur de noir ou blanc (fig. 2.2). Cette opération permet d'augmenter la qualité des textes et donc leur détection par les programmes de transcription automatique. Un autre avantage est

13. *Ibid.*

qu'elle supprime les annotations manuscrites, par exemple celles ajoutées par les agents des bibliothèques et relatives au classement du volume.

Dès lors que le corpus était nettoyé, les opérations de transcription et d'encodage pouvaient débuter.

Chapitre 3

Un encodage automatique

3.1 Choix de l'automatisation

Le programme *Time Us* a choisi de procéder lui-même à la transcription des treize volumes des *Ouvriers des deux mondes* et à l'encodage du texte brut ainsi obtenu.

Cela n'était pas envisagé ainsi au départ, un budget ayant été prévu pour externaliser cette étape. Néanmoins, l'équipe ALMAAnCH d'Inria a fait valoir les risques inhérents à une telle opération, et notamment la difficulté à rédiger un cahier des charges suffisamment rigide pour pouvoir contrôler les différentes étapes de réalisation et les orientations scientifiques décidées lors de celles-ci.

Le choix s'est donc porté sur une réalisation en interne avec la possibilité de la conduire manuellement (un humain réalise les transcriptions avant de les encoder) ou automatiquement (un humain écrit un programme qui prend les images en entrée et fournit des fichiers XML en sortie)¹. La solution du travail manuel avait été envisagée dans un premier temps, notamment par le Centre Maurice-Halbwachs, grâce au recrutement d'un ingénieur d'études à temps plein pour six mois².

ALMAAnCH a cependant proposé de prendre en charge l'écriture du programme permettant d'automatiser l'opération, en insistant sur les apports méthodologiques et scientifiques que cela représenteraient pour le programme³. Le succès d'une telle opération lui permettrait en effet de développer ses compétences dans l'acquisition et l'encodage automatique de texte imprimé.

Ajoutons que cette proposition correspond à la dynamique interne dont nous avons constaté la prégnance sur les membres d'ALMAAnCH dès le début du stage. Une réflexion y est systématiquement menée pour étudier l'opportunité d'automatiser les tâches répétitives dans un objectif évident de gain de temps, mais également de développement

1. Id., *Constituer un corpus pour la fouille de texte — de la transcription des documents d'archives à l'annotation....*, p. 52.

2. *Ibid.*

3. *Ibid.*

expérimental de nouvelles techniques de traitement des fichiers.

Pour mener à bien cette phase du programme, deux contrats d'ingénieur d'étude de trois mois chacun (octobre à décembre 2018, février à mai 2019) ont été financés. La fiche de poste comptait sept missions⁴ :

- Réaliser un benchmark de toutes les solutions d'OCR disponibles et adaptées pour le projet ;
- Établir une méthodologie de traitement des documents ;
- Télécharger les images, les contrôler et les pré-traiter ;
- Réaliser l'OCR ;
- Analyser la structure logique des enquêtes pour identifier les indices textuels et typographiques permettant la structuration des transcriptions ;
- Établir un schéma TEI traduisant la structure logique ;
- Rédiger un programme cherchant ces indices et implémentant le schéma TEI.

3.2 Le script LSE-OD2M

La transcription automatique d'un document imprimé nécessite l'usage d'un logiciel d'OCR (*Optical Character Recognition*, en français *reconnaissance optique de caractères*). Quatre étapes sont nécessaires afin d'obtenir un résultat satisfaisant : préparer les images en amont afin de faciliter la reconnaissance des caractères, segmenter l'image en zones d'information, extraire le texte et enfin contrôler et corriger le rendu⁵.

Nous avons déjà évoqué la préparation des images dans le chapitre précédent, la correction correspondant quant à elle au travail produit pendant notre stage. Nous nous arrêterons donc sur les deuxième (segmentation) et troisième (transcription automatique) phases dans le propos qui suit, en y adjoignant la structuration du texte sous forme de fichiers XML-TEI.

Sur ces trois étapes, les deux dernières ont été entièrement prises en charge par un script Python dénommé *Logical Structure Extraction from Les Ouvriers des Deux Mondes* (LSE-OD2M), écrit par Alix Chagué de février à mai 2019.

3.2.1 Segmentation

Le but d'une segmentation est d'identifier dans l'image d'un texte des zones qui correspondent à des sous-ensembles logiques de ce texte. Il peut s'agir de paragraphes, de lignes ou de mots, la granularité pouvant être poussée jusqu'aux caractères⁶. Ce moment est crucial pour l'OCR, puisque la reconnaissance des caractères s'effectuera sur les

4. Informations transmises par Alix Chagué.

5. Romain Karpinski et Abdel Belaid, *Rapport Evaluation des OCR*, Research Report, LORIA - Université de Lorraine, 2016, URL : <https://hal.inria.fr/hal-01356824>, p. 1.

6. *Ibid.*, p. 3.

zones ainsi segmentées⁷. Il est identifié depuis de nombreuses années⁸ comme une source majeure des erreurs présentes dans le résultat final⁹.

Plusieurs caractéristiques des *Ouvriers des deux mondes* facilitent leur transcription. D'une part, il s'agit d'un imprimé du XIX^e siècle, ce qui garantit un texte relativement régulier dont l'alphabet correspond au nôtre (les *s* sont ronds et non longs). D'autre part, les pages sont organisées en une seule colonne. Les erreurs « de fusion », par exemple un texte dont les deux colonnes seraient identifiées comme un seul bloc, ne peuvent pas survenir¹⁰.

Les Ouvriers des deux mondes ne contiennent cependant pas que des chaînes de caractères organisées en lignes. On y trouve des « objets » : outre des reproductions de photographies occupant tout ou partie de la page dans les monographies des deuxième et troisième séries, il s'agit principalement de tableaux (fig. 3.1). Les paragraphes 14 à 16 consistent ainsi en des tableaux de budget qui s'étendent sur plusieurs pages. Ils suscitent un intérêt particulier pour le programme *Time Us*, dans la mesure où ces budgets permettent de reconstituer les différents postes de dépenses des familles — et notamment le textile — et les ressources qui leur sont allouées. Des tableaux peuvent également se trouver dans les autres sections, entre deux paragraphes.

La segmentation en elle-même a été réalisée à l'aide du logiciel ABBYY *FineReader*¹¹, en tant que service proposé par le logiciel *Transkribus*¹². Ce dernier est une plate-forme de transcription et d'annotation développée par le programme européen READ (*Recognition and Enrichment of Archival Documents*)¹³.

Un total de 6 668 images issues des *Ouvriers des deux mondes* a été chargé dans *Transkribus*. *FineReader* a ensuite analysé leur mise en page et fourni un résultat sous la forme d'un fichier XML ALTO¹⁴. ALTO (*Analyzed Layout and Text Object*) est un format maintenu par la Bibliothèque du Congrès qui « conserve les informations de mise en page et le texte reconnu par OCR des pages de tout type de documents imprimés »¹⁵. Les

7. Richard Casey et Eric Lecolinet, « A Survey of methods and strategies in character segmentation », *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18 (août 1996), p. 690-706, DOI : 10.1109/34.506792, p. 3.

8. *Ibid.*, p. 4.

9. Khaoula Elagouni, Christophe Garcia, Franck Mamalet et Pascale Sébillot, « Combining Multi-Scale Character Recognition and Linguistic Knowledge for Natural Scene Text OCR », dans *10th IAPR International Workshop on Document Analysis Systems, DAS*, Gold Coast, Queensland, Australia, 2012, p. 120-124, URL : <https://hal.archives-ouvertes.fr/hal-00753908>, p. 120.

10. R. Karpinski et A. Belaid, *Rapport Evaluation des OCR...*, p. 5.

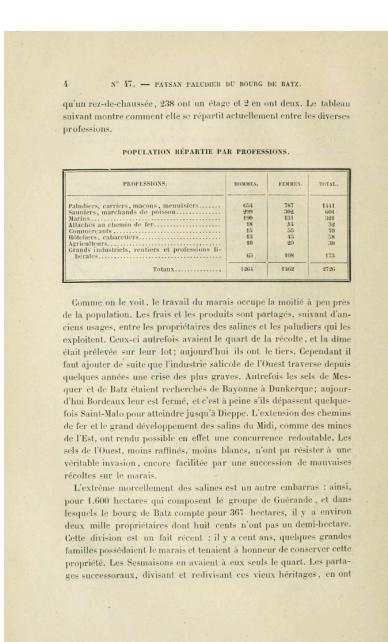
11. *FineReader* produit également une transcription, mais celle-ci est de trop mauvaise qualité pour pouvoir être exploitée

12. A. Chagué, *Constitution d'un corpus textuel sur les monographies de Le Play...*

13. Présentation du logiciel : <https://readcoop.eu/transkribus/> (consulté le 21 septembre 2020).

14. Abdel Belaïd, Yves Rangoni et Ingrid Falk, *Représentation des données en XML pour l'analyse d'images de documents*, juil. 2007, URL : <http://lodel.irevues.inist.fr/cide/index.php?id=147>.

15. « *ALTO stores layout information and OCR recognized text of pages of any kind of printed documents* » : « *ALTO Principles* », Bibliothèque du Congrès (<https://www.loc.gov/standards/alto/description.html>, consulté le 21 septembre 2020).



Comme on le voit, le travail du marin occupe la moitié à peu près de la population. Les fruits et les produits sont partagés, suivant d'anciens usages, entre les propriétaires des salines et les paludiers qui les exploitent. Cenc-ci autrefois avaient le quart de la récolte, et la dîme était prélevée sur leur lot; aujourd'hui ils ont le tiers. Cependant il faut ajouter de suite que l'industrie salicole de l'Ouest traverse depuis quelques années une crise des plus graves. Autrefois les sels de Mesquer et de Batz étaient recherchés par la marine anglaise, mais depuisfois Saint-Malo pour atteindre jusqu'à Béring. L'extension des chemins de fer et le grand développement des salines du Midi, comme des mines de l'Est, ont rendu possible en effet une concurrence redoutable. Les sels de l'Ouest, moins raffinés, moins blancs, n'ont pas résister à une véritable invasion, encore facilitée par une succession de mauvaises récoltes sur le marais.

L'extrême morcellement des salines est un autre embarras; ainsi, pour 1.600 hectares qui composent le groupe de Guérande, et dans les environs de 1.000 autres, il existe quelque 1.500 propriétaires, dont deux mille propriétaires dont huit cent n'ont pas un demi-hectare. Cette division est un fait récent : il y a cent ans, quelques grandes familles possédaient le marais et tentaient à honneur de conserver cette propriété. Les Sennassons en avaient à eux seuls le quart. Les partages successoraux, divisant et rédivisant ces vieux héritages, en ont

ELEMENTS DIVERS DE LA CONSTITUTION SOCIALE. 173

EXTRAIT DES ACTIVITÉS DU OCTOBRE 1895

PROFESSIONS	NOMBRE DE PERSONNES OCCUPÉES AU TRAVAIL DES INDUSTRIES		TYP
	ROISSES	FEMMES	
Paysans, agriculteurs, maraîchers, bûcherons, marchands de poisson.....	1.154	1.156	—
Mariés.....	65	65	—
Marieuses au chemin de fer.....	52	52	—
Marchands.....	23	23	—
Hôteliers, cabaretiers.....	55	55	—
Grands cultivateurs, restaurateurs et professions liées à l'agriculture.....	1.154	1.156	—
Total.....	3.073	3.073	—

EXTRAIT DES ACTIVITÉS DU OCTOBRE 1895

PROFESSIONS	NOMBRE DE PERSONNES OCCUPÉES AU TRAVAIL DES INDUSTRIES		TYP
	ROISSES	FEMMES	
Paysans, agriculteurs, maraîchers, bûcherons, marchands de poisson.....	1.154	1.156	—
Mariés.....	65	65	—
Marieuses au chemin de fer.....	52	52	—
Marchands.....	23	23	—
Hôteliers, cabaretiers.....	55	55	—
Grands cultivateurs, restaurateurs et professions liées à l'agriculture.....	1.154	1.156	—
Total.....	3.073	3.073	—

EXTRAIT DES ACTIVITÉS DU OCTOBRE 1895

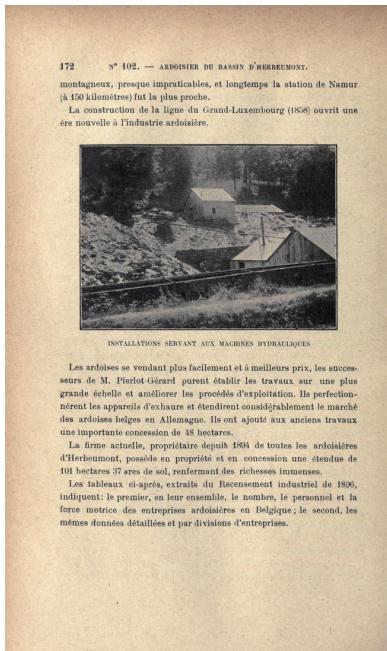
PROFESSIONS	NOMBRE DE PERSONNES OCCUPÉES AU TRAVAIL DES INDUSTRIES		TYP
	ROISSES	FEMMES	
Paysans, agriculteurs, maraîchers, bûcherons, marchands de poisson.....	1.154	1.156	—
Mariés.....	65	65	—
Marieuses au chemin de fer.....	52	52	—
Marchands.....	23	23	—
Hôteliers, cabaretiers.....	55	55	—
Grands cultivateurs, restaurateurs et professions liées à l'agriculture.....	1.154	1.156	—
Total.....	3.073	3.073	—

EXTRAIT DES ACTIVITÉS DU OCTOBRE 1895

PROFESSIONS	NOMBRE DE PERSONNES OCCUPÉES AU TRAVAIL DES INDUSTRIES		TYP
	ROISSES	FEMMES	
Paysans, agriculteurs, maraîchers, bûcherons, marchands de poisson.....	1.154	1.156	—
Mariés.....	65	65	—
Marieuses au chemin de fer.....	52	52	—
Marchands.....	23	23	—
Hôteliers, cabaretiers.....	55	55	—
Grands cultivateurs, restaurateurs et professions liées à l'agriculture.....	1.154	1.156	—
Total.....	3.073	3.073	—

(a) s. 2 vol. 1, n° 47, p. 4.

(b) s. 3 vol. 1, n° 102, p. 173.



(c) s. 2 vol. 1, n° 47, p. 3.

(d) s. 3 vol. 1, n° 102, p. 172.

FIGURE 3.1 – Exemples d'objets graphiques dans les pages des *Ouvriers des deux mondes*.

fichiers ALTO contiennent « des balises *TextBlock*/*TextLine* renvoyant les coordonnées des lignes de texte, ainsi que des balises *GraphicalElement*, pour les éléments graphiques, qui indiquent notamment les éléments de type *table* »¹⁶.

Il était dès lors possible de passer à la reconnaissance des caractères.

3.2.2 Reconnaissance des caractères

La reconnaissance des caractères est la deuxième étape de l'OCR. Elle nécessite d'établir un modèle à partir d'une « vérité terrain » (*ground truth*), c'est-à-dire la transcription parfaite d'un échantillon du corpus. Le logiciel calcule ensuite le taux de similarité entre les formes qu'il détecte et celles qui se trouvent dans le modèle. Pour ce faire, il détecte le premier caractère de la zone segmentée, identifie ses signes distinctifs et, à partir de ceux-ci, l'associe au symbole du modèle fourni avec qui il partage le plus de points communs¹⁷. Cette séquence est répétée jusqu'à ce que plus aucun caractère ne soit détecté dans le segment, le logiciel réitérant alors l'opération sur le segment suivant jusqu'à arriver au dernier du document.

Dans le cadre du projet *Time Us*, étant donné les choix techniques déjà opérés pour le traitement des autres documents, deux solutions étaient envisageables pour la reconnaissance des caractères des *Ouvriers des deux mondes*. La première était l'usage de *Transkribus*, la seconde utilisait *Kraken*.

Kraken est un logiciel libre développé en Python par Benjamin Kiessling¹⁸ à partir du système *OCRopus*¹⁹. Dans le processus d'OCR, *Kraken* se passe de l'étape de la segmentation des mots et des caractères grâce à un « réseau neuronal artificiel » qu'il utilise « pour analyser l'image d'une ligne de texte, la séquence d'entrée, en une suite de caractères, la séquence de sortie »²⁰. Ce point est l'apport principal de *Kraken* par rapport à tous les logiciels similaires tels que *Transkribus* : grâce à un apprentissage profond (*deep learning*), il peut analyser un nombre important d'écritures de toute époque, latines et même arabes, et en proposer une transcription²¹.

« En 2018 », note cependant Alix Chagué, « il n'était pas possible d'obtenir une

16. A. Chagué, *Constitution d'un corpus textuel sur les monographies de Le Play...*

17. R. Casey et E. Lecolinet, « A Survey of methods and strategies in character segmentation »..., p. 3.

18. Chercheur du laboratoire d'humanités numériques « Alexander von Humboldt » de l'Université de Leipzig (*Leipzig University's Alexander von Humboldt Chair for Digital Humanities*). Benjamin Kiessling a également fait partie de l'équipe ALMAñACH de 2018 à 2019.

19. Benjamin Kiessling, *Kraken — a Universal Text Recognizer for the Humanities*, rapp. tech., Université PSL et Université de Leipzig, 2019, DOI : 10.34894/Z9G2EX.

20. « [Kraken's] recognition engine operates as a segmentation-less sequence classifier using an artificial neural network to map an image of a single line of text, the input sequence, into a sequence of characters, the output sequence : *Ibid*.

21. B. Kiessling, Matthew Thomas Miller, Maxim G. Romanov et Sarah Bowen Savant, « Important New Developments in Arabographic Optical Character Recognition (OCR) », *Al-‘Usur al-Wusta : The Journal of Middle East Medievalists*, 25 (2017), p. 1-13, DOI : 10.17613/M6TZ4R.

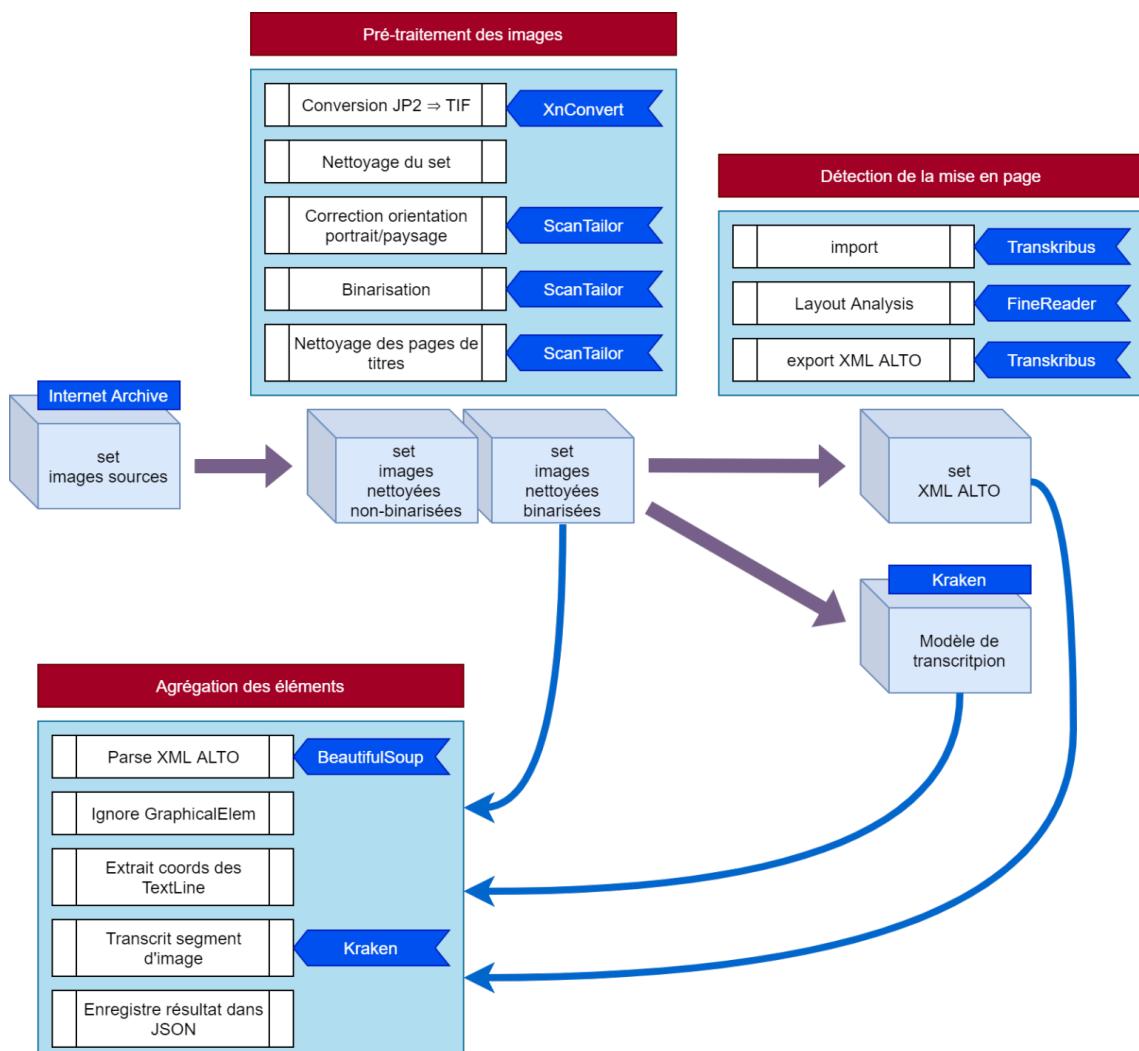


FIGURE 3.2 – Schématisation des étapes suivies pour l'extraction du texte et des informations de mise en page (source : Alix Chagué, <https://timeus.hypotheses.org/626>).

analyse de la mise en page complète avec *Kraken* »²². Mais le logiciel permettait déjà d'entraîner un modèle pour l'OCR. Cela était également possible avec *Transkribus*, mais le fichier modèle ne peut pas être exporté depuis cette plate-forme, alors que l'opération est possible avec *Kraken*, qui donne de meilleurs résultats, sous la forme de fichiers `.mlmodel`. L'équipe ALMAnaCH a donc choisi de recourir à *Kraken* pour la phase de transcription (entraînement du modèle puis reconnaissance des caractères)²³.

Les fichiers « vérité terrain » étaient composés des transcriptions de 1300 lignes (fichiers texte) et de leurs images (fichiers image). Pour s'entraîner, *Kraken* itère sur chaque couple de fichiers, transcrit le fichier image et compare le résultat au contenu du fichier texte. Le passage en revue de l'ensemble des données (ici, les 1300 paires) constitue une *epoch*. Un taux de réussite (*accuracy report*) est calculé à la fin de chaque itération. Il se compose de trois données numériques : le taux de réussite (positif ou négatif), le

22. A. Chagué, *Constitution d'un corpus textuel sur les monographies de Le Play...*

23. *Ibid.*

```
$ ketos train output_dir/*.png
Building training set [#####
Building validation set [#####
[270.2364] alphabet mismatch {'9', '8', '!', '3', ' ', '4', '1', '7', '5'
Initializing model ✓
Accuracy report (0) -1.5951 3680 9550
epoch 0/-1 [#####] 788/788
Accuracy report (1) 0.0245 3504 3418
epoch 1/-1 [#####] 788/788
Accuracy report (2) 0.8445 3504 545
epoch 2/-1 [#####] 788/788
Accuracy report (3) 0.9541 3504 161
epoch 3/-1 [-----] 13/788 0d 00:22:09
...

```

FIGURE 3.3 – Exemple d'un entraînement opéré par *Kraken*. Trois *epochs* (n° 0 à 2) sont arrivées à leur terme, la quatrième étant en cours de chargement (n° 3). Le taux de réussite augmente à chaque nouvelle *epoch* : de -1.5951 pour la première (9550 fautes sur... 3680 caractères détectés), il atteint 0.8445 pour la troisième (545 fautes sur 3504 caractères détectés). Capture d'écran du site de *Kraken* (<http://kraken.re/training.html>, consulté le 21 septembre 2020).

nombre de caractères contenus dans l'*epoch* et les faux-positifs (c'est-à-dire le nombre de caractères mal reconnus). Plusieurs dizaines d'*epochs* sont nécessaires pour parvenir à un taux de réussite optimal et donc à un modèle fiable (fig. 3.3).

Une fois ce modèle obtenu, la phase de transcription automatique pouvait commencer. *Kraken* pouvant être utilisé comme une librairie du langage Python, il a pu être envisagé de fédérer les étapes de transcription et d'encodage des textes obtenus sous l'égide d'un même script, LSE-OD2M. Son fonctionnement débute ainsi :

1. Il prend pour paramètres les fichiers images des *Ouvriers des deux mondes* et les fichiers ALTO produit par *Transkribus* ;
2. Chaque image est associé à son fichier ALTO ;
3. Les coordonnées des lignes de texte sont extraits de l'ALTO et le script lance la transcription de ces lignes à l'aide de *Kraken* et du modèle précédemment obtenu.
4. Les transcriptions de chacune des 6 608 images sont stockées dans autant de fichiers XML-TEI.

Ces transcriptions relèvent du texte brut, c'est-à-dire que la donnée n'y est pas structurée.

3.2.3 Structuration

C'est à ce stade que l'analyse de la structure logique menée précédemment est prise en compte (Annexe A.2).

1. Les limites des chapitres (monographies et éléments de paratexte) sont identifiées en fonction des pages de titres qui marquent le début de celui en cours et la fin du précédent.
2. Les transcriptions sont rassemblées par chapitre.
3. Si un chapitre est identifié comme étant une monographie, alors le script y lance une recherche pour détecter les titres de la structure logique. Le script est rendu tolérant par une distance de Levenshtein (ou distance d'édition) souple afin de reconnaître un titre en dépit d'une transcription partiellement fautive. La distance de Levenshtein mesure l'écart entre deux chaînes de caractères, ici entre le modèle (le titre standard) et sa transcription.
4. Un arbre XML est créé en respectant les différents niveaux de la structure logique.
5. En sortie, le script renvoie :
 - Treize fichiers « source » avec le contenu de chaque volume ;
 - Deux cent vingt-trois fichiers correspondant à autant de chapitres détectés.

3.3 Les fichiers XML-TEI

Ces fichiers XML sont structurés avec un schéma TEI.

XML (*Extensible Markup Language*, en français *langage de balisage extensible*) est un format de données conçu pour la description des documents textuels, ne possédant pas de jeu de balises prédéfini²⁴.

La TEI (*Text Encoding Initiative*) est un schéma de données XML commencé en 1987 dont le but est de « fournir des recommandations pour la création et la gestion sous forme numérique de tout type de données créées et utilisées par les chercheurs en sciences humaines »²⁵. Ses trois caractéristiques principales sont qu'il « s'intéresse au sens du texte plutôt qu'à son apparence », qu'il est « indépendant de tout environnement logiciel particulier » et qu'il a été conçu par une communauté scientifique qui le maintient toujours aujourd'hui à travers le *TEI Consortium* fondé en 2000²⁶.

Ce format conjugue ainsi une certaine « flexibilité et un intérêt centré sur les besoins scientifiques »²⁷. L'utiliser pour encoder les transcriptions des *Ouvriers des deux mondes*, c'est donc décrire de la meilleure manière possible l'organisation originale du texte dans les volumes imprimés tout en assurant la pérennité de la donnée ainsi produite.

Un document XML peut être représenté sous une forme arborescente, « l'arbre XML » (fig. 3.4). Il débute par un élément racine (`<TEI>`) qui contient tous les autres.

24. Définition d'Ariane Pinche : https://github.com/ArianePinche/coursTNAH_XML-TEI/blob/master/seance01/InitiationXML.md (consulté le 21 septembre 2020).

25. Lou Burnard, *Qu'est-ce que la Text Encoding Initiative ?, Comment ajouter un balisage intelligent aux ressources numériques*, Marseille, 2015, DOI : 10.4000/books.oep.1237, p. 9.

26. *Ibid.*

27. *Ibid.*, p. 10.

Les trois grandes sections de l’arbre sont le `<teiHeaer>`, les `<facsimile>` et le `<text>`. Le `<teiHeaer>` est l’en-tête où se trouvent les métadonnées du document. Il est ici réduit à son minimum, n’étant composé que d’une description bibliographique (`<fileDesc>`) : le titre et ses auteurs (`<titleStmt>`), l’organisation responsable de la publication, la date et le lieu de celle-ci (`<publicationStmt>`), et enfin une description sommaire des *Ouvriers des deux mondes* (`<sourceDesc>`)²⁸. L’en-tête contient également un `<encodingDesc>` où le lien est fait avec le programme ANR ; l’action du script LSE-OD2M est également mentionnée²⁹.

Le document continue avec une succession de balises `<facsimile>` dont le but est de conserver le lien entre le texte et son segment dans l’image d’origine (fig. 3.5)³⁰. L’élément suivant, `<surface>`, possède quatre attributs : `@ulx` et `@uly` qui ont pour valeurs l’abscisse *x* et l’ordonnée *y* du coin supérieur gauche du bloc de segmentation (la page), et `@urx` et `@ury` contenant les valeurs de l’abscisse *x* et de l’ordonnée *y* du coin inférieur droit. Ces coordonnées sont directement reprises des fichiers XML ALTO obtenus avec *FineReader*. On trouve ensuite un élément `<graphic>` avec un attribut `@url` ayant pour valeur le chemin local vers l’image de la page, et un élément `<zone>` contenant lui-même un ensemble de balises `<zone>`. Ces dernières contiennent les coordonnées des blocs de segmentation à l’intérieur du bloc supérieur défini par `<surface>` ; ces sous-blocs sont catégorisés par un attribut `@rendition` ayant pour valeur `paragraph` lorsqu’il s’agit de texte ou `table` pour un tableau.

Le `<teiHeaer>` et les éléments `<facsimile>` montrent que la structuration d’un texte brut issu d’une OCR ne consiste pas simplement en l’encodage de la transcription. Il s’agit aussi de produire des métadonnées permettant de décrire le contenu du fichier.

La dernière section de l’arbre, `<text>`, contient les transcriptions structurées au sein de son élément `<body>`. Notons que l’implémentation de la structure logique par LSE-OD2M n’a pas donné de résultats optimaux et a fait l’objet d’une reprise majeure lors du stage. Ces résultats restent néanmoins satisfaisants au regard du temps alloué pour l’écriture et le fonctionnement du script (trois mois).

Le principe retenu est le suivant : chaque division de la structure est enchaînée dans un élément `<div>` pour lequel un attribut `@type` vient préciser son niveau. De manière assez classique, cet encodage traduit ainsi la hiérarchie sémantique interne aux monogra-

28. « Les *Ouvriers des deux mondes* sont des recueils d’enquêtes sociologiques publiées pendant la deuxième moitié du XIX^e siècle. Ces enquêtes ont été rassemblées sous la forme de 3 séries de plusieurs volumes. Accès aux fichiers images sources : http://timeusage.paris.inria.fr/mediawiki/index.php/Aperçu_des_états#Les_Ouvriers_des_Mondes ».

29. « Cette édition numérique des *Ouvriers des deux mondes* a été réalisée dans le cadre du projet ANR *Time Us* (<http://larhra.ish-lyon.cnrs.fr/anr-time-us>). Elle est générée automatiquement grâce au programme LSE-OD2M (<https://gitlab.inria.fr/almanach/time-us/LSE-OD2M>), développé par Alix Chagué, équipe Inria ALAMAnaCH ».

30. *TEI Guidelines, 11 Representation of Primary Sources, 11.1 Digital Facsimiles* (<https://tei-c.org/release/doc/tei-p5-doc/fr/html/PH.html#PHFAX>, consulté le 21 septembre 2020).

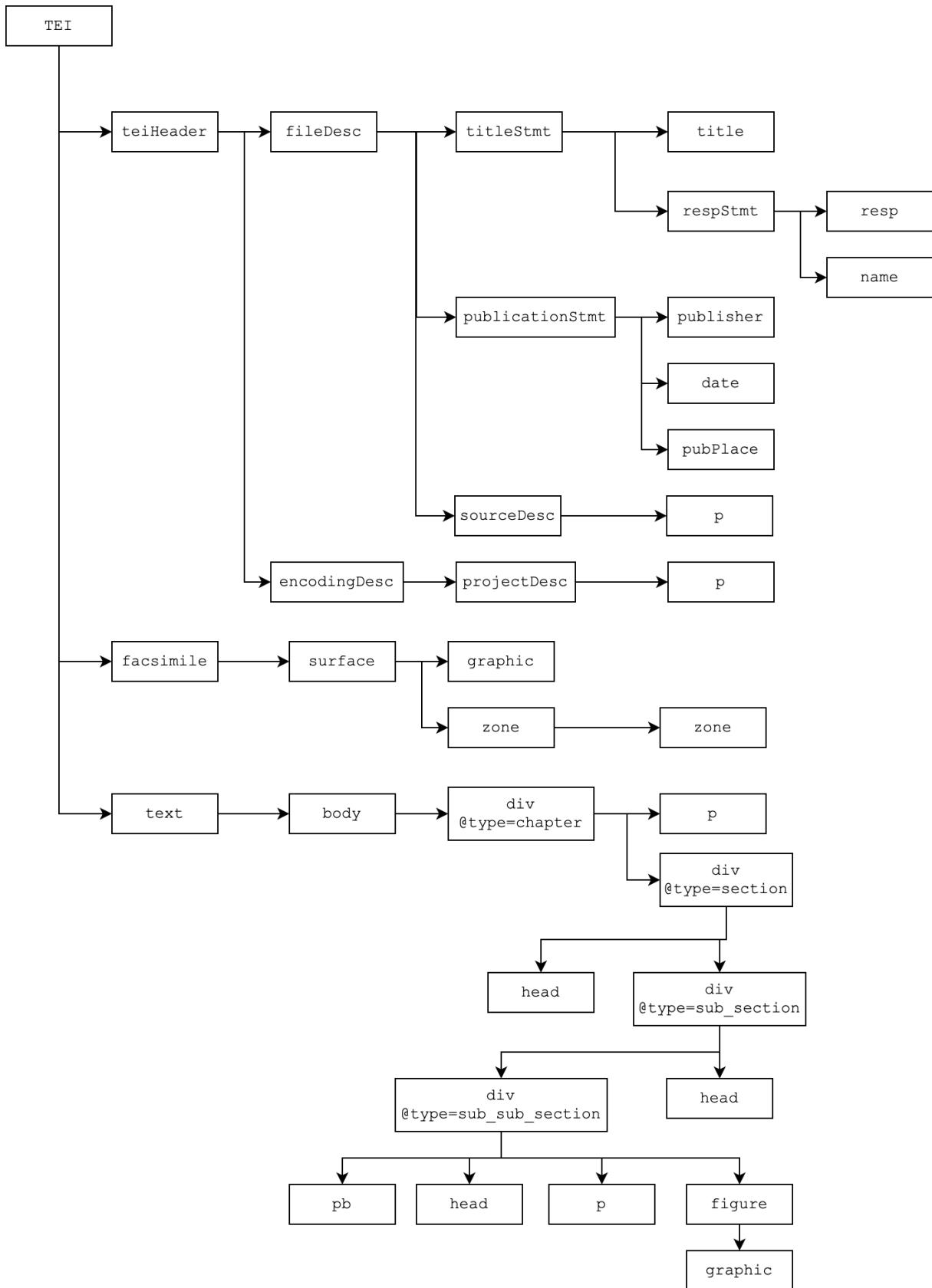


FIGURE 3.4 – Représentation de l’arbre XML utilisé par le script LSE-OD2M pour encoder les transcriptions des monographies des *Ouvriers des deux mondes*.

```

<facsimile xml:id="facs_464">
  <surface lrx="2151" lry="3478" urx="0" ury="0">
    <graphic url="../images/bin/s2lesouvriersdes03sociuoft_0509.tif"/>
    <zone rendition="printspace">
      <zone lrx="1833" lry="1570" rendition="table" ulx="0" uly="379" xml:id="facs_464_g_1"/>
      <zone lrx="1753" lry="1731" rendition="paragraph" ulx="22" uly="1618" xml:id="facs_464_p_2"/>
      <zone lrx="1828" lry="1876" rendition="paragraph" ulx="0" uly="1758" xml:id="facs_464_p_3"/>
      <zone lrx="1104" lry="2281" rendition="paragraph" ulx="0" uly="2203" xml:id="facs_464_p_4"/>
      <zone lrx="1099" lry="2397" rendition="paragraph" ulx="19" uly="2363" xml:id="facs_464_p_5"/>
      <zone lrx="1324" lry="2008" rendition="paragraph" ulx="1167" uly="1978" xml:id="facs_464_p_6"/>
      <zone lrx="1837" lry="2724" rendition="paragraph" ulx="0" uly="2607" xml:id="facs_464_p_7"/>
    </zone>
  </surface>
</facsimile>

```

FIGURE 3.5 – Exemple d'un ensemble `<facsimile>` dans le fichier source du troisième volume de la deuxième série.

phie grâce à une imbrication de `<div>` jusqu'à arriver au texte lui-même³¹. Les valeurs de l'attribut `@type` sont calquées sur les divisions des documents du langage de composition LATEX : `\chapter{}`, `\section{}`, `\subsection{}`, `\subsubsection{}`. Cela a pour but de faciliter une éventuelle transformation des documents XML-TEI au format LATEX³².

La monographie correspond à une `<div>` de type `chapter`. Les ensembles de niveau A (page de titre, *Observations préliminaires* et *Notes*) sont de type `section`, les titres de niveau I `sub_section` et les paragraphes `sub_sub_section` (*cf.* Annexe A.2). Dans chaque niveau, le titre en lui-même est encodé au sein d'un élément `<head>`, le texte se trouvant ensuite dans des `<p>` (fig. 3.6).

Pour les éléments de paratexte, tels que les avertissements liminaires de chaque volume, seul le niveau `chapter` a été utilisé, l'ensemble du texte se trouvant dans des `<p>`.

Le script LSE-OD2M a été lancé sur le *cluster* (ou *grappe de serveurs*) RIOC d'Inria, où il a fallu un peu plus de quinze heures pour traiter l'ensemble des *Ouvriers des deux mondes*. Un *cluster* est un groupe d'ordinateurs inter-connectés où une « reine » (le *master node*) répartit grâce à un « planificateur » (*scheduler*) les tâches qui lui sont demandées à des « ouvrières » (les *slave nodes*)³³. Quatre des « ouvrières » de RIOC possèdent des GPU (*graphics processing unit*). Équipées d'une carte graphique, ces processeurs sont capables de réaliser du calcul vectoriel et peuvent ainsi traiter les images — par exemple les pages des *Ouvriers des deux mondes* — très rapidement. Cette architecture en réseau permet d'augmenter de façon significative la vitesse de calcul et donc l'exécution des tâches.

Au final, les treize fichiers source créés par LSE-OD2M représentent 25,5 Mo. Ils ont été divisés en 223 fichiers (25,7 Mo) nommés de la façon suivante : `s[numéro de la série]t[numéro du volume]_chapt_[numéro du chapitre dans le volume].xml` (Annexe A.1).

31. Thibault Clérice, « Les outils CapiTainS, l'édition numérique et l'exploitation des textes », *Médiévales*, 73 (automne 2017), p. 115-131, DOI : 10.4000/medievales.8211, p. 117.

32. A. Chagué, *Constitution d'un corpus textuel sur les monographies de Le Play...*, p. 52.

33. RIOC Architecture, https://sed-paris.gitlabpages.inria.fr/rioc/rioc_architecture/rioc_architecture.html (consulté le 21 septembre 2020).

```

<div n="001" type="section">
  <div n="001" type="sub_section">
    <div n="001" type="sub_sub_section">
      <head facs="#fac350_p_8" type="section" xml:id="para_31_325_8">
        OBSERVATIONS PRÉLIMINAIRES
      </head>
      <p facs="#fac350_p_9" xml:id="para_31_325_9">
        DÉFINISSANT LA CONDITION DES DIVERS MEMBRES DE LA FAMILLE.
      </p>
    </div>
  </div> A
<div n="002" type="sub_section">
  <div n="001" type="sub_sub_section">
    <head facs="#fac350_p_10" type="sub_section" xml:id="para_31_325_10">
      Définition du lieu, de l'organisation industrielle et de la famille
    </head>
    <p facs="#fac350_p_11" xml:id="para_31_325_11">
      § 1
    </p>
  </div> B
<div n="002" type="sub_sub_section">
  <head facs="#fac350_p_12" type="sub_sub_section" xml:id="para_31_325_12">
    § 1er. - ÉTAT DU SOL, DE L'INDUSTRIE ET DE LA POPULATION.
  </head>
  <p facs="#fac350_p_13" xml:id="para_31_325_13">
    La famille habite à Paris, sur la rive gauche de la Seine, vers la limite du faubourg Saint-
  </p>
  <pb facs="#fac351" n="326" source="s2lesouvriersdes03sociuoft_0386.xml" topMargin="283"/>

```

FIGURE 3.6 – Exemple de la structuration du début des *Observations préliminaires* de la monographie n° 70 par le script LSE-OD2M (*s2t3_chapt_31.xml*). L'encodage est fautif dans la mesure où des niveaux supplémentaires ont été rajoutés : *Observations...* ne devrait être que dans une **section** (les lignes des repères A n'ont pas lieu d'être), *Définitions du lieu...* dans une **sub_section** (les lignes des repères B sont fautives). On remarque également une erreur de transcription, le numéro de paragraphe (§1) étant répété. La balise finale, **<pb>**, marque un changement de page.

Au moment de commencer notre stage, nous avons donc trouvé un ensemble de 223 fichiers XML contenant les transcriptions structurées avec un schéma TEI des treize volumes des *Ouvriers des deux mondes*. Les fichiers contenaient des enquêtes sociologiques (les monographies) et des éléments de paratexte. Les chercheurs du programme ANR *Time Us* n'étaient intéressés que par les monographies, mais Inria avait la volonté de valoriser l'ensemble des fichiers.

Plusieurs points avaient été identifiés comme posant problème pour une future valorisation sous la forme d'une publication ou d'une réutilisation dans des projets de traitement automatique du langage (TAL).

Nous allons à présent exposer les différentes reprises que nous avons menées.

Deuxième partie

Une structuration à reprendre

Chapitre 4

Outils, méthodologie et gestion du projet

4.1 Outils de développement

4.1.1 *GitLab*

La gestion quotidienne du programme *Time Us* se fait grâce à un dépôt sur *GitLab*, un logiciel libre permettant aux entités comme Inria de créer une plate-forme interne de développement informatique. Sur *GitLab* se trouve le dépôt central, organisé en plusieurs dossier. La technologie *git* permet d'administrer les différentes versions du projet, sur une échelle à la fois verticale (les anciennes versions, appelées *commits*, restant accessibles à travers un historique) et horizontale (le travail peut s'effectuer sur des *branches* divergentes de la branche principale dite *master* sans affecter l'état de cette dernière). Un commentaire est ajouté par l'utilisateur à chacun de ses *commits* pour résumer ses modifications.

Chaque participant peut rapatrier le dépôt *GitLab* en local pour travailler dessus ; ce rapatriement est un *pull*. Il peut ensuite effectuer l'opération inverse, un *push*, consistant à mettre ses *commits* en ligne. Son équipe peut ainsi prendre connaissance des dernières avancées de la branche sur laquelle il travaille.

Une fois le travail dans une branche achevé, celle-ci peut être fusionnée avec *master*. *GitLab* propose une interface pour effectuer des *merge requests*. Il s'agit d'une demande de fusion, où *GitLab* affiche l'historique de la branche. Les utilisateurs peuvent ainsi contrôler l'ensemble des *commits* de la branche locale et vérifier qu'ils ne vont pas corrompre *master* en créant des conflits (*merge conflicts*).

GitLab permet enfin à ses utilisateurs d'exposer un élément posant problème, d'effectuer des suggestions ou encore de porter à l'attention de leur équipe une ressource utile à travers des *issues* (ou *tickets*). Les *issues* et les *merge requests* sont numérotées (#\d et !\d¹), écrire leurs numéros dans *GitLab* ou dans le commentaire de modification d'un

1. Par exemple, #5 fait référence à la cinquième *issue* et !5 à la cinquième *merge request*.

commit permettant de faire automatiquement référence à elles à travers un lien interne.

L'équipe ALMAnaCH possède un espace sur la plate-forme *GitLab* d'Inria, où un dossier (en accès restreint) est réservé au programme *Time Us*. C'est ici, dans un sous-ensemble, que se trouve le dépôt des *Ouvriers des deux mondes*. Le script LSE-OD2M est déposé dans un dossier externe.

La branche *master* compte quatre sous-dossiers :

- **source** contient les treize fichiers XML-TEI des volumes des *Ouvriers des deux mondes* ;
- **script** contient les scripts développés afin d'automatiser le traitement des fichiers ;
- **files** contient les fichiers XML-TEI générés à partir des fichiers sources et modifiés automatiquement ou à la main en vue de leur publication (monographies et fichiers de paratexte) ;
- **metadata** contient des fichiers de métadonnées sur le projet.

Ce dossier était notre espace de travail principal, notre première action ayant consisté en son rapatriement au niveau local. Notre méthodologie était la suivante :

1. Une branche était créée pour chaque mission ;
2. Des *commits* étaient effectués en local sur cette branche ;
3. Les *commits* d'une journée étaient mis en ligne sur *GitLab* par un *push* ;
4. Lorsque la mission était achevée, une *merge request* était ouverte ;
5. La *merge request* était acceptée et la branche fusionnée avec *master*.

Le *linter Pylint* était également implanté dans le dépôt ; il s'agit d'un système de vérification de code Python. Son rôle est de contrôler à chaque *push* la qualité du code des scripts, notamment la longueur des lignes, les intitulés des variables ou encore la documentation des fonctions. Le résultat est affiché dans une console sous forme de messages désignant le fichier concerné, la ligne de l'erreur et une désignation standardisée de celle-ci (*line too long*, *final newline missing*, etc. : fig. 4.1).

4.1.2 *PyCharm* et *Oxygen*

Pour manipuler les fichiers XML, développer et activer les scripts Python, nous usions des logiciels *Oxygen* (éditeur de code XML sous licence propriétaire) et *PyCharm* (environnement de développement intégré pour la programmation en Python, une version est sous licence libre et une seconde sous licence propriétaire).

Plusieurs fonctionnalités d'*Oxygen* ont facilité le traitement du corpus, notamment la possibilité de rassembler l'ensemble des fichiers dans un « projet ». Par ce biais, des opérations — par exemple effectuer une recherche ou remplacer une expression par une autre — peuvent être menées sur le corpus sans requérir l'ouverture successive de chaque

```

### ----- LINTER ----- ###
***** Module basic_iteration
script/basic_iteration.py:41:34: E0001: invalid syntax (<unknown>, line 41) (syntax-error)
WARNING: no score parsed from Pylint output

-----
Your code has been rated at 10.00/10 (previous run: 10.00/10, +0.00)

-----
Your code has been rated at 10.00/10 (previous run: 10.00/10, +0.00)

***** Module delete_tags_and_attrs
script/delete_tags_and_attrs.py:40:18: W1401: Anomalous backslash in string: '\d'. String constant might be missing an r prefix.
(anomalous-backslash-in-string)
script/delete_tags_and_attrs.py:40:0: C0301: Line too long (112/100) (line-too-long)
script/delete_tags_and_attrs.py:45:0: C0304: Final newline missing (missing-final-newline)
script/delete_tags_and_attrs.py:15:0: R0914: Too many local variables (16/15) (too-many-locals)
script/delete_tags_and_attrs.py:34:12: C0103: Variable name "p" doesn't conform to snake_case naming style (invalid-name)

-----
Your code has been rated at 8.08/10 (previous run: 8.08/10, +0.00)

***** Module make_index
script/make_index.py:13:0: R0914: Too many local variables (26/15) (too-many-locals)

-----
Your code has been rated at 9.67/10 (previous run: 9.67/10, +0.00)

***** Module make_index_comments
script/make_index_comments.py:60:12: W1401: Anomalous backslash in string: '\d'. String constant might be missing an r prefix.
(anomalous-backslash-in-string)
script/make_index_comments.py:60:0: C0301: Line too long (109/100) (line-too-long)
script/make_index_comments.py:62:0: C0301: Line too long (102/100) (line-too-long)
script/make_index_comments.py:16:0: R0914: Too many local variables (22/15) (too-many-locals)
script/make_index_comments.py:57:0: R1721: Unnecessary use of a comprehension (unnecessary-comprehension)

-----
Your code has been rated at 8.65/10 (previous run: 8.65/10, +0.00)

```

FIGURE 4.1 – Exemple d'un contrôle de code par *Pylint*, qui donne une note à chaque script (*module*) et détaille ensuite les erreurs détectées.

fichier. En outre, le logiciel dispose d'un système de validation du schéma du code, qui là encore peut être appliqué à tout le corpus à travers l'outil « projet ».

PyCharm est équipé d'un outil de type *linter* comparable à *Pylint*, qui permet de s'assurer de la validité et de la lisibilité du code et facilite la programmation. *Pylint* est cependant plus exigeant que le *linter* natif de *PyCharm*, un double contrôle était donc nécessaire.

4.2 Espaces de discussion

La situation de confinement d'avril à mai et le maintien de la fermeture aux stagiaires des locaux d'Inria de juin à juillet a conduit à la mise en place d'outils de discussion.

4.2.1 *Mattermost*

Mattermost est un logiciel de discussion instantanée dont le code, écrit à l'origine sous un format propriétaire, a été publié en *open source*² en 2015³. Auto-hébergé — Inria

2. Consultable sur *Github* (<https://github.com/mattermost/mattermost-server>, consulté le 21 septembre 2020).

3. Lindsay Brock, *Open source Slack-alternative reaches 1.0 : Self-host ready, Slack-compatible, MIT licensed*, 2 octobre 2015, <https://mattermost.com/blog/mattermost-3-4-16/> (consulté le 21

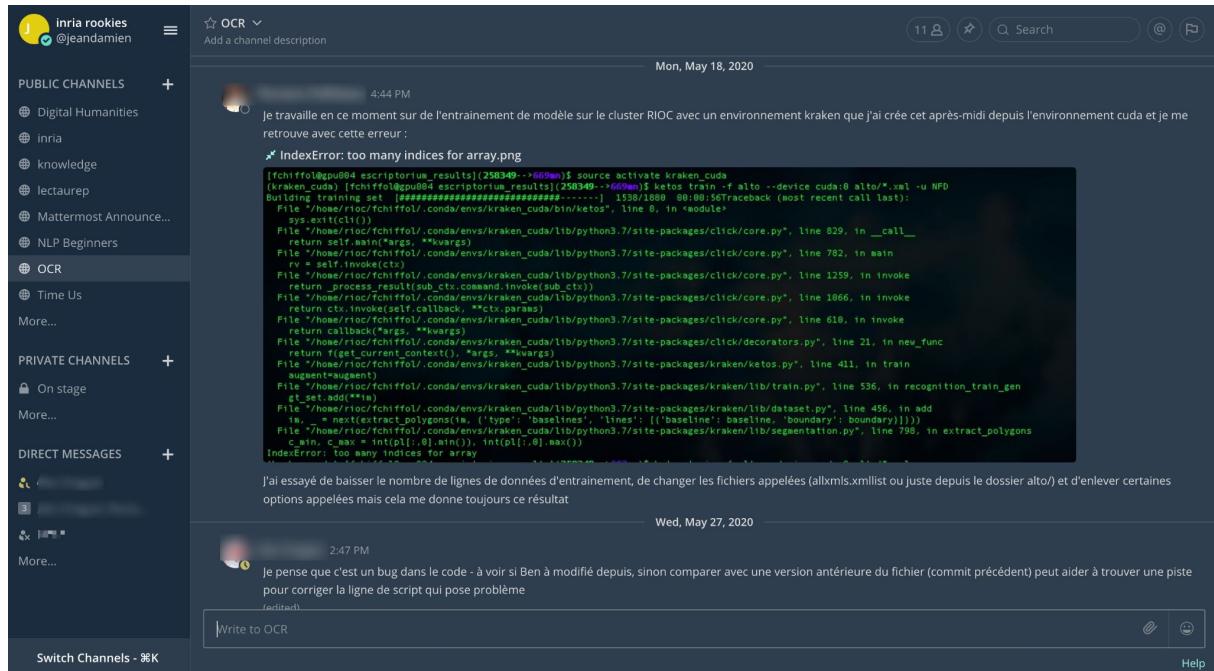


FIGURE 4.2 – Exemple de messages sur la chaîne « OCR » du *Mattermost* d’Inria : une utilisatrice demande de l'aide au sujet d'une erreur dans l'exécution de *Kraken*. Sur le volet gauche se trouve la liste des chaînes disponibles.

stocke le code dans ses propres installations et n'a pas recours à un serveur distant —, il s'agit de l'espace de discussion principal des agents d'Inria.

Composé de plusieurs « chaînes » (*public channels*) organisées de façon thématique, il leur permet d'échanger sur les différents projets et de suivre leur avancement, mais aussi d'exposer les difficultés techniques qu'ils rencontrent dans leur travail quotidien afin d'obtenir de l'aide. La chaîne « OCR » est ainsi fréquemment utilisée par les utilisateurs de *Kraken* (fig. 4.2).

4.2.2 Issues et merge requests sur *GitLab*

Le programme *Time Us* possède également une chaîne ; cependant, pour les échanges afférents à nos missions, nous usions des espaces de discussion de *GitLab*.

Les *issues* et les *merge requests* n'ont en effet pas pour seule utilité de permettre aux utilisateurs d'exposer des problèmes ou de demander des fusions de branches. Il s'agit d'espaces dynamiques qui participent pleinement à la gestion de projet en offrant à ses participants la possibilité de donner leur avis ou d'apporter des solutions, par exemple pour résoudre un *merge conflict*.

Cet aspect est facilité par l'usage du langage à balises *Markdown*⁴. Il permet une mise en forme légère (listes, liens hypertexte, diagrammes), ainsi qu'une intégration d'échan-

septembre 2020).

4. *GitLab* use de sa propre version du *Markdown*, le *GitLab Flavored Markdown* : <https://docs.gitlab.com/ee/user/markdown.html> (consulté le 21 septembre 2020).

Closed Tester la conformité du schéma

Jean-Damien Généro @jgenero1 mentioned in commit [26ff8c5c](#) 3 months ago

Jean-Damien Généro @jgenero1 · 3 months ago

Pour l'erreur évoquée ci-dessous <graphic type=illustration>, je te fais deux propositions :

- 1 Ajouter un élément <desc> dans <graphic> avec comme valeur "illustration":

```
<figure>
  <graphic facts="#fac5_124_g_1" url="#" xml:id="illu_5_121_4">
    <desc>Illustration</desc>
  </graphic>
</figure>
```

- 2 Remonter le @type au niveau supérieur de <figure> :

```
<figure type="illustration">
  <graphic facts="#fac5_124_g_1" url="#" xml:id="illu_5_121_4"/>
</figure>
```

Aucun attribut autorisé dans <graphic> ne me semble correspondre à l'utilisation de @type. Si tu tiens vraiment à laisser l'information dans la balise et que la solution avec <desc> ne convient pas il est éventuellement possible de détourner l'attribut @style, qui normalement sert à spécifier des éléments CSS ([lien](#)).

Edited by Jean-Damien Généro 2 hours ago

Alix Chagué @achague · 3 months ago

On n'est pas obligé de garder le @type au niveau de graphic, je pense que basculer sur l'élément figure est très logique. Va pour l'option [2](#) !

Jean-Damien Généro @jgenero1 mentioned in commit [491b2e62](#) 3 months ago

To Do Add a To Do

0 Assignees None - assign yourself

Milestone None

Time tracking No estimate or time spent

Due date None

Labels mission

Confidentiality Not confidential

Lock issue Unlocked

2 participants

Notifications

Reference: almanach/time-us/...

FIGURE 4.3 – Exemple de messages échangés avec une coloration syntaxique d'un code Python dans une *issue* sur *GitLab*.

tillons de code. Si ceux-ci ne sont pas fonctionnels, le *Markdown* permet de leur appliquer une coloration syntaxique, les rendant ainsi plus facile à lire (fig. 4.3).

Les échanges sur ces espaces restent accessibles après la clôture des *issues* ou des *merge requests*, constituant ainsi un historique de l'avancement du projet.

4.3 Feuille de route

4.3.1 Les missions du stage

Au commencement de notre stage, huit *issues* étaient ouvertes sur le *GitLab* des *Ouvriers des deux mondes*. Chacune correspondait à un problème ou à un point qui n'avait pas encore pu être développé. Elles constituaient donc notre « feuille de route », c'est-à-dire l'exposé des missions que nous avions à réaliser (Annexe B.1). Il nous a été demandé de les utiliser pour poser nos questions ou proposer nos solutions.

Les missions qui nous ont été confiées étaient de deux ordres.

Une première moitié consistait à contrôler les résultats du script LSE-OD2M, tant au niveau du corpus qu'à celui de chaque fichier, et à effectuer des reprises si nécessaire. Tout d'abord, il s'agissait de détecter les erreurs de découpage des fichiers de volume et d'opérer les fusions ou les séparations nécessaires. Ceci avait pour but de donner à chaque fichier un identifiant unique après s'être assuré de l'unité de son contenu (*issue 1*). Dans un second temps, nous devions nous intéresser à des éléments particuliers dans chaque fichier, à l'instar du contenu des balises <facsimile> (*issue 2*), de la qualité des transcriptions

(*issue* 6) ou encore de la validité du schéma TEI (*issue* 3) ; le contrôle de l'implémentation de la structure logique ayant constitué notre occupation principale.

L'autre moitié de nos missions avait pour but de valoriser les données des *Ouvriers des deux mondes* en menant des actions ciblées. La principale consistait à identifier les individus enquêtés en liant les informations onomastiques du deuxième paragraphe (*État civil de la famille*) à un tableau prosopographique établi par Stéphane Baciocchi, ingénieur de recherche du Centre de recherches historiques de l'EHESS (*issue* 4). De plus, nous devions implémenter dans les fichiers un système permettant la citation de passages précis afin de faciliter les études des chercheurs du programme en leur offrant la possibilité d'établir un lien direct entre leur travail et les données contenues dans les fichiers XML (*issue* 5).

Il nous a enfin été demandé de publier sur le carnet de recherche en ligne de *Time Us* des billets rendant compte de notre travail⁵.

4.3.2 Une gestion de projet ?

Nous n'avons pas utilisé d'outil véritablement dédié à la gestion de projet, détournant plutôt un outil courant de *GitLab*, les *issues*.

Des fonctionnalités de *GitLab* sont pourtant dédiées à une gestion plus fine. La principale est l'outil « tableau de bord » (*boards*). Par défaut, les *issues* sont affichées sous la forme d'une liste présentant toutes celles qui sont ouvertes, deux onglets permettant d'accéder à celles qui ont été résolues ou bien de les afficher toutes. Avec l'outil « tableau de bord », l'utilisateur a accès à un tableau de quatre colonnes, la première listant les issues ouvertes, la seconde affichant une liste de tâches (*todo list*), la troisième présentant les tâches en cours (*doing*) et la dernières les *issues* terminées.

Au-delà de la présentation des tâches, le tableau de bord est également interactif. Ainsi, sélectionner une *issue* affiche ses métadonnées (agent en charge de sa résolution, labels, temps estimé, échéance). Un système de labels — des étiquettes thématiques — permet de classer les *issues* et les tâches de la *todo list*. Le tableau de bord offre donc une vision globale du projet selon une organisation chronologique, tout en permettant des vues synthétiques ciblées.

Le fonctionnement des *boards* de *GitLab* est comparable à celui d'autres applications, comme par exemple *Trello*. Il s'agit d'un organisateur de tâche participatif, également sous forme de tableaux, qui est utilisé par ALMAnaCH et les Archives nationales pour coordonner le projet de lecture automatique des répertoires du Minutier central des notaires de Paris, Lectaurep.

Pour autant, ces outils ne sont utilisés ni dans le cadre général du programme *Time Us*, ni dans le cadre particulier des *Ouvriers des deux mondes*. Plusieurs raisons

5. Carnet consultable à cette adresse : <https://timeus.hypotheses.org/>.

peuvent l'expliquer. En premier lieu, *Time Us* s'appuie sur une documentation dont le caractère disparate — tant géographique que chronologique et typologique — se heurte à toute volonté de gestion centralisée, d'autant que chaque membre est chargé de la gestion de sa documentation locale.

Un tel outil doit également être constamment maintenu à jour afin de garantir un gain de productivité. À l'échelle des *Ouvriers des deux mondes* et de notre stage, le nombre relativement faible de missions (8) et d'intervenants (2) ne justifiaient pas la mise en place du tableau de bord *GitLab* ou la création d'un *Trello*. L'affichage basique des *issues* sous forme d'une liste se suffisait à lui-même.

Chapitre 5

Contrôle du découpage des fichiers source

5.1 Les différents niveaux d'encodage

L'encodage d'un texte brut s'effectue sur plusieurs niveaux, depuis une échelle documentaire surplombante jusqu'à celle plus fine de l'analyse scientifique. Ces niveaux sont décrits dans un document intitulé *Best Practices for TEI in Libraries*, édité par le Consortium TEI et disponible en ligne¹.

Le premier niveau est celui du découpage documentaire, soit la constitution d'un fichier qui reproduit le texte brut d'une unité codicologique et lui associe des métadonnées. Dans le cas des *Ouvriers des deux mondes*, il s'agit du découpage des treize fichiers des volumes, qui ne s'est pas fait sans erreur, et de la constitution du <teiHeaer>.

Le second niveau est celui dit de l'encodage « minimal », dont l'objectif est d'améliorer la navigation dans le document. Il s'agit d'identifier les paragraphes par des balises <p> et de lier celles-ci aux ensembles <facsimile> des images d'origine par un identifiant, tout en marquant les changements de page par des éléments <pb>.

Le troisième niveau s'intéresse au découpage éditorial, c'est-à-dire à l'identification et à la reproduction de la structure hiérarchique du texte, en l'occurrence la structure initiée par Frédéric Le Play.

Le quatrième niveau est celui de l'encodage sémantique, destiné à mettre en valeur les éléments internes au texte afin d'en faire une production électronique autonome. Dans les fichiers des *Ouvriers des deux mondes*, cela s'est traduit par l'élimination des éléments de mise en page des volumes tels que les en-têtes ou les numéros de page.

Le cinquième et dernier niveau est celui de l'annotation scientifique. Il s'agit par

1. Kevin Hawkins, Michelle Dalmau, Elli Mylonas et Syd Bauman, *Best Practices for TEI in Libraries, a guide for mass digitization, automated workflows, and promotion of interoperability with XML using the TEI*, Consortium TEI, 2018 (septembre), URL : <https://tei-c.org/extra/teiinlibraries/4.0.0/bpt1-driver.html>, 4.2. Encoding Levels.

exemple de repérer et de mettre en valeur les éléments d'onomastique ou d'effectuer un traitement particulier pour les objets graphiques

L'ensemble de ces niveaux a été contrôlé lors du stage à travers les différentes missions. Lorsqu'une correction était nécessaire, nous devions favoriser son automatisation par le biais d'un script Python. Cependant, ceci n'a pas toujours été possible, nous conduisant à engager des actions manuelles plus d'une fois.

Nous allons maintenant étudier le contrôle de ces différents niveaux. Notre travail au cours du stage ne s'est pas fait d'une manière aussi linéaire, et de fait nous allons, dans les pages qui suivent, nous détacher totalement de la chronologie que nous avons suivie.

5.2 Vérification de la cohérence documentaire

Le découpage des fichiers source a donné des résultats étonnantes pour certain volume. Ainsi, plus de trente fichiers avaient résulté du troisième volume de la deuxième série (Annexe A.1.8).

Le contrôle a été opéré manuellement, par une ouverture successive des fichiers. Les monographies devaient commencer par la reproduction de l'en-tête, point de repère du script LSE-OD2M pour le découpage. Deux erreurs majeures ont été constatée.

5.2.1 Fission horizontale lors de la segmentation

La première consistait en une erreur de découpage dans les précis n° 48 *bis*², 66 *bis*³ et 66 *ter*⁴. Le premier était scindé en dix fichiers contenant deux pages chacun (`s2t1_chapt_6.xml` à `s2t1_chapt_15.xml`), le second et le troisième en respectivement sept et quatorze fichiers contenant quatre (`s2t3_chapt_7.xml` et `s2t3_chapt_8.xml`; `s2t3_chapt_14.xml`) et deux pages (`s2t3_chapt_9.xml` à `s2t3_chapt_13.xml`; `s2t3_chapt_15.xml` à `s2t3_chapt_27.xml`). Lorsqu'il y avait deux pages, il s'agissait toujours d'un recto et d'un verso, et de deux rectos et deux versos lorsqu'il y en avait quatre.

Mis à part les premiers fichiers, qui commençaient par le titre du précis, les `<body>` de ces fichiers avaient un point commun : ils débutaient tous par les deux mêmes lignes, `<p>PRÉCIS DE MONOGRAPHIE</p> \n <p>[numéro de la page]</p>`. Or ces deux informations — le rappel du titre du chapitre et le numéro de la page courante — se trouvent sur une seule et même ligne dans les images. Une erreur de fission horizontale

2. Alexandre de Peretz, « Précis d'une monographie de l'armurier des manufactures impériales de Toula (Grande-Russie), Ouvrier propriétaire et chef de métier dans le système du travail sans engagements, par le général A. Peretz (de Saint-Petersbourg) [1886] », dans *Les Ouvriers des deux mondes*, Paris, 1887 (série 2e (1)), p. 113-132, URL : <https://archive.org/details/s2lesouvriersdes01sociuoft/page/n143>.

3. F. Escard, « Précis d'une monographie du pêcheur-côtier du Finmark (Laponie - Norvège) »...

4. S. M. Coronel, « Précis d'une monographie d'un tisserand d'Hilversum (Hollande septentrionale - Pays-Bas) »...

résultant d'une mauvaise segmentation s'était donc produite au moment de l'OCR⁵.

Dans son comportement normal, LSE-OD2M est programmé pour détecter tout ce qui relève de l'en-tête ou du pied de page et le retirer afin de permettre une reconstitution optimale des paragraphes au moment de la transformation des fichiers XML pour une éventuelle édition. Ici, le script avait été dupé par l'erreur de segmentation : il n'avait pas vu des rappels de titre et des numéros de page, mais des titres et des numéros de chapitre. En conséquence, il avait opéré une coupure à cet endroit, c'est-à-dire achevé le fichier en cours pour en amorcer un nouveau.

Une question demeure néanmoins : pourquoi trois des fichiers contenaient-ils quatre pages, à l'inverse des autres qui n'en contenaient que deux ? Le fait est que les en-têtes des troisièmes pages avaient été détectés convenablement, et donc retirés du flux du texte. Aussi la séparation n'était-elle pas nécessaire.

L'erreur a été résolue par un transfert manuel du contenu de chaque fichier particulier dans un fichier global (`s2t1_chapt_6.xml`, `s2t3_chapt_7.xml` et `s2t3_chapt_14`). Le choix a été fait de ne pas modifier la numérotation des fichiers suivants, notamment afin de pouvoir effectuer un suivi sur la durée. C'est la raison pour laquelle, dans la deuxième série, le n° 6 est suivi du n° 16 dans le premier volume et, dans le troisième, le n° 7 par le n° 14, lui-même précédant le n° 28 (Annexe A.1.6 et A.1.8, p. 102).

Cette opération a conduit à la suppression de vingt-huit fichiers. Ajoutons à cela le retrait de `s2t1_chapt_23.xml`, doublon de `s2t2_chapt_5.xml` : un total de vingt-neuf fichiers a été supprimé, ramenant le corpus de 223 à 194 unités.

5.2.2 Défaut de transcription

La seconde erreur majeure à laquelle nous avons été confrontés est le défaut d'une partie de la transcription dans six monographies (Annexe B.1.1). Le titre et l'ensemble de la partie *Observations préliminaires* manquent (n° 30⁶, 33⁷, 37⁸, 44⁹, 45¹⁰ et 46¹¹).

5. R. Karpinski et A. Belaid, *Rapport Evaluation des OCR...*, p. 5-6.

6. Léon Donnat et Ouang-Tching-Yong, « Paysans en communauté du Ning-Po-Fou (province de Tché-Kian - Chine) », dans *Les Ouvriers des deux mondes*, Paris, 1862 (série 1 (4)), chap. 30, p. 83-158, URL : <https://archive.org/details/lesouvriersdesde04sociuoft/page/82>.

7. Alexis-Félix Badier, « Compositeur-typographe de Paris (Seine - France) », dans *Les Ouvriers des deux mondes*, Paris, 1862 (série 1 (4)), chap. 33, p. 241-282, URL : <https://archive.org/details/lesouvriersdesde04sociuoft/page/240>.

8. Samuël-Mozes Coronel et F Allan, « Pêcheur côtier, maître de barques, de Marken (Hollande septentrionale - Pays-Bas) », dans *Les Ouvriers des deux mondes*, Paris, 1862 (série 1 (4)), chap. 37, p. 405-460, URL : <https://archive.org/details/lesouvriersdesde04sociuoft/page/404>.

9. U. Guérin, « Paysan-résinier de Lévignacq (Landes - France) », dans *Les Ouvriers des deux mondes*, Paris, 1885 (série 1 (5)), chap. 44, p. 315-386, URL : <https://archive.org/details/lesouvriersdesde05sociuoft/page/n347>.

10. Félicien Pariset, « Bûcheron usager de l'ancien Comté de Dabo (Lorraine allemande) », dans *Les Ouvriers des deux mondes*, Paris, 1885 (série 1 (5)), chap. 45, p. 387-458.

11. Vincent Darasse, « Paysans en communauté et colporteurs émigrants de Tabou-Douchd-El-Baar (Grande Kabylie - Province d'Alger) », dans *Les Ouvriers des deux mondes*, Paris, 1885 (série 1 (5)), chap. 46, p. 459-502, URL : <https://archive.org/details/lesouvriersdesde05sociuoft/page/n499>.

Il ne s'agit pas d'une erreur de découpage, puisque déjà présente dans les treize fichiers source. C'est néanmoins l'opération de contrôle du découpage qui a permis de s'en rendre compte.

Il ne s'agit pas d'un déficit total. L'analyse de la mise en page et la segmentation se sont effectuées de manière convenable, comme en atteste la présence de `<facsimile>` entre le `<teiHeader>` et le `<text>`, repris des éléments `<TextBlock>` des fichiers ALTO d'origine. Néanmoins, les paragraphes 1 à 16, qui peuvent contenir des figures ou des tableaux mais sont principalement composés de textes, ont été considérés comme des éléments graphiques. En conséquence, chaque page est représentée par un élément `<figure>` contenant une balise `<graphic>` que l'attribut `@fac`s relie à un `<facsimile>` ; `<pb>` venant signifier le changement de page (fig. 5.1).

```
<pb facs="#facs_116" n="113" source="lesouvriersdesde04sociuoft_0119.xml"/>
<figure type="illustration">
<graphic facs="#facs_116_g_1" url="#" xml:id="illu_8_113_35"/>
</figure>
<pb facs="#facs_117" n="114" source="lesouvriersdesde04sociuoft_0120.xml"/>
<figure type="illustration">
<graphic facs="#facs_117_g_1" url="#" xml:id="illu_8_114_36"/>
</figure>
<figure type="illustration">
<graphic facs="#facs_117_g_2" url="#" xml:id="illu_8_114_37"/>
</figure>
<figure type="illustration">
<graphic facs="#facs_117_g_3" url="#" xml:id="illu_8_114_38"/>
</figure>
<figure>
<graphic facs="#facs_117_g_4" url="#" xml:id="tble_8_114_39"/>
</figure>
<figure type="illustration">
<graphic facs="#facs_117_g_5" url="#" xml:id="illu_8_114_40"/>
</figure>
<pb facs="#facs_118" n="115" source="lesouvriersdesde04sociuoft_0121.xml"/>
<figure type="illustration">
<graphic facs="#facs_118_g_1" url="#" xml:id="illu_8_115_41"/>
</figure>
</div>
<div n="003" type="section">
<pb facs="#facs_119" n="116" source="lesouvriersdesde04sociuoft_0122.xml"/>
<head type="section" xml:id="para_8_116_1">
    NOTES
</head>
<p xml:id="para_8_116_2">
    FAITS IMPORTANTS D'ORGANISATION SOCILE, PARTICULARITÉS REMARQUABLES ; APPRÉCIATIONS GÉNÉRALES ; CONCLUSIONS.
</p>
<div n="001" type="sub_section">
<div n="001" type="sub_sub_section">
<head type="sub_sub_section" xml:id="para_8_116_3">
    (A) SUR LE RESPECT DES CHINOIS POUR L'AUTORITÉ PATERNELLE.
</head>
<p xml:id="para_8_116_4">
    Ce qui frappe surtout à l'aspect de la civilisation chinoise, ce qui semble la caractériser, c'est la prépon-
```

FIGURE 5.1 – Dans six fichiers, le contenu de la section *Observations préliminaires* n'a pas été considéré comme du texte mais comme des figures. À l'inverse, la section *Notes* a bien été prise en compte comme du texte. Exemple du fichier `s1t4_chapt_8.xml`.

Que s'est-il passé ? Le problème provient d'une fonction de LSE-OD2M, `where_do_budgets_start`. Son rôle est de déterminer l'index de la page où commencent les tableaux de budget et de placer cette donnée dans la variable `budget_start`. Pour cela, il cherche la ligne « BUDGET DES RECETTES DE L'ANNÉE », qui se trouve entre l'en-tête de la page

5.2. VÉRIFICATION DE LA COHÉRENCE DOCUMENTAIRE

49

N° 30. — PAYSANS EN COMMUNAUTÉ DU XING-PO-FOU.		BUDGET DES RECETTES DE L'ANNÉE.	
		N° 33. — COMPTOEUR-TYPGRAPHIE DE PARIS.	
		BUDGET DES RECETTES DE L'ANNÉE.	
SOURCES DES RECETTES.		SOURCES DES RECETTES.	
SECTION I^e. Propriétés possédées par la famille.		SECTION I^e. Propriétés possédées par la famille.	
ART. 1 ^e . — Propriétés immobilières.		ART. 1 ^e . — Propriétés immobilières.	
MÉTIER : Maison d'un rez-de-chaussée et d'un premier étage avec hangar... IMMOBILS BIENFAIT : Champs de riz, d'orge, etc. ('84 ares)... Autre pêcherie située à la maison ('3 ares)... ART. 2. — VALEURS MOBILIÈRES.		(La famille ne possède aucun propriété de ce genre)... ART. 2. — VALEURS MOBILIÈRES.	
ANIMAUX DOMESTIQUES entretenus toute l'année : 2 porcs : valeur calculée... 1 bœuf : valeur calculée... 1 vache : valeur calculée... ANIMAUX DOMESTIQUES entretenus seulement une partie de l'année : 2 porcs : valeur calculée... 1 bœuf : valeur calculée... 1 vache : valeur calculée... MATÉRIEL SPÉCIAL des travaux et industries : Pour la culture des champs... Pour la culture des vignes... Pour l'exploitation des bois et autres... Pour la fabrication des articles de toilette... Pour la fabrication des articles de ménage... Pour la pêche... ART. 3. — DROITS SUR ALLOCATIONS DE SOCIÉTÉ D'ASSURANCES MUTUELLES. (Il n'existe dans le pays aucun société de ce genre)... VALEUR TOTALE des propriétés... SECTION II. Subventions reçues par la famille.		ART. 1 ^e . — PROPRIÉTÉS IMMOBILIÈRES. AGENT : Fonds public (fond) reçus 4 1/2 pour 100 évalué au cours de 1897/98... Fonds placés à la Courre Générale... Souscrire gérant dans le magasin pour besoins imprévus... MATÉRIEL SPÉCIAL des travaux et industries : Outils pour les réparations et l'entretien du mobilier... Utiles pour l'exploitation du magasin et des vêtements... — pour l'entretien... ART. 1. — DROITS SUR ALLOCATIONS DE SOCIÉTÉ D'ASSURANCES MUTUELLES. Droits sur allocations de la caisse de sécurité stable dans l'atelier de l'entrepôt... — de la société de prévoyance... — de Saint-François-Xavier... VALEUR TOTALE des propriétés... SECTION III. Subventions reçues par la famille.	
ART. 1 ^e . — PROPRIÉTÉS IMMOBILIÈRES.		ART. 1 ^e . — PROPRIÉTÉS IMMOBILIÈRES.	
Moulin communal... ART. 2. — DROITS D'USAGE SUR LES PROPRIÉTÉS VOISINES. Droits sur l'arbre et les arbustes plantés sur les chemins... — sur les terrains loués par les habitants... — sur les terrains de la commune... — sur les propriétés des étrangers, etc., étrangers... ART. 2. — ALLOCATIONS VIVRETS ET DE SERVICE. (La famille ne reçoit aucune allocation de ce genre)... VALEUR TOTALE à attribuer au capital des subventions... SECTION IV. Subventions reçues par la famille.		ART. 1 ^e . — PROPRIÉTÉS IMMOBILIÈRES. (La famille ne reçoit aucune propriété en usufruit)... ART. 2. — DROITS D'USAGE SUR LES PROPRIÉTÉS VOISINES. (La famille ne jouit d'aucun droit de ce genre)... ART. 3. — ALLOCATIONS VIVRETS ET DE SERVICES. ALLOCATIONS concernant la nourriture... — les vêtements des enfants... — les réfrigérations... — les besoins ménagers... VALEUR TOTALE à attribuer au capital des subventions... SECTION V. Subventions reçues par la famille.	
ART. 1 ^e . — PROPRIÉTÉS IMMOBILIÈRES.		ART. 1 ^e . — PROPRIÉTÉS IMMOBILIÈRES.	
ART. 2. — DROITS D'USAGE SUR LES PROPRIÉTÉS VOISINES.		ART. 2. — DROITS D'USAGE SUR LES PROPRIÉTÉS VOISINES.	
ART. 3. — ALLOCATIONS VIVRETS ET DE SERVICES.		ART. 3. — ALLOCATIONS VIVRETS ET DE SERVICES.	
(La famille ne reçoit aucune allocation de ce genre)... VALEUR TOTALE à attribuer au capital des subventions...		(La famille ne reçoit aucune allocation de ce genre)... VALEUR TOTALE à attribuer au capital des subventions...	

(a) n° 30, p. 104.

(b) n° 33, p. 260.

(c) n° 37, p. 426.

FIGURE 5.2 – Segmentations dans *Transkribus* des pages liminaires des budgets des monographies n° 30, 33 et 37. Dans les monographies 30 et 33, seul la ligne d'en-tête est considérée comme du texte.

et la ligne supérieure délimitant le tableau de budget. Une fois cette information connue, le script sait que la page correspondante et les suivantes jusqu'au commencement de la section *Notes* (index placé dans la variable `budget_stop`) devront être considérées comme des objets graphiques.

Or les trois monographies concernées dans le quatrième volume (n° 30, 33 et 37) présentent un problème au niveau de la segmentation de la première page du budget. En effet, soit l'en-tête a été considéré comme du texte et le reste comme une illustration (fig. 5.2a et 5.2b), soit l'ensemble de la page a été considéré comme une illustration (fig. 5.2c). *Kraken*, à qui les zones de segmentation sont adressées, n'a donc rien à transcrire (à l'exception de la ligne d'en-tête dans deux cas), et LSE-OD2M n'obtient pas de résultat quand il cherche la ligne « BUDGET DES RECETTES DE L'ANNÉE ». La variable `budget_start` est équivalente à `false` et LSE-OD2M considère que les budgets débutent à la première page de la monographie. Il remplace alors toutes les transcriptions par des objets graphiques jusqu'aux *Notes* qu'il traite normalement, et nous obtenons un fichier partiellement transcrit.

Ces fichiers requièrent une nouvelle OCR et une nouvelle structuration ; une correction par le biais d'une transcription manuelle ne fait en effet pas partie des possibilités acceptables pour l'équipe ALMAnaCH. Cette opération a été repoussée et nous n'avons pas eu à la mener. La reprise de ces six fichiers pourra cependant servir à valider les modifications apportées à LSE-OD2M en tenant compte de l'ensemble des axes d'amélioration relevés au cours du stage et présentés dans ce mémoire. Notamment, l'éventualité de l'absence de valeur pour `budget_start` pourrait être envisagée et traduite par un message

d'erreur informant l'opérateur que le calibrage de la segmentation doit être changé.

5.2.3 Identification, cartographie et inclusion

Au-delà d'une vérification de la cohérence du découpage, cette étape de vérification avait pour but de reconnaître le contenu des fichiers, de leur associer un identifiant unique et de produire un fichier de cartographie ou « *mapping* » du corpus. Une partie des identifiants provenaient d'une bibliographie réalisée au format BibTeX¹² par Stéphane Baciocchi à partir de son exemplaire personnel du corpus.

Une partie seulement, car le fichier bibliographique ne s'intéressait qu'aux monographies, là où ALMAnaCH avait la volonté de travailler sur l'ensemble du corpus, monographies et fichiers de paratexte compris. Il a donc été nécessaire d'établir de nouveaux identifiants sur le modèle de ceux des monographies. Ces derniers se composent d'une série de trois chiffres suivie d'une lettre, **a** pour une monographie, **b** pour le précis *bis* et **c** pour le *ter* (`\d\d\d[a-c]`). Pour poursuivre dans ce sens, nous avons choisi de donner au premier fichier de paratexte le numéro **401a** et de continuer jusqu'au dernier.

Trois versions de la cartographie ont été réalisées : la première contient uniquement les monographies, la seconde le paratexte et la troisième l'ensemble des fichiers¹³. Ces versions sont enregistrées au format CSV : en texte brut, elles se présentent sous la forme d'une succession de lignes contenant des données (identifiant, intitulé puis libellé du fichier), ces données étant séparées par des virgules. Ces fichiers CSV peuvent ensuite être interprétés comme des tableaux, chaque ligne correspondant à une ligne tabulaire et chaque virgule à une séparation entre deux colonnes. Ce format, outre son extrême simplicité, a l'avantage d'être libre et de pouvoir être parsé par le langage Python¹⁴. Les données peuvent donc être manipulées par un script.

Une fois les CSV constitués, il a fallu implanter les identifiants dans les fichiers TEI : il s'agissait de la première opération que nous avons pu automatiser. Après avoir relevé les identifiants et les libellés de fichier dans le CSV général pour constituer un dictionnaire Python (`dict_xml = {'401a': 's1t1_chapt_1.xml', etc}`), le script ouvrait chaque fichier un par un. À l'aide de la librairie BeautifulSoup, il analysait ensuite l'arbre XML (fig. 3.4, p. 30) et implantait l'identifiant comme valeur de l'attribut `@xml:id` de `<TEI>` et de l'attribut `@ana` de la première `<div>`.

La première exécution de ce script a permis de faire remonter une erreur dans la constitution des identifiants : la grammaire TEI n'autorise qu'un `@xml:id` unique commençant par une lettre dans chaque document, ce qui n'était pas le cas ici¹⁵. Or il n'était

12. BibTeX est le logiciel de gestion de bibliographie du langage LATEX.

13. La troisième cartographie a été scindée en fonction des volumes afin de constituer l'annexe A.1 du présent mémoire.

14. Parser un fichier consiste à le lire et à interpréter son contenu afin d'en extraire certains éléments.

15. « *Values for the @xml:id attributes must be unique within a single document, and @xml:id values must begin with a letter* » : TEI Guidelines, 3.10.2 Creating New Reference Systems (<https://www.tei-c.org/ns/1.0/@xml:id.html>)

```

<?xml version="1.0" encoding="utf-8"?>
<master xml:id="master_od2m" xmlns:xi="http://www.w3.org/2001/XInclude">
  <xi:include href="s1t1_chapt_1.xml"/>
  <xi:include href="s1t1_chapt_2.xml"/>
  <xi:include href="s1t1_chapt_3.xml"/>
  <xi:include href="s1t1_chapt_4.xml"/>
  <xi:include href="s1t1_chapt_5.xml"/>
  <xi:include href="s1t1_chapt_6.xml"/>
  <xi:include href="s1t1_chapt_7.xml"/>
  <xi:include href="s1t1_chapt_8.xml"/>
  <xi:include href="s1t1_chapt_9.xml"/>
  <xi:include href="s1t1_chapt_10.xml"/>
  <xi:include href="s1t1_chapt_11.xml"/>
  <xi:include href="s1t1_chapt_12.xml"/>
  <xi:include href="s1t1_chapt_13.xml"/>
  <xi:include href="s1t1_chapt_14.xml"/>
  <xi:include href="s1t1_chapt_15.xml"/>
  <xi:include href="s1t1_chapt_16.xml"/>
  <xi:include href="s1t1_chapt_17.xml"/>
  <xi:include href="s1t2_chapt_1.xml"/>

```

FIGURE 5.3 – Premières lignes du fichier `master.xml`. Dans la balise `<master>`, l’attribut `@xmlns:xi` contient l’adresse de l’espace de nom XInclude.

pas possible de revenir sur ces identifiants, utilisés par d’autres chercheurs. La seule solution trouvée a été d’ajouter le préfixe ID- devant chaque identifiant afin de satisfaire aux règles de la TEI.

ALMAAnCH souhaitait enfin pouvoir disposer d’un fichier XML regroupant l’ensemble des 194 fichiers. Pour cela, nous avons eu recours à un mécanisme d’inclusion grâce au langage XInclude.

Le principe est de constituer un fichier — ici, `master.xml` — dont le seul contenu est une balise `<master>` englobant une succession d’éléments `<xi:include>`. Ceux-ci possèdent un attribut `@href` contenant l’URI d’un fichier XML de paratexte ou de monographie. Il s’agit de l’identifiant uniforme d’une ressource (*Uniform Resource Identifier*), c’est-à-dire une séquence de caractères qui localise et nomme une ressource de manière pérenne¹⁶. Dans notre corpus, les URI des fichiers ne sont pas leurs identifiants mais leurs libellés : au moment d’analyser `master.xml`, le logiciel — par exemple, *Oxygen* — va lui adjoindre le contenu de la ressource localisée par la première URI, c’est-à-dire le fichier `s1t1_chapt_1.xml`, qui contient la page de titre du premier volume, puis passer à l’URI suivante (le premier paratexte du premier volume), et ainsi de suite jusqu’au dernier fichier du troisième volume de la troisième série.

¹⁶//www.tei-c.org/release/doc/tei-p5-doc/fr/html/CO.html#CORS2, consulté le 21 septembre 2020).

16. « A URI is an identifier consisting of a sequence of characters matching the syntax rule named `<URI>`. It enables uniform identification of resources via a separately defined extensible set of naming schemes » : RFC 3986, *Uniform Resource Identifier (URI) : Generic Syntax*, IETF, janvier 2005 (<https://tools.ietf.org/html/rfc3986>, consulté le 21 septembre 2020).

Ce fichier `master.xml` a été là encore constitué de manière automatique avec un script qui rassemble tous les libellés des fichiers dans une liste et, pour chacun d'eux, écrit la ligne `<xi:include href="[libellé]" />` (fig. 5.3).

5.3 Vérification de l'encodage minimal

La mise en forme minimale des documents a été correctement exécutée par LSE-OD2M, aucune partie de la transcription n'ayant été laissée sans encodage.

```

<div n="005" type="sub_section">
<div n="001" type="sub_sub_section">
<pb facs="#facs_41" n="16" source="lesouvriersdesd02soci_0048.xml"/>
<head type="sub_sub_section" xml:id="para_5_16_126">
    § 14.
    <lb/>
    BUDGET DES RECETTES DE L'ANNÉE.
</head>
<figure>
    <graphic facs="#facs_41_g_2" url="#" xml:id="tble_5_16_1"/>
</figure>
<p xml:id="para_5_16_127">
    ART. 3. – DROITS AUXN ALLOCATIONS DE SOCIÉTÉS DE SECOURS MUTUELS.
</p>
<p xml:id="para_5_16_128">
    DRoIr aux allocations de la Société de secours mutuels des Mécaniciens réunis..
</p>
<p xml:id="para_5_16_129">
    VALEUR TOTALE des propriétés..... ::....:
</p>
<p xml:id="para_5_16_130">
    SECTION II.
</p>
<p xml:id="para_5_16_131">
    SUBVENTIONS RECUES PAR LA FAMILLE.
</p>

```

FIGURE 5.4 – Encodage du début du §14 de la monographie 56.

Les objets graphiques ont pu néanmoins poser problème au script. Nous avons vu dans la section précédente que dans six fichiers, toute une partie du texte avait été représentée par des éléments `<figure>`. L'inverse s'est également produit, c'est-à-dire que des objets graphiques véritables — à l'instar des tableaux de budget des paragraphes 14 à 16 — ont souvent été considérés comme du texte. Ainsi, dans la monographie n° 56¹⁷, le tiers du premier tableau du paragraphe 14 est encodé dans un élément `<figure>` (articles 1 à 2), le reste, à partir de l'article 3, étant transcrit et placé dans des `<p>` (fig. 5.4).

Pour autant, ces faits sont moins gênants que les cas inverses, dans la mesure où il s'agit de surplus de transcription. Les `<div>` encadrant ces sections de budgets étant

17. U. Guérin, « Tourneur-mécanicien des usines de la Société Cockerill, de Seraing (Belgique) »...

en place, il est envisageable d'effacer automatiquement les balises `<head>` et `<p>` qu'elles contiennent et de n'ajouter que des `<graphic>` avec l'adresse de l'image de la page.

Cependant, toute mesure d'ajout ou de suppression d'élément `<p>` ou `<head>` a pour conséquence de rompre la logique interne du document. En effet, ces balises possèdent un attribut `@facsimile` dont la valeur renvoie à l'identifiant (`@xml:id`) d'un élément `<zone>` dans les `<facsimile>`.

Pousser la précision de ces derniers jusqu'au niveau des paragraphes devient inutile puisque la correspondance entre les `<zone>` et les `<p>` ou les `<head>` ne peut plus être garantie. De fait, le niveau de référence devient la page (`<surface>`) et non le paragraphe. Les nombreuses modifications que nous avons menées sur ces balises nous ont donc conduit à retirer l'ensemble des éléments `<zone>` et des attributs `@facsimile` des paragraphes et des titres par le biais d'un script. Ce dernier a effectué un ensemble de 138263 suppressions de ligne dans le corpus. Dans l'arbre XML de départ (fig. 3.4 p. 30), le seul descendant de l'élément `<surface>` est désormais `<graphic>`.

Chapitre 6

Contrôle des découpages éditorial et sémantique

Le contrôle du découpage éditorial, c'est-à-dire de l'implémentation de la structure logique, a donné lieu à la reprise la plus importante. Il s'agit d'un point crucial pour les fichiers des *Ouvriers des deux mondes*, du fait de l'importance accordée par l'école leplaysienne à une structure logique qui seule permet de transformer les « faits observés » sur le terrain en des « faits décrits » par la monographie¹.

6.1 Découpage éditorial

6.1.1 Niveau des titres (`<head>`)

Pour commencer, nous avons cherché à évaluer l'étendue des corrections à effectuer à l'aide d'un script. Il comparait une structure idéale — c'est-à-dire l'idée que chaque fichier de monographie devait contenir dans des `<head>` deux titres de niveau section (B et C), quatre titres de sous-section (I à IV) et seize titres de paragraphes (§1 à §16, cf. Annexe A.2) — à la structure réelle contenue les fichiers. Les titres qui n'avaient pas été trouvés étaient listés en sortie.

Le script a fait remonter plus de quatre cent cinquante erreurs. Ce chiffre comptait de nombreux faux positifs en raison de la non-prise en compte de la qualité des transcriptions. Certains titres étaient en effet présents mais leurs transcriptions étaient fautives. Les erreurs principales pouvaient néanmoins être identifiées dès cette étape :

- Des distances d'édition trop faible entre deux titres avaient causé le remplacement de l'un par l'autre :
- (II.) « Mode d'existence de la famille » par (III.) « Moyens d'existence de la famille ».
- « § 11. – Récréations » par « § 7. – Subventions ».

1. S. Baciocchi et J. David, « IV. Ramifications épistémologiques. »..., p. 87.

- La monographie n° 44² est la dernière à posséder des titres de paragraphe sur une seule ligne, un retour à la ligne étant ensuite systématiquement effectué entre le numéro et le libellé du paragraphe. Cet élément n'avait pas été compris par LSE-OD2M, qui a pu répéter ou non le numéro du paragraphe sur la ligne du libellé.
- En cas de répétition, le numéro original était placé dans un `<p>` et le titre avec le numéro ajouté dans un `<head>` (fig. 3.6 p. 32).
- En cas de non répétition, le titre n'était pas toujours détecté et les deux lignes étaient encodées dans des `<p>`.
- Des titres transcrits avec une distance trop grande par rapport au modèle n'avaient pas été détectés. Ce point concerne l'ensemble des titres, sans prévalence particulière.
- Les titres des tableaux de budgets étaient souvent transcrits plus d'une fois.
- Les cas particuliers n'avaient pas été pris en compte :
 - La monographie porte sur un ouvrier ou une ouvrière et non pas une famille, ce qui conduit à un changement dans la terminologie (par exemple, « État civil de la famille » devient « État civil de l'ouvrier ») : concerne notamment `s1t2_chapt_4.xml`, `s1t3_chapt_7.xml` et 10 ;
 - Certains titres de paragraphe n'étaient pas séparés du texte mais imprimés en italique au début de celui-ci (Annexe B.1.3) ;
 - Le titre n'est pas imprimé (Annexe B.1.2, B.1.3 et B.1.4) ;
 - La monographie possède une structure qui lui est propre (Annexe B.1.5).
 - La sous-section des budgets est titrée dans huit monographies, mais ce titre n'a jamais été détecté comme tel par LSE-OD2M.

6.1.2 Niveau des divisions (`<div>`)

L'implémentation de la structure logique ne s'est pas uniquement traduite par une mise en forme des titres : chaque `<head>` est en effet attaché à une `<div>` (fig. 5.1 p. 48). Tous deux possèdent un identifiant `@type` qui définit leur niveau (`section`, `sub_section` ou `sub_sub_section`), la `<div>` étant en plus nantie d'un numéro (`@n`).

Trois erreurs ont été constatées à ce niveau :

- Lorsqu'un titre n'était pas détecté, la `<div>` correspondante n'était pas créée et le paragraphe était intégré dans la division précédente, invalidant de fait l'ensemble de la numérotation ;
- Des `<div>` surnuméraires ont été implantées (fig. 3.6 p. 32) ;
- Au cours du stage, il a été décidé de placer le titre de la monographie dans une `<div>` de type `section`, ce qui impliquait une reprise dans l'ensemble des fichiers.

2. U. Guérin, « Paysan-résinier de Lévignacq (Landes - France) »...

6.1.3 *Tabula rasa*

Le nombre élevé de cas particuliers rendait impossible une automatisation complète de cette reprise. Nous allons d'abord exposer la méthode suivie pour le traitement des cas particuliers avant de rendre compte de celle utilisée pour le reste du corpus.

Cas particuliers

Les monographies d'ateliers (`s3t1_chapt_2.xml`³ et `s3t2_chapt_10.xml`⁴) imposaient une reprise manuelle du fait de leur structure unique. De plus, dans ces fichiers, ainsi que dans certaines monographies de familles, plusieurs titres n'étaient pas détachés des paragraphes. Le recours à la balise `<head>` n'était pas possible pour les traiter, dans la mesure où elle constitue normalement un bloc séparé des `<p>`. Plusieurs solutions, toutes insatisfaisantes, ont été envisagées.

La plus légère consistait à signaler la mise en forme de ces titres (italique ou gras) par des balises de mise en valeur telles que `<emph style="italic">` ou `<hi rend="i">`. Nous n'avons pas opté pour ce choix en raison des règles de la TEI qui ne prévoient pas d'attributs pour caractériser ces balises (notamment, `@type`). Aussi n'aurait-il pas été possible de les différencier d'autres passages possédant une mise en forme similaire dans le corps du texte lors de la transformation des fichiers XML en d'autres formats.

Une deuxième solution, beaucoup plus lourde pour le fichier, aurait consisté en l'imbrication de `<div>`. Le `@type` assigné à la division aurait permis de savoir que le `<head>` et le `<p>`, eux-même placés dans une sous-division, ne constituaient qu'un seul paragraphe dans le volume original (fig. 6.1). La contre-partie est que chaque paragraphe aurait dû être placé dans une sous-division.

Au final, *Time Us* n'a opté pour aucune solution, considérant que le nombre de cas était suffisamment bas pour ne pas avoir à mettre en place une solution particulière. Pour autant, cette décision n'a pas eu pour effet d'arrêter le niveau de granularité des titres à la sous-sous-section.

En effet, les fichiers `s1t5_chapt_9.xml`⁵ et `10`⁶ comprennent des précis de monographie dotés de leur propre structure logique et intégrés dans le corps de leurs sections *Notes*. Les titres étant détachés des paragraphes, il a été possible de mobiliser un niveau inférieur à la sous-sous-section, le `paragraph`, afin de rendre compte de cette structuration.

Reconstruction de la structure

Ces cas particuliers réglés, il restait à prendre en charge l'essentiel du corpus, c'est-à-dire une soixantaine de monographies. Une reprise entièrement manuelle n'était pas

3. P. du Maroussem, « La société générale des papeteries du Limousins »...

4. Id., « Usine hydraulique d'éclairage et de transport de force »...

5. U. Guérin, « Ouvrier cordonnier de Malakoff (Seine - France) »...

6. J. Reviers de Mauny, « Serrurier-forgeron de Paris (Seine - France) »...

```

<body>
  <div type="chapter">
    <div type="section">
      <div type="sub_section">
        <div type="sub_sub_section">
          <div type="paragraph">
            <div>
              <head type="paragraph"> titre du paragraphe </head>
              <p> texte du paragraphe </p>
            </div>
            <div>
              <p> texte du paragraphe suivant </p>
            </div>
            <div>
              <p> ainsi de suite </p>
            </div>
          </div>
        </div>
      </div>
    </div>
  </div>
</body>

```

FIGURE 6.1 – Représentation de la deuxième solution, dont le résultat est de produire une imbrication malvenue de `<div>`.

envisageable en raison du temps trop important qu'elle aurait nécessité. À l'inverse, une automatisation totale aurait nécessité une longue étude et un relevé minutieux des aléas de la détection et de la transcription des titres.

Nous avons donc choisi d'implanter la structure logique par des cascades d'expressions régulières appliquées sur un lot de fichiers à l'aide d'un script. Les expressions régulières sont des motifs qui permettent de décrire des chaînes de caractères dans un texte. Le module `re` de Python permet de mobiliser ces expressions dans un script, et notamment d'effectuer des substitutions grâce à la fonction `re.sub()`.

Du fait de la désorganisation massive des `<div>`, la première expression régulière consistait à faire table rase de ces balises. Le script reconstruisait ensuite la structure en progressant selon ses niveaux. Il commençait par planter la `<div>` de type `chapter` et celle de la première section (la page de titre). Venaient ensuite les `<div>` des deux sections restantes (*Observations préliminaires* et *Notes*), puis celles des sous-sections, etc.

Les expressions régulières recherchaient les libellés des titres lorsqu'il s'agissait de sections ou de sous-sections, car la plupart étaient convenablement transcrits. En revanche, il recherchait le numéro des sous-sous-sections, qui avaient tous pour point commun de commencer par le symbole typographique §. Les paragraphes *Notes* des quatre-vingt-quatre premières monographies ayant pour particularité d'être numérotés par une lettre entre parenthèses, le motif des sous-sous-sections des premiers lots était légèrement diffé-

rent.

Une dernière commande, qui affichait la liste des balises `<head>` de chaque fichier, nous permettait de contrôler l'intégrité de la nouvelle structure. Si une erreur était constatée, nous intervenions manuellement afin de la résoudre. Pour chaque lot (qui correspondait *grosso modo* à un volume), jusqu'à 50% d'erreur était constaté. Cela était dû à des titres qui ne correspondaient pas aux motifs en raison de l'absence d'un symbole ou d'une transcription trop éloignée de la vérité terrain.

L'opération s'est conclue par la constitution automatique d'un tableau synoptique, le contenu de chaque balise `<head>` étant placé dans une cellule. Son étude a permis, d'une part, de corriger les dernières erreurs résiduelles, et d'autre part de rédiger une note exposant l'ensemble des défauts restant (Annexe B.1).

6.2 Découpage sémantique

L'opération de reconstruction de la structure logique a mis en évidence une erreur qui n'avait pas été détectée jusque là : des en-têtes (rappel du type de chapitre et de son titre, numéro de page ou de la monographie) et des bas de page (numéro du cahier) étaient toujours présents dans le flux du texte. L'élément révélateur a été la mise en forme par LSE-OD2M de plusieurs rappels de titre comme s'il s'agissait de titres à part entière.

Pour les supprimer, nous avons là encore utilisé des expressions régulières. L'automatisation n'était pas possible pour les numéros de page, de cahier ou de monographie — par exemple, `<p>3</p>` — car certaines transcriptions de tableau avaient produit des données numériques semblables qu'il fallait conserver.

Pour résoudre ce problème, nous avons utilisé la fonction de recherche de l'outil « projet » d'*Oxygen*, qui permet de rechercher une expression dans l'ensemble des fichiers du corpus. Les expressions étaient toutes construites par rapport à la balise `<pb/>` qui marque un changement de page, et toute occurrence d'une donnée numérique et d'un titre était systématiquement contrôlée (fig. 6.2 et 6.3).

Cette opération a permis de garantir que la reconstitution des paragraphes serait très peu perturbée par des éléments externes. Des contrôles visuels au cours d'autres actions ont cependant permis de déceler des numéros subsistants, placés à la fin de certains paragraphes ou bien transcrits par des lettres (fig. 6.4).

6.3 Validation du schéma XML-TEI

Les erreurs de structuration et de transcription traitées dans les quatre premiers niveaux d'encodage, nous pouvions envisager de formaliser le schéma des fichiers. Il s'agit d'établir un document définissant quelles balises peuvent être utilisées dans les fichiers,

1973	1973	<head facs="#fac_374_p_2" type="sub_sub_section" xml:id="para_20_357_126">
1974	1974	§ 7. <lb/> SUBVENTIONS.
1975	1975	</head>
...	...	@@ -3949,9 +3949,6 @@
3949	3949	<p facs="#fac_406_p_45" xml:id="para_20_389_772">
3950	3950	Totaux..... (15. SP 1I1)
3951	3951	</p>
3952	-	<p facs="#fac_406_p_46" xml:id="para_20_389_773">
3953	-	28
3954	-	</p>
3955	3952	<pb facs="#fac_407" n="390" source="s2lesouvriersdes01sociuoft_0446.xml"/>
3956	3953	<p facs="#fac_407_p_1" xml:id="para_20_390_774">
3957	3954	VALEUnS

FIGURE 6.2 – Suppression du numéro de page 28 dans le fichier s2t1_chapt_20.xml (image du commit sur *GitLab* — monographie n° 53).

3040	3040	<p facs="#fac_250_p_5" xml:id="para_10_221_598">
3041	3041	Déjà l'argent a trop perdu de sa valeur, et quelle que soit depuis quelques années la hausse des salaires, des grèves continues persistent à les faire monter. L'ouvrier ne peut se procurer avec un salaire double, la somme de bien-être obtenue il y a trente ans à moitié prix. Ajoutez une vanité absurde et un besoin devenu impérieux de cercles, de plaisirs, de spectacles. Ainsi l'alimentation, le nécessaire, renchérit toujours en raison directe de l'avilissement du prix
3042	3042	</p>
3043	-	<p facs="#fac_250_p_6" xml:id="para_10_221_599">
3044	-	16
3045	-	</p>
3046	3043	<pb facs="#fac_251" n="222" source="lesouvriersdesd02soci_0266.xml"/>
3047	3044	<p facs="#fac_251_p_1" xml:id="para_10_222_600">
3048	3045	des produits manufacturés. Au plus fort de ces crises alimentaires, qu'un peu de travail aurait largement atténuées sinon prévenues, on rencontre sur les ponts, dans tous les carrefours et sur les

FIGURE 6.3 – Suppression du numéro de cahier 16 dans le fichier s2t2_chapt_10.xml (image du commit sur *GitLab* — monographie n° 59, page 221).

1036	1036	<p facs="#fac_558_p_16" xml:id="para_16_529_1056">
1037	-	BONNES MOEURS. – Mal sauvegardées parmi les ouvriers de la grande industrie manufacturière, à Seraing (Belgique), 6, 1t3, 14. – Maintenues par des habitudes d'une décence scrupuleuse chez les familles de la Grande-abylie, 60, 69. – Compromises durant l'apprentissage des mousses pêcheurs d'Ileyst (Belgique), puis restaurées par le mariage, 113. – Mieux conservées chez les pêcheurs normands que chez les ouvriers des villes, 156. – Maintenues sans garanties pour l'avenir chez le paysan métayer de la basse Provence, 179. – En décadence marquée chez les paysans trop peu religieux de la Marche (France) ; 235. – tactes, sauf exceptions dans les ménages accumulés des cités ouvrières, chez les 36
	+	BONNES MOEURS. – Mal sauvegardées parmi les ouvriers de la grande industrie manufacturière, à Seraing (Belgique), 6, 1t3, 14. – Maintenues par des habitudes d'une décence scrupuleuse chez les familles de la Grande-abylie, 60, 69. – Compromises durant l'apprentissage des mousses pêcheurs d'Ileyst (Belgique), puis restaurées par le mariage, 113. – Mieux conservées chez les pêcheurs normands que chez les ouvriers des villes, 156. – Maintenues sans garanties pour l'avenir chez le paysan métayer de la basse Provence, 179. – En décadence marquée chez les paysans trop peu religieux de la Marche (France) ; 235. – tactes, sauf exceptions dans les ménages accumulés des cités ouvrières, chez les
1038	1038	</p>
1039	1039	<pb facs="#fac_559" n="530" source="lesouvriersdesd02soci_0596.xml"/>
1040	1040	<p facs="#fac_559_p_1" xml:id="para_16_530_1057">

FIGURE 6.4 – Suppression du numéro de cahier 36 dans le fichier s2t2_chapt_10.xml après un contrôle visuel (image du commit sur *GitLab* — table alphabétique et analytique du deuxième volume de la deuxième série, page 529).

régulant l'ordre de leur imbrication et arrêtant le nombre et la nature des attributs qui peuvent leur être attachés.

Nous avons choisi d'établir un fichier ODD, entièrement écrit dans une syntaxe XML, qui permet d'établir très précisément les éléments, les séquences d'éléments (nombre et ordre) et leurs attributs. *Oxygen* est capable d'en générer un après avoir analysé un ensemble de fichiers TEI, puis de le transformer en un schéma relaxNG. Ce document relaxNG contient le schéma servant à la validation du fichier TEI, l'ODD étant un fichier XML de documentation et de description de ce schéma. Le document relaxNG est finalement associé au document TEI (fig. 6.6), une des fonctionnalités d'*Oxygen* permettant de contrôler à tout moment la validité du code par rapport au schéma.

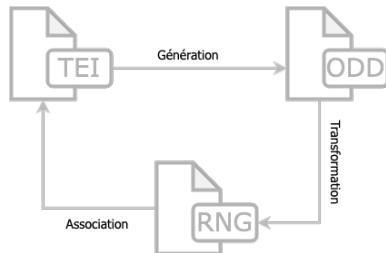


FIGURE 6.5 – Schématisation du processus de validation à l'aide d'une ODD.

Le document ODD est généré automatiquement par *Oxygen* grâce au scénario de transformation du Consortium TEI intitulé *oddbyexample*. Il s'agit d'une feuille de style XSL qui, appliquée à un document TEI, analyse l'ensemble des balises et leurs attributs pour les comparer aux préconisations de la TEI⁷. Puis il génère une ODD où les modules TEI sont chargés, les balises et les attributs non-utilisés étant supprimés d'office et les valeurs des attributs collectées et inscrites dans des listes.

L'utilisateur peut modifier le document obtenu en imposant des règles plus restrictives ou au contraire en relâchant certaines ; il peut également ajouter des valeurs d'attribut, rendre des attributs obligatoires ou contraindre l'enchaînement de certaines balises. Par exemple, nous avons rendu l'attribut `@type` obligatoire pour les balises `<head>` et `<div>`, et clos la liste de ses valeurs possibles (fig. 6.6).

Une ODD peut également être documentée, c'est-à-dire que son rédacteur peut justifier ses choix d'encodage. Disposer d'un schéma et de sa documentation est un élément extrêmement important pour un projet en humanités numériques : c'est en effet une assurance pour la pérennité des données qu'il a rassemblées et structurées. Celles-ci ne sont plus dépendantes des ingénieurs ou des structures institutionnelles et peuvent être réutilisées par d'autres projets⁸.

7. The XSLT stylesheet which traverses a nominated directory tree looking for *.xml files which have `<TEI>` or `<teiCorpus>` root elements. It analyzes the collection of elements and attributes in the resulting corpus, and compares that to the whole of TEI P5 : *oddbyexample* sur le TEI Wiki (<https://wiki.tei-c.org/index.php/Oddbyexample>, consulté le 21 septembre 2020).

8. Vincent Jolivet, « Éditions ou données ? API et (re)publications », dans *Actes royaux et princiers*

```

<elementSpec ident="div" mode="change">
  <attList>
    <attDef ident="fac" mode="delete"/>
    <attDef ident="source" mode="delete"/>
    <!-- @type obligatoire pour <div> -->
    <attDef ident="type" mode="change" usage="req">
      <valList mode="add" type="closed">
        <valItem ident="chapter"/>
        <valItem ident="section"/>
        <valItem ident="sub_section"/>
        <valItem ident="sub_sub_section"/>
        <valItem ident="paragraph"/>
      </valList>
    </attDef>
  </attList>
</elementSpec>

```

FIGURE 6.6 – Extrait de l’ODD des *Ouvriers des deux mondes* concernant la balise `<div>`. Les attributs `@fac` et `@source` sont supprimés (`mode="delete"`), `@type` est requis (`usage="req"`) et sa liste de valeurs est close (`type="closed"`).

Il existe plusieurs possibilités pour permettre la vérification du document TEI. La première et la plus courante est la spécification d’une instruction de traitement au début du document sous la forme suivante : `<?xml-model href="#" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>`, l’attribut `@href` devant être complété par l’adresse du schéma. Cela suppose bien évidemment que ce schéma soit hébergé, par exemple sur *GitLab*. La ligne peut ensuite être ajoutée automatiquement à l’ensemble des fichiers du corpus grâce à la fonction de recherche et de remplacement.

Une autre possibilité est une validation externe, c’est-à-dire qu’un logiciel ou un script vérifie la compatibilité des fichiers par rapport à un schéma qui lui est fourni par l’utilisateur à l’instant `t` de la validation. C’est une des fonctionnalités du mode « projet » d’*Oxygen. Time Us* a cependant souhaité se détacher de tout logiciel, et *a fortiori* d’un logiciel propriétaire, afin de se réserver la possibilité de contrôler la validité de ses fichiers à tout moment.

Le module `etree` de la librairie Python `lxml`, qui permet d’analyser un arbre XML, possède un outil répondant parfaitement à ce besoin. En effet, un schéma RNG peut être chargé avec la méthode `.RelaxNG()`, `validate()` contrôlant ensuite la validité de tout fichier lui étant donné comme argument par rapport à ce schéma.

La spécificité de ce script est qu’il s’applique à un corpus entier, ce qui avait pour effet d’allonger démesurément son exécution sur notre ordinateur (environ une demi-heure pour traiter une vingtaine de fichiers). Nous avons donc charger l’ensemble des fichiers, le schéma et le script sur le *cluster* RIOC d’Inria, ce qui a considérablement rétréci le temps d’exécution (vingt-quatre minutes au total).

à l’ère du numérique (*Moyen Âge — Temps moderne*), dir. Olivier Canteaut, Olivier Guyotjeannin et Olivier Poncelet, Pau, 2020, p. 59-68, URL : <https://www.nakala.fr/nakala/data/11280/eae11732>, p. 61.

La structuration automatique des fichiers des *Ouvriers des deux mondes* par le script LSE-OD2M a donc été contrôlée selon différents niveaux d'encodage.

Sur le plan méthodologique, ce contrôle s'est traduit par plus d'interventions automatiques que manuelles, illustrant le gain de temps que la technologie peut apporter dans la production d'un corpus numérique. Sur le plan scientifique, ces problèmes montrent l'importance de la modélisation d'un corpus physique avant toute entreprise de rédaction d'un script de traitement. Nombre d'erreurs ont en effet été causées par une souplesse très faible d'LSE-OD2M face aux cas particuliers du corpus. Cet effort de modélisation ne peut être mené que si un temps lui est alloué dans le projet.

La modélisation permet également de s'assurer de la pérennité des données. En effet, les fichiers des *Ouvriers des deux mondes* n'existent pas dans le seul objectif de produire une édition numérique, c'est-à-dire d'être mis en scène à travers une interface de consultation. Le processus de valorisation des données de la recherche peut emprunter deux voies différentes. L'une consiste à inscrire ces données dans le temps long et à favoriser leur ré-utilisation grâce à une structuration standardisée et documentée. L'autre entraîne leur publication dans un cadre institutionnel et budgétaire conjoncturel qui, s'il peut produire des résultats visibles rapidement, n'est pas assuré d'être reconduit et pérennisé.

Les fichiers des *Ouvriers des deux mondes* se trouvent face à cette injonction qui, sans être réellement paradoxale, conjugue une obligation — publier avant la fin du programme ANR *Time Us* — et une ambition — assurer la pérennité d'un corpus numérique.

Plusieurs voies sont envisagées pour répondre à ces nécessités. Aucune n'a pu être mise en place au cours de notre stage, mais nous avons mené plusieurs réflexions et quelques actions en leur sens. Nous allons d'abord présenter celles qui concernent les objets graphiques, puis nous intéresser à l'état des transcriptions, avant de conclure sur les différents types d'édition possibles et ce qu'ils impliquent pour les données issues du traitement automatique des fichiers des *Ouvriers des deux mondes*.

Troisième partie

Des données à valoriser

Chapitre 7

Données graphiques, données chiffrées

7.1 Le lien entre le texte et les images du texte

Dans chaque élément `<facsimile>`, une balise `<surface>` définit le bloc de segmentation supérieur (la page). L'élément suivant, `<graphic>`, indique dans son attribut `curl` la localisation de l'image de la page segmentée. L'ensemble des images étant stocké dans l'espace alloué au programme *Time Us* au sein du service *ShareDocs* de la TGIR Huma-Num, la localisation consiste en un chemin relatif (fig. 3.5 p. 31).

Or ces images sont déjà hébergées par *Internet Archive*. Elles ont été téléchargées sur le *ShareDocs* dans le but de permettre leur segmentation par *FineReader* et *Transkribus* puis leur OCR par LSE-OD2M. Leur conservation après l'obtention des fichiers XML n'est plus aussi pertinent que celle des prises de vues des registres prud'homaux effectuées dans des dépôts d'archives, qui n'existent sous aucune autre forme. Rappelons également qu'*Internet Archive* se donne pour objectif d'être un centre stable et durable d'archives digitales ; aussi est-il peu probable que les images des *Ouvriers des deux mondes* disparaissent de ses serveurs. Il nous a donc été demandé de substituer le chemin local par l'adresse de l'image sur *Internet Archive* (Annexe B.1, *issue 2*) ; cette mission a donné lieu à la publication d'un billet sur le carnet de recherche du programme *Time Us*¹.

Nous avons apporté deux solutions : l'une employait les URLs basiques des images, l'autre mobilisait les manifestes IIIF des volumes. Toutes deux ont donné lieu à des scripts Python dont les fonctionnements sont similaires.

Les URLs se trouvaient dans des fichiers JSON renseignés dans le code source des pages d'*Internet Archive*. Un fichier JSON contient des informations représentées de manière structurée ; il s'agissait ici de métadonnées concernant le volume numérisé. Parmi celles-ci, une sous-section intitulée `data` contenait des métadonnées (longueur, largeur,

1. Jean-Damien Généro, *Les ouvriers des deux mondes : des images aux urls*, carnet de recherche de *Time Us*, URL : <https://timeus.hypotheses.org/645>.

etc.) et l'URI de chaque image. Ces images étaient au format JPEG et correspondaient à celles déposées sur le *ShareDocs*.

Le problème posé par cette solution est que les adresses n'étaient pas stables. En effet, au bout d'un certain temps, elles ne fonctionnaient plus car les informations avaient changé dans le fichier JSON source.

Aussi avons-nous exploré la piste du IIIF. L'*International Image Interoperability Framework* permet d'afficher une image avec ses métadonnées dans le contexte d'une application web directement depuis le serveur où elle est stockée (ici, *Internet Archive*). Les URLs de cette deuxième solution se trouvent dans les « manifestes IIIF » des volumes : il s'agit des documents JSON contenant leurs métadonnées et référençant les points d'accès aux images (c'est-à-dire leurs URIs dans le protocole IIIF).

Si ces images possèdent le même format que celles de la première solution, leur qualité est bien supérieure et permet d'effectuer des agrandissements d'une très grande profondeur. Du reste, le IIIF permet également de naviguer dans un volume en passant de page en page. Cette seconde solution est ainsi plus intéressante pour le projet *Time Us* dans la mesure où elle lui permet d'accéder d'une manière relativement simple à un document contenant un ensemble de données et de métadonnées qui pourront être valorisées au moment d'une édition en ligne.

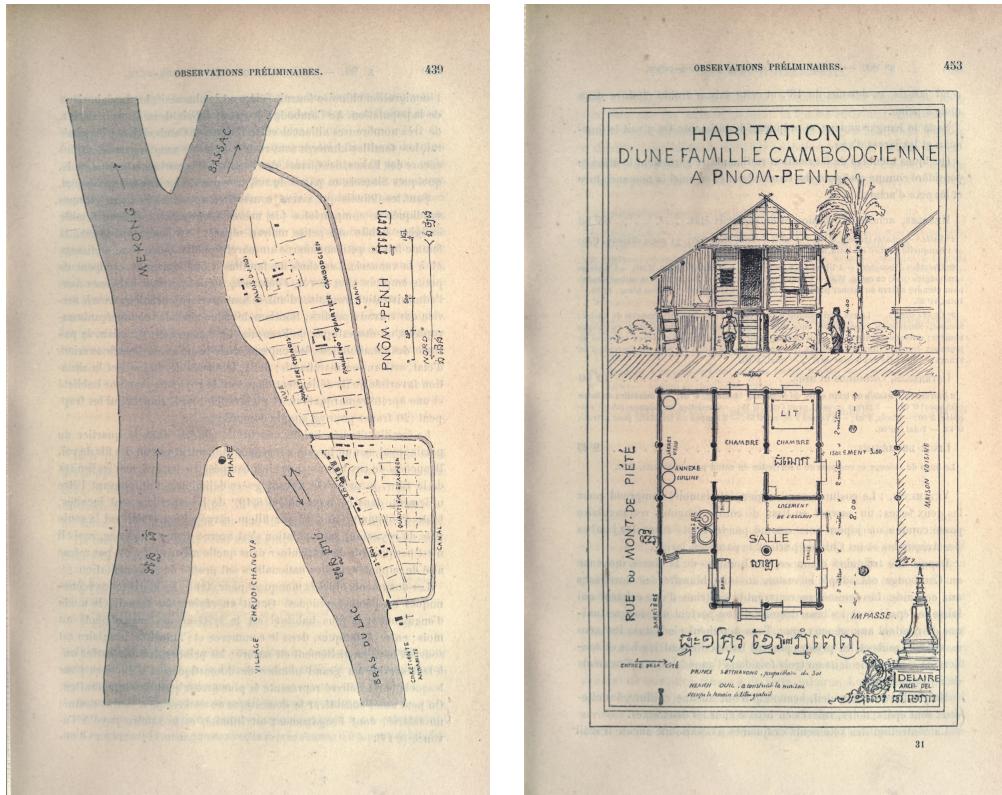
Le script dont il est ici question, s'il se limite à insérer les URIs des images dans le code XML, commence par effectuer une requête pour lire le contenu des manifestes IIIF des *Ouvriers des deux mondes*. Pour cela, il lit un CSV où nous avons enregistré les identifiants donnés par *Internet Archive* aux numérisations, et les utilise pour compléter l'adresse des manifestes (<https://iiif.archivelab.org/iiif/<itemid>/manifest.json>). Ces lignes de code peuvent être réutilisées, probablement sous la forme d'une fonction, pour obtenir tout type de métadonnées issues des manifestes. En plus de rationaliser le stockage sur le *ShareDocs*, ce script prépare donc l'étape de la publication en ligne.

7.2 Les données graphiques dans le flux textuel

Un des principaux apport du IIIF pourrait être le traitement des objets graphiques qui, nous l'avons vu, sont nombreux dans les pages des *Ouvriers des deux mondes* (fig. 3.1 p. 24). En particulier, les photographies, les figures ou les cartes pourraient bénéficier des fonctionnalités d'agrandissement afin de permettre à l'utilisateur de pleinement les prendre en considération.

Pour rendre cela possible, il faudrait disposer d'une évaluation du taux de détection des figures. Ce chiffre n'est pas connu, mais des relevés aléatoires laissent penser qu'il risque d'être inférieur à 60%. Les tableaux sont les figures qui ont posé le plus de problème au script, nous y reviendrons.

Parmi les autres, plusieurs n'ont été que partiellement détectées. La carte de la



(a) Page 439 : détection optimale de la carte.

(b) Page 453 : détection en double et transcription du titre.

FIGURE 7.1 – Figures dans la monographie n° 90.

page 439 de la monographie n° 90² est ainsi parfaitement détectée et transposée dans un élément <figure> (fig. 7.1a). À l'inverse, le plan d'une « habitation cambodgienne à Phnom-Penh » est détecté en double — peut-être en raison de la superposition d'une vue de face de l'habitation et de son plan — et le titre de la figure est transcrit comme du texte (fig. 7.1b). Dans la monographie suivante³, la photographie pleine page « Cambodgiens et amanites », insérée entre les pages 484 et 485, a été retirée du texte et n'est pas présente dans les fichiers sous quelque forme que ce soit.

Ces erreurs de détection ne peuvent pas être résolues de manière automatique : aucune table des figures n'est présente dans les volumes. Un tel outil eut permis de cibler les pages à contrôler et de ne pas avoir à chercher visuellement les figures. Une intervention manuelle, possiblement longue, s'impose donc.

La présence de ces figures est également un élément à garder à l'esprit au moment de la reconstitution des paragraphes. En effet, certaines sont insérées en plein milieu de ces derniers. Pour ne pas perturber la reconstitution, il sera nécessaire d'ajouter une condition spécifiant que les balises <p> commençant pas une minuscule ne constituent pas

2. E. A. Delaire, « Petit fonctionnaire de Phnom-Penh (Cambodge) », dans *Les Ouvriers des deux mondes*, Paris, 1899 (série 2e (5)), p. 437-483, URL : <https://archive.org/details/2serlesouvriersde05sociuoft/page/n487>.

3. Id., « Précis d'une monographie d'un manœuvre-coolie de Phnom-Penh (Cambodge) »...

une unité indépendante, et ce même si elles sont précédées d'une figure. Le problème de cette méthode est qu'elle presuppose que LSE-OD2M a bien fait la différence entre une majuscule et une minuscule en début de phrase, ce dont on ne peut pas être totalement certain.

Ajoutons enfin que les figures de type carte, croquis ou photographie ont une importance relative du point de vue de la donnée pure. Elles ne sont pas utiles aux chercheurs qui s'intéressent au vocabulaire du textile ou du monde ouvrier, ni à ceux qui souhaitent recueillir des données chiffrées. Dans le document XML, elles ne sont que des repères dont la nécessité ne se révélera qu'au moment de la transformation des fichiers pour une interface de visualisation.

7.3 Les tableaux : images ou données ?

À l'inverse, les tableaux qui jalonnent le corpus ont une importance considérable en raison des informations statistiques qu'ils contiennent. Plus particulièrement, les budgets des paragraphes 14 à 16 constituent « la pièce centrale » des monographies leplaysiennes⁴. Les *Observations préliminaires* sont en effet entièrement tournées vers leur établissement ; le monographe n'arrête son enquête que lorsqu'il a établi un budget équilibré offrant une vue globale des recettes et des dépenses de la famille. En effet, dans la démarche de Frédéric Le Play, « l'argent est pris comme unité de mesure de la vie sociale, le budget étant la quantification, en termes de revenus et dépenses, de l'ensemble des activités de la famille »⁵.

D'un point de vue technologique, il n'est cependant pas simple de reproduire à l'identique ces tableaux. La TEI comporte un ensemble de balises destiné à cet effet — `<table>`, `<row>` et `<cell>` — mais le problème est ici fonction de la segmentation et de la compréhension par un script de la mise en page du tableau. Or cette mise en page peut se révéler sophistiquée. Ainsi, le texte est centré lorsqu'il s'agit de titre d'article ou de section, aligné à droite pour les postes de dépenses ou les recettes, aligné à gauche pour le total. Lorsque plusieurs postes de dépense successifs partagent une même formulation dans leurs intitulés, celle-ci n'est pas répétée mais signifiée par des traits de rappel.

Ces modulations permises par l'imprimé ne sont pas envisageables dans un tableau numérique, où toute donnée chiffrée doit correspondre à un libellé défini et non suggéré afin d'être rendu exploitable. Face à ce type d'information, un algorithme idéal devrait être capable de reproduire le processus cognitif de reconstruction effectué par le cerveau humain. En l'état actuel de l'avancée technique, faire en sorte que la machine différencie un trait de rappel d'une ligne de séparation entre deux cellules n'est pas aisé. Cela constitue un sujet de recherche à part entière et ne peut pas être mené dans le temps imparti par

4. A. Savoye, « Les continuateurs de Le Play au tournant du siècle »..., p. 317.

5. *Ibid.*

l'ANR au programme *Time Us*.

Le traitement des tableaux doit donc être pensé à l'aune de ces difficultés techniques et temporelles. Dans un essai de mise en scène des fichiers XML au format HTML, Alix Chagué a ainsi choisi de transformer les balises <figure> — signifiant notamment la présence des tableaux — en des icônes cliquables qui redirigent vers les images des pages où ces figures se trouvent. Le procédé est habile mais se fonde sur l'idée que toutes les figures ont été détectées, ce qui n'est pas encore le cas. Du reste, cela ne permet pas d'exploiter directement les informations des tableaux.

Pour autant, cette idée du tableau comme une image est sérieusement étudiée par le programme *Time Us*. L'effort d'ingénierie d'études se concentrerait alors sur la modélisation et l'implémentation d'une indexation fine des différents tableaux.

Il serait tout d'abord nécessaire d'effectuer un recensement exhaustif des tableaux des *Ouvriers des deux mondes*. Ensuite, une étape de modélisation commencerait par l'analyse de leur structuration afin de faire ressortir les points communs et *in fine* de constituer des catégories ; en parallèle, les chercheurs devront définir leurs besoins – sont-ils intéressés par l'ensemble des informations, ou bien seules celles ayant trait au textile ? Une fois la typologie arrêtée et les besoins déterminés, il sera possible d'envisager l'implémentation d'une nouvelle couche d'encodage scientifique dans les fichiers XML.

Celle-ci pourrait se traduire par deux actions. D'une part, un identifiant @xml:id pourrait être donné à chaque titre de paragraphe, et ensuite référencé dans un attribut @ana (*analytic*). Peu intéressants pour les paragraphes de budget qui contiennent systématiquement des tableaux, ces attributs permettraient de signaler ceux que l'on peut rencontrer dans les paragraphes 6 à 10 (*Propriétés, Subventions, Travaux et industries, Aliments et repas, Habitation, mobilier et vêtement*). D'autre part, un recensement des objets des tableaux — dépenses de nourriture, dépenses pour l'habitation, dépenses pour le textile, etc. — pourrait fournir une liste de valeurs pour un attribut @type. Au-delà de permettre un traitement rationnel des tableaux au regard des impératifs scientifiques et temporels, cette méthode pourrait servir de base à l'établissement de vues logiques ou de facettes de recherche conduisant à l'affichage des tableaux sur une base thématique.

La valorisation des objets graphiques n'est donc pas chose aisée. Elle peut se restreindre à leur simple reproduction doublée d'une indexation qui, si elle oriente le chercheur, ne le dispense pas d'effectuer lui-même l'extraction des données chiffrées ou textuelles contenues dans la figure. Il y a ici une limite au projet de numérisation du corpus.

Chapitre 8

Données textuelles, données scientifiques

8.1 La qualité de l'ocr

L'extrême majorité des données des *Ouvriers des deux mondes* est composée de texte. Ce texte peut être le support de différentes études, qui se basent toutes sur la lecture. La lecture par l'œil humain — le corpus n'est pas réellement important et un chercheur peut en prendre intégralement connaissance —, mais aussi la lecture par une machine, c'est-à-dire l'analyse automatique. Ces deux opérations ne nécessitent pas un même niveau de qualité de l'OCR et ne s'adressent pas au même public — la première est accessible à un large panel, la seconde requiert l'assistance technique d'un ingénieur et l'exécution par une machine.

La qualité d'une OCR peut être mesurée par différents indices, au niveau du caractère (*character error rate*, CER) ou du mot (*word error rate*, WER). Le CER est obtenu grâce la formule suivante :

$$CER = \frac{S + D + I}{N}$$

où S représente le nombre de substitutions (caractères dont la reconnaissance n'est pas correcte), D le nombre de suppressions (*deletions*), et I le nombre d'insertions, c'est-à-dire les caractères qui ne sont pas présents dans la vérité terrain que l'on trouve pourtant dans l'OCR. La somme de ces trois chiffres est divisée par le nombre total de caractères dans le fichier de vérité terrain (N).

À partir d'une vérité terrain de 1300 lignes, Alix Chagué avait calculé un CER de 2,2% pour l'OCR des *Ouvriers des deux mondes*¹. Ce taux très faible est le signe d'une

1. A. Chagué, V. Le Fournier, M. Martini et Éric Villemonte de La Clergerie, « Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ? », dans *Colloque DHNord 2019 "Corpus et archives numériques"*, MESHS Lille Nord de

transcription de bonne qualité. Où se situent les erreurs restantes ?

Dans les *Ouvriers des deux mondes* comme dans de nombreux autres corpus², elles se concentrent sur les entités nommées. Au premier rang de ces dernières figurent les patronymes et les toponymes (fig. 8.4) ; les « expressions temporelles et les expressions de quantité » ou encore « les nombres, les formules chimiques, les unités monétaires » sont également concernés³. Ces entités sont très peu représentées dans les fichiers de vérité terrain sur lesquels le logiciel s’entraîne ; pour augmenter leur détection, un entraînement spécifique et focalisé serait nécessaire.

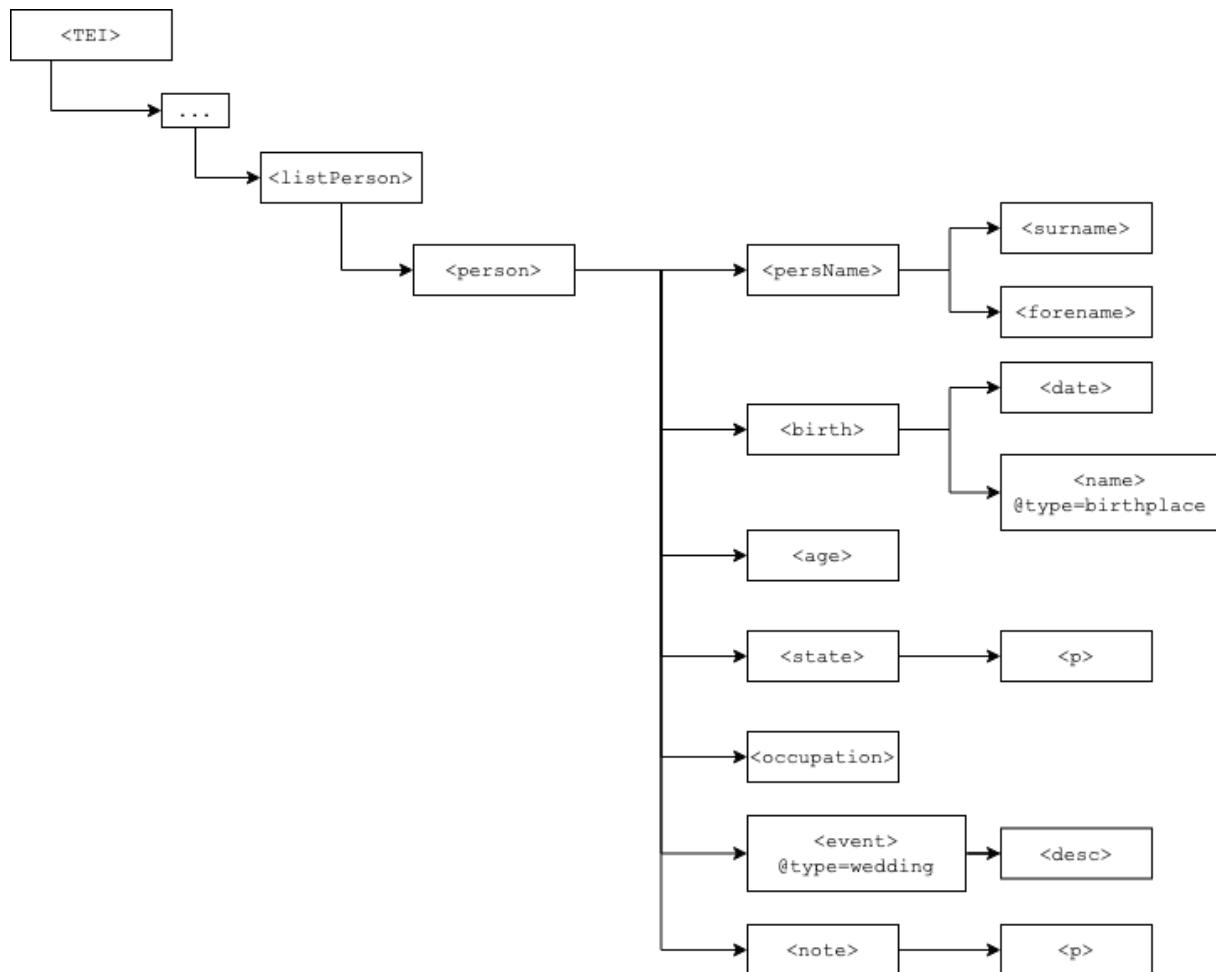


FIGURE 8.1 – Arbre XML simplifié montrant l’organisation d’une entrée d’index.

France, Lille, France, 2019, URL : <https://hal.inria.fr/hal-02448921>, slide 16.

2. Benoît Sagot et Kata Gábor, « Détection et correction automatique d’entités nommées dans des corpus OCRisés », dans *Traitement Automatique du Langage Naturel 2014*, Marseille, France, 2014, URL : <https://hal.inria.fr/hal-01022378>, p. 1.

3. Les entités nommées sont définies dans le domaine du traitement automatique du langage comme des « unités faisant référence à une entité unique et concrète et réalisées par des noms propres (noms de personnes, d’organisations, d’artefacts ou de lieux) » : *Ibid.*, p. 4

8.2 Indexer les individus

Cette défaillance dans la détection des entités nommées pose problème dans la mesure où les monographies de familles prennent pour sujet un groupe d'individus (les enquêtés) sur un territoire précis ; si leurs noms sont souvent anonymisés (*M***, F****, etc.), les prénoms sont écrits en toute lettre.

Un projet de recherche mené par le CMH et le CHR concerne ces enquêtés et souhaite les indexer en relevant les informations biographiques données par les monographies. Ce travail d'indexation des familles enquêtées est mené par Alain Cottreau et Stéphane Baciocchi (CRH, EHESS), avec Anne Lhuissier (CMH, INRAE). Ils cherchent notamment à lever l'anonymat des individus et à compléter (voire corriger) les informations biographiques délivrées dans les monographies, grâce à l'exploitation des fonds d'archives départementaux ou municipaux.

Anne Lhuissier et Stéphane Baciocchi se sont ainsi rendus aux Archives départementales de l'Isère pour trouver des informations sur le gantier Théodore G., sujet principal de la monographie n° 55⁴. Théodore G. est présenté comme un gantier vivant à Biviers, à 9km de Grenoble. Son père, mort avant le début de l'enquête, était un riche cultivateur ; les frères de Théodore cultivent toujours, au moment de l'enquête, les terres familiales. Le monographe explique que Théodore, doté « d'une forte constitution », a lui-même « pris part jusqu'à l'âge de vingt-cinq ans aux travaux de l'agriculture »⁵.

Son service militaire l'amène ensuite à participer au siège de Sébastopol lors de la guerre de Crimée en 1855⁶. Cependant, son père « l'a racheté et rappelé auprès de lui », c'est-à-dire qu'il a payé un remplaçant qui a effectué la fin de son service militaire à sa place⁷. C'est alors que, en dépit du fait qu'il « passait à juste titre pour un des ouvriers les plus intelligents et les plus certains de réussir » dans l'agriculture, relate le monographe, « il n'a pu résister au désir d'apprendre la profession de gantier »⁸. Théodore est ainsi le seul homme de sa fratrie à ne pas être devenu agriculteur.

Or Anne Lhuissier et Stéphane Baciocchi, en exploitant les archives des séries E (archives notariales), P (cadastre) et de la sous-série 3Q (enregistrement et timbre), ont pu démontrer que Théodore avait reçu une part si faible de la succession paternelle, qu'il n'était pas envisageable pour lui de poursuivre dans l'agriculture. Ses frères avaient en effet fait valoir que le paiement d'un remplaçant pour son service militaire devait être considéré comme une avance sur l'héritage, et donc déduit de celui-ci. Théodore n'a reçu aucune partie des terres paternelles mais seulement une maison ; Anne Lhuissier et Stéphane

4. Ernest de Toytot, « Gantier de Grenoble (Isère - France) », dans *Les Ouvriers des deux mondes*, Paris, 1887 (série 2e (1)), p. 465-520, URL : <https://archive.org/details/s2lesouvriersdes01sociuoft/page/n520/>.

5. *Ibid.*, p. 471.

6. *Ibid.*

7. *Ibid.*

8. *Ibid.*

Baciocchi font l'hypothèse que cette absence de terre à cultiver a été déterminante dans sa volonté de devenir ouvrier et donc de rompre avec le travail traditionnel de sa famille⁹. Cet exemple illustre les possibilités de valorisation que le programme *Time Us* envisage pour les enquêtés des *Ouvriers des deux mondes*.

Une tableau prosopographique au format CSV était déjà établi au moment de notre stage. Huit cent quarante-deux individus y étaient identifiés et décrits selon leur état civil et des critères sociaux. Les chercheurs nous ont demandé de procéder, dans un premier temps, à la transformation de ce tableau en un index XML et, dans un second temps, à la liaison entre chaque entrée et l'individu correspondant dans les monographies (Annexe B.1, *issue 4*). Du point de vue de l'ingénierie, cela se traduisait par la constitution automatique de l'index puis par l'implémentation dans les fichiers TEI des identifiants des individus cités dans le deuxième paragraphe.

Un fichier d'index au format XML repose sur un `<body>` dans lequel une ou plusieurs listes de personnes (`<listPerson>`) sont établies, les individus y étant indexés au sein d'éléments `<person>`. Nous avons constitué une liste de personnes unique, car aucun besoin particulier ne nous a été transmis quant à cette fonctionnalité de la TEI. On pourrait néanmoins imaginer des listes en fonction des sexes et, à l'intérieur de celles-ci, des sous-listes relatives aux rôles sociaux des individus.

Nous nous sommes concentrés sur les différentes catégories à faire figurer dans l'index et sur les balises pouvant les traduire (fig. 8.1).

L'identifiant de l'individu — formé de celui de la monographie, de la lettre *E* pour *enquêté* et du numéro d'apparition — constitue la valeur de l'attribut `xml:id` de la balise `<person>`. L'état civil se trouve ensuite dans un ensemble `<persName>` où le nom figure dans `<surname>` et le prénom dans `<forename>`.

D'autres balises pourraient venir compléter cette section, notamment `<addName>` pour les surnoms ou les prénoms surnuméraires. Une reprise du tableau serait néanmoins nécessaire pour rendre l'usage de cet élément possible : dans l'état actuel, le surnom ou le second prénom se trouvent dans la même cellule que le prénom, ici entre parenthèses, là après un tiret et ailleurs introduits par le transitif *dit*. Une telle reprise, consistant donc à répartir en trois cellules (prénom, prénom surnuméraire, surnom) les informations contenues en une pourrait être réalisée sur la plate-forme *Dataiku* qui permet la manipulation de données à grande échelle, par exemple grâce à des expressions régulières.

Les informations de naissance (date et lieu) sont placées dans un ensemble `<birth>`, suivi de l'âge (`<age>`).

La situation matrimoniale est décrite par la balise `<desc>` dans un ensemble `<event>` (évènement). La TEI conseille de décrire l'évènement de manière normalisée par un `<label>` placé avant `<desc>`. Cependant, dans la mesure où cet index ne compte qu'un

9. Ces lignes sont écrites avec l'aimable autorisation d'Anne Lhuissier et de Stéphane Baciocchi ; leurs conclusions, encore inédites, devraient donner lieu à une prochaine publication.

seul évènement, nous avons choisi de simplifier la structure et d'en préciser la nature grâce à un attribut @type au niveau de <event>. Suivent deux indicateurs contenant le positionnement de l'individu dans la cellule familiale (<state> : chef de famille, femme, fille) et son activité, qu'il s'agisse d'un apprentissage, d'un métier ou d'un travail à la tâche ou à la journée (<occupation>). Dans une <note> finale figurent la date de référence pour les calculs des dates de naissance et de mariage et le rappel du titre de la monographie où apparaît l'individu.

8.3 Corriger les transcriptions ?

Dans chaque volume, le paragraphe « État civil de la famille » commence par une liste standardisée des membres de la famille. Les figures 8.3 et 8.4 montrent une comparaison entre le texte d'origine de la monographie 56, sa transcription et son encodage par LSE-OD2M¹⁰. Les erreurs concernant la reconnaissance des entités nommées y apparaissent clairement : sur quatre prénoms, tous sont transcrits avec une distance d'édition importante et le nom de famille « B*** » n'est absolument pas reconnu.

§ 2. — ÉTAT CIVIL DE LA FAMILLE.

La famille comprend quatre personnes, savoir :

1. ANTOINE F***, <i>chef de famille</i> , né à L*** (Charente-Inférieure), marié en secondes noces depuis 15 ans.....	58 ans.
2. MARIE P***, sa femme, née à M*** (Charente-Inférieure).....	50 —
3. Étienne F***, leur fils unique, né à la G***.....	13 —
4. Anne P***, mère de la femme, née à M***.....	70 —

L'ouvrier avait eu un enfant de son premier mariage, contracté

FIGURE 8.2 – Liste d'individus et son encodage (série 1, volume 3, monographie n° 23, page 209).

Ici, ces erreurs ne posent pas qu'un problème de compréhension : elles empêchent la réalisation de la deuxième mission qui nous avait été confiée et qui consistait à lier les individus à leurs identifiants. En effet, cette opération se serait normalement traduite par l'implémentation d'une balise <persName> autour de chaque ensemble prénom et nom, avec un attribut @ref ayant pour valeur l'identifiant donné dans le fichier d'index.

L'implémentation automatique aurait pu s'effectuer de plusieurs manières, et notamment par la recherche du prénom : cela n'est pas possible du fait des distances d'édition trop grandes. Certaines monographies présentent également des listes numérotées (fig.

10. U. Guérin, « Tourneur-mécanicien des usines de la Société Cockerill, de Seraing (Belgique) »..., p. 4.

LOUIS-JOSEPH B***, père de famille.....	36 ans.
FÉLICITÉ-JOSÉPHE B***, mère de famille.....	45 —
GASPARDINE-MARGUERITE B***, leur fille.....	11 —
LOUISE-FÉLICITÉ B***, leur fille.....	11 —

Les deux sœurs sont jumelles. Une autre fille a été enlevée à l'âge

FIGURE 8.3 – Liste d’individus (série 2, volume 2, monographie n° 56, page 4).

```

<div n="002" type="sub_sub_section">
<head type="sub_sub_section" xml:id="para_5_4_27">
  § 2.
  <lb/>
  ÉTAT CIVIL DE LA FAMILLE.
</head>
<p xml:id="para_5_4_28">
  La famille comprend quatre personnes.
</p>
<p xml:id="para_5_4_29">
  LOUIS-JOSEPH p***, père de famille. . . . . 36 ans.
</p>
<p xml:id="para_5_4_30">
  FÉLICITÉ-JOSÉPH * mère de famille. . . . . 45 —
</p>
<p xml:id="para_5_4_31">
  GASPARDINE-MARGUERITE leur fille. . . . . 11 —
</p>
<p xml:id="para_5_4_32">
  LOUISE-FÉLICITÉ * leur fille. . . . . 11 —
</p>
<p xml:id="para_5_4_33">
  Les deux sœurs sont jumelles. Une autre fille a été enlevée à l'âge de six mois par une

```

FIGURE 8.4 – Encodage de la liste d’individus, monographie n° 56 (s2t2_chapt_5.xml).

8.2) : là encore les numéros auraient pu être utiles, si seulement ils avaient été convenablement détectés.

Une autre solution aurait procédé de l’observation que, les paragraphes deuxièmes commençant toujours par une phrase de type « La famille comprend X personnes », « La famille comprend » ou « Les membres de la famille sont », les lignes qui suivent correspondent aux membres de la famille. Il eut été dès lors possible, grâce à l’ordre d’apparition noté dans le CSV, de considérer que la première de ces lignes est celle du premier individu, et ainsi de suite. Ce procédé repose cependant sur l’idée que les items des listes se trouvent sur une seule ligne, ce qui n’est pas toujours le cas (fig. 8.2). Du reste, placer une balise `<persName>` sur l’ensemble d’une ligne serait revenu à corrompre son usage, normalement réservé à « un nom propre ou une expression nominale se référant à une personne »¹¹.

11. TEI element `persName` (*personal name*), TEI Guidelines : <https://tei-c.org/release/doc/>

Vérité terrain	ocr	Correction
LOUIS	LOUIS	—
JOSEPH	JosEP	joseph
FÉLICITÉ	FLICITÉ	félicité
JOSÈPHE	JosÉPH	joseph
GASPARDINE	GAPARDINE	gagarine
MARGUERITE	MARGUERIT	marguerite
LOUISE	ILOIE	loin
FÉLICITÉ	FÉLCITÉ	félicité

TABLE 8.1 – Comparaison entre la vérité terrain, la transcription effectuée par l’OCR et la proposition de correction de `pyspellchecker` pour les prénoms de la monographie n° 56 (série 2, volume 2, page 4 — `s2t2_chapt_5.xml`).

Ainsi, dans tous les cas, la réalité des transcriptions empêchait l’implémentation automatique de la balise et nous confinait à une opération de correction. Or celle-ci ne peut pas être menée sans une intervention humaine, qu’elle soit entièrement manuelle ou semi-automatisée grâce à un logiciel de suggestion de correction¹². Ces logiciels — par exemple, *Antidote*¹³ — s’appuient sur un lexique et, « pour chaque mot inconnu, [cherchent] des candidats proches (par exemple en termes de distance d’édition) qui figurent dans le lexique et choisissent en prenant en compte la fréquence des candidats, le contexte, ou éventuellement un poids associé au type d’erreur présumé »¹⁴. Une telle intervention, quels que soient les moyens choisis pour sa réalisation, serait probablement très longue et dans tous les cas coûteuse¹⁵.

Nous nous sommes donc intéressés à la librairie Python `pyspellchecker`¹⁶. Elle utilise un lexique, dont l’utilisateur spécifie la langue, et la distance de Levenshtein pour identifier des cacographies et proposer, selon les paramètres, soit une correction soit plusieurs candidats. La méthode `.unknown` affiche les mots du texte analysé qui ne se trouvent pas dans le lexique, et `.known()` ceux qui s’y trouvent¹⁷. `Pyspellchecker` analyse ensuite les mots faux et peut, selon les paramètres, proposer une solution unique grâce à

`tei-p5-doc/fr/html/ref-persName.html` (consulté le 21 septembre 2020).

12. B. Sagot et K. Gábor, « Détection et correction automatique d’entités nommées dans des corpus OCRisés »..., p. 1.

13. Présentation : <https://www.antidote.info/fr/> (consulté le 21 septembre 2020).

14. *Ibid.*, p. 1-2.

15. Damien Nouvel, Jean-Yves Antoine, Nathalie Friburger et Arnaud Soulet, « Fouille de règles d’annotation pour la reconnaissance d’entités nommées », dir. Sophia Ananiadou, Nathalie Friburger et Rosset Sophie, *Traitemen Automatique des Langues, Entités Nommées*, 54-2 (), p. 13-41, URL : <https://www.atala.org/content/fouille-de-r%C3%A9gles-d%C5%80%C80%C99annotation-pour-la-reconnaissance-d%C2%A9%C80%C99entit%C3%A9%C80%C99s-nomm%C3%A9%C80%C99s>, p. 14.

16. Déposé sur *GitHub* :<https://github.com/barrust/pyspellchecker> (consulté le 21 septembre 2020).

17. `Pyspellchecker` ne permet pas que de faire des corrections orthographiques; en affichant les mots connus avec `.known()`, l’utilisateur peut calculer leur fréquence grâce à la méthode `.word_probability()`.

Vérité terrain	ocr	Correction
MOHAMMED	MloAMMED	mohammed
KADIDJA	bADIDJA	badidja
RHKEÏMA	Rhleima	rhlleima
AHSOÛN	Ahsouùn	ahsoka
AÏCHA	Aicha	aisha
AROUÇI	AroOuci	adouci
KOUKA	oul	oui

TABLE 8.2 – Comparaison entre la vérité terrain, la transcription effectuée par l’OCR et la proposition de correction de `pyspellchecker` pour les prénoms de la monographie n° 25 (série 1, volume 3, page 286 — `s1t3_chapt_10.xml`).

la méthode `.correction()` ou un choix entre plusieurs candidats grâce à la méthode `.candidates()`. Par défaut, la distance est de 2, elle peut être ramenée à 1. Cela signifie que selon la valeur de la propriété `distance`, la librairie va considérer qu'une ou deux opérations de suppression, de substitution ou d'insertion ont pu avoir lieu.

Un test sur les cinq premiers `<p>` de la figure 8.4 a donné de bons résultats (tabl. 8.1). Le script en question extrait le texte des balises `<p>` grâce à la librairie `BeautifulSoup` et le nettoie en enlevant les signes de ponctuation. Ces derniers induisent en effet le correcteur en erreur : un mot suivi d'un point sera ainsi considéré comme un seul ensemble et jugé faux¹⁸. Le correcteur analyse ensuite chaque mot et retourne les mots faux ; nous l'avons paramétré de manière à ce qu'il propose une correction et des candidats de correction.

Le tableau 8.1 montre que, sur sept transcriptions fausses, `pyspellchecker` permet d'en corriger quatre de manière exacte (*Félicité* à deux reprises, *Joseph* et *Marguerite*). La vérité terrain est presque approchée pour *Josèphe* : le *e* final et l'accent grave manquent. En revanche, lorsque le même test est effectué sur une monographie où les prénoms ne se trouvent pas forcément dans un lexique français, par exemple la n° 25 consacrée à un parfumeur du bazar El Attharin-el-kebar de Tunis¹⁹, le taux chute à un prénom corrigé sur sept erreurs (tabl. 8.2). La graphie de la vérité terrain *Aïcha* figure bien dans la liste des candidats, mais c'est la forme *Aisha* qui est choisie comme correction. On voit également que le dernier prénom, *Kouka*, est transcrit avec une distance d'édition si importante (*oul*, soit trois suppressions et une insertion) qu'il est improbable que le script puisse trouver le mot d'origine.

Si ce test pourrait être transformé en un script d'aide à la correction, la présence

18. Floriane Chiffolleau, *Vers un alignement de traductions et d'éditions à partir d'un lexique et à travers un corpus multilingue, Travail sur Dei Delitti e delle Pene du marquis de Beccaria*, dir. Thibault Clérice, mémoire du Master « Technologies numériques appliquées à l'histoire », École nationale des chartes, 2019, URL : <https://github.com/FloChiff/memoire-M2>, p. 60.

19. N. Cotte et Soliman el Haraïri, « Parfumeur de Tunis (Régence de Tunis - Afrique) du bazar appelé : El Attharin-el-kebar (les grands parfumeurs) », dans *Les Ouvriers des deux mondes*, Paris, 1861 (série 1 (3)), chap. 25, p. 285-326, URL : <https://archive.org/details/lesouvriersdesde03sociuoft/page/284>.

d'un opérateur humain pour valider ou non les propositions de **pyspellchecker** reste indispensable. C'est la méthode utilisée par Floriane Chiffoleau (prom. 2019) lors de son stage de fin d'étude au sein du projet MetaLEX de l'EHESS, dont l'objectif est « d'élaborer un système d'informations métalexicographiques, concernant le vocabulaire portant sur les langues historiques du droit en Europe »²⁰. Elle a plus précisément travaillé sur l'édition numérique du *Dei Delitti e delle Pene* de Cesare Beccaria (1764), ce qui l'a conduit à utiliser **pyspellchecker** pour corriger les résultats de son OCR. Elle produit dans son mémoire un retour d'expérience détaillé sur l'utilisation de cette librairie, en insistant sur la division des tâches et la constitution de dictionnaires Python²¹.

Cinq étapes sont décrites. La première consiste à nettoyer et mettre en forme automatiquement le texte en supprimant les sauts de ligne, les espaces en trop et les signes de ponctuation²². Les erreurs sont ensuite relevées par **pyspellchecker** et réparties en trois ensembles : une liste avec les mots non-reconnus mais justes, un dictionnaire Python avec les versions non-usuelles du XVII^e siècle (par ex. *paroître* à la place de *paraître*²³) et un dernier dictionnaire contenant les erreurs véritables et leurs corrections²⁴. Lors de la troisième étape, le texte est cette fois-ci récupéré en entier, avec sa mise en forme et sa ponctuation²⁵. Le dictionnaire d'erreurs est passé en revue lors de l'étape suivante afin de contrôler et d'amender si nécessaire les propositions de **pyspellchecker**²⁶. Les opérations de correction sont enfin opérées à l'aide d'une fonction de recherche et de remplacement à partir du dictionnaire d'erreurs : chaque mot faux est recherché et ses occurrences sont remplacées par sa correction²⁷.

Pyspellchecker a fait ses preuves dans plusieurs projets de recherche : MetaLEX, comme nous venons de le voir, mais également DAHN pour lequel Floriane Chiffoleau travaille aujourd'hui en tant qu'ingénierie de recherche et de développement de l'équipe ALMAnaCH. Son expérience et ses scripts peuvent être ré-utilisés pour les fichiers des *Ouvriers des deux mondes*, avec quelques modifications dues notamment au fait que nos fichiers ne contiennent pas de mots en ancien français²⁸. Nous pouvons également observer que la majorité du processus est automatisé, mis à part la correction du dictionnaire d'erreurs.

Une autre librairie Python, **pygrammalecte**, peut être utilisée pour générer un rap-

20. F. Chiffoleau, *Vers un alignement de traductions et d'éditions à partir d'un lexique et à travers un corpus multilingue...*, p. 9.

21. *Ibid.*, chap. 8, *Établir une correction orthographique et une annotation linguistique semi-automatique*, p. 57-66.

22. *Ibid.*, p. 54 et 60.

23. *Ibid.*, p. 57.

24. *Ibid.*, p. 59-60.

25. *Ibid.*, p. 61.

26. *Ibid.*, p. 61-63.

27. *Ibid.*, p. 63-64.

28. Les scripts utilisés par Floriane Chiffoleau lors de son stage sont déposés sur *GithHub* : <https://github.com/FloChiff/memoire-M2/tree/master/Livrable%20technique/Scripts/20-Nettoyage%20de%20texte> (consulté le 21 septembre 2020).

port d’identification des erreurs. La différence avec la librairie précédente est qu’elle fonctionne avec un dictionnaire et est capable, en plus de détecter les erreurs, de les répartir dans une typologie (orthographe, grammaire, ponctuation). Cette librairie est utilisée par le service humanités numériques de l’École nationale des chartes pour évaluer la qualité des OCR des positions des thèses pour le diplôme d’archiviste paléographe, et permet de générer un rapport d’erreur sous la forme d’un fichier JSON²⁹.

On le voit, il est possible de développer des solutions en interne pour corriger une OCR. Si elles n’automatisent pas totalement cette correction, elles permettent de réduire son coût budgétaire en épargnant le recours à un prestataire externe ou l’achat d’un logiciel dédié. Elle peuvent néanmoins se traduire par le recours à un vacataire pour le temps de la correction, ce qui représente tout de même un poste de dépense.

Durant notre stage, nous avons contourné le problème en ajoutant, au début des deuxièmes paragraphes, un commentaire contenant les identités des individus cités et leurs identifiants. Un commentaire est dans un fichier informatique une section du code destinée à la lecture par un humain et non par une machine, qui de fait l’ignore au moment de l’interprétation du fichier. En XML, le commentaire est délimité par les symboles <!-- et -->. Dans la monographie n° 56, cela donne cette ligne :

```
<!-- Individus cités : Louis-Joseph B*** (n° 056aE1) - Félicité-Josèphe
B*** (n° 056aE2) - Gaspardine-Marguerite B*** (n° 056aE3) - Louise-
Félicité B*** (n° 056aE4) -->
```

Il s’agit d’une solution d’attente qui n’a pas vocation à demeurer dans la version définitive des fichiers.

29. Voir notamment la fonction `ocrquality` dans le script `encpos_control.py`, ligne 72 : https://github.com/chartes/encpos/blob/38ad0277a03467642dd8a329a37237c9e663f4d1/utils/encpos_control.py#L72 (consulté le 21 septembre 2020).

Chapitre 9

Une valorisation plurielle

9.1 Édition papier, édition numérique

Un postulat commun veut qu'une édition numérique offre à son utilisateur plus de possibilités qu'une édition papier n'en offre à son lecteur¹. Le numérique permet en effet de concevoir des plate-formes sur mesure pour la consultation des textes, et les outils des humanités numériques permettent à « de nombreuses données non interrogables jusqu'à présent [d'être] l'objet d'enquêtes »². « Des structures cachées, des faits de système difficilement décelables à l'œil et à la main » deviennent ainsi accessibles³.

Dans le programme *Time Us*, l'édition numérique des *Ouvriers des deux mondes* a éveillé l'intérêt d'au moins trois participants. Le LARHRA de l'université de Lyon 2, dans le strict respect des objectifs du programme, souhaite utiliser les informations économiques fournies par les tableaux de budget et celles d'ordre prosopographique contenues dans le paragraphe « §2 — État civil de la famille ». L'équipe ALMAnaCH d'Inria a pour sa part la volonté de puiser dans les champs lexicaux des mondes ouvrier et industriel afin d'alimenter des algorithmes de traitement automatique du langage (TAL). Enfin, le Centre de recherches historiques (CRH) veut fournir à la communauté scientifique une édition numérique des *Ouvriers des deux mondes* qui intégrerait des informations matérielles sur la constitution du corpus et le travail de la SESS.

On le voit, les objectifs poursuivis par ces entités sont très divers. Les matériaux qui permettront de les réaliser le sont tout autant : le LARHRA et ALMAnaCH ont besoin de données brutes issues des tableaux statistiques (des chiffres) et du texte (des mots), tandis que le CRH agrège des métadonnées inédites qui proviennent de plusieurs corpus ;

1. « Often these descriptions glance at their print predecessors, usually with expressions of how much more these digital editions can contain than ever could be included in print editions, and how much more the reader can do with them » : Peter Robinson, « Towards a Theory of Digital Editions », *Variants, The Journal of the European Society for Textual Scholarship*, 10 (2013), p. 105-131, p. 105-106.

2. Frédéric Duval, « Pour des éditions numériques critiques. L'exemple des textes français », *Médiévales*, 73 (automne 2017), p. 13-29, DOI : 10.4000/medievales.8165, p. 20.

3. *Ibid.*

il dispose notamment d'un exemplaire des deuxième et troisième séries encore à l'état de fascicules non reliés. Le LARHRA a également besoin d'une transcription de qualité pour s'assurer de la viabilité des informations prosopographiques du second paragraphe.

Cette pluralité de directions illustre la tension qui traverse les éditions numériques : elles portent avant tout sur un document et non sur une œuvre⁴. Cette distinction est issue de la triade document, texte, œuvre (*document, text, work*) qui désigne les dimensions matérielle, linguistique et intellectuelle d'un écrit⁵. De fait, l'encodage que nous avons décrit dans la partie précédente fait la part belle au document *Ouvriers des deux mondes* de l'Université de Toronto, digitalisé par *Internet Archive*. *Transkribus* et le script LSE-OD2M mobilisent la TEI pour conduire sa reproduction fidèle grâce à des ensembles <facsimile>⁶. L'œuvre n'est pas pour autant oubliée. Elle se rencontre dans la structure logique leplaysienne, là encore reproduite fidèlement par le biais des divisions et des titres.

Cette coexistence apparente et inégale n'a cependant pas vocation à durer : dans la vision de *Time Us*, c'est bien l'œuvre et non le document qui doit prendre le dessus. Il n'est pas question de concevoir un support de consultation qui présenterait des échantillons successifs correspondant au contenu d'une page. Cela reviendrait à reproduire l'interface de visualisation d'*Internet Archive*, tout en donnant une réalité concrète à l'hypertexte qui est « en gestation dans les tables, index et diverses aides à la lecture de consultation déjà présentes » dans les volumes⁷.

Ces observations montrent qu'une édition numérique se doit d'être équilibrée et de rendre compte à la fois du document-texte et de l'œuvre-texte⁸, sans quoi elle risque de perdre tant ses lecteurs⁹ que ses utilisateurs¹⁰.

Le premier essai d'édition des fichiers XML, mené par Alix Chagué, consiste ainsi en un document HTML où l'intégralité du texte du premier volume est reproduit, organisé

4. « *Two decades of making digital editions, and recent papers about digital editions, have moved the needle away from the “work” to the “document”, to the point where we might need only think of “documents”* » : P. Robinson, « Towards a Theory of Digital Editions »..., p. 107

5. « *Work* désigne le texte de l'auteur, éventuellement le texte correspondant à la volonté de l'auteur, et implique la notion d'authenticité ; *text* dénomme la séquence linguistique attestée dans un document transmettant l'œuvre ; enfin *document* est une manifestation physique d'un *text* » : F. Duval, « Pour des éditions numériques critiques. L'exemple des textes français »..., p. 15-16.

6. P. Robinson, « Towards a Theory of Digital Editions »..., p. 124.

7. F. Duval, « Pour des éditions numériques critiques. L'exemple des textes français »..., p. 19.

8. « *A scholarly edition must, so far as it can, illuminate both aspects of the text, both text-as-work and text-as-document* » : P. Robinson, « Towards a Theory of Digital Editions »..., p. 123.

9. « *But there are dangers here. (...) Facsimile editions in print form are of very little use to the reader, or even to scholars, whose interest (...) is likely to be in questions of how the received text changed over time, how it was received, how it was altered, transformed, passed into different currencies. If we make only digital documentary editions, we will distance ourselves and our editions from the readers* » : *Ibid.*, p. 127.

10. « La lisibilité des éditions électroniques n'a rien à envier à celle des éditions papier. (...) Parfois, les interfaces ne sont pas intuitives et requièrent une longue familiarisation ; d'autres fois, des aides à la lecture systématiquement présentes dans les éditions papier disparaissent » : F. Duval, « Pour des éditions numériques critiques. L'exemple des textes français »..., p. 21.

en fonction de la structure logique et non des zones de segmentation¹¹. Elle pourrait cependant être améliorée avec des informations issues du document, à commencer par la traduction, par exemple entre crochets droits, des balises <pb> à chaque changement de page.

Time Us a identifié deux voies possibles pour une publication effective. La première consiste à déposer les textes sur *Wikisource*. Il s'agit d'une bibliothèque numérique maintenue par la fondation Wikimédia, gratuite, en *open access*, contenant des *open data* (aucune restriction sur l'usage des données) et ouverte à tous ceux qui souhaitent y contribuer. Le problème est que les fichiers XML ne peuvent pas être versés en l'état : ils doivent être modifiés voire transformés dans la syntaxe de la Wikimedia, le wikicode. Cette opération peut donc s'avérer coûteuse et le transfert de la totalité des informations n'est pas garanti ; les dépôts git sont de ce point de vue de bien meilleures solutions.

Cependant, le principe de *Wikisource* pourrait être conservé tout en contournant les règles de sa communauté. Le CRH souhaite en effet utiliser l'un de ses wikis sémantiques, *Ethnocomptabilités*, pour publier les monographies et leurs métadonnées entendues au sens large (volumes, enquêtés, enquêteurs)¹². Un wiki sémantique est un wiki — c'est-à-dire une application web où des pages peuvent être créées et modifiées par des utilisateurs — augmenté de renseignements sémantiques permettant de qualifier les relations entre les pages et, *in fine*, de manipuler les données qui s'y trouvent¹³. Le programme *Time Us* y est libre de publier ses données sous la forme qu'il souhaite ; cette publication doit néanmoins s'accompagner d'un effort de référencement et de communication afin que le wiki soit consulté et utilisé par la recherche.

9.2 Retrouver les fascicules dans le volume

Au premier abord, il est aisément de considérer les *Ouvriers des deux mondes* comme une œuvre. Mais plus on progresse dans l'histoire de cette entreprise, plus cette idée première se fragilise. En effet, si ce corpus s'incarne aujourd'hui dans des volumes, ceux-ci étaient autrefois des fascicules. Il est en outre constitué de monographies réparties en trois séries, tout en formant un ensemble cohérent dont les membres « prennent sens les uns par rapport aux autres »¹⁴. D'une certaine manière, la volonté du CRH d'explorer l'histoire matérielle du corpus pour compléter les métadonnées des fichiers XML remet en question l'existence de « l'œuvre *Ouvriers des deux mondes* ».

11. Cette démonstration est visible à cette adresse : http://demo-leplay.herokuapp.com/volume_parsed_test.html (consulté le 21 septembre 2020).

12. *Ethnocomptabilités* est hébergé à cette adresse : <http://ethnocompta.huma-num.fr/> (consulté le 21 septembre 2020).

13. Gautier Poupeau, *Et le wiki devint sémantique*, Les Petites cases, URL : <https://www.lespetitescases.net/et-le-wiki-devint-semantique>.

14. A. Chenu, « Préface »..., p. 5.

Les enjeux de l'édition numérique d'un texte imprimé aux XIX^e et XX^e siècles divergent de ceux sous-entendus par l'édition de texte imprimé lors de siècles antérieurs. Aucune ré-impression n'est ici attestée : le texte est le même d'un exemplaire à l'autre, et ce jusque dans ses imperfections. C'est la raison pour laquelle l'encodage ne fait aucun effort de lématisation et qu'il s'appuie, de fait, sur les seules numérisations des exemplaires de Toronto.

Dans l'encodage d'un texte imprimé tel que *Les Ouvriers des deux mondes*, le défi se situe non pas au niveau du texte et de ses différentes versions, mais bien dans la restitution de la génétique matérielle qui a amené à la constitution des volumes. Comment traduire dans l'encodage les stratégies mises en place par les différents relieurs pour fondre les fascicules dans le volume ? Dans les exemplaires de la Bibliothèque nationale de France, les feuillets liminaires des fascicules ont été placés après les tables, là où ailleurs ils ont été conservés dans le flux du texte. Il y a ici une subtilité dont l'encodage de niveau « document » ne se préoccupe pas, et qui pourtant est essentiel pour les deux autres niveaux.

Les observations relevées par le CRH peuvent être insérées dans la partie <sourceDesc> du <teiHeader>. Le document encodé y est décrit dans une section <msDesc>, où une unité codicologique distincte de l'unité documentaire principale peut être elle-même décrite grâce à la sous-section <msPart> (fig. 9.1).

Plusieurs types de métadonnées sont spécifiés dans le <msDesc>. Le volume est tout d'abord identifié dans le <msIdentifier> par l'exposé de ses situations géographique (pays et ville de conservation) et institutionnelle (institution et département de conservation, cote), et la liste de ses copies, telles que la numérisation sur *Internet Archive*. Il est ensuite titré et daté dans une partie <head>, avant que le <msContents> ne donne des informations sur sa structure : l'auteur, la pagination, la langue, l'*incipit* et l'*explicit*. Les caractéristiques physiques sont ensuite inscrites dans le <physDesc> (support, dimensions, mise en page, état). La reliure du volume et les marbrures des gardes sont décrites dans le <bindingDesc>.

L'histoire du volume est enfin détaillée dans la partie <history>, grâce à des éléments sur son origine (contexte d'édition), sa provenance (événements entre l'origine et l'acquisition) et son acquisition (contexte d'acquisition). La pertinence d'une telle partie n'est pas certaine, tant pour les *Ouvriers des deux mondes* que pour le corpus de fascicules non-reliés à la disposition du CRH et du CMH, car leurs histoires ne sont pas connues. Quelques relevés peuvent permettre d'éclaircir des points très spécifiques, par exemple la dédicace d'un monographe dans le volume 5¹⁵.

Le <msDesc> se conclurait par un <msPart> centré sur le fascicule, divisé selon le

15. « M. Auguste Moussel, hommage en cordial confraternité. Urbain Guérin » : U. Guérin, « Fileur en peigné et régleur de métier de la Manufacture du Val-des-Bois (Marne - France) », dans *Les Ouvriers des deux mondes*, Paris, (série 2e (5)), p. 73-136, URL : <https://archive.org/details/2serlesouvriersde05sociuoft/page/n98/>.

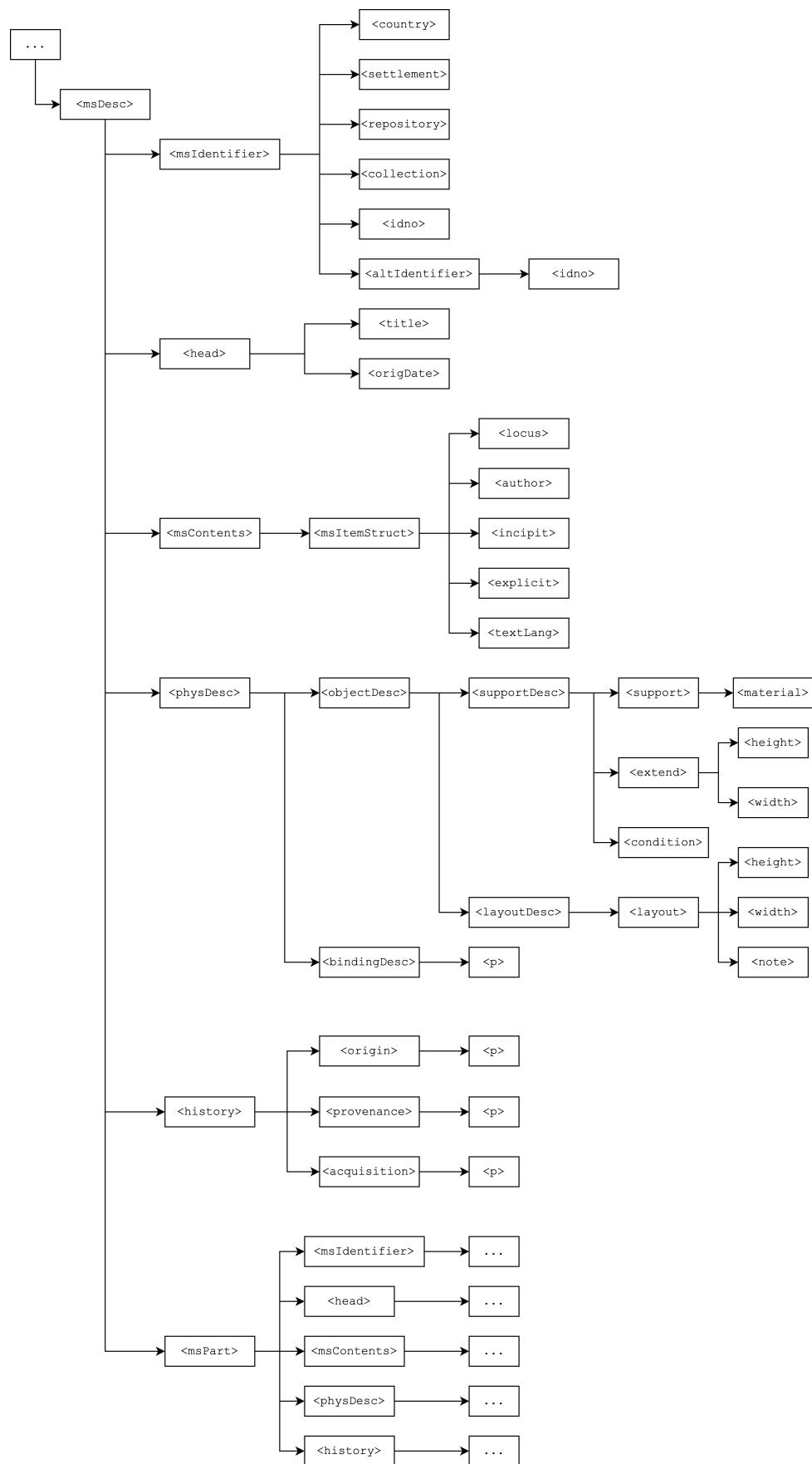


FIGURE 9.1 – Représentation de la section <msDesc> du <teiHeader>.

même plan.

Ces apports se bornent aux métadonnées, l'encodage produit par LSE-OD2M et amélioré par notre reprise n'étant pas affecté. Rappelons que celui-ci n'est pas pour autant fixé. La place des objets graphiques reste à définir. Si la structure logique est en place, le système des renvois entre les monographies n'a pas été valorisé ; les informations prosopographiques doivent également être repérées et signalées après une correction des transcriptions.

9.3 *Quid de la donnée ?*

Une question demeure au sujet des fichiers des *Ouvriers des deux mondes* : quel avenir pour eux, non pas dans une mise en scène quelconque, mais en tant que données brutes ? Le choix de recourir au format XML-TEI montre que *Time Us* entend assurer la conservation des fichiers. La TEI est en effet maintenue par une communauté active. En rédigeant une cartographie du corpus et surtout une ODD, nous avons documenté la pratique éditoriale et favorisé sa compréhension par d'autres chercheurs ou projets qui souhaiteraient réutiliser l'encodage. Le dépôt en ligne sur le *GitLab* de l'Inria assure la conservation de l'historique de nos interventions.

Néanmoins, le dépôt *GitLab* est aujourd'hui en accès restreint, aussi le corpus n'est-il pas en accessible librement (*open access*). Cette restriction est bien évidemment due au fait que le travail n'est pas achevé et sera levée à la fin du programme ANR. Pour autant, *Time Us* souhaite démultiplier les espaces de conservation en clonant le dépôt *GitLab* sur *Github*. Ces deux sites offrent des services d'hébergement utilisant la technologie git, à la différence que *Github* est une entreprise commerciale possédée par Microsoft, aussi ne dépend-t-elle pas d'une institution publique dont les orientations budgétaires peuvent changer. Des procédures de migration existent entre les deux plates-formes et concernent les *commits* comme les *issues* et les *merge requests*.

Le CRH veut également verser les fichiers XML et peut-être, par sécurité, les images, dans l'entrepôt de données de l'EHESS, *Didomena*¹⁶. Cette application traite les données sur deux niveaux. Le logiciel *Hyrax* reçoit d'abord les fichiers, les stocke et leur agrège des métadonnées dont les descriptions respectent le standard *Portland Common Data Model*¹⁷ ; elles sont ainsi assurées d'être interopérables. La seconde dimension est la couche applicative, soutenue par le logiciel *Solr*, qui organise des vues logiques et permet des recherches à facettes dans les données comme dans les métadonnées. Le point faible est que la consultation n'est pas libre : il faut disposer d'un compte dans un établissement

16. *Le projet Didomena : une plate-forme moderne au service de la gestion des données de la recherche en sciences humaines et sociales* (<https://didomena.ehess.fr/informations>, consulté le 21 septembre 2020).

17. Présentation accessible à <https://github.com/duraspace/pcdm/wiki> (consulté le 21 septembre 2020).

d'enseignement supérieur ou de recherche pour se connecter sur la plate-forme.

Ces différentes pistes montrent que les données des *Ouvriers des deux mondes*, déjà standardisées, documentées et pérennisées, bientôt en *open access*, sont assurées d'être indépendantes de toutes les mises en scène auxquelles elles pourraient se prêter¹⁸.

Ces dernières années, l'écosystème de la donnée s'est néanmoins tourné vers la mise en relation de celles-ci¹⁹. Une des conditions pour réaliser cela est l'interopérabilité. Les fichiers des *Ouvriers des deux mondes* satisfont-ils à cette demande ? En d'autres termes, la question est de déterminer si les données qu'ils contiennent peuvent être enrichies par des informations venues d'autres formats, ou bien si elles peuvent être réutilisées par des programmes de recherche ou avec d'autres corpus dont le standard n'est peut-être pas la TEI, ou qui n'ont pas le même usage de la TEI. C'est là où le bâton blesse : la TEI, du fait de son extrême souplesse, permet des pratiques éditoriales qui peuvent être très éloignées. Ce que les corpus gagnent en précision et en fidélité par rapport au(x) document(s) d'origine, ils le perdent en interopérabilité²⁰.

18. V. Jolivet, « Éditions ou données ? API et (re)publications »..., p. 63.

19. *Ibid.*, p. 66.

20. *Ibid.*, p. 61-62.

Conclusion

Blbla

Annexes

A. *Les Ouvriers des deux mondes*

A.1 Liste des monographies et fichiers correspondant

La date de publication du volume est indiquée entre parenthèses, tandis que les dates de parution des fascicules le sont entre crochets droits.

A.1.1 Série 1, vol. 1 (1857).

URL sur *Internet Archive* :

<https://archive.org/details/lesouvriersdesde01sociuoft>.

Id	Intitulé	Fichier
401a	Page de titre	s1t1_chapt_1.xml
402a	Avertissement. Considérations générales sur la Société internationale des études pratiques d'économie sociale. Son but et ses moyens d'action.	s1t1_chapt_2.xml
403a	Institution. Société internationale des études pratiques d'économie sociale. Fondation et premiers travaux.	s1t1_chapt_3.xml
404a	Définitions par ordre alphabétiques des termes à employer dans les monographies, pour désigner les ouvriers, leurs moyens d'existence, et les rapports qui les unissent soit entre eux, soit avec les autres classes.	s1t1_chapt_4.xml
405a	Explications des signes de renvoi et des abréviations.	s1t1_chapt_5.xml
001a	Charpentier de Paris (Seine - France), de la Corporation des compagnons du Devoir	s1t1_chapt_6.xml
002a	Manœuvre-Agriculteur de la Champagne pouilleuse (Marne - France)	s1t1_chapt_7.xml
003a	Paysans en communauté du Lavedan (Hautes-Pyrénées - France)	s1t1_chapt_8.xml
004a	Paysan du Labourd (Basses-Pyrénées - France)	s1t1_chapt_9.xml

005a	Métayer de la banlieue de Florence (Grand-Duché de Toscane)	s1t1_chapt_10.xml
006a	Nourrisseur de vaches de la banlieue de Londres (Middlesex - Angleterre)	s1t1_chapt_11.xml
007a	Tisseur en châles de la fabrique urbaine collective de Paris (Seine - France)	s1t1_chapt_12.xml
008a	Manœuvre-agriculteur du comté de Nottingham (Angleterre)	s1t1_chapt_13.xml
009a	Pêcheur côtier, maître de barque de Saint-Sébastien (Guipuscoa - Espagne)	s1t1_chapt_14.xml
406a	Tables alphabétique et analytique des matières contenues dans ce tome premier.	s1t1_chapt_15.xml
407a	Liste des monographies destinées aux prochaines publications de la société d'économie sociale.	s1t1_chapt_16.xml
408a	Tables des matières contenues dans ce tome premier	s1t1_chapt_17.xml

A.1.2 Série 1, vol. 2 (1858).

URL sur *Internet Archive* :

<https://archive.org/details/lesouvriersdesde02sociuoft>.

Id	Intitulé	Fichier
409a	Page de titre	s1t2_chapt_1.xml
410a	Avertissement	s1t2_chapt_2.xml
010a	Ferblantier, couvreur et vitrier d'Aix-les-Bains (Savoie - États Sardes)	s1t2_chapt_3.xml
011a	Carrier des environs de Paris (Seine - France)	s1t2_chapt_4.xml
012a	Menuisier-charpentier (Nedjar) de Tanger (Province de Tanger - Maroc)	s1t2_chapt_5.xml
013a	Tailleur d'habits de Paris (Seine - France)	s1t2_chapt_6.xml
014a	Compositeur-typographe de Bruxelles (Brabant - Belgique)	s1t2_chapt_7.xml
015a	Décapeur d'outils en acier de la fabrique d'Hérimoncourt (Doubs - France)	s1t2_chapt_8.xml
016a	Monteur d'outils en acier de la fabrique d'Hérimoncourt (Doubs - France)	s1t2_chapt_9.xml

017a	Porteur d'eau de Paris (Seine - France)	s1t2_chapt_10.xml
018a	Paysans en communauté et en polygamie de Bousrah (Esky Cham), dans le pays de Haouran (Syrie - Empire Ottoman)	s1t2_chapt_11.xml
019a	Débardeur et piocheur de craie de la banlieue de Paris (Seine - France)	s1t2_chapt_12.xml
411a	Tables alphabétique et analytique des matières contenues dans ce tome second	s1t2_chapt_13.xml
412a	Errata	s1t2_chapt_14.xml
413a	Tables des matières contenues dans ce tome second	s1t2_chapt_15.xml

A.1.3 Série 1, vol. 3 (1861).

URL sur *Internet Archive* :

<https://archive.org/details/lesouvriersdesde03sociuoft>.

Id	Intitulé	Fichier
414a	Page de titre	s1t3_chapt_1.xml
415a	Avertissement	s1t3_chapt_2.xml
416a	Rapport. Société d'économie sociale. Travaux de 1859-1860	s1t3_chapt_3.xml
417a	Liste générale des membres de la Société internationale des études pratiques d'économie sociale	s1t3_chapt_4.xml
020a	Brodeuses des Vosges (Vosges - France)	s1t3_chapt_5.xml
021a	Paysan et savonnier de la Basse-Provence (Bouches-du-Rhône - France)	s1t3_chapt_6.xml
022a	Mineur des Placers du comté de Mariposa (Californie - États-Unis)	s1t3_chapt_7.xml
023a	Manœuvre-vigneron de l'Aunis (Charente-inférieure - France)	s1t3_chapt_8.xml
024a	Lingère de Lille (Nord - France)	s1t3_chapt_9.xml
025a	Parfumeur de Tunis (Régence de Tunis - Afrique) du bazar appelé : El Attharin-el-kebar (les grands parfumeurs)	s1t3_chapt_10.xml
026a	Instituteur primaire d'une commune rurale de la Normandie (Eure - France)	s1t3_chapt_11.xml

027a	Manœuvre à famille nombreuse de Paris (Seine - France)	s1t3_chapt_12.xml
028a	Fondeur de plomb des Alpes Apuanes (Toscane - Italie)	s1t3_chapt_13.xml
418a	Tables alphabétique et analytique des matières contenues dans ce tome troisième	s1t3_chapt_14.xml
419a	Errata de ce tome troisième	s1t3_chapt_15.xml
420a	Tables des matières contenues dans ce tome troisième	s1t3_chapt_16.xml

A.1.4 Série 1, vol. 4 (1862).

URL sur *Internet Archive* :

<https://archive.org/details/lesouvriersdesde04sociuoft>.

Id	Intitulé	Fichier
421a	Page de titre	s1t4_chapt_1.xml
422a	Explications des signes de renvoi et des abréviations.	s1t4_chapt_2.xml
423a	Avertissement	s1t4_chapt_3.xml
424a	Rapport. Société d'économie sociale. Travaux de 1860-1861	s1t4_chapt_4.xml
425a	Instruction. Méthode d'observation des monographies de famille propre à l'ouvrage intitulé Les ouvriers européens	s1t4_chapt_5.xml
426a	Histoire de la famille. Prix fondé par M. le baron de Damas. Par la société d'économie sociale	s1t4_chapt_6.xml
029a	Paysan d'un village à banlieue morcelée du Laonnais (Aisne - France)	s1t4_chapt_7.xml
030a	Paysans en communauté du Ning-Po-Fou (province de Tché-Kian - Chine)	s1t4_chapt_8.xml
031a	Mulâtre affranchi de l'Ile de la Réunion (Océan Indien)	s1t4_chapt_9.xml
032a	Manœuvre-vigneron de la Basse-Bourgogne (Yonne - France)	s1t4_chapt_10.xml
033a	Compositeur-typographe de Paris (Seine - France)	s1t4_chapt_11.xml
034a	Auvergnat brocanteur en boutique à Paris (Seine - France)	s1t4_chapt_12.xml

035a	Mineur de la Maremme de Toscane (Toscane - Italie)	s1t4_chapt_13.xml
036a	Tisserand des Vosges (Haut-Rhin - France)	s1t4_chapt_14.xml
037a	Pêcheur côtier, maître de barques, de Marken (Hollande septentrionale - Pays-Bas)	s1t4_chapt_15.xml
427a	Société internationale des études pratiques d'économie sociale. Officiers composants les conseils d'administration et de surveillance pour la session 1863-1864.	s1t4_chapt_16.xml
428a	Liste générale des membres de la Société internationale des études pratiques d'économie sociale au 1er août 1863	s1t4_chapt_17.xml
429a	Tables alphabétique et analytique des matières contenues dans ce tome quatrième	s1t4_chapt_18.xml
430a	Errata de ce tome quatrième	s1t4_chapt_19.xml
431a	Tables des matières contenues dans ce tome quatrième	s1t4_chapt_20.xml

A.1.5 Série 1, vol. 5 [1875, 1883, 1884] (1885).

URL sur *Internet Archive* :

<https://archive.org/details/lesouvriersdesde05sociuoft>

Id	Intitulé	Fichier
432a	Page de titre	s1t5_chapt_1.xml
433a	Oeuvres de F. Le Play	s1t5_chapt_2.xml
434a	Sommaire. Monographies de familles publiées dans ce volume	s1t5_chapt_3.xml
435a	Avertissement	s1t5_chapt_4.xml
436a	Explications des signes de renvoi et des abréviations employés dans le cours de cet ouvrage	s1t5_chapt_5.xml
038a	Fermiers à communauté taisable du Nivernais (Saône-et-Loire - France)	s1t5_chapt_6.xml
039a	Paysan de Saint-Irénée (Bas-Canada - Amérique du Nord)	s1t5_chapt_7.xml
040a	L'Ouvrier éventailiste de Sainte-Geneviève (Oise - France)	s1t5_chapt_8.xml

041a	Ouvrier cordonnier de Malakoff (Seine - France)	s1t5_chapt_9.xml
041b	Précis d'une monographie ayant pour objet un chifonnier instable et, par alternance, mégissier fumiste et brossier de Paris (France - Seine)	<i>In supra.</i>
042a	Serrurier-forgeron de Paris (Seine - France)	s1t5_chapt_10.xml
042b	Précis d'une monographie ayant pour objet le monteur en bronze de Paris	<i>In supra.</i>
043a	Brigadier de la Garde républicaine de Paris (Seine - France)	s1t5_chapt_11.xml
044a	Paysan-résinier de Lévignacq (Landes - France)	s1t5_chapt_12.xml
045a	Bûcheron usager de l'ancien Comté de Dabo (Lorraine allemande)	s1t5_chapt_13.xml
046a	Paysans en communauté et colporteurs émigrants de Tabou-Douchd-El-Baar (Grande Kabylie - Province d'Alger)	s1t5_chapt_14.xml
437a	Société d'économie sociale. Conseil d'administration pour l'année 1885	s1t5_chapt_15.xml
438a	Liste générale des membres de la Société d'économie sociale au 15 mars 1885	s1t5_chapt_16.xml
439a	Tables alphabétique et analytique des matières contenues dans ce tome cinquième	s1t5_chapt_17.xml
440a	Tables des matières contenues dans ce tome cinquième	s1t5_chapt_18.xml

A.1.6 Série 2, vol. 1 [1885-1887] (1887).

URL sur *Internet Archive* :

<https://archive.org/details/s2lesouvriersdes01sociuoft>

Id	Intitulé	Fichier
441a	Page de titre	s2t1_chapt_1.xml
442a	Sommaire des monographies de familles publiées dans ce volume	s2t1_chapt_2.xml
443a	Avertissement sur ce premier volume, deuxième série des Ouvriers des deux mondes	s2t1_chapt_3.xml
047a	Paysan-paludier du Bourg de Batz (Loire-Inférieure - France)	s2t1_chapt_4.xml

048a	Bordiers émancipés en communauté rurale de la Grande-Russie	s2t1_chapt_5.xml
048b	Précis d'une monographie de l'armurier des manufactures impériales de Toula (Grande-Russie)	s2t1_chapt_6.xml
049a	Charron des forges et fonderies de Montataire (Oise - France)	s2t1_chapt_16.xml
050a	Faienciers de Nevers (Nièvre - France)	s2t1_chapt_17.xml
051a	Cultivateur-maraîcher de Deuil (Seine-et-Oise - France)	s2t1_chapt_18.xml
052a	Pêcheur-côtier, maître de barque, de Martigues (Bouches-du-Rhône - France)	s2t1_chapt_19.xml
053a	Métayer à famille-souche du pays d'Horte (Landes - France)	s2t1_chapt_20.xml
054a	Arabes pasteurs nomades de la tribu des Larbas (Région saharienne de l'Algérie)	s2t1_chapt_21.xml
055a	Gantier de Grenoble (Isère - France)	s2t1_chapt_22.xml
444a	Tables alphabétique et analytique des matières contenues dans le présent volume	s2t1_chapt_24.xml
445a	Table des matières dans ce tome premier (deuxième série)	s2t1_chapt_25.xml

A.1.7 Série 2, vol. 2 [1887-1889] (1890).

URL sur *Internet Archive* :

<https://archive.org/details/s2lesouvriersdes02sociuoft>.

Id	Intitulé	Fichier
446a	Page de titre	s2t2_chapt_1.xml
447a	Page de titre	s2t2_chapt_2.xml
448a	Sommaire des monographies de familles publiées dans ce volume	s2t2_chapt_3.xml
449a	Avertissement sur ce deuxième tome de la deuxième série des Ouvriers des deux mondes	s2t2_chapt_4.xml
056a	Tourneur-mécanicien des usines de la Société Cocke-rill, de Seraing (Belgique)	s2t2_chapt_5.xml
057a	Bordier (Fellah) berbère de la Grande-Kabylie (Province d'Alger)	s2t2_chapt_6.xml

057b	Précis d'une monographie du paysan colon du Sahel (Algérie)	s2t2_chapt_7.xml
058a	Pêcheur côtier d'Heyst (Flandre occidentale - Belgique)	s2t2_chapt_8.xml
058b	Précis d'une monographie du pêcheur côtier, maître de barque, d'Étretat (Seine-Inférieure - France)	s2t2_chapt_9.xml
059a	Paysan-métayer de la Basse Provence (Bouches-du-Rhône - France)	s2t2_chapt_10.xml
059b	Précis d'une monographie du paysan et maçon émigrant de la Marche (Creuse - France)	s2t2_chapt_11.xml
060a	Mineur silésien du bassin houiller de la Ruhr (Prusse rhénane - Allemagne)	s2t2_chapt_12.xml
061a	Mineur des soufrières de Lercara (Province de Palerme - Sicile)	s2t2_chapt_13.xml
062a	Tailleur de Silex et vigneron de l'Orléanais (Loir-et-Cher - France)	s2t2_chapt_14.xml
063a	Vigneron précariste et métayer de Valmontone (Province de Rome - Italie)	s2t2_chapt_15.xml
064a	Paysans corses en communauté, porchers-bergers des montagnes de Bastelica	s2t2_chapt_16.xml
450a	Tables alphabétique et analytique des matières contenues dans le présent tome, avec index explicatif des mots employés dans un sens propre à l'économie sociale	s2t2_chapt_17-1.xml
450b	Table des matières dans ce tome deuxième (deuxième série)	s2t2_chapt_17-2.xml

A.1.8 Série 2, vol. 3 [1890-1892] (1892).

URL sur *Internet Archive* :

<https://archive.org/details/s2lesouvriersdes03sociuoft>.

Id	Intitulé	Fichier
451a	Page de titre	s2t3_chapt_1.xml
452a	Page de titre	s2t3_chapt_2.xml
453a	Sommaire des monographies de familles publiées dans ce volume	s2t3_chapt_3.xml

454a	Avertissement sur ce troisième tome de la deuxième série des Ouvriers des deux mondes	s2t3_chapt_4.xml
065a	Métayers en communauté du Confolentais (Charente - France)	s2t3_chapt_5.xml
066a	Vignerons de Ribeaupillé (Alsace)	s2t3_chapt_6.xml
066b	Précis d'une monographie du pêcheur-côtier du Finmark (Laponie - Norvège)	s2t3_chapt_7.xml
066c	Précis d'une monographie d'un tisserand d'Hilversum (Hollande septentrionale - Pays-Bas)	s2t3_chapt_14.xml
067a	Tisserand de la fabrique collective de Gand (Flandre orientale - Belgique)	s2t3_chapt_28.xml
068a	Paysan agriculteur de Torremaggiore (Province de Foggia - Italie)	s2t3_chapt_29.xml
069a	Tanneur de Nottingham (Angleterre)	s2t3_chapt_30.xml
070a	Charpentier indépendant de Paris (Seine - France)	s2t3_chapt_31.xml
071a	Conducteur-typographe de l'agglomération bruxelloise (Brabant - Belgique)	s2t3_chapt_32.xml
072a	Coutelier de la fabrique collective de Gembloux (Province de Namur - Belgique)	s2t3_chapt_33.xml
455a	Tables alphabétique et analytique des matières contenues dans le présent tome, avec index explicatif des mots employés dans un sens propre à l'économie sociale	s2t3_chapt_34.xml
456a	Table des matières dans ce tome troisième (deuxième série)	s2t3_chapt_35.xml

A.1.9 Série 2, vol. 4 [1892-1895] (1895).

URL sur *Internet Archive* :

<https://archive.org/details/s2lesouvriersdes04sociuoft>.

Id	Intitulé	Fichier
457a	Page de titre	s2t4_chapt_1.xml
458a	Page de titre	s2t4_chapt_2.xml
459a	Sommaire des monographies de familles publiées dans ce volume	s2t4_chapt_3.xml

460a	Avertissement sur ce quatrième volume de la deuxième série	s2t4_chapt_4.xml
073a	Ajusteur-surveillant de l'usine de Guise (Aisne - France)	s2t4_chapt_5.xml
074a	Ébéniste parisien de haut luxe (Seine - France)	s2t4_chapt_6.xml
075a	Métayer de l'Ouest du Texas (États-Unis d'Amérique)	s2t4_chapt_7.xml
076a	Ouvrière mouleuse en cartonnage d'une fabrique collective de jouets parisiens (Seine - France)	s2t4_chapt_8.xml
077a	Savetier de Bâle (Suisse)	s2t4_chapt_9.xml
078a	Ouvrier-employé de la fabrique coopérative de papiers d'Angoulême (Charente - France)	s2t4_chapt_10.xml
079a	Tisseur de San Leucio (Province de Caserte - Italie)	s2t4_chapt_11.xml
080a	Fermiers montagnards du Haut-Forez (Loire - France)	s2t4_chapt_12.xml
081a	Allumeur de réverbères de Nancy (Meurthe-et-Moselle - France)	s2t4_chapt_13.xml
461a	Tables alphabétique et analytique des matières contenues dans le présent tome, avec index explicatif des mots employés dans un sens propre à l'économie sociale	s2t4_chapt_14.xml
462a	Table des matières contenues dans ce tome quatrième (deuxième série)	s2t4_chapt_15.xml

A.1.10 Série 2, vol. 5 [1895-1899] (1899).

URL sur *Internet Archive* :

<https://archive.org/details/2serlesouvriersde05sociuoft>.

Id	Intitulé	Fichier
463a	Page de titre	s2t5_chapt_1.xml
464a	Société d'économie sociale [nota : liste des publications]	s2t5_chapt_2.xml
465a	Sommaire des monographies de familles publiées dans ce volume	s2t5_chapt_3.xml
466a	Avertissement sur ce cinquième tome de la deuxième série	s2t5_chapt_4.xml

082a	Ouvrier garnisseur de canons de fusils de la fabrique collective d'armes à feu de Liège (Liège - Belgique)	s2t5_chapt_5.xml
083a	Fileur en peigné et réglleur de métier de la Manufacture du Val-des-Bois (Marne - France)	s2t5_chapt_6.xml
084a	Cordonnier d'Iseghem (Flandre Occidentale - Belgique)	s2t5_chapt_7.xml
085a	Paysan métayer (Contadino mezzajuolo) de Rocca-sancasciano (Romagne Toscane - Italie)	s2t5_chapt_8.xml
085b	Précis d'une monographie d'un ouvrier agriculteur de la campagne de Ravenne (Romagne - Italie)	s2t5_chapt_9.xml
086a	Mineur des mines de houille du Pas-de-Calais (France)	s2t5_chapt_10.xml
087a	Agriculteur du Pas-de-Calais (France)	s2t5_chapt_11.xml
088a	Serrurier-forgeron du quartier de Picpus, à Paris (France)	s2t5_chapt_12.xml
088b	Précis d'une monographie du serrurier poseur de persiennes en fer de Paris	s2t5_chapt_13.xml
089a	Piqueur sociétaire de la Mine aux Mineurs de Montlieux (Loire - France)	s2t5_chapt_14.xml
090a	Petit fonctionnaire de Pnom-Penh (Cambodge)	s2t5_chapt_15.xml
090b	Précis d'une monographie d'un manœuvre-coolie de Pnom-Penh (Cambodge)	s2t5_chapt_16.xml
091a	Métayer de Corrèze (Bas Limousin - France)	s2t5_chapt_17.xml
467a	Tables alphabétique et analytique des matières contenues dans le présent tome, avec index explicatif des mots employés dans un sens propre à l'économie sociale	s2t5_chapt_18-1.xml
467b	Table des matières dans ce tome cinquième	s2t5_chapt_18-2.xml

A.1.11 Série 3, vol. 1 [1900-1904] (1904).

URL sur *Internet Archive* :

<https://archive.org/details/lesouvriersdesde0108sociuoft/>.

Id	Intitulé	Fichier
468a	[Fichier sans texte]	s3t1_chapt_1.xml

472a	La société générale des papeteries du Limousins	s3t1_chapt_2.xml
092a	Fermier normand de Jersey	s3t1_chapt_3.xml
092b	Précis d'une monographie d'un pêcheur-côtier, maître de barques, de l'archipel Chusan (Chine)	s3t1_chapt_4.xml
093a	Aveugle accordeur de pianos de Levallois-Perret (Seine - France)	s3t1_chapt_5.xml
094a	Bouilleur de cru du Bas-Pays de Cognac (Charente - France)	s3t1_chapt_6.xml
095a	Mineur du bassin houiller du Couchant de Mons (Borinage - Belgique)	s3t1_chapt_7.xml
096a	Fellah de Karnak (Haute-Egypte)	s3t1_chapt_8.xml
097a	Tisserand d'usine de Gladbach (Prusse rhénane)	s3t1_chapt_9.xml
098a	Décoreuse de porcelaine de Limoges (Haute-Vienne - France)	s3t1_chapt_10.xml
099a	Cantonnier-poseur de voie du chemin de fer du Nord à Paris	s3t1_chapt_11.xml

A.1.12 Série 3, vol. 2 [1904-1908] (1908).

URL sur *Internet Archive* :

<https://archive.org/details/lesouvriersdesde916sociuoft/>.

Id	Intitulé	Fichier
469a	Page de titre	s3t2_chapt_1.xml
100a	Cordonnier de la fabrique collective de Binche (Province de Hainaut - Belgique)	s3t2_chapt_2.xml
101a	Compositeur typographe de Québec (Canada - Amérique du Nord)	s3t2_chapt_3.xml
102a	Ardoisier du bassin d'Herbeumont (Belgique)	s3t2_chapt_4.xml
103a	Commis à l'administration centrale des chemins de fer de l'État belge (Schaerbeek-Bruxelles - Belgique)	s3t2_chapt_5.xml
104a	Teinturier de ganterie et gantières de Saint-Junien (Haute-Vienne - France)	s3t2_chapt_6.xml
105a	Jardinier-plantier de Gasseras (Commune de Montauban, Tarn-et-Garonne - France)	s3t2_chapt_7.xml
106a	Corsetière du Raincy (banlieue de Paris - France)	s3t2_chapt_8.xml

107a	Étameur sur fer-blanc des usines de Commentry (Allier - France)	s3t2_chapt_9.xml
473a	Usine hydraulique d'éclairage et de transport de force	s3t2_chapt_10.xml

A.1.13 Série 3, vol. 3 [1908-1913] (1913).

URL sur *Internet Archive* :

<https://archive.org/details/lesouvriersdesde17sociuoft>.

Id	Intitulé	Fichier
470a	Page de titre	s3t3_chapt_1.xml
108a	Paysan cultivateur du Ruvo di Puglia (Province de Bari - Italie, 1903)	s3t3_chapt_2.xml

A.1.14 Série 3, vol. 3bis [1928-1930] (1930).

Id	Intitulé	Fichier
471a	Page de titre	s3t3-bis_chapt_1.xml

A.2 Structure logique

- A. *Titre.*
- B. *Observations préliminaires définissant la condition des divers membres de la famille.*
 - I. *Définition du lieu, de l'organisation industrielle et de la famille.*
 - § 1. *État du sol, de l'industrie et de la population.*
 - § 2. *État civil de la famille.*
 - § 3. *Religion et habitudes morales.*
 - § 4. *Hygiène et services de santé.*
 - § 5. *Rang de la famille.*
 - II. *Moyens d'existence de la famille.*
 - § 6. *Propriétés.*
 - § 7. *Subventions.*
 - § 8. *Travaux et industries.*
 - III. *Mode d'existence de la famille.*
 - § 9. *Aliments et repas.*
 - § 10. *Habitation, mobilier et vêtements.*
 - § 11. *Récréations.*
 - IV. *Histoire de la famille.*
 - § 12. *Phases principales de l'existence.*
 - § 13. *Mœurs et institutions assurant le bien-être physique et moral de la famille.*
 - V. (*Budget domestique annuel*²¹).
 - § 14. *Budget des recettes de l'année.*
 - § 15. *Budget des dépenses de l'année.*

Comptes annexés aux budgets (n° 1 à 84) puis § 16. Comptes annexés aux budgets.
- C. *Notes* (n° 1 à 84) puis *Éléments divers de la constitution sociale.*
 - (A) (*titre du paragraphe*) (n° 1 à 84) puis § 17. (*titre du paragraphe*).
 - (B) (*titre du paragraphe*) (n° 1 à 84) puis § 18. (*titre du paragraphe*).
 - etc.*

21. Cette section ne possède un titre que dans huit monographies

A.3 Numérisations de *Google Books*

Les *Ouvriers des deux mondes* sont disponibles en version numérisée sur *Google Books*. Nous listons ci-dessous les volumes en accès libre en fonction de leur lieu de conservation (les URL du domaine `hdl.handle.net` renvoient vers la *HathiTrust Digital Library*). Les volumes sont numérotés ainsi : *série (numéro du volume)*.

A.3.1 Bibliothèque de l'université de Californie

Volume	URL
1 (1)	https://books.google.fr/books?id=eNOWAAAAYAAJ
1 (1)	https://hdl.handle.net/2027/uc1.b4577103
1 (2)	https://books.google.fr/books?id=4GJwAAAAIAAJ
1 (3)	https://hdl.handle.net/2027/uc1.b4577105
1 (4)	https://books.google.fr/books?id=Y2hwAAAAIAAJ
1 (4)	https://hdl.handle.net/2027/uc1.b4577106

A.3.2 Bibliothèque nationale centrale de Florence

Volume	URL
1 (1)	https://books.google.fr/books?id=rqZexB0-3V8C

A.3.3 Bibliothèque de l'université Harvard

Volume	URL
1 (1)	https://hdl.handle.net/2027/hvd.32044079431714
2 (2)	https://hdl.handle.net/2027/hvd.32044018834879
2 (5)	https://hdl.handle.net/2027/hvd.32044100859230

A.3.4 Bibliothèque municipale de la ville de Lyon

Volume	URL
1 (1)	https://books.google.fr/books?id=3r3hsfUlRYoC
1 (2)	https://books.google.fr/books?id=JSHDEpeveFgC
1 (3)	https://books.google.fr/books?id=3GWA_Kz5AW0C

A.3.5 *Bayerische Staatsbibliothek* de Munich

Volume	URL
1 (1)	https://books.google.fr/books?id=6I9LAAAACAAJ
1 (2)	https://books.google.fr/books?id=apBLAAAACAAJ
1 (3)	https://books.google.fr/books?id=5pBLAAAACAAJ
1 (4)	https://books.google.fr/books?id=J5FLAAAACAAJ

A.3.6 *New York State College of Agriculture at Cornell University*

Volume	URL
1 (3)	https://books.google.fr/books?id=10FBAAAAYAAJ

A.3.7 Bibliothèque de l'Université de Princeton

Volume	URL
1 (1)	https://hdl.handle.net/2027/njp.32101064529090
1 (2)	https://hdl.handle.net/2027/njp.32101064529108
1 (3)	https://books.google.fr/books?id=fTooAAAAYAAJ
1 (3)	https://hdl.handle.net/2027/njp.32101064529116
1 (4)	https://hdl.handle.net/2027/njp.32101064529124
1 (5)	https://hdl.handle.net/2027/njp.32101064529132
2 (1)	https://hdl.handle.net/2027/njp.32101064529140
2 (2)	https://hdl.handle.net/2027/njp.32101064529157
2 (4)	https://hdl.handle.net/2027/njp.32101064529173
2 (5)	https://hdl.handle.net/2027/njp.32101064529181
3 (1)	https://hdl.handle.net/2027/njp.32101064529199
3 (2)	https://hdl.handle.net/2027/njp.32101064529207
3 (3)	https://hdl.handle.net/2027/njp.32101064529215

A.3.8 Université de Rome — *Instituto de filosofia del diritto*

Volume	URL
1 (2)	https://books.google.fr/books?id=HjqApJuPx0gC

B. Feuille de route et typologie des erreurs

B.1 Feuille de route

Cette liste reprend le texte des *issues* ouvertes dans le *GitLab* des *Ouvriers des deux mondes* au commencement du stage.

1. *Trier et renommer les fichiers :*
 - Identifier les fichiers qui correspondent aux monographies et ceux qui correspondent au paratexte ;
 - Donner un identifiant aux fichiers de monographie en fonction des identifiants déjà existants dans le fichier de référence ;
 - Créer un identifiant pour les fichiers de paratexte ;
 - Ajouter ces identifiants aux @xml:id de chaque fichier ;
 - Créer un mapping de l'ensemble des fichiers sous la forme d'un CSV avec le nom du fichier et son identifiant ;
 - Créer un fichier master.xml contenant des renvois vers les autres fichiers grâce à des <xi:includes>.
2. *Mettre à jour l'attribut @url dans la balise <graphic> :*
 - Les images des pages sont stockées localement sur Humanum, mais également en ligne sur Internet Archives.
 - Trouver l'url de chaque page de chaque volume sur *Internet Archive* ;
 - Remplacer automatiquement le chemin local par l'url de l'image dans chaque fichier.
3. *Tester la conformité du schéma :*
 - Écrire un script pour tester la validité des arbres XML de chaque fichier ;
 - Corriger les erreurs qui seraient signaler par ce script.
4. *Intégrer les métadonnées des « enquêtés » :*
 - Créer un fichier référentiel de personnes (XML) à partir du fichier CSV de prosopographie ;

- Ajouter aux paragraphes 2 des monographies ("§2. - Etat civil de la famille") des @refs au référentiel de personnes.
5. *Intégrer modèle de citation dans les chapitres et créer un système de référence bibliographique :*
- Pour chaque niveau de la structure d'une monographie ;
 - Pour chaque chapitre (sans descendre dans les niveaux) ;
 - Par exemple en utilisant DTS.
6. *Corrections des transcriptions :*
- Paragraphe par paragraphe, implémenter une correction automatique des transcriptions.
7. *Corriger le passage de source vers split :*
- Corriger le script python de transformation des fichiers sources en une suite de fichiers XML TEI.
 - Éliminer les traces de teiCorpus.
8. *Simplifier l'implémentation de la structure logique :*
- Aplatir la structure des monographies :
 - Est-il vraiment nécessaire d'utiliser les div enchâssées ? Une structure à plat avec des marqueurs signalant le début d'une nouvelle section ne suffirait-elle pas ?
 - Comment gérer l'articulation entre l'arbre principal (un arbre pour l'ensemble des monographies) et les sous-arbre (un sous-arbre par monographie) ? L'ensemble du corpus n'a pas de front/back mais chaque volume a un front et un back et chaque monographie a potentiellement un front et un back.
 - Avec quelles restrictions peut-on créer une structure similaire à celle d'un fichier LATEX avec ses imports ?

B.1 Relevé des erreurs dans la structure logique

Jean-Damien Généro, 8 juillet 2020²².

B.1.1 Déficit dans la transcription

- *Déficit partiel : s1t2_chapt_11 et s1t3_chapt_10 (manque début §7), s1t2_chapt_4 (manquent quatre pages de la note A), s2t3_chapt_14 (manquent vingt lignes).*

22. Reproduction d'une synthèse mise en ligne sur une issue *GitLab*.

- *Déficit majeur* : s1t4_chapt_8, 11, 15 et s1t5_chapt_13 et 14 (seules les notes ont été transcrites, le reste se trouve dans des <figure>), s1t5_chapt_12 (transcription à partir du §14, le reste dans des <figure>).
-

B.1.2 Titres manquants parce que non-imprimés dans les exemplaires d'*Internet Archive*

- s1t1_chapt_12 : manque le titre de la s/section II.3 “Mode d’existence...” ;
 - s2t1_chapt_5 : manque le titre de la s/section II.4 “Histoire de...” ;
 - s2t1_chapt_20 : manque les titres des s/sections II.3 “Mode d’existence...” et II.4 “Histoire de...” (mais ce titre est repris dans l’intitulé du §12 qui suit) ;
 - s2t4_chapt_9 : manque les titres de la s/section II.1 “Définition du lieu...” ;
 - s3t1_chapt_3 : manque le titre des s/sections II.2 “Moyens d’existence...”, II.3 “Mode d’existence...” et II.4 “Histoire de...” (mais ce titre est repris dans l’intitulé du §12 qui suit) ;
 - s2t5_chapt_15 : manque le titre de la s/section II.4 “Histoire de...”.
-

B.1.3 Structure allégée

Titres de s/s/section intégrés au début de paragraphe + certains titres non-utilisés (spec. §16).

- s2t2_chapt_7 (précis) : pas de s/s/section §16 et de titre pour la partie Notes (composée d’une seule *nota*) ;
 - s2t2_chapt_9 (précis) : pas de titre de section II “Observations préliminaires...” et pas de s/s/section §16 ;
 - s2t2_chapt_11 (précis) : pas de titre de section II “Observations préliminaires...”, pas de titre pour les s/s/sections (sauf les paragraphes de budget §14 et §15), pas de s/s/section §16, pas de section Notes.
 - s2t3_chapt_7 (précis) : pas de titre de section II “Observations préliminaires...”, pas de s/s/section §16, les titres des s/s/section sont en début de paragraphe ;
 - s2t5_chapt_9 (précis) : dans la section II, tous les titres des s/s/sections sont en début de paragraphe (à l’exception des paragraphes de budget §14 et §15), il n’y a pas de s/s/section §16 ;
 - s2t5_chapt_16 (précis) : pas de s/s/section dans la s/section II.4 “Histoire de...”, pas de s/s/section §16, la section Notes est occupée par la traduction d’un document.
-

B.1.4 Pas de s/s/section §16 et pas de section *Notes*

- (voir *supra* s2t2_chapt_11) ;
 - s2t1_chapt_6 et s2t5_chapt_12.
-

B.1.5 Remarques et cas particuliers

- Dans s1t5_chapt_9, 10 et 11, une section des notes est intitulée “Précis de monographie” et organisée comme un précis (càd avec des sections et des titres). Néanmoins il ne s’agit pas de précis tels qu’on en trouve dans les séries 2 et 3, puisqu’ils ne constituent pas un ensemble logique indépendant de la monographie (il y a des paragraphes avant et des paragraphes après). Donc le “précis” est considéré comme une sub_sub_section et ses divisions comme des sub_sub_sub_section.
- s2t5_chapt_17 et 18 découpés chacun en deux fichiers (table analytique et table des matières).
- s3t1_chapt_2 (“Société générale...”) et s3t2_chapt_10 (“Usine hydraulique...”) sont des cas particuliers avec des structures totalement différentes des autres monographies. En plus de l’en-tête, le premier se divise en trois grandes sections : observations préliminaires (sorte d’intro), première partie, deuxième partie, troisième partie, conclusion et appendices ; le second est plus succinct avec l’en-tête, une grande section avec des s/sections numérotées et ensuite des appendices.
- Dans s3t2_chapt_9 se trouve une partie “Note sur l’état de la famille en 1905” entre le §13 et §14, que j’ai structurée comme une s/s/section.
- La s/section II.5, consacrée aux budgets (§14, §15, §16), n’est pas titrée sauf dans les précis de monographie s2t1_chapt_6, s2t2_chapt_9 et 11, s2t3_chapt_7 et 14, s2t5_chapt_9 et 13, s3t1_chapt_4 (le titre est “Budget domestique annuel”).

Table des figures

1.1	Comparaison des exemplaires de Toronto et de Paris	7
1.2	Tableau des sept situations principales que les ouvriers peuvent occuper successivement	10
2.1	Images exclues du lot à transcrire	17
2.2	Exemple d'une binarisation	18
3.1	Exemples d'objets graphiques	24
3.2	Schématisation des étapes suivies pour l'extraction du texte et des informations de mise en page	26
3.3	Exemple d'un entraînement opéré par <i>Kraken</i>	27
3.4	Représentation de l'arbre XML des fichiers TEI	30
3.5	Exemple d'un ensemble <facsimile>	31
3.6	Exemple de la structuration du début de la monographie n° 70	32
4.1	Exemple d'un contrôle de code par <i>Pylint</i>	39
4.2	<i>Mattermost</i>	40
4.3	Messages dans une <i>issue GitLab</i>	41
5.1	Exemple d'un déficit de transcription	48
5.2	Segmentation défectueuse dans <i>Transkribus</i>	49
5.3	Premières lignes du fichier <code>master.xml</code>	51
5.4	Encodage du début du §14 de la monographie 56.	52
6.1	Proposition pour un encodage des paragraphes.	58
6.2	Suppression d'un numéro de page	60
6.3	Suppression d'un numéro de cahier (1)	60
6.4	Suppression d'un numéro de cahier (2)	60
6.5	Schématisation du processus de validation à l'aide d'une ODD.	61
6.6	Extrait de l'ODD concernant la balise <div>	62
7.1	Figures dans la monographie n° 90.	69

8.1	Arbre XML simplifié montrant l'organisation d'une entrée d'index.	74
8.2	Liste d'individus au début du paragraphe 2 (n° 23)	77
8.3	Liste d'individus au début du paragraphe 2 (n° 56)	78
8.4	Encodage de la liste d'individus (n° 56)	78
9.1	Représentation de la section <msDesc> du <teiHeader>.	87

Table des matières

Résumé	iii
Liste des sigles et abréviations	v
Introduction	ix
Bibliographie	xix
Bibliographie générale	xxii
<i>Les Ouvriers des deux mondes</i>	xxv
I Un corpus déjà structuré	1
1 Un corpus d'imprimés	5
1.1 Une longue publication	5
1.2 La structure logique des monographies	8
1.3 Continuité et discontinuité	10
2 Des numérisations multiples	15
2.1 Les volumes de la Bibliothèque nationale de France	15
2.2 Les volumes d' <i>Internet Archive</i>	16
2.3 Choix du corpus à numériser	17
3 Un encodage automatique	21
3.1 Choix de l'automatisation	21
3.2 Le script LSE-OD2M	22
3.2.1 Segmentation	22
3.2.2 Reconnaissance des caractères	25
3.2.3 Structuration	27
3.3 Les fichiers XML-TEI	28

II Une structuration à reprendre	35
4 Outils, méthodologie et gestion du projet	37
4.1 Outils de développement	37
4.1.1 <i>GitLab</i>	37
4.1.2 <i>PyCharm</i> et <i>Oxygen</i>	38
4.2 Espaces de discussion	39
4.2.1 <i>Mattermost</i>	39
4.2.2 <i>Issues</i> et <i>merge requests</i> sur <i>GitLab</i>	40
4.3 Feuille de route	41
4.3.1 Les missions du stage	41
4.3.2 Une gestion de projet ?	42
5 Contrôle du découpage des fichiers source	45
5.1 Les différents niveaux d'encodage	45
5.2 Vérification de la cohérence documentaire	46
5.2.1 Fission horizontale lors de la segmentation	46
5.2.2 Défaut de transcription	47
5.2.3 Identification, cartographie et inclusion	50
5.3 Vérification de l'encodage minimal	52
6 Découpages éditorial et sémantique	55
6.1 Découpage éditorial	55
6.1.1 Niveau des titres (<i><head></i>)	55
6.1.2 Niveau des divisions (<i><div></i>)	56
6.1.3 <i>Tabula rasa</i>	57
6.2 Découpage sémantique	59
6.3 Validation du schéma XML-TEI	59
III Des données à valoriser	65
7 Données graphiques, données chiffrées	67
7.1 Le lien entre le texte et les images du texte	67
7.2 Les données graphiques dans le flux textuel	68
7.3 Les tableaux : images ou données ?	70
8 Données textuelles, données scientifiques	73
8.1 La qualité de l'OCR	73
8.2 Indexer les individus	75
8.3 Corriger les transcriptions ?	77

TABLE DES MATIÈRES	123
9 Une valorisation plurielle	83
9.1 Édition papier, édition numérique	83
9.2 Retrouver les fascicules dans le volume	85
9.3 <i>Quid de la donnée ?</i>	88
Conclusion	93
Annexes	97
A. <i>Les Ouvriers des deux mondes</i>	97
A.1 Liste des monographies et fichiers correspondant	97
A.1.1 Série 1, vol. 1 (1857)	97
A.1.2 Série 1, vol. 2 (1858)	98
A.1.3 Série 1, vol. 3 (1861)	99
A.1.4 Série 1, vol. 4 (1862)	100
A.1.5 Série 1, vol. 5 [1875, 1883, 1884] (1885)	101
A.1.6 Série 2, vol. 1 [1885-1887] (1887)	102
A.1.7 Série 2, vol. 2 [1887-1889] (1890)	103
A.1.8 Série 2, vol. 3 [1890-1892] (1892)	104
A.1.9 Série 2, vol. 4 [1892-1895] (1895)	105
A.1.10 Série 2, vol. 5 [1895-1899] (1899)	106
A.1.11 Série 3, vol. 1 [1900-1904] (1904)	107
A.1.12 Série 3, vol. 2 [1904-1908] (1908)	108
A.1.13 Série 3, vol. 3 [1908-1913] (1913)	109
A.1.14 Série 3, vol. 3bis [1928-1930] (1930)	109
A.2 Structure logique	110
A.3 Numérisations de <i>Google Books</i>	111
A.3.1 Bibliothèque de l'université de Californie	111
A.3.2 Bibliothèque nationale centrale de Florence	111
A.3.3 Bibliothèque de l'université Harvard	111
A.3.4 Bibliothèque municipale de la ville de Lyon	111
A.3.5 <i>Bayerische Staatsbibliothek</i> de Munich	111
A.3.6 <i>New York State College of Agriculture at Cornell University</i>	112
A.3.7 Bibliothèque de l'Université de Princeton	112
A.3.8 Université de Rome — <i>Instituto de filosofia del diritto</i>	112
B. Feuille de route et typologie des erreurs	113
B.1 Feuille de route	113
B.1 Relevé des erreurs dans la structure logique	114

B.1.1	Déficit dans la transcription	114
B.1.2	Titres non-imprimés	115
B.1.3	Structure allégée	115
B.1.4	Pas de s/s/section §16 et pas de section <i>Notes</i>	116
B.1.5	Remarques et cas particuliers	116
Table des figures		117
Table des matières		121

