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
— 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)
— READ : *Recognitionand Enrichment of Archival Documents*
— TGIR : Très grande infrastructure de recherche
— TELEMMe : Temps, Espaces, Langages, Europe Méridionale-Méditerranée (Université d'Aix-Marseille)
— UMR : Unité mixte de recherche

—

Informatique et nouvelles technologies

— ALTO : *Analyzed Layout and Text Object*
— CSV : *Comma-separated values*
— GPU : *Graphics processing unit*
— IETF : *Internet Engineering Task Force*
— JPEG : *Joint Photographic Experts Group*
— JP2 : *JPEG 2000*
— LSE-OD2M : *Logical Structure Extraction from Les Ouvriers des Deux Mondes*
— OCR : *Optical Character Recognition*
— PDF : *Portable Document Format*

— RFC : *Request for comments*
— TEI : *Text Encoding Initiative*
— URI : *Uniform Resource Identifier*
— XML : *Extensible Markup 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 XVIIe siècle au début du XXe 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 XIXe et XXe 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 Scientifique [3], d'une équipe d'accueil (EA) [4] de l'Université de Paris [5] et de l'équipe projet ALMAnaCH [6] de l'Institut national de recherche en informatique et en automatique (Inria [7] Paris). 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, ingénieure de recherche et de développement au sein de l'équipe ALMAnaCH.

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 proces-

1. Présentation du programme sur le site de l'ANR (`https://anr.fr/Projet-ANR-16-CE26-0018`, consulté le 7 septembre 2020).

2. Présentation du programme sur le site du LARHA (`http://larhra.ish-lyon.cnrs.fr/anr-time-us`, consulté le 7 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 formée par 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/charte-logo-inria/charte`, consulté le 7 septembre 2020).

sus 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 femmes [9]. 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-Drouet [12].

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 Lyon [13]. 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 relayés 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 restant l'étant dans les années 1840 (*L'Écho des ouvriers*, *L'Écho de la Fabrique de 1841*, *L'Écho de l'industrie*, *L'Avenir*, *La Tribune Lyonnaise*) [14].

8. Alix Chagué, Manuela Martini, Victoria Le Fourner 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 municipales 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 7 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 7 septembre 2020).

13. Le corpus se trouve à l'adresse `https://collections.bm-lyon.fr/PER003` (consulté le 7 septembre 2020).

14. Présentation de ce corpus sur le wiki du programme : `http://timeusage.paris.inria.fr/mediawiki/index.php/Documentation_régionale_-_Presse_lyonnaise` (consulté le 7 septembre 2020).

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 langues 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*.

⋆

Initiées 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 sociale [16] de 1857 à 1928. Répartie en trois séries comptant un total de cent vingt-six monographies [17], 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 1855 [18]. Les enquêtes sont couramment désignées sous le vocable de « monographies de famille » 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 nombreux [19]. *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 de part et d'autre de la Méditerranée, et parfois plus loin encore. Se succèdent ainsi une

15. Liste complète sur le wiki du programme : `http://timeusage.paris.inria.fr/mediawiki/index.php/Aperçu_des_états#Imprimés_divers` (consulté le 7 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*–n° 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*–n° 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%A9l%C3%A9ments_d'%C3%A9pist%C3%A9mologie_et_de_science_sociale` (visité le 11/08/2020), p. 193.

enquête consacrée à un charpentier de Paris [21], à un métayer de la banlieue de Florence [22] et à un menuisier-charpentier de Tanger [23].

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 ce programme, d'autant que les monographies se focalisent sur le budget et son usage [24]. 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 domestiques [25] et parfois les esclaves [26]. 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 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

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*–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. Alain Chenu, *op. cit.*, p. 5.

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 Fourner (prom. 2019) y ont effectué leur stage de fin d'étude. Alix Chagué a travaillé sur la collecte et le traitement de la documentation [29], tandis que Victoria Le Fournier s'est intéressée à la structuration de cette documentation en prenant pour exemple la juridiction prud'homale [30]. Leurs mémoires correspondent à deux phases successives du projet, et la présente étude entend décrire sa phase finale correspondant à la 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 Fourner 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 compte rendus d'audience publiés dans la presse. Le présent mémoire ne cherche ni à replacer les *Ouvriers des deux mondes* au sein de la documentation globale du programme, ni à décrire des opérations de traitement et de mise en forme de ce corpus.

En effet, 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 prédécesseures — 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 à mener 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 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

28. Antoine Savoye, « Éditorial », *Les Études sociales, Les monographies de famille de l'École de Le Play*–n° 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*, mémoire du Master « Technologies numériques appliquées à l'histoire », École nationale des chartes, 2018, URL : `https://github.com/alix-tz/M2TNAH-memoireDeStage`.

30. V. Le Fourner, *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* XIXe *siècle du projet Time Us*, mémoire du Master « Technologies numériques appliquées à l'histoire », École nationale des chartes, 2019, URL : `https://github.com/Victorialf/M2TNAH-memoireDeStage`.

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 du corpus à celui, plus fin, des fichiers.

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 IIIF [31]). 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 machine et à l'humain de les utiliser [32].

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

32. Ce mémoire est entièrement rédigé avec le langage LaTeX. Les fichiers qui le constituent sont disponibles sur le GitHub de l'auteur, consultables et téléchargeables à cette adresse : `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%A9l%C3%A9ments_d'%C3%A9pist%C3%A9mologie_et_de_science_sociale (visité le 11/08/2020).

— « 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%A9l%C3%A9ments_d'%C3%A9pist%C3%A9mologie_et_de_science_sociale (visité le 11/08/2020).

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 (visité le 24/08/2020).

— *Archiving the Internet*, 1996, URL : http://web.archive.org/web/19971011050140/http://www.archive.org/sciam_article.html (visité le 24/08/2020).

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/ntn.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*–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*, mémoire du Master « Technologies numériques appliquées à l'histoire »,

École nationale des chartes, 2018, URL : https://github.com/alix-tz/M2TNAH-memoireDeStage.

Chagué (Alix), Martini (Manuela), Le Fourner (Victoria) 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 ?*, DH Nord 2019, 2020.

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

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.

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/bptl-driver.html.

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-ᶜUsur al-Wusta : The Journal of Middle East Medievalists*–n° 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* XIXᵉ *siècle du projet Time Us*, 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*–n° 131-132 (2000), p. 93-181.

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 (visité le 24/08/2020).

Savoye (Antoine), « Éditorial », *Les Études sociales, Les monographies de famille de l'École de Le Play*–n° 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*–n° 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.

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), « 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.

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.

REVIERS 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.

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'autre version 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 possible [2].

Une tentative de reprise avait déjà été effectuée en 1875 où « un premier fascicule composé de trois monographies » avait paru [3]. Deux autres paraissent en 1883 et 1884, et les trois sont finalement rassemblés en 1885 dans le cinquième volume de la première série [4]. 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 volumes [5]. 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 matières.

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*[6]. »

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 lui[7], 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*[8]. »

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'homme[9], la seconde se trouve à l'Université de Toronto et a été numérisée par *Internet Archive*[10].

Des différences de composition existent. Ainsi, l'exemplaire de Toronto (fig. 1.1a, 1.1b, 1.1c) comporte deux feuillets qui ne sont pas présents dans celui de Paris (fig. 1.1d). Le premier porte au recto le titre des *Ouvriers des deux mondes* et au verso l'*Avertissement* que nous avons cité, le second 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.

Cet exemple illustre le fait que laisser aux acheteurs le soin de relier les fascicules pour composer le volume final revient à rompre l'unité codicologique de ce dernier, chaque nouvelle opération de reliure donnant naissance à un nouvel exemplaire. Les relieurs ont

6. 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.

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

8. « 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.

9. Consultable sur *Gallica* : `https://gallica.bnf.fr/ark:/12148/bpt6k54465138` (consulté le 7 septembre 2020).

10. Consultable à cette adresse : `https://archive.org/details/s2lesouvriersdes01sociuoft` (consulté le 7 septembre 2020).

(a) (b)

(c) (d)

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.

en effet pu choisir de conserver ou de ne pas conserver les pages propres aux fascicules, 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 XIXe siècle une « étude scientifique, minutieuse et détaillée, portant sur un objet ou un phénomène circonscrit » [11]. La démarche monographique cherche à embrasser à travers l'observation directe des phénomènes inatteignables pour la démarche statistique, à laquelle elle s'oppose [12].

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 (1re éd.) puis cinquante-sept (2e éd.) monographies issues d'enquêtes menées entre 1833 et 1853 [13]. 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 européens* : *Instruction sur la méthode d'observation dite des monographies de familles, propre à l'ouvrage intitulé* Les ouvriers européens [14].

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 ouvrière [15], érigée en « clef de compréhension des organisations sociales » [16]. La phase d'observation est ensuite divisée en étapes [17], savoir l'observation des faits, l'interrogatoire de l'ouvrier et celui des personnes tiers [18].

La troisième partie est celle où Le Play expose la structure logique qu'il utilise à

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

12. 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.

13. 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.

14. 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/lesouvriersdesde04sociuoft/page/14/`.

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

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

17. 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.

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

partir de 1855 et que ses continuateurs 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*[19].

Le miroir est cependant légèrement déformant, afin de satisfaire au « hiatus » propre à la démarche leplaysienne, qui différencie les « faits observés » des « faits décris »[20]. 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 »[21]. 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 recourt au deuxième niveau et ne contient que des paragraphes.

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

19. 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.

20. 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%A9l%C3%A9ments_d'%C3%A9pist%C3%A9mologie_et_de_science_sociale` (visité le 11/08/2020), p. 87.

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

occupe dans l'organisation sociale caractérisée par cet engagement » [22]. 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 monographes (fig. 1.2 [23]).

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*.

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* [24].

Ce même volume est l'occasion pour la SESS d'agréger aux *Ouvriers des deux mondes* un type de monographie déjà introduite dans la ré-édition des *Ouvriers euroépens* (1877-1879), le précis [25]. Elle le définit comme un document qui, « n'ayant pas la forme complète

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

23. *Ibid.*, p. 21 .

24. 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`.

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

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 » [26]. 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 » [27].

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 réduit 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 section [28]. 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 six [29]).

Les libellés des titres peuvent également être modifiés. Celui de la deuxième partie (*Observations préliminaires*) est absent des précis n° 58 bis [30] et 66 bis [31]. Ceux des paragraphes sont également placés en italique au début du texte dans quatre précis (n° 58 bis, 66 bis, 85 bis [32], 90 bis [33]).

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° 41 [34], 42 [35] et 43 [36]) mais sont traitées comme une section de la partie *Notes*. Il n'y a ni grandes parties

26. « 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.

27. A. Chenu, « Préface »..., p. 7.

28. 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/lesouvriersdesde0108sociuoft/page/n127`, p. 78-81.

29. 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.

30. 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.

31. 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.

32. 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`.

33. 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`.

34. Urbain 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/lesouvriersdesde05sociuoft/page/n165`, p. 188-196.

35. 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/lesouvriersdesde05sociuoft/page/n225`, 246-258.

36. 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/`

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 Limousins (1904 [37]) et la seconde à une usine hydraulique d'éclairage et de transport de force de la Loire (1908 [38]). La monographie d'atelier est théorisée par Émile Cheysson, membre de la SESS, pour qui le « foyer cesse d'être le centre unique de notre activité et cède à l'atelier une partie de ses attributions primitives » [39].

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 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 (titre, développement et notes renommées *Appendices*), un second niveau signalé par une numérotation en romain et des paragraphes qui peuvent être titrés ou non.

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.

`details/lesouvriersdesde05sociuoft/page/n289`, p. 300-305.

37. 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/lesouvriersdesde0108sociuoft/page/n9`.

38. 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/lesouvriersdesde0108sociuoft/page/n9`.

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

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 d'accéder 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 2007 [1].

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 7 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 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 2001 [3] 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 serveurs [5].

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 2005 [6]. La digitilasation est faite page par page au moyen de scanners [7]. Les livres ainsi numérisés sont

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` (visité le 24/08/2020) (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` (visité le 24/08/2020), 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` (visité le 24/08/2020), 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.*

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 *Google Books*[8]. Par rapport à *Gallica*, *Internet Archive* offre beaucoup plus de formats pour le téléchargement de ses fichiers[9].

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 maximale[10]. 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 Go[11].

L'ensemble a été versé sur un espace de stockage créé grâce au service *ShareDocs*de la TGIR Huma-Num. Lancé 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

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*, juin 2020, URL : `https://timeus.hypotheses.org/626`.
11. *Ibid.*
12. *Ibid.*, p. 33-34.

ShareDocs permet également de centraliser l'ensemble des fichiers image du programme *Time Us*.

Choisir le corpus d'*Internet Archive* n'est pas sans conséquence, notamment en raison de son 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 XXe 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 **??**), 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 donc intéressant directement le programme *Time Us* sont bel et bien 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 imperfections[13]. 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).

N° 47.

PAYSAN PALUDIER

DU BOURG DE BATZ (LOIRE-INFÉRIEURE),

TENANCIER-PROPRIÉTAIRE,

DANS LE SYSTÈME DES ENGAGEMENTS VOLONTAIRES PERMANENTS,

d'après les

RENSEIGNEMENTS RECUEILLIS SUR LES LIEUX EN 1883,

PAR

M. A. DELAIRE.

OBSERVATIONS PRÉLIMINAIRES

DÉFINISSANT LA CONDITION DES DIVERS MEMBRES DE LA FAMILLE.

DÉFINITION DU LIEU, DE L'ORGANISATION INDUSTRIELLE ET DE LA FAMILLE.

§ 1.

ÉTAT DU SOL, DE L'INDUSTRIE ET DE LA POPULATION.

La commune de Batz, qu'habite la famille, est située à l'embouchure de la Loire, à 27 kilomètres O. de Saint-Nazaire, par 4° 42' long. O. et 47° 17' lat. N.

Toute la région de la basse Loire offre, au sud du fleuve et au nord jusqu'à Pont-Château, des collines peu élevées, disposées sans ordre apparent, séparées par des vallées étroites et peu profondes où coulent de faibles ruisseaux. C'est le Bocage : il forme environ les deux tiers du département; des roches cristallines, gneiss et granites, en constituent le sol, et partout la vigne y est cultivée. Au delà, vers le nord, l'aspect change tout à coup; aux molles ondulations succèdent de grandes lignes courant de l'est à l'ouest : ce sont les terrains anciens, cambriens et siluriens, fortement plissés et redressés; la vigne dispa-

N° 47.

PAYSAN PALUDIER

DU BOURG DE BATZ (LOIRE-INFÉRIEURE),

TENANCIER-PROPRIÉTAIRE,

DANS LE SYSTÈME DES ENGAGEMENTS VOLONTAIRES PERMANENTS,

d'après les

RENSEIGNEMENTS RECUEILLIS SUR LES LIEUX EN 1883,

PAR

M. A. DELAIRE.

OBSERVATIONS PRÉLIMINAIRES

DÉFINISSANT LA CONDITION DES DIVERS MEMBRES DE LA FAMILLE.

DÉFINITION DU LIEU, DE L'ORGANISATION INDUSTRIELLE ET DE LA FAMILLE.

§ 1.

ÉTAT DU SOL, DE L'INDUSTRIE ET DE LA POPULATION.

La commune de Batz, qu'habite la famille, est située à l'embouchure de la Loire, à 27 kilomètres O. de Saint-Nazaire, par 4° 42' long. O. et 47° 17' lat. N.

Toute la région de la basse Loire offre, au sud du fleuve et au nord jusqu'à Pont-Château, des collines peu élevées, disposées sans ordre apparent, séparées par des vallées étroites et peu profondes où coulent de faibles ruisseaux. C'est le Bocage : il forme environ les deux tiers du département; des roches cristallines, gneiss et granites, en constituent le sol, et partout la vigne y est cultivée. Au delà, vers le nord, l'aspect change tout à coup; aux molles ondulations succèdent de grandes lignes courant de l'est à l'ouest : ce sont les terrains anciens, cambriens et siluriens, fortement plissés et redressés; la vigne dispa-

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

13. *Ibid.*

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 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 ALMAnaCH 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.

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 mois [2].

ALMAnaCH 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 programme [3]. 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'ALMAnaCH dès le début du stage. Une réflexion y est systématiquement menée pour étudier l'opportunité d'automatiser les tâches dans un objectif évident de gain de temps, mais également de développement expérimental de nouvelles techniques de traitement des fichiers.

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.*

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 missions :

— Réaliser un benchmark de toutes les solutions d'OCRdisponibles 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 rendu [4].

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.

De 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 de 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 caractères [5]. Il s'agit d'un moment crucial pour l'OCR, puisque la reconnaissance des caractères s'effectuera sur les

4. 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.
5. *Ibid.*, p. 3.

zones ainsi segmentées [6]. Il est identifié depuis de nombreuses années [7] comme une source majeure des erreurs présentes dans le résultat final [8].

Plusieurs caractéristiques des *Ouvriers des deux mondes* facilitent leur transcription. D'une part, il s'agit d'un imprimé du XIXe 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 survenir [9].

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 de la troisième série, 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.

Les performances du logiciel d'OCR dans la restitution de la mise en page des tableaux ne sont pas optimales [10]. Inria a donc pris la décision de seulement détecter leur présence dans une page afin de pouvoir situer leur emplacement et d'ajouter à cet endroit un lien pointant vers l'image de la page. Il s'agit d'une solution d'attente en prévision d'un traitement spécifique de ces tableaux, dont l'intérêt premier est de les localiser dans le flux du texte pour savoir où l'intervention future aura lieu.

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

6. 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.

7. *Ibid.*, p. 4.

8. 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.

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

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

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

12. *Ibid.*

13. Présentation du logiciel : `https://readcoop.eu/transkribus/` (consulté le 7 septembre 2020).

4 N° 47. — PAYSAN PALUDIER DU BOURG DE BATZ.

qu'un rez-de-chaussée, 238 ont un étage et 2 en ont deux. Le tableau suivant montre comment elle se répartit actuellement entre les diverses professions.

POPULATION RÉPARTIE PAR PROFESSIONS.

PROFESSIONS.	HOMMES.	FEMMES.	TOTAL.
Paludiers, carriers, maçons, menuisiers	654	787	1441
Sauniers, marchands de poisson	279	302	581
Marins	190	131	321
Attachés au chemin de fer	18	14	32
Commerçants	15	55	70
Hôteliers, cabaretiers	13	45	58
Agriculteurs	10	29	30
Grands industriels, rentiers et professions libérales	65	108	173
Totaux	1264	1462	2726

Comme on le voit, le travail du marais occupe la moitié à peu près de la population. Les frais et les produits sont partagés, suivant d'anciens usages, entre les propriétaires des salines et les paludiers qui les exploitent. Ceux-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 de Bayonne à Dunkerque; aujourd'hui Bordeaux leur est fermé, et c'est à peine s'ils dépassent quelquefois Saint-Malo pour atteindre jusqu'à Dieppe. L'extension des chemins de fer et le grand développement des salins 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 pu 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 lesquels le bourg de Batz compte pour 367 hectares, il y a environ deux mille propriétaires dont huit cents n'ont pas un demi-hectare. Cette division est un fait récent : il y a cent ans, quelques grandes familles possédaient le marais et tenaient à honneur de conserver cette propriété. Les Semaisons en avaient à eux seuls le quart. Les partages successoraux, divisant et redivisant ces vieux héritages, en ont

(a) s. 2 vol. 1, n° 47, p. 4.

ÉLÉMENTS DIVERS DE LA CONSTITUTION SOCIALE. 173

ENTREPRISES EN ACTIVITÉ LE 31 OCTOBRE 1896

PROVINCES	EXPLOITATIONS PRINCIPALES	COMMUNES par ARRONDISSEMENT ADMINISTRATIF	TOTAL DES ENGINS A VAPEUR	TOTAL DES ENTREPRISES	NOMBRE D'ENTREPRISES DE MINES	NOMBRE DE PERSONNES OCCUPÉES AU SIÈGE DES ENTREPRISES: PERSONNEL OUVRIER (Fond et surface réunis)			PERSONNEL SUPÉRIEUR (Y compris les membres du conseil d'administration, ingénieurs, employés, etc.)			ENSEMBLE DU PERSONNEL			NOMBRE DE CHEVAUX-VAPEUR PAR ENTREPRISE
						Hommes.	Femmes.	Total.	Hommes.	Femmes.	Total.	Hommes.	Femmes.	Total.	
Liége.	Carrières souterraines (Fond et surface).	*Arrondiss. de Verviers.*	—	—	—	—	—	—	—	—	—	—	—	—	—
		Liége.	—	—	—	—	—	—	—	—	—	—	—	—	—
		La province.	—	—	—	—	—	—	—	—	—	—	—	—	—
Luxembourg.	Carrières souterraines (Fond et surface réunis).	*Arrondiss. de Bastogne.*	14	14	17	23	—	23	176	—	176	199	—	199	1
		Virton.	04	04	1	—	—	—	—	—	—	—	—	—	—
		Arrondiss. de Neufchâteau.													
		Bertrix.	04	04	1	22	—	22	176	—	176	199	—	199	1
		Graide-Station.	04	04	1	—	—	—	—	—	—	—	—	—	—
		Neufchâteau.	04	04	1	—	—	—	—	—	—	—	—	—	—
		Rochefort.	04	04	1	—	—	—	—	—	—	—	—	—	—
		Ville-Pommeroeul.	04	04	1	—	—	—	—	—	—	—	—	—	—
		Saint-Médard.	04	04	1	—	—	—	—	—	—	—	—	—	—
		La province.	14	14	22	23	—	23	176	—	176	199	—	199	1
Namur.	Carrières souterraines (Fond et surface réunis).	*Arrondiss. de Dinant.*	04	04	1	—	—	—	—	—	—	—	—	—	—
		Ath.	04	04	1	—	—	—	—	—	—	—	—	—	—
		La province.	04	04	1	—	—	—	—	—	—	—	—	—	—
Belgique.	Carrières souterraines (Fond et surface réunis).	Province de Liége.	04	04	1	6	—	6	114	—	114	120	—	120	1
		Prov. de Luxembourg.	04	04	1	—	—	—	—	—	—	—	—	—	—
		Province de Namur.	04	04	1	—	—	—	—	—	—	—	—	—	—
		La Belgique.	18	18	25	29	—	29	290	—	290	319	—	319	2
		Prov. de Liége.	15	15	21	48	6	54	114	3	117	162	9	171	2
		Prov. de Luxembourg.	14	14	22	23	—	23	176	—	176	199	—	199	1
		Prov. de Namur.	04	04	1	6	—	6	114	3	117	120	3	123	1
		La Belgique.	33	33	44	77	6	83	404	3	407	481	9	490	4
		Prov. de Liége.	15	15	21	54	6	60	404	—	404	459	6	465	3
		Prov. de Luxembourg.	14	14	22	23	—	23	176	—	176	199	—	199	1
		Prov. de Namur.	04	04	1	6	—	6	114	3	117	120	3	123	1
		La Belgique.	33	33	44	77	6	83	694	3	697	778	9	787	5
		La Belgique.	19	19	25	56	2	58	974	3	977	1,033	5	1,038	177

N. B. — La ligne ‘ figurant aux tableaux A et B indique que pour éviter des doubles emplois, un certain nombre de personnes ou de chevaux-vapeur compté à une ligne ne doivent pas être comptés à une autre ligne.

N. B. Les lignes 1 à 4 du tableau A ne contiennent que des entreprises présentant certaines des catégories distinctes dans le tableau B, de façon à montrer le nombre total de personnes et de chevaux-vapeur employés par les entreprises qui comportent plusieurs délégations.

(b) s. 3 vol. 1, n° 102, p. 173.

OBSERVATIONS PRÉLIMINAIRES. 3

réunirent les négociateurs du traité de 1365, « pour ce que, dit le chroniqueur, on était lors en caresme et qu'on y trouvait du poisson frais ».

A l'ouest, c'est le port du Croisic, qui, au quinzième siècle, avait 4.000 habitants, deux foires par an et trois cents navires tant pour le cabotage que pour la grande pêche.

Le Croisic
Le Croisic
Le Croisic
Batz
Guérande
Mesquer
Piriac
Pornichet
Saint-Nazaire
Le Pouliguen
Le Croisic
Le Croisic
Le Croisic

la forme d'un fichier XML ALTO[14]. 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 fichiers ALTO contenaient « 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. Il s'agit, pour le logiciel, de détecter le premier caractère de la zone segmentée, d'en reconnaître les signes distinctifs et, à partir de ceux-ci, de l'associer au symbole du modèle fourni avec qui il partage le plus de points communs[17]. 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.

Cette étape nécessite donc 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.

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 Kiessling[18] 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

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 7 septembre 2020).

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 ALMAnaCH 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*

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`).

tous les logiciels similaires tels que *Transkribus* : il peut analyser un nombre important d'écritures de toute époque, latines et même arabes [21].

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

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-ᶜUsur al-Wusta : The Journal of Middle East Medievalists*–n° 25 (2017), p. 1-13, DOI : `10.17613/M6TZ4R`.

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

23. *Ibid.*

```
$ ketos train output_dir/*.png
Building training set  [####################################]  100%
Building validation set  [##################################]  100%
[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 7 septembre 2020).

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 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 marque le début de celui en cours et la fin du précédent.
2. Les transcription 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-deux 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 prédéfini [24].

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 2000 [26].

24. Définition d'Ariane Pinche : `https://github.com/ArianePinche/coursTNAH_XML-TEI/blob/master/seance01/InitiationXML.md` (consulté le 7 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.*

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.

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 mentionnée [29].

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 dernière contiennent les coordonnées des blocs de segmentation à l'intérieur du bloc supérieur définit 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

27. *Ibid.*, p. 10.

28. « Les *Ouvriers des deux mondes* sont des recueils d'enquêtes sociologiques publiées pendant la deuxième moitié du XIXe 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 7 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.

LSE-OD2M n'a pas donné de résultats optimals et a fait l'objet d'une reprise majeure lors du stage. Ces résultats restent néanmoins satisfaisant 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 enchâssée 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 monographie grâce à une imbrication de `<div>` jusqu'à arriver au texte lui-même [31]. 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 LaTeX [32].

La monographie correspond à une `<div>` de type `chapter`. Les ensemble 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

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 7 septembre 2020).

```
<div n="001" type="section">
 <div n="001" type="sub_section">
  <div n="001" type="sub_sub_section">   A
   <head facs="#facs_350_p_8" type="section" xml:id="para_31_325_8">
    OBSERVATIONS PRÉLIMINAIRES
   </head>
   <p facs="#facs_350_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">   B
   <head facs="#facs_350_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="#facs_350_p_11" xml:id="para_31_325_11">
    § 1
   </p>
  </div>   B
  <div n="002" type="sub_sub_section">
   <head facs="#facs_350_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="#facs_350_p_13" xml:id="para_31_325_13">
    La famille habite à Paris, sur la rive gauche de la Seine, vers la li¬mite du faubourg Saint·
   </p>
   <pb facs="#facs_351" 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.

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).

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 : des tâches manuelles et semi-automatiques

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 dossiers, dont un est réservé aux *Ouvriers des deux mondes*. 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 `!\d`[1]), écrire leurs numéros dans *GitLab* ou dans le commentaire de modification d'un *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 de travail 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 mise 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 (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

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

```
### ----- 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 vérification du code, qui là encore peut être appliqué à tout le corpus à travers l'outil « projet ».

PyCharm est également 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 2015 [3]. Auto-hébergé — Inria

2. Consultable sur *GitHub* (`https://github.com/mattermost/mattermost-server`, consulté le 7 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 7 septembre

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 aux 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-

2020).

4. *GitLab* use de sa propre version du *Markdown*, le *GitLab Flavored Markdown* : `https://docs.gitlab.com/ee/user/markdown.html` (consulté le 7 septembre 2020).

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, chercheur 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 blog Hypothèses de *Time Us* des billets rendant compte de notre travail [5].

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, label, temps estimé, échéance). Un système de labels — des étiquettes thématiques — permet également 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, 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 peuvent l'expliquer. En premier lieu, *Time Us* s'appuie sur une documentation dont le

5. Consultable à cette adresse : `https://timeus.hypotheses.org/`.

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 ligne [1].

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 le seul but 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, tant 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 et reprise par ses continuateurs.

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. Pour l'équipe

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/bptl-driver.html`, *4.2. Encoding Levels.*

ALMAnaCH, il s'agissait de repérer les éléments d'onomastique et d'implanter un système permettant la citation de passage précis selon une granularité qui restait à définir.

L'ensemble de ces niveaux a été contrôlé lors du stage. 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.

5.2 Vérification de la cohérence documentaire

Nous avons commencé par contrôler le découpage des fichiers source, dont les résultats étaient étonnants 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 est 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 a deux pages, il s'agit toujours d'un recto et d'un verso, et de deux rectos et deux versos lorsqu'il y en a quatre.

Mis à part les premiers fichiers, qui commencent par le titre du précis, les `<body>` de ces fichiers ont un point commun : ils débutent 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 résultant d'une mauvaise segmentation s'est donc produite au moment de l'OCR[5].

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

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) »…

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

optimale des paragraphes au moment de la transformation des fichiers XML pour une éventuelle édition. Dans le cas présent, le script a été dupé par l'erreur de segmentation : il n'a pas vu des rappels de titre et des numéros de page, mais des titres et des numéros de chapitre. En conséquence, il a 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 contiennent-ils quatre pages, à l'inverse des autres qui n'en contiennent que deux ? Le fait est que les en-têtes des troisième pages ont 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. 64).

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` (*cf.* 1.1, p. 5) : un total de vingt-neuf fichiers ont été supprimés, le corpus passe donc de 223 à 194 fichiers.

5.2.2 Défaut de transcription

La seconde erreur majeure à laquelle nous avons été confronté 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° 30 [6], 33 [7], 37 [8], n° 44 [9], 45 [10] et 46 [11]). 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.

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`.

```
<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 GÉNERALES ; 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
   </p>
```

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`.

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 d'ensemble `<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 `@facs` relie à un `<facsimile>` ; `<pb>` venant signifier le changement de page (fig. 5.1).

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

104 N° 30. — PAYSANS EN COMMUNAUTÉ DU NING-PO-FOU.

BUDGET DES RECETTES DE L'ANNÉE.

SOURCES DES RECETTES. — Évaluation approximative des sources des recettes.

SECTION Ire. — Valeur des propriétés.

Propriétés possédées par la famille.

Art. 1er. — Propriétés immobilières.

Habitation :
Maison d'un rez-de-chaussée et d'un premier étage avec hangars 200f 00

Immeubles ruraux :
Champs de riz, d'orge, de fèves, etc. (94 ares) 2,800 00
Jardin potager attenant à la maison (5 ares) 250 00

Art. 2. — Valeurs mobilières.

Animaux domestiques entretenus toute l'année :
2 bœufs de labour et de transport 200 00
2 poules et 6 poulets 16 00

Animaux domestiques entretenus seulement une partie de l'année :
2 porcs : valeur calculée ..... (§ 6) 45 00
2 chèvreaux : valeur calculée ..... (§ 6) 3 56

Matériel spécial des travaux et industries :
Pour la culture des champs 481 00
Pour la culture du jardin potager 8 25
Pour l'exploitation des bêtes à cornes 13 25
Pour l'exploitation de la basse-cour 10 25
Pour la fabrication des étoffes de coton 21 75
Pour le blanchissage du linge 9 75
Pour la pêche 1 25

Art. 3. — Droits aux allocations des sociétés d'assurances mutuelles.

(Il n'existe dans le pays aucune société de ce genre) ..... »
Valeur totale des propriétés 4,059 45

SECTION II. — Évaluation du capital des subventions.

Subventions reçues par la famille.

Art. 1er. — Propriétés reçues en usufruit.

Moulin communal 500 00

Art. 2. — Droits d'usage sur les propriétés voisines.

Droit sur l'herbe coupée pour les bœufs sur les chemins 172 00
— — broutée par les bœufs sur les chemins 48 00
— sur la poisson de la rivière 83 70
— sur les grenouilles, les escargots, les salades 70 00

Art. 3. — Allocation d'objets et de services.

(La famille ne reçoit aucune allocation de ce genre) ..... »
Valeur totale à attribuer au capital des subventions 873 70

(a) N° 30, p. 104.

260 N° 33. — COMPOSITEUR-TYPOGRAPHE DE PARIS.

BUDGET DES RECETTES DE L'ANNÉE.

SOURCES DES RECETTES. — Évaluation approximative des sources des recettes.

SECTION Ire. — Valeur des propriétés.

Propriétés possédées par la famille.

Art. 1er. — Propriétés immobilières.

(La famille ne possède aucune propriété de ce genre) ..... »

Art. 2. — Valeurs mobilières.

Argent :
Fonds publics français (rente 4 1/2 pour 100) évalués au cours de 105f 55 ..... 2,715f 00
Fonds placés à la Caisse d'épargne 900 00
Somme gardée dans le ménage pour besoins imprévus 50 00

Matériel spécial des travaux et industries :
Outils de compositeur-typographe 21 05
— — pour les réparations et l'entretien du mobilier 20 30
Ustensiles pour le blanchissage du linge et des vêtements 19 20
— — pour l'entretien — — 7 10

Art. 3. — Droits aux allocations de sociétés d'assurances mutuelles.

Droit aux allocations de la caisse de secours établie dans l'atelier de l'ouvrier ..... »
— — — de la société typographique parisienne ..... »
— — — de Saint-François-Xavier ..... »

Valeur totale des propriétés 3,729 65

SECTION II. — Évaluation du capital des subventions.

Subventions reçues par la famille.

Art. 1er. — Propriétés reçues en usufruit.

(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 d'objets et de services.

Allocations concernant la nourriture 175 00
— — les vêtements des enfants 340 00
— — les récréations 50 00
— — — — 40 00
— — les besoins moraux 800 00

Valeur totale à attribuer au capital des subventions 1,445 00

(b) N° 33, p. 260.

426 N° 37. — PÊCHEUR-CÔTIER DE MARKEN.

BUDGET DES RECETTES DE L'ANNÉE.

SOURCES DES RECETTES. — Évaluation approximative des sources des recettes.

SECTION Ire. — Valeur des propriétés.

Propriétés possédées par la famille.

Art. 1er. — Propriétés immobilières.

Habitation :
Maison ..... 2,555f 1[illegible]

Immeubles ruraux :
1 hectare de prairies naturelles ..... 1,255 [illegible]

Art. 2. — Valeurs mobilières.

Argent :
Somme prêtée à des pêcheurs pour l'achat de barques ..... 3,194 [illegible]
— placée sur les fonds publics (dette nationale des Pays-Bas) ..... 2,157 [illegible]

Matériel spécial des travaux et industries :
Barque avec tous ses accessoires pour la pêche du hareng ..... 3,270 [illegible]
Petite barque pour la pêche des anguilles ..... 170 [illegible]
Bateau pour le transport des foins, des ustensiles de pêche et des provisions de ménage. 319 [illegible]
Engins de pêche ..... 2,710 [illegible]
Ustensiles pour le blaze du chanvre et du lin ..... 21 [illegible]
Outils pour l'exploitation de la pêche ..... 56 [illegible]
Ustensiles pour le blanchissage du linge et des vêtements ..... 17 [illegible]

Art. 3. — Droits aux allocations de sociétés d'assurances mutuelles.

(La famille ne fait partie d'aucune société de ce genre) ..... »
Valeur totale des propriétés ..... 10,703 [illegible]

SECTION II. — Évaluation du capital des subventions.

Subventions reçues par la famille.

Art. 1er. — Propriétés reçues en usufruit.

(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 d'objets et de services.

(La famille ne reçoit aucune allocation de ce genre) ..... »
Valeur totale à attribuer au capital des subventions ..... »

(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.

comme des objets graphiques.

Or les trois monographies concernées du 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 BibTeX [12] 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]`). Nous avons donc choisi de donner au premier fichier de paratexte le numéro `401a` et de continuer jusqu'au dernier.

Trois versions de la cartographie ont été réalisée : la première contient uniquement les monographies, la seconde le paratexte et la troisième l'ensemble des fichiers [13]. 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 interprété par le langage Python. Les données peuvent donc être manipulées par un script.

Une fois les CSV constitué, il a fallu implanté 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. 28) 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 et qui commence par une lettre dans chaque document [14]. Or il n'était 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.

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. « *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/release/doc/tei-p5-doc/fr/html/CO.html#CORS2`, consulté le 7 septembre 2020).

```
<?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*.

ALMAnaCH 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 sera 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 pérenne [15]. 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.

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).

15. « *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 7 septembre 2020).

```
<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.

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.

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° 56 [16], 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 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 au document. En effet, ces balises possèdent un

16. U. 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`.

attribut `@facs` 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 `@facs` des paragraphes et des titres par le biais d'un script. Dans l'arbre XML de départ (fig. 3.4 p. 28), le seul descendant de l'élément `<surface>` est désormais `<graphic>`.

Chapitre 6

Contrôle du découpage éditorial

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 terrains en des « faits décris » par la monographie [1].

6.1 Erreurs constatées dans l'implémentation

6.1.1 Niveau des titres (`<head>`)

Pour commencer, nous avons tenter de réaliser 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. En sortie, les titres qui n'avaient pas été trouvés étaient listés.

Le script a fait remonter plus de quatre cent cinquante erreurs. Ce chiffre était largement gonflé par la non-prise en compte de la qualité de la transcription : certains titres étaient présents mais avaient été transcrits avec des fautes. Les erreurs principales pouvaient néanmoins être identifiées dès cette étape :

- Des distances d'édition trop faible entre deux titres ont 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 ».
- La monographie n° 44 [2] est la dernière à posséder des titres de paragraphe sur

1. S. Baciocchi et J. David, « IV. Ramifications épistémologiques. »…, p. 87.
2. U. Guérin, « Paysan-résinier de Lévignacq (Landes - France) »…

une seule ligne, un retour à la ligne étant ensuite systématiquement effectué entre le numéro et le libellé du paragraphe. Cet élément n'a 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 est placé dans un `<p>` et le titre avec le numéro ajouté dans un `<head>` (fig. 3.6 p. 30).
  - En cas de non répétition, le titre n'est pas toujours détecté et les deux lignes sont encodées dans des `<p>`.
- Des titres transcrits avec une distance trop grande par rapport au modèle n'ont pas été détectés. Ce point concerne l'ensemble des titres, sans prévalence particulière.
- Les titres des tableaux de budgets sont transcrits plus d'une fois.
- Les cas particuliers n'ont pas été pris en compte :
  - La monographie concerne un ouvrier ou une ouvrière et non 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` ;
  - Les titres du paragraphes n'étaient pas séparés du texte mais imprimé en italique au début de celui-ci ;
  - 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).

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. 46). 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`).

Deux 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. 30) ;
- Au cour du stage, il a été décidé de placer la titre de la monographie dans une `<div>` de type `section`, ce qui impliquait une reprise dans l'ensemble des fichiers.

6.1.3 Méthodologie : *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 de ces

cas avant de rendre compte de celle utilisée pour le reste du corpus.

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. Ces fichiers partageaient également avec certaines monographies de familles le fait d'avoir des titres qui ne sont pas détachés des paragraphes qu'ils concernent. Le recours à la balise `<head>` n'était pas possible en l'état, 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 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 n'autorisent pas pour ces balises les attributs qui auraient permis de caractériser ces titres (notamment, `@type`). Aussi ne serait-il pas possible de les différencier d'autres passages ayant la même mise en forme dans le corps du texte. 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.

```
<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>`.

Au final, *Time Us* n'a opté pour aucune solution, considérant que le nombre de cas

3. P. du Maroussem, « La société générale des papeteries du Limousins »…
4. Id., « Usine hydraulique d'éclairage et de transport de force »…

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.

Les cas de ces quelques fichiers 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 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 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 d'utiliser ces expressions dans un script, et notamment d'effectuer des substituions grâce à la fonction `re.sub()`. Une dernière commande, qui affichait la liste des balises `<head>` de chaque fichier, nous permettait ensuite de contrôle l'intégrité de la nouvelle structure. Si une erreur était constatée, nous intervenions manuellement afin de la résoudre.

Du fait de la désorganisation massive des `<div>`, la première expressions régulière consistait à faire table rase de ces balises. Pour reconstruire la structure logique, le script suivait la structure. Ainsi, il commençait par implanter au niveau supérieur la `<div>` de type `chapter` et celle de la première section (la page de titre). Venaient ensuite les `<div>` des deux autres section, (*Observations préliminaires* et *Notes*)

5. U. Guérin, « Ouvrier cordonnier de Malakoff (Seine - France) »…
6. J. Reviers de Mauny, « Serrurier-forgeron de Paris (Seine - France) »…

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 taisible 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 éventailliste 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 chiffonnier instable et, par alternance, mégissier fumiste et brossier de Paris (France - Seine)	*In supra.*
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	*In supra.*
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-maraicher 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 Cockerill, 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 Ribeauvillé (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 régleur 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 Roccasancasciano (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 Monthieux (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***[7]).

§ 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.

7. 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=eN0WAAAAYAAJ`
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 2020 [8].

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).

8. 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 matières

Résumé **iii**

Liste des sigles et abréviations **v**

Introduction **ix**

Bibliographie **xvii**

Bibliographie générale xix

Les Ouvriers des deux mondes xxi

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 7

1.3 Continuité et discontinuité 10

2 Des numérisations multiples **13**

2.1 Les volumes de la Bibliothèque nationale de France 13

2.2 Les volumes d'*Internet Archive* 14

2.3 Choix du corpus à numériser 15

3 Un encodage automatique **19**

3.1 Choix de l'automatisation 19

3.2 Le script LSE-OD2M 20

3.2.1 Segmentation 20

3.2.2 Reconnaissance des caractères 23

3.2.3 Structuration 26

3.3 Les fichiers XML-TEI 26

II Une structuration à reprendre : des tâches manuelles et semi-automatiques 33

4 Outils, méthodologie et gestion du projet 35
4.1 Outils de développement . . . 35
4.1.1 *GitLab* . . . 35
4.1.2 *PyCharm* et *Oxygen* . . . 36
4.2 Espaces de discussion . . . 37
4.2.1 *Mattermost* . . . 37
4.2.2 *Issues* et *merge requests* sur *GitLab* . . . 38
4.3 Feuille de route . . . 39
4.3.1 Les missions du stage . . . 39
4.3.2 Une gestion de projet ? . . . 40

5 Contrôle du découpage des fichiers source 43
5.1 Les différents niveaux d'encodage . . . 43
5.2 Vérification de la cohérence documentaire . . . 44
5.2.1 Fission horizontale lors de la segmentation . . . 44
5.2.2 Défaut de transcription . . . 45
5.2.3 Identification, cartographie et inclusion . . . 48
5.3 Vérification de l'encodage minimal . . . 50

6 Contrôle du découpage éditorial 53
6.1 Erreurs constatées dans l'implémentation . . . 53
6.1.1 Niveau des titres (`<head>`) . . . 53
6.1.2 Niveau des divisions (`<div>`) . . . 54
6.1.3 Méthodologie : *tabula rasa* . . . 54

Annexes 59

A. *Les Ouvriers des deux mondes* 59
A.1 Liste des monographies et fichiers correspondant . . . 59
A.1.1 Série 1, vol. 1 (1857). . . . 59
A.1.2 Série 1, vol. 2 (1858). . . . 60
A.1.3 Série 1, vol. 3 (1861). . . . 61
A.1.4 Série 1, vol. 4 (1862). . . . 62
A.1.5 Série 1, vol. 5 [1875, 1883, 1884] (1885). . . . 63
A.1.6 Série 2, vol. 1 [1885-1887] (1887). . . . 64
A.1.7 Série 2, vol. 2 [1887-1889] (1890). . . . 65
A.1.8 Série 2, vol. 3 [1890-1892] (1892). . . . 66

A.1.9 Série 2, vol. 4 [1892-1895] (1895). 67
A.1.10 Série 2, vol. 5 [1895-1899] (1899). 68
A.1.11 Série 3, vol. 1 [1900-1904] (1904). 69
A.1.12 Série 3, vol. 2 [1904-1908] (1908). 70
A.1.13 Série 3, vol. 3 [1908-1913] (1913). 71
A.1.14 Série 3, vol. 3bis [1928-1930] (1930). 71
A.2 Structure logique 72
A.3 Numérisations de *Google Books* 73
A.3.1 Bibliothèque de l'université de Californie 73
A.3.2 Bibliothèque nationale centrale de Florence 73
A.3.3 Bibliothèque de l'université Harvard 73
A.3.4 Bibliothèque municipale de la ville de Lyon 73
A.3.5 *Bayerische Staatsbibliothek* de Munich 73
A.3.6 *New York State College of Agriculture at Cornell University* 74
A.3.7 Bibliothèque de l'Université de Princeton 74
A.3.8 Université de Rome — *Instituto de filosofia del diritto* 74

B. Feuille de route et typologie des erreurs 75
B.1 Feuille de route 75
B.1 Relevé des erreurs dans la structure logique 76
B.1.1 Déficit dans la transcription 76
B.1.2 Titres non-imprimés 77
B.1.3 Structure allégée 77
B.1.4 Pas de s/s/section §16 et pas de section *Notes* 78
B.1.5 Remarques et cas particuliers 78

Table des matières 79