

ÉCOLE NATIONALE DES CHARTES

Alix Chagué

Licenciée ès histoire

Licenciée ès histoire de l'art

Diplômée de master histoire de l'art

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 pour le diplôme

« Technologies numériques appliquées à l'histoire »

2018

Résumé

Ce mémoire a été réalisé en vue de l'obtention du diplôme de Master 2 « Technologies numériques appliquées à l'histoire » de l'École nationale des chartes. Il a été rédigé à la suite d'un stage de quatre mois au sein de l'équipe ALMAnaCH d'Inria, et dont le déroulé s'est inscrit dans le cadre du projet de recherche ANR pluri-institutionnel intitulé « Time Us ». Ce projet de recherche porte sur l'histoire de l'industrie du textile en France (fin XVII^e-début XX^e siècles) et sur la reconstitution des budget temps des ouvriers et ouvrières du textile. Il explore également l'utilisation d'outils informatique pour réaliser cette recherche. Ce mémoire vise à reconstituer l'ensemble des étapes de traitement de sources d'archives permettant d'aboutir à la création de transcriptions annotées sous la forme de fichiers XML-TEI. Il s'agit d'une analyse critique des enjeux, stratégies et résultats envisagés dans le cadre du projet Time Us autant que du stage, dont le but est de rendre compte d'un exemple de projet et de développement s'inscrivant dans le cadre des humanités numériques.

Mots-clefs : reconnaissance automatique d'écriture ; transcription collaborative ; annotation sémantique ; TAL ; Transkribus ; XML-TEI ; XSLT ; industrie du textile ; histoire économique et sociale.

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

Remerciements

Je tiens à remercier toutes les personnes qui m'ont apporté leur soutien durant cette année d'étude et au cours du stage qui en a été l'aboutissement.

Je souhaite en particulier remercier Ms Vincent Jolivet, Éric de la Clergerie et Mme Manuela Martini, pour le rôle de tuteurs et tutrice qu'ils et elle ont joué durant ce stage.

J'adresse également mes remerciements à Charles Riondet et Marie Puren pour leur encadrement quotidien, ainsi qu'à l'ensemble des membres de l'équipe ALMAnaCH d'Inria, pour les conseils que j'ai reçus lorsque j'en avais besoin ; je remercie en particulier Ms Benoît Sagot, Yoann Dupont et Lionel Tadonfouet ainsi que Mmes Mathilde Regnault et Tanti Kristanti.

Je remercie aussi l'ensemble des membres du projet « Time Us » pour leur accueil et la confiance qui m'a été accordée.

Enfin, j'adresse un grand merci à ma famille et à mes ami·es pour leurs encouragements et leurs remarques, qui ont été précieux tout au long de mon année scolaire et pendant la réalisation de ce stage autant que durant la rédaction du présent mémoire.

Liste des sigles et abréviations

- A.D. : Archives Départementales
- A.M. : Archives Municipales
- A.R. : Archives Régionales
- B.M. : Bibliothèque Municipale

*

- ALMAaCH : *Automatic Language Modelling and Analysis & Computational Humanities*
- CMH : Centre Maurice Halbwachs
- CNRS : Centre National de Recherche Scientifique
- ENS : École Normale Supérieure
- EPC : Équipe-Projet Commune
- EPI : Équipe-Projet Inria
- EHESS : École des Hautes Études en Sciences Sociales
- HDR : Habilitation à diriger des recherches
- IRHIS : Institut de Recherches Historiques du Septentrion
- LARHRA : Laboratoire de Recherche Historique Rhône-Alpe
- LLF : Laboratoire de Linguistique Formelle
- NSF : *National Science Foundation*
- TAL : Traitement Automatique des Langues
- UMR : Unité Mixte de Recherche

*

- API : *Application Programming Interface*
- CER : *Character Error Rate*
- CMS : *Content Management System*

- EXIF : *Exchangeable Image file Format*
- JSON : *JavaScript Object Notation*
- PAGE : *Page Analysis and Ground-truth Elements*
- PDF : *Portable Document Format*
- REST : *Representational State Transfer*
- TEI : *Text Encoding Initiative*
- WADL : *Web Application Description Language*
- WER : *Word Error Rate*
- XML : *eXtensible Markup Language*
- XSLT : *eXtensible Stylesheet Language Transformations*

Table des figures

1.1	Graphe résultant d'une analyse syntaxique avec FrMG (schéma <i>depxml</i>).	17
1.2	Capture d'écran d'une page de wikisource.org, en août 2018.	19
1.3	Aperçu de l'interface graphique de Transkribus.	21
2.1	Schématisation du protocole de nommage dans ShareDocs.	35
2.2	Schématisation de la chaîne de traitement des documents.	39
3.1	Comparaison des transcriptions manuelles et automatiques de la page 292.	49
3.2	Schématisation de la structure logique du fichier TEI, sans annotation.	51
4.1	Schématisation du processus d'annotation manuelle et automatique.	57
4.2	<i>Tags</i> et équivalents TEI en juillet 2018.	67
4.3	Capture d'écran : affichage des info-bulles par TEITags.	69
4.4	Extraits du fichier TEItags.body.php.	70
4.5	Extrait du fichier ext.teitags.css.	70
5.1	Extraits de modifTag.xsl.	76
5.2	Exemple d'annotation segmentée.	76
5.3	Schématisation de la reconstruction des annotations segmentées.	77
5.4	Exemple d'annotations segmentées se chevauchant.	77
5.5	Exemple de métadonnées obtenues à l'export depuis Transkribus.	78
5.6	Schématisation de l'interaction Client-Serveur.	79
5.7	Capture d'écran de l'interface de création de requêtes de Postman.	80
5.8	Capture d'écran de l'interface de consultation en ligne de Postman.	81
5.9	Extrait de la réponse à la requête « collections/list ».	83
5.10	Extrait de la réponse à la requête « collections/8097/list ».	83
5.11	Extrait de la réponse à la requête « collections/8097/41459/fulldoc ».	84
5.12	Configuration minimale du dossier pour exécuter les <i>scripts</i>	85
5.13	Extrait du fichier config.py (l.1 à 19).	86
5.14	Modélisation du <i>script</i> requestingTranskribus.py.	88
5.15	Extrait du fichier config.py (l.22 à 35).	89

5.16 Comparaison d'extraits des fichiers page2tei, original (l.32 à 35) et adapté (l.30 à 33)	91
5.17 Comparaison d'extraits des fichiers PAGE, original et adapté.	92
5.18 Extrait du fichier page2tei_TU.xsl (l.71 à 89).	93
5.19 Comparaison des métadonnées créées par Transkribus et par page2tei_TU. .	94
5.20 Extrait d'un fichier XML-PAGE exporté de Transkribus.	95
5.21 Extrait du fichier page2tei-0.xsl de Dario Kampkaspar (l.434 à 441). . . .	95
5.22 Extrait du fichier page2tei_TU.xsl (l.512 à 523).	95
5.23 Extrait du fichier page2tei_TU.xsl (l.650 à 672).	96
5.24 Capture d'écran de l'affichage par défaut de TEI Boiler Plate.	100
5.25 Capture d'écran de l'affichage après les consignes de Charles Riondet. . . .	101
5.26 Extrait de la structure HTML interprétée par le navigateur.	102
5.27 Capture d'écran de l'affichage après nos modifications.	103

Introduction

L'outil informatique s'est durablement implanté dans les infrastructures de recherche en sciences humaines, en particulier en histoire, depuis la fin des années 1950¹. Il a permis notamment de faciliter « l'accès aux méthodes statistiques de l'analyse des données »² et a rendu possible le maniement de masses d'informations. Les analyses statistiques ont elle-mêmes été révolutionnées par le développement de l'informatique, et ont donné lieu à une approche nouvelle de l'exploitation informatisée de corpus de données massifs : le *data mining*.

« Le data mining est l'art d'extraire des informations, voire des connaissances, à partir des données. »³

Le *data mining*, aussi appelé « fouille de données », est incarné par un ensemble de méthodes, d'outils et de concepts dont la montée en puissance est allée de pair avec l'accroissement des ressources disponibles pour l'analyse. Les usages et champs d'application de la fouille de données sont extrêmement variés et dépassent très largement le domaine des sciences humaines. La fouille de données peut être appliquée à des données de nature textuelle, qualitatives, auquel cas on parle de *text mining*, ou « fouille de texte » :

« Le *text mining* est l'ensemble des techniques et méthodes destinées au *traitement automatique de données textuelles en langage naturel*, disponibles sous forme informatique, en assez grande quantité, en vue d'en *dégager et structurer le contenu, les thèmes* dans une perspective d'analyse rapide (non littéraire), de découverte d'informations cachées ou de prise automatique de décision. »⁴

La fouille de texte allie fouille de données et lexicométrie. Elle représente un outil aux enjeux considérables pour les historiens qui peuvent y trouver le moyen de réaliser plus efficacement le dépouillage, alors automatique, de corpus jusqu'alors impossibles à traiter à une échelle humaine.

En étroite collaboration avec Inria, et son équipe ALMAaCH, le projet de recherche soutenu par l'Agence Nationale de la Recherche (ANR) intitulé « Time Us » entend interroger cet enjeu, sa faisabilité et son intérêt pour la recherche en histoire, dans le cadre d'une étude portant sur l'industrie du textile en France de la fin du XVII^e siècle au début du XX^e. Six laboratoires de recherche français sont impliqués dans le projet.

C'est dans ce cadre qu'un stage de quatre mois m'a été confié, encadré par l'équipe ALMAaCH à Inria, sous la direction de Monsieur Éric de la Clergerie et supervisé par Madame Manuela Martini, et pris en charge administrativement par le laboratoire ICT

1. Jean-Philippe Genet, « Informatique et Histoire », *Le bulletin de l'EPI*–49 (mars 1988), p. 12.

2. *Ibid.*, p. 133.

3. Stéphane Tufféry, *Data mining et statistique décisionnelle - 4ème édition*, 4e édition, Paris, 2012, p. 2.

4. *Ibid.*, p. 739.

de l'Université Paris-Diderot. Il s'est agi d'accompagner la réflexion menée sur les aspects de standardisation et de préparation des corpus soumis à la fouille de données.

Sachant que plusieurs outils avaient déjà été mis en place, dont une plate-forme de type « wiki » et le logiciel de transcription collaborative Transkribus, les missions présentées dans le cahier des charges ont consisté à :

- Proposer des améliorations pour ces plate-formes, en particulier pour une gestion plus fine des annotations TEI nécessaires à l'analyse des textes ;
- Établir un schéma TEI pour l'ensemble du corpus, couvrant à la fois la représentation des textes à transcrire, les métadonnées et les annotations sémantiques spécifiques au projet Time Us ;
- Participer à la collecte, à la mise en forme et à la transcription des données en collaboration avec les équipes partenaires.

Concrètement, les résultats attendus étaient :

- La création d'un guide de bonnes pratiques pour la constitution, l'analyse et l'annotation du corpus ;
- Le développement d'un schéma TEI et son instanciation, notamment pour le corpus des monographies de Le Play⁵ ;
- L'évolution des outils de transcription collaborative et de visualisation (Transkribus et le wiki).

C'est principalement sur les conseils d'Éric de la Clergerie, de Marie Puren et de Charles Riondet que ces missions ont été réalisées. Leur exécution nous a permis d'établir un modèle de chaîne de traitement des sources, depuis les documents d'archives papiers jusqu'aux fichiers nativement numériques mis en forme en vue de leur analyse par des outils de fouille de texte. La question à laquelle nous avons essayé de répondre est donc de savoir quels moyens et quels processus de traitement doivent être mis en oeuvre pour rendre possible la création d'un tel corpus. Nous avons en outre mener constamment une réflexion sur les stratégies à adopter pour garantir des environnements de travail homogènes pour des partenaires dispersés géographiquement dont les équipements n'étaient pas identiques.

Le présent mémoire retrace l'ensemble de la chaîne de traitement mise en place dans le cadre du projet Time Us. En effet, le cahier des charges invitait à participer à la réflexion soutenant chacune d'entre elles.

Étant données les ambitions du projet de recherche et le nombre important de ses collaborateurs, nous nous intéresserons dans une première partie à la contextualisation précise du stage et du projet de recherche. Nous établirons d'abord la liste des équipes

5. Il s'agit d'un corpus d'enquêtes sociologiques menées entre 1857 et 1930 sous l'impulsion de Frédéric Le Play, sur lequel nous reviendrons plus loin.

rassemblées ainsi que le détail de leur implication dans le projet. Nous poursuivrons en détaillant le corpus extrêmement varié à partir duquel le travail a été mené, ainsi que les outils déjà implémentés pour leur traitement.

Dans une deuxième partie, nous retracerons les différentes étapes de traitement du corpus, nous pointerons les problématiques techniques et méthodologiques soulevées et les moyens mis en oeuvre pour y répondre. Nous verrons tout d'abord comment a été menée la campagne de transcription des sources, nécessaire pour la récupération des données textuelles, structurées grâce à des schémas TEI. Celle-ci a ensuite donné lieu à l'annotation du corpus numérique de textes, pour lequel il a été nécessaire de créer un modèle d'annotation ainsi qu'un guide pour assurer son implémentation. Enfin, l'ensemble de ce processus ayant été réalisé grâce à la plate-forme Transkribus, nous avons dû mettre en place des outils pour extraire et normaliser toutes les données annotées. C'est ce que nous décrirons dans un dernier temps.

Première partie

Un projet ambitieux

Le stage qui a donné lieu à la rédaction de ce mémoire s'est déroulé dans les locaux d'Inria, à Paris, au sein de l'équipe ALMAncH. Bien que mené en étroite collaboration avec les membres du laboratoire lyonnais LARHRA, le contexte institutionnel de ce stage a joué un rôle dans la prise en compte des différents enjeux liés à la constitution d'un corpus propre et standardisé en vue de sa mise à disposition pour la communauté des chercheur·ses en histoire et en sciences humaines.

Chapitre 1

Un contexte pluri-institutionnel

1.1 Inria et Almanach : soutiens du projet

1.1.1 Présentation générale d’Inria

Inria, l’institut national de recherche en informatique et automatique, est un établissement public à caractère scientifique et technologique (EPST) placé sous la double tutelle du ministère en charge de la recherche et de celui en charge de l’industrie. Il est le seul organisme de recherche français dédié aux sciences du numérique¹. Sa mission principale est d’oeuvrer pour le transfert technologique vers l’industrie, en répondant aux enjeux sociaux actuels et futurs et en décloisonnant les relations entre le monde de la recherche et celui de l’industrie et des entreprises. Créé à Rocquencourt en 1967, Inria est aujourd’hui implanté en France dans 8 centres de recherche (à Paris, Saclay, Lille, Nancy, Grenoble, Sophia Antipolis, Bordeaux et Rennes), et fait partie de 6 laboratoires communs dans le monde : au Chili, au Sénégal, en Chine, en Suisse et aux États-Unis². La recherche à Inria est mise en place dans le cadre flexible et agile des « équipes projet », encadrées par la Direction Générale déléguée à la Science (DGD-S) mais placées sous la responsabilité directe de chaque centre Inria.

Les équipes-projet³ sont au fondement du fonctionnement de la recherche à Inria depuis sa création. Elles sont composées de dix à trente personnes, coordonnées par une personnalité scientifique de haut niveau qui établit les objectifs scientifiques de l’équipe, en accord avec les différents services de direction qui l’encadrent. D’un point de vue financier, les équipes-projet sont relativement autonomes : leur budget est constitué de ressources fournies par Inria ou obtenues grâce à leurs objectifs de recherche. Elles sont de deux

1. *Plan stratégique scientifique*, Inria, URL : <https://www.inria.fr/institut/strategie/plan-strategique> (visité le 12/08/2018).

2. *Politique européenne & internationale*, Inria, URL : <https://www.inria.fr/europe-international/politique-europeenne-internationale> (visité le 12/08/2018).

3. *Le modèle « Équipe-projet »*, Inria, Inria, URL : <https://www.inria.fr/recherches/structures-de-recherche/modele-equipe-projet> (visité le 12/08/2018).

types : les équipes-projet Inria (EPI), composées uniquement de personnels Inria, ou les équipes-projet communes (EPC), associées à des établissements partenaires (universités, écoles ou encore centres de recherche, par exemple). Les équipes-projet sont évaluées tous les 4 ans sur leur capacité à atteindre leurs objectifs. Elles ont une durée de vie de 8 ans en moyenne, mais peuvent être reconduites deux fois au maximum, soit pour une durée totale de 12 ans. Elles ont deux obligations : la communication des résultats scientifiques et la participation au transfert des connaissances et des technologies acquises vers les utilisateur·rices, y compris vers l'industrie. Cela se fait sous la forme de brevets, de formations, de licences ou encore de partenariats scientifiques, etc. Inria promeut par ailleurs l'ouverture et le partage des données, et les logiciels libres.

En 2018, alors que 35 équipes-projet dépendent du site de Paris, Inria en compte près de 180 au total, rassemblant environ 1400 chercheur·ses et enseignant·es chercheur·ses, plus de 1200 doctorant·es et près de 300 post-doctorant·es ou contractuel·les. Toutes ces équipes-projet oeuvrent dans l'un des cinq domaines de recherche définis par Inria :

- mathématiques appliquées, calcul et simulation ;
- algorithmique, programmation, logiciels et architectures ;
- réseaux, systèmes et services, calcul distribué ;
- perception, cognition, interaction ;
- santé, biologie et planète numériques.

1.1.2 ALMANaCH

ALMANaCH est une EPC d'Inria dont l'acronyme signifie : *Automatic Language Modelling and Analysis & Computational Humanities*⁴ et qui s'inscrit dans le domaine « perception, cognition, interaction » et son axe « langue, parole et audio ». Elle a été créée en janvier 2017, mais son histoire est plus ancienne. Elle est en effet née de la refondation de l'équipe ALPAGE⁵, une Unité Mixte de Recherche (UMR) commune à Inria et l'université Paris-Diderot (UMR-I 001), créée sous forme d'EPI en 2007, devenue UMR en 2009. ALPAGE était principalement dédiée à la recherche en Traitement Automatique de la Langue (TAL), en particulier pour l'analyse syntaxique automatique, le développement de ressources lexicales et le traitement du discours pour la langue française. ALPAGE rassemblait informaticiens, linguistes et « TAL-istes » dépendants d'Inria ou de l'UFR de linguistique de l'Université Paris-Diderot⁶. ALPAGE a donné naissance à ALMANnaCH d'une part, mais aussi, d'autre part, au Laboratoire de Linguistique Formelle (LLF), commun à l'université Paris-Diderot et au Centre National de la Recherche Scientifique

4. Modélisation et analyse automatique du langage et humanités computationnelles

5. Analyse Linguistique Profonde À Grande Échelle

6. Site web Alpage, URL : <https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=Accueil> (visité le 12/08/2018).

(CNRS)⁷.

Benoît Sagot est le responsable scientifique d'ALMAAnaCH. Il est accompagné de six membres permanents : Marc Bui (EPHE), Laurent Romary (Inria), Djamé Seddah (Paris-Sorbonne), Daniel Stöckl Ben Ezra (EPHE) et Éric de la Clergerie (Inria). En juillet 2018, ALMAAnaCH rassemblait par ailleurs douze doctorant·es ou pré-doctorant·es, dix post-doctorant·es ou ingénieur·es, ainsi qu'un nombre variable de stagiaires.

La particularité du projet ALMAAnaCH est d'avoir souhaité élargir ses domaines d'application pour inclure les humanités numériques et les humanités computationnelles⁸. À la différence des humanités numériques, où l'informatique est envisagée comme un outil de recherche, les humanités computationnelles intègrent l'informatique comme un champ de recherche à part entière, où il s'agit de faire avancer l'état des connaissances autant dans un ou des champs de sciences humaines, qu'en informatique. La linguistique demeure l'un des domaines d'application et de recherche principaux de l'équipe, ainsi qu'en témoigne ses trois axes de recherche : « Analyse linguistique automatique améliorée par les informations contextuelles », « Modélisation computationnelle de la variation linguistique » et « Modélisation et développement de ressources linguistiques ».

Les projets menés par les différents membres d'ALMAAnaCH sont souvent le fruit de collaborations avec d'autres équipes ou institutions, à l'échelle de la France ou à l'international. ALMAAnaCH intervient ainsi dans cinq projets européens, dont trois s'inscrivent dans le programme Horizon 2020 (H2020)⁹. C'est le cas des projets Parthenos¹⁰, EHRI¹¹ ou encore Iperion CH¹². ALAMAnaCH participe en outre à 6 projets financés par l'ANR : PARSITI¹³, SoSweet¹⁴, PARSE-ME¹⁵, PROFITEROLE¹⁶, MCM-NL (un projet international financé par l'ANR et par son équivalent américain, la *National Science Foundation* (NSF))¹⁷, et enfin le projet Time Us¹⁸.

7. *Laboratoire de linguistique formelle*, URL : <http://www.llf.cnrs.fr/> (visité le 12/08/2018).

8. Accueil, ALMAAnaCH, URL : <https://team.inria.fr/almanach/fr/> (visité le 12/08/2018).

9. Horizon 2020, URL : <http://www.horizon2020.gouv.fr/> (visité le 12/08/2018). - Horizon 2020 est un programme de financement européen pour la recherche et l'innovation. Il est doté d'un fond de 79 milliards d'euros pour la période 2014-2020. Il vise à financer des projets interdisciplinaires capables de répondre aux défis sociaux et économiques identifiés pour l'Europe dans le domaine de l'innovation.

10. PARTHENOS Project, URL : <http://www.parthenos-project.eu/> (visité le 12/08/2018).

11. European Holocaust Research Infrastructure, URL : <https://ehri-project.eu/> (visité le 12/08/2018).

12. Iperion CH, URL : <http://www.iperionch.eu/> (visité le 12/08/2018).

13. Projet PARSITI, ANR, URL : [http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2\[CODE\]=ANR-16-CE33-0021](http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2[CODE]=ANR-16-CE33-0021) (visité le 12/08/2018).

14. SoSweet, URL : <http://sosweet.inria.fr/> (visité le 12/08/2018).

15. Syntactic Parsing and Multiword Expressions in French, URL : <http://parsemefr.lif.univ-mrs.fr/doku.php> (visité le 12/08/2018).

16. Projet PROFITEROLE, ANR, URL : [http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2\[CODE\]=ANR-16-CE38-0010](http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2[CODE]=ANR-16-CE38-0010) (visité le 12/08/2018).

17. Benoit Sagot, *Un projet ANR-NSF sur le développement d'outils informatiques de modélisation de données neurolinguistiques*, ALMAAnaCH, URL : <https://team.inria.fr/almanach/fr/anr-nsf-project-to-develop-computational-tools-for-modeling-neurolinguistic-data/> (visité le 12/08/2018).

18. Pour la liste complète : Projets, ALMAAnaCH, URL : <https://team.inria.fr/almanach/fr/>

En raison des liens rapprochant ALMAAnaCH et le laboratoire ICT, Éric de la Clergerie et Benoît Sagot, spécialistes du traitement automatique des langues, ont participé à la mise en place du projet Time Us. Ils en sont membres permanents. En décembre 2016, Charles Riondet et Marie Puren, ingénieur·es de recherche spécialistes des humanités numériques, ont rejoint le projet en tant que membres non permanents.

1.2 ANR Time US : le projet de recherche

Le programme de recherche Time Us devait être lancé en octobre 2016, mais il a officiellement débuté en janvier 2017. Son titre complet est : « Rémunération et usages du temps des femmes et des hommes dans l’industrie du textile en France de la fin du XVII^e siècle au début du XX^e siècle ». Il est coordonné par Manuela Martini, professeure d’histoire contemporaine à l’Université Lumière-Lyon 2, et membre du Laboratoire de Recherche Historique Rhône-Alpes (LARHRA).

Le programme rassemble plusieurs équipes, basées à Aix, Marseille, Lyon, Paris ou encore Lille. Il porte sur l’étude des rémunérations et des budgets temps des ouvrièr·es de l’industrie du textile en France pendant les premières phases de l’industrialisation. A partir du cas de quatre grandes villes industrielles (Paris, Lille, Marseille et Lyon), il s’agit d’analyser les activités extra-domestiques autant que domestiques, qu’elles soient rémunérées ou non, des hommes et des femmes participant à l’industrie du textile. L’objectif étant de produire des données permettant de mieux comprendre l’évolution des différences en terme de rémunération au sein de cette industrie¹⁹, et d’explorer des questions d’ordre méthodologique propres aux humanités numériques.

1.2.1 Les équipes de Time Us

Outre ALMAAnaCH, cinq équipes ont fondé le projet Time Us : elles sont issues de laboratoires ou de centres de recherche. Elles prennent en charge l’exploration et l’inventaire des sources propres à leurs domaines et/ou leur cadre géographique de spécialité : le LARHRA pour Lyon, TELEMM pour Lyon et Marseille, l’IRHIS pour Lille, et ICT pour Paris. Le CMH, quant à lui, s’attache à un corpus d’enquêtes sociologiques réalisées au XIX^e siècle sur lesquelles nous aurons l’occasion de revenir : les monographies de Le Play.

projects/ (visité le 12/08/2018).

19. N’est prise en compte que la partie « fabrication » de l’industrie du textile : de la filature à l’assemblage de pièces. La distribution du textile, elle, est laissée de côté.

1.2.1.1 LARHRA

Le Laboratoire de Recherche Historique Rhône-Alpes²⁰ est une UMR du CNRS (UMR 5190) créée en 2003. Elle rassemble les universités Lumière-Lyon 2, Jean-Moulin-Lyon 3 et Grenoble Alpes, ainsi que l’École Normale Supérieure de Lyon, l’Institut des Sciences de l’Homme et la Maison des Sciences de l’Homme - Alpes. Les recherches de ce laboratoire portent sur l’histoire des périodes moderne et contemporaine, notamment dans une perspective d’histoire sociale. Elles sont organisées autour de 6 axes : « Art, images, sociétés », « Genre », « Savoirs, acteurs, dynamiques, espaces », « Religions et croyances », « Territoires, économie, enjeux sociaux », « Action publique et mondes urbains ». S’y ajoutent deux ateliers transversaux, « Images, sons, mémoires » et « Sociétés en guerre », ainsi qu’un pôle « Histoire numérique », témoin de l’engagement du laboratoire dans la recherche historique supportée par les outils informatiques.

Manuela Martini et Pierre Vernus sont des membres permanents de l’équipe pour le projet Time Us. Marie Lauricella et Tiphaine Gaumy interviennent également dans le projet en tant que membres non-permanents. S’y ajoutent plusieurs stagiaires et vacataires²¹.

1.2.1.2 TELEMMé

TELEMMé (Temps, Espaces, Langages, Europe Méridionale - Méditerranée)²² est une UMR du CNRS (UMR 7303) créée en 1994. Elle est basée à Aix-Marseille Université. Elle est la seule UMR en sciences humaines dont les axes de recherche portent spécifiquement sur l’Europe méditerranéenne, une aire géographique prise en compte depuis la péninsule ibérique jusqu’aux Balkans. Elle rassemble une équipe pluridisciplinaire d’historiens, historiens de l’art et hispanistes, spécialistes des périodes médiévale, moderne et contemporaine.

Anne Montenach coordonne l’action de TELEMMé au sein de Time Us, elle est à ce titre, ainsi que Xavier Daumalin, membre permanent du projet. Stéphane Kronenberger, post-doctorant, ainsi que deux étudiants de master dirigés l’un par Anne Montenach, l’autre par Xavier Daumalin ont également participé au projet.

1.2.1.3 IRHIS

L’Institut de Recherches Historiques du Septentrion²³ est un laboratoire de recherche pluridisciplinaire, créé en 2006, rassemblant historien·nes et historien·nes de l’art. Il fait partie d’une UMR du CNRS (UMR 8529) et est basé à l’Université de Lille. Les

20. *LARHRA* : Accueil, URL : <http://larhra.ish-lyon.cnrs.fr/> (visité le 03/08/2018).

21. Pour le détail, voir : *Participants au projet*, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Participants_au_projet (visité le 03/08/2018).

22. *TELEMMé* : Accueil, URL : <http://telemme.mmsh.univ-aix.fr/> (visité le 03/08/2018).

23. *IRHiS* : Accueil, URL : <https://irhis.univ-lille.fr/> (visité le 03/08/2018).

recherches de ce laboratoire portent sur l'histoire, l'histoire de l'art et la culture visuelle, du Moyen-Âge à la période contemporaine.

Matthieu de Oliveira est le coordinateur des actions menées par l'IRHIS dans le cadre du projet Time Us, il en est donc membre permanent. C'est sous sa supervision qu'interviennent plusieurs étudiant·es de master.

1.2.1.4 Laboratoire ICT

Le laboratoire ICT (Identité - Cultures - Territoires)²⁴ est une équipe d'accueil (EA 337) basée à l'Université Paris Diderot. Ses recherches portent sur l'histoire et les civilisations, du Moyen-Âge à la période contemporaine, dans une approche interdisciplinaire et interculturelle. Elles sont développées selon trois axes : « Territoires, mobilités, pouvoirs », « Genre et diversités » et « Savoirs, circulations et représentations ».

Liliane Hilaire Pérez et Anaïs Albert sont membres permanents du projet Time Us. Patricia Heisert, Audrey Millet et Kamila Adja y ont également pris part.

1.2.1.5 Centre Maurice Halbwachs

Le Centre Maurice Halbwachs (CMH)²⁵, a été créé en 2006 sur le campus de l'École Normale Supérieure de Paris. C'est une UMR placée sous la tutelle du CNRS, de l'ENS et de l'EHESS (UMR 8097). Il s'agit d'un centre de recherche en sociologie organisé autour de quatre équipes : ERIS (Équipe de recherche sur les inégalités sociales), ETT (Enquêtes, Terrains, Théorie), GRECO (Groupe de recherche sur la cohésion et la justice), et PRO (Profession, Réseaux, Organisations).

Coordonnée par Anne Lhuissier (ETT-INRA), l'équipe est également composée d'Alain Cotterau et de Stéphane Baciocchi (EHESS).

1.2.2 Un projet de trois ans en plusieurs étapes

Le projet Time Us tel qu'il a été soumis à l'ANR en 2016 prévoit plusieurs étapes de travail, aussi appelées « *work packages* ».

La première d'entre elles consiste en l'établissement d'un inventaire des sources à partir desquelles extraire des données pour le projet. Dès le début, cette tâche a été envisagée sur deux plans : régional et général. Il s'agit d'une part de produire un inventaire utile aux futurs projets de recherche menés sur un sujet similaire. D'autre part, conscients de la quantité de documents susceptibles d'être inventoriés, il est question d'établir une sélection de documents, de qualité particulière, à soumettre au traitement automatique de la langue après numérisation.

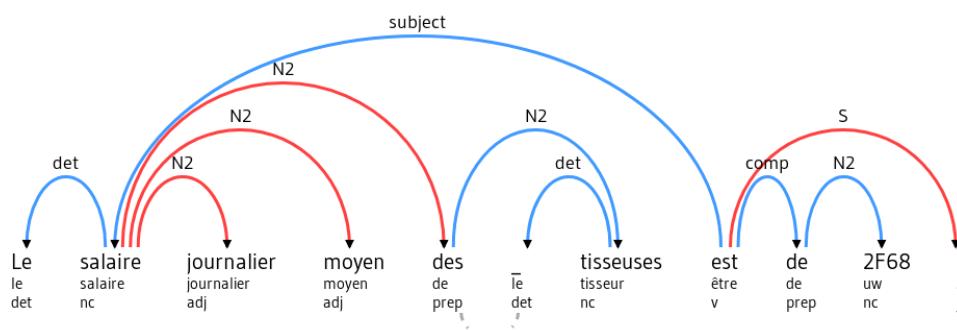
24. *Laboratoire ICT : Accueil*, URL : <http://www.ict.univ-paris-diderot.fr/> (visité le 03/08/2018).

25. *Centre Maurice Halbwachs : Accueil*, URL : <https://www.cmh.ens.fr/> (visité le 03/08/2018).

Après la création d'un inventaire et la sélection d'une partie des ressources identifiées vient la numérisation des documents. Celle-ci s'opère en deux étapes : d'une part la numérisation à proprement parler, c'est-à-dire la copie digitale des documents papiers ; et d'autre part, après transcription, l'encodage des contenus textuels extraits. En 2016, le projet prévoyait que cette transcription soit réalisée manuellement, en ayant recours à des contrats courts (5 à 6 par équipe sur la durée totale du projet).

Une fois numérisés, les documents font l'objet d'une troisième phase de traitement, à l'aide d'outils de traitement automatique créés par ALMAnaCH. Il s'agit en particulier de l'outil FrMG²⁶. Cette troisième phase inclut l'adaptation et l'étalonnage des outils de traitement automatique aux sources et aux objectifs du projet Time Us. Ces objectifs sont : la création à données sérielles contextualisées et l'extraction d'informations. Les outils de traitement automatiques doivent être capables de comprendre la grande variété des syntaxes et grammaires utilisées dans les documents.

FIGURE 1.1 – Graphe résultant d'une analyse syntaxique avec FrMG (schéma *depxml*).



L'analyse de ces données conduit à la production de connaissances scientifiques. Celles-ci sont diffusées par le biais de différents mécanismes éditoriaux : articles scientifiques, conférences, plate-forme de consultation en ligne, etc. Elles sont aussi produites à partir d'approches plus traditionnelles des sources, sans passer par le numérique comme outil d'exploitation.

Toutes les équipes participent à la visibilité du projet en cours et des résultats obtenus progressivement sous la forme de conférences ou d'articles publiés dans des ouvrages dédiés à la thématique²⁷. Le LARHRA a mis en place un blog sur la plate-forme Hy-

26. FrMG est une métagrammaire développée dans le cadre du projet ALPAGE depuis 2004. Elle est implémentée dans un analyseur syntaxique qui détecte la structure grammaticale d'une simple phrase ou de corpus complets. ALPAGE, *FRMG Wiki*, URL : <http://alpage.inria.fr/frmgwiki/> (visité le 03/08/2018) ; cf. figure 1.1.

27. Anaës Albert, « Consumption as "hidden transcript" in the 1917 midinettes' strike in Paris », *European Social Science History Conference*, Belfast, Irlande du Nord (4 avr. 2018) ; Manuela Martini, « Gendered division of work and wage conflicts in the Lyon silk trades at the end of the 19th century », *European Social Science History Conference*, Belfast, Irlande du Nord (4 avr. 2018) ; Anne Montenach, « "The Bazaar economy of trades" : gender and wage systems in the late seventeenth and eighteenth century Lyon textile industry », *European Social Science History Conference*, Belfast, Irlande du Nord

pothèses²⁸; il fonctionne comme un carnet de recherche en ligne. Dans le cadre de mes missions, j'ai participé à ce travail de communication des résultats. J'ai en effet été chargée de préparer la présentation données par l'équipe ALMAnaCH au congrès international *Digital Humanities* de 2018 qui s'est tenu à Mexico.²⁹.

1.2.3 Les principaux outils numériques du projet

Depuis 2016, et avec l'arrivée de Charles Riondet et Marie Puren, la question de la méthodologie et des outils employés pour la numérisation du corpus et leur traitement avant l'intervention des traitements automatiques a été approfondie. Elle a notamment conduit à opérer des choix de logiciels, de formats et de standards et à l'établissement de bonnes pratiques numériques à mettre en oeuvre.

1.2.3.1 MediaWiki

MediaWiki est un *Content Management System* (CMS), ou système de gestion de contenu, qui permet d'éditer et de publier des sites internet. Le CMS MediaWiki est développé dans le cadre des mouvements libre³⁰ et *open-source*³¹, en lien étroit avec le projet WikiPédia. La modularité du CMS est garantie par l'existence de très nombreuses extensions créées et maintenues par une communauté d'utilisateur·rices active.

Le projet Time Us a opté pour ce CMS dès 2017, avec l'intention de créer une plate-forme similaire à celle créée dans le cadre du projet anglais *Transcribe Bentham*³². Cette dernière est basée sur MediaWiki et se présente sous la forme d'une plate-forme de transcription collaborative appelée « Transcription Desk »³³. Cette configuration du CMS fait appel aux extensions **TEITags**³⁴ et **SemanticMediaWiki**³⁵ qui ont été installées à la création du site internet du projet Time Us³⁶. Toutefois, d'autres outils ont finalement été adopté pour réaliser la transcription collaborative des documents et le site tend a

(4 avr. 2018).

28. Une plate-forme pour la création de blogs académiques développée par OpenEdition. *Hypotheses*, URL : <https://hypotheses.org/> (visité le 12/08/2018).

29. Marie Puren, Alix Chagué, M. Martini, Éric de la Clergerie et Charles Riondet, « Creating gold data to understand the gender gap in the French textile trades (17th - 20th centuries) », *Digital Humanities*, Mexico City, Mexique (28 juin 2018). - cf. Annexes H.1.

30. Par opposition aux logiciels propriétaires.

31. Un mouvement qui vise à garantir la possibilité de distribuer librement des logiciels, d'accéder à leur code source et de créer des logiciels dérivés de ces codes sources.

32. University College London, *UCL Transcribe Bentham*, URL : <http://blogs.ucl.ac.uk/transcribe-bentham/> (visité le 04/06/2018).

33. *Transcription Desk - Technical Requirements*, Transcribe Bentham, URL : http://www.transcribe-bentham.da.ulcc.ac.uk/td/Technical_Requirements (visité le 12/08/2018).

34. *MediaWiki / Extension :TEITags*, URL : <https://www.mediawiki.org/wiki/Extension:TEITags> (visité le 12/08/2018).

35. *Semantic MediaWiki*, Semantic MediaWiki, URL : https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki (visité le 12/08/2018).

36. *Version*, TimeUsage, URL : <http://timeusage.paris.inria.fr/mediawiki/index.php/Sp%C3%A9cial:Version> (visité le 12/08/2018).

perdre cet usage initial. Il doit également servir de plate-forme de documentation du projet. Il a donc été conçu comme un « wiki », c'est-à-dire une application web, dont le contenu peut être édité par les visiteurs, ce qui permet de créer et modifier les pages de manière collaborative.

Enfin, le site doit faire l'objet de développements supplémentaires pour permettre son utilisation comme interface de consultation des documents numériques produits grâce à la transcription et à l'annotation. Dans cette optique, le module **DjVuLibre**³⁷ a été installé. Il permet la prise en charge des fichiers aux format DjVu. Sur le modèle d'un service comme Wikisource³⁸, il est envisagé d'organiser cette interface de consultation à partir de l'alignement de la transcription annotée des documents avec leur image.

FIGURE 1.2 – Capture d'écran d'une page de [wikisource.org](https://wikisource.org/w/index.php?title=Pierre_de_Coubertin_Souvenirs_d_Oxford_et_de_Cambridge_1887.djvu&oldid=10000000), en août 2018.

The screenshot shows a Wikipedia-like page for a scanned document titled "Pierre de Coubertin Souvenirs d Oxford et de Cambridge 1887.djvu/7". A green banner at the top indicates the page has been validated by two contributors. The main content area displays the scanned text of the book's title page, which reads "SOUVENIRS D'OXFORD ET DE CAMBRIDGE". Below this, the scanned text of the first page of the book is shown, containing French text about the universities of Oxford and Cambridge. A sidebar on the left contains various navigation links such as Accueil, Index des auteurs, Portails thématiques, and Aide. At the bottom of the page, there is a category link "Catégorie : Page validée" and a note about the last modification date.

Le wiki sert avant tout à publier des éléments de documentation sur et pour le projet en cours en produisant des articles pour les visiteur·ses extérieur·es mais aussi à l'attention des membres du projet Time Us. Y sont donc publiés les guides produits pour Time Us et établissent les bonnes pratiques et les consignes, comme nous aurons l'occasion d'y revenir.

37. *DjVuLibre*, URL : <http://djvu.sourceforge.net/> (visité le 12/08/2018).

38. Cf. figure 1.2.

Si la création de pages et de contenu est relativement accessible à n’importe quel·le utilisateur·rice initié·e au « wikicode »³⁹, la configuration du CMS requiert en revanche une maintenance et des compétences avancées, notamment en PHP.

1.2.3.2 TEI

La TEI (*Text Encoding Initiative*) est un schéma pour XML mis au point et maintenu depuis 1987 par le TEI Consortium. Son objectif est de permettre la description des textes et leur encodage de manière standardisée. La TEI permet de décrire les éléments de représentation du document, des éléments d’analyse du texte, et de rédiger une documentation sur le document numérique sous la forme de métadonnées. Les principes et le vocabulaire mis en oeuvre dans le cadre de ce standard sont exprimés dans des *guidelines* publiées sur le site internet de la TEI⁴⁰.

L’utilisation de ce standard pour l’édition numérique des documents dans le cadre de projets comme Time Us présente plusieurs avantages. S’agissant d’un standard largement appliqué et reconnu par la communauté, un tel usage garantit un haut degré de pérennité des documents ainsi que leur interopérabilité. Assurer cela signifie que les documents créés pour répondre aux objectifs du projet pourront aussi être ré-appropriés par d’autres équipes, avec des objectifs similaires ou tout à fait différents. Ainsi, Time Us contribue d’autant plus à la communauté des sciences humaines, que le projet crée des corpus pérennes et de qualité.

1.2.3.3 Transkribus

Transkribus est l’outil principal utilisé pour le projet Time Us. Il prend en charge la transcription et l’annotation des documents. Il s’agit d’une plate-forme développée dans le cadre du projet européen READ (*Recognition and Enrichment of Archival Documents*)⁴¹, une infrastructure financée par la Commission Européenne⁴², et dont l’objectif est de mettre en place des outils pour améliorer l’accès au contenu des objets archivistiques. Transkribus est le point de convergence de toutes les recherches menées par READ⁴³.

La plate-forme n’est disponible qu’en anglais. Ses principales fonctionnalités de traitement automatique sont la reconnaissance optique de caractères (OCR), la reconnaissance d’écriture manuscrite (HTR), l’analyse de la structure (*Layout Analysis*) et le repérage de mots (*Word Spotting*).

39. Il s’agit de la syntaxe utilisée pour rédiger et structurer les textes sur les sites supportés par MediaWiki.

40. TEI Consortium, *P5 : Guidelines for Electronic Text Encoding and Interchange*, 23 juil. 2018, URL : <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/> (visité le 03/08/2018).

41. *Transkribus*, READ Project, URL : <https://read.transkribus.eu/transkribus/> (visité le 06/06/2018).

42. Dans le cadre du programme Horizon 2020.

43. Voir en particulier : *Transkribus - Interfaces Maps*, URL : https://read.transkribus.eu/wp-content/uploads/2017/07/Interfaces_Map_v4.0.pdf (visité le 06/06/2018).

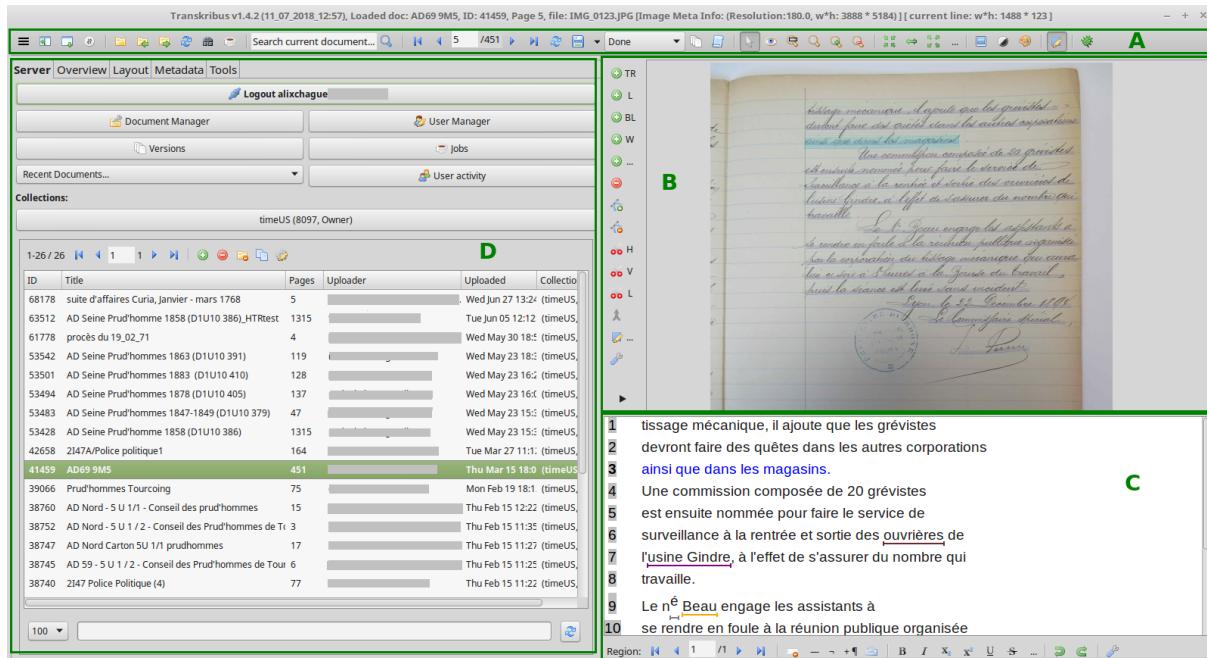


FIGURE 1.3 – Aperçu de l’interface graphique de Transkribus.

Lorsqu’un·e utilisateur·rice crée un compte, ses données sont synchronisées avec le serveur de la Transkribus. Cela rend possible le travail collaboratif et privé sur un ensemble de corpus de textes. Un·e utilisateur·rice propriétaire (*owner*) d’une collection peut inviter d’autres utilisateur·rices, en leur attribuant un rôle de propriétaire ou de lecteur (*reader*). Un·e utilisateur·rice propriétaire peut modifier la collection en ajoutant, modifiant ou supprimant des documents. Chaque document correspond à un versement. Il dispose d’un titre, d’un identifiant, d’un·e utilisateur·rice associé·e comme télèverseur·se (*uploader*), et peut être composé de plusieurs pages. L’interface ne permet pas de modifier le contenu d’une sous-collection ou d’en fusionner plusieurs.

L’interface graphique de Transkribus est composée de quatre zones.

1. **Une barre de menu (A)** en haut de la fenêtre permet d’accéder aux fonctionnalités principales, notamment le menu principal, la gestion du profil d’utilisation, l’import et l’export de fichiers ou encore l’actualisation des pages.
2. **Le canevas (B)**, où s’affiche l’image à transcrire. Lorsqu’un document comporte plusieurs images, elles sont affichées une par une. Le canevas permet de réaliser manuellement la segmentation du texte ou de l’éditer. La barre de menu est complétée d’outils de navigation au sein de la sous-collection et de boutons de zoom.
3. **L’éditeur de texte (C)**, où s’affiche la transcription du texte. Chaque ligne d’une zone de texte est numérotée et correspond à un segment de l’image. Une barre de menu en bas permet de modifier le style du texte (texte en gras, italique, souligné ou barré, en exposant ou en indice, etc).

4. Les **onglets** (D)⁴⁴ dans le panneau latéral gauche permettent de réaliser un très grand nombre de tâches : 1) gérer l'accès à la collection et à son contenu (*Server*), 2) avoir un aperçu du statut de chaque élément dans la sous-collection en cours de consultation (*Overview*), 3) gérer la structure et la segmentation de la page en cours de consultation (*Layout*), 4) gérer les métadonnées de la page en cours (*Metadata*), ou encore, 5) accéder aux différents outils de traitement automatique disponibles (*Tools*). Parmi ces outils, on trouve « *Layout Analysis* », pour le repérage automatique des zones et lignes de texte, « *Text Recognition* », pour la reconnaissance automatique du texte (HTR), ou encore « *Compute Accuracy* », pour l'analyse automatique des taux d'erreur d'une transcription automatique. Chacun de ces cinq onglets principaux dispose d'onglets ou d'options avancées.

La plate-forme propose également d'exporter la totalité ou une portion d'une sous-collection sous plusieurs formes :

- Sous forme de paquets liant fichiers de métadonnées (standard METS⁴⁵), fichiers de texte au format XML (standard ALTO⁴⁶ ou PAGE⁴⁷) et images (JPG).
- Sous forme de PDF, avec ou sans l'alignement de la transcription avec l'image, avec ou sans coloration des annotations.
- Sous forme de fichiers Word (DOCX), avec la possibilité de développer ou non les abréviations, de conserver ou non les coupures de lignes, et d'inclure ou non les éléments d'annotation comme les passages incertains.
- Sous forme de fichiers XML selon le standard de la TEI⁴⁸ avec la possibilité de paramétrier certains aspects du fichier de sortie, parmi lesquels la mise en forme des lignes (<1b/> ou <1>...</1>) ou encore le choix d'intégrer ou non les coordonnées des zones et lignes de texte.

Quel que soit le format d'export, Transkribus permet de paramétrier les fichiers de sortie en précisant les pages du document courant à exporter, ou encore en appliquant un traitement d'anonymisation des données signalées comme sensibles.

En définitive, l'un des principaux avantages de Transkribus est de proposer une seule et même interface pour réaliser 1) la segmentation des zones à transcrire, 2) la transcription –qu'elle soit manuelle ou automatique–, accompagnée d'un alignement simultanée de

44. Cf. Annexes B.1.

45. *Metadata Encoding Transmission Schema*, un standard créé pour conserver les métadonnées et la structure hiérarchique d'objets faisant partie d'une collection numérique, ainsi que les liens vers ces objets.

46. *Analyzed Layout and Text Object*, un standard pour le stockage des données techniques de description de la structure d'un document ayant fait l'objet d'un OCR. Il est généralement articulé avec un fichier METS.

47. *Page Analysis and Ground truth Elements*, un format pour le stockage des données de description de la structure d'un document OCRisé, ainsi que la transcription (idéale) associée à chacune des zones de texte ; Stefan Pletschacher et Apostolos Antonacopoulos, « The PAGE (Page Analysis and Ground-truth Elements) format framework », dans, 2010, p. 257–260, DOI : [10.1109/ICPR.2010.72](https://doi.org/10.1109/ICPR.2010.72).

48. *Text Encoding Initiative*, un standard pour la représentation numérique de textes.

l'image et du texte, ou encore 3) gérer les métadonnées. Enfin, le fait que l'ensemble des éléments ainsi produits peut être exporté dans une très grande variété de formats garantit la compatibilité de l'export avec la grande majorité des besoins des utilisateur·rices de Transkribus.

L'aspect « couteau suisse » de Transkribus est cependant aussi un inconvénient car l'utilisateur·rice se perd, dès lors qu'il·elle n'est pas techniquement compétent·e sur tous ces aspects de traitement des documents. La documentation fournie pour Transkribus est à la fois extrêmement prolixie et très limitée. Transkribus propose en effet plusieurs canaux pour mieux comprendre le logiciel. Le principal vecteur d'informations est un wiki⁴⁹. Dans l'ensemble, il rassemble des explications pour la prise en main de Transkribus par les débutant·es ou par des utilisateur·rices plus avancé·es sur la question de la reconnaissance automatique d'écriture et qui chercheraient, par exemple, à entraîner leur propre modèle. L'un des inconvénients de ce wiki est sa tendance à proposer une entrée vers la documentation par le biais des réponses plutôt que des questions. Il dispose également d'une section « FAQ »⁵⁰, mais celle-ci revient sur des aspects d'installation et de prise en main déjà évoqués dans le site. Il s'agit avant tout de répondre à des besoins de débugage pour des cas particuliers et de faire la promotion de Transkribus par le biais de questions du type « Comment créer des données d'entraînement dans Transkribus ?⁵¹ », « Quelle est la précision de l'HTR ?⁵² », etc. Ce sont des éléments auxquels Transkribus apporte déjà des réponses détaillées par un autre biais que le wiki : les « *How to guides* ». Ces guides sont des documents au format PDF qui fonctionnent comme des tutoriels pour guider l'utilisateur·rice à travers les différentes phases de traitement des documents, par exemple : « Comment utiliser Transkribus en 10 étapes », « Comment transcrire des documents avec Transkribus », « Comment entraîner un modèle de reconnaissance d'écriture manuscrite avec Transkribus », ou encore « Comment exporter des documents depuis Transkribus »⁵³. Il est parfois nécessaire d'avoir recours à un moteur de recherche pour accéder à des versions antérieures de ces guides car il arrive qu'ils contiennent des informations toujours valables qui ont disparu dans les versions les plus actuelles.

En dépit de la quantité apparente de documentation fournie par Transkribus, celle-ci est très répétitive et se concentre presque exclusivement sur les aspects de traitement automatique des documents et de mise en place de projet de transcription. Deux outils très utiles comme l'export des données/documents et la gestion des métadonnées sont très

49. *Home*, Transkribus, URL : https://transkribus.eu/wiki/index.php/Main_Page (visité le 12/08/2018).

50. *Questions and Answers*, Transkribus, URL : https://transkribus.eu/wiki/index.php/Questions_and_Answers (visité le 12/08/2018).

51. *How do I create training data in Transkribus?*

52. *How accurate is the HTR?*

53. Dans l'ordre : « *How To use Transkribus - in 10 steps (or less)* », « *How To Transcribe Documents with Transkribus* », « *How To Train a Handwritten Text Recognition Model in Transkribus* », « *How To Export documents from Transkribus* ».

peu abordés. Lorsqu'ils le sont, il n'est pas question d'expliciter les standards ou les choix techniques qui sous-tendent leur mise en place. En outre, dans l'ensemble, les choix de terminologie et le *design* de la plate-forme ne permettent pas de compenser le manque de documentation.

En conclusion, Transkribus est un outil très prometteur, et qui réalise convenablement un certain nombre de tâches de traitement du texte. Mais il est d'autant plus frustrant à utiliser que plusieurs fonctionnalités dont on apprécie l'existence sont mal documentées.

Chapitre 2

Les sources du projet

Au printemps 2018, une part importante de la phase de sondage et d'inventaire des sources était réalisée. Elle a été menée par l'ensemble des équipes impliquées dans le projet, pour les aires géographiques ou les corpus qui les concernaient.

L'équipe de l'IRHIS a ainsi travaillé sur les fonds des Archives départementales du Nord et l'équipe ICT sur ceux des Archives de la ville de Paris. TELEMM a dépouillé les fonds des Archives municipales de Marseille et des A. D. des Bouches-du-Rhône, mais également ceux des A. M. de Lyon. Enfin, l'équipe du LARHRA a étudié ceux des A. D. du Rhône, ceux des A. M. de Lyon, mais aussi ceux du Musée des tissus et de la Bibliothèque Municipale de Lyon.

Plusieurs types de documents ont été identifiés et inventoriés.

2.1 Typologie des sources

2.1.1 Archives des Conseils de Prud'hommes

Les sources judiciaires constituent la majeure partie des sources inventoriées dans le cadre du projet. Il s'agit en premier lieu de documents produits dans le cadre de l'exercice des Conseils de Prud'hommes des différentes villes et de leurs équivalents antérieurs, pour la période moderne.

Les Conseils de Prud'hommes contemporains sont apparus en France, tout d'abord à Lyon, à partir de 1806, dans le but de régler les nombreux contentieux dans le secteur du textile, un secteur qui domine les activités de la ville au début du XIX^e siècle. Dans l'exposition virtuelle créée par le LARHRA à l'occasion du bicentenaire de la création du premier Conseil de Prud'hommes¹, plusieurs périodes de création de Conseils de Prud'hommes dans les autres villes françaises sont définies pour la Franche. La première

1. LARHRA, Anne-Catherine Marin, François Robert et Pierre Vernus, *Bicentenaire du premier conseil des prud'hommes ; Lyon 1806-2006*, avec la coll. de Clémentine Breed, Delphine Digout et Jean-luc Bouville, URL : <http://larhra.ish-lyon.cnrs.fr/cdeprudhomme2/index.htm> (visité le 17/07/2018).

s’achève en 1848. Ces créations concernent alors avant tout les centres textiles. Le Nord fait partie de la première vague, un Conseil de Prud’hommes est créé à Tourcoing dès 1821. A Paris, il n’est créé qu’en 1847 pour l’industrie des tissus, alors qu’il l’est dès 1845 pour l’industrie des métaux. Il existe des formes plus anciennes d’institutions équivalentes aux Conseils de Prud’hommes dans de nombreuses villes, notamment à Marseille et à Lyon. On en trouve des archives sous la forme de comptes-rendus de jugements donnés par les tribunaux de police, et par le biais de registres de contraventions.

Les comptes-rendus de conciliations prud’homales sont des sources intéressantes car ils contiennent de très nombreuses informations sur les rémunérations et l’organisation du travail. Les récits des conflits et la présentation du contexte judiciaire d’un jugement présentent l’identité des protagonistes de l’affaire, ainsi que leur statut social et/ou matrimonial, leur relations et la nature de l’engagement qui les lie, le temps estimé nécessaire à la réalisation d’une pièce ou d’une tâche et celui effectif, etc. En outre, les jugements induisent également que le Conseil définit une valeur monétaire pour certains aspects non rémunérés de la vie des travailleur·ses du textile par le biais des régimes d’indemnité. La richesse et la précision des informations contenues dans ces archives doivent être relativisées car ces compte-rendus sont toujours le résultat de situations exceptionnelles dans les relations de travail. En effet, ce sont toujours des situations de conflits qui conduisent à un jugement rendu par le Conseil de Prud’hommes, et la majeure partie de ces affaires est consignée avant d’en arriver au jugement. Les jugements prud’homaux ne révèlent donc que des situations de travail conflictuelles et extrêmes.

Les archives du Conseil de Prud’hommes de Tourcoing² courent de 1821 jusqu’à la fin de la période concernée par le projet, soit le début de la Première Guerre mondiale. Elles ont été entièrement dépouillées par l’IRHIS. Cette équipe a fait le choix de trier les affaires pour ne conserver que celles qui concernent au moins une femme en tant que protagoniste.

La stratégie adoptée par l’ICT pour les jugements des Prud’hommes de Paris est différente. L’ensemble des compte-rendus de jugements prud’homaux pour l’industrie des tissus pour l’année 1858³ a été dépouillé, sans trier les jugements. La quantité de jugements rendus à Paris a conduit l’équipe à se concentrer sur deux années : 1858 et une année au tournant des années 1870⁴. Il s’agit de se doter de points précis de comparaison entre deux périodes économiques différentes à Paris, mais aussi avec les situations des autres aires géographiques étudiées à la même époque.

L’équipe de TELEMMé a réalisé plusieurs sondages, aux A.D. des Bouches-du-Rhône et aux A.M. de Marseille, parmi les archives des notaires et des tribunaux de commerce

2. Lille, A. D. du Nord, 5U1-1 à -4, Minutes des jugements prud’homaux [sic], 1823-1914.

3. Paris, Archives de Paris, D1U10 1 - 1053, Conseil de Prud’hommes de la Seine puis de Paris, 1844-1940.

4. Celle-ci n’a pas encore été déterminée par l’équipe, mais cette période correspond à une deuxième phase d’industrialisation du secteur textile.

pour la période moderne, à partir de 1701 et jusqu'à 1792. Ces sondages ont conduit l'équipe à abandonner le dépouillement de ces sources car les femmes ne s'y trouvaient pas assez présentes compte tenu des ambitions du projet. L'équipe a toutefois dépouillé des fonds juridiques pour la période moderne, aux A.M. de Lyon, tirés des registres de contraventions aux règlements des métiers⁵, en particulier, à ce stade du projet, pour les années 1670, 1770, 1776 et 1780⁶.

A Lyon, il ne subsiste pas d'archives des Conseils de Prud'hommes pour l'industrie de la soie et du textile pour tout le XIX^e siècle. Elles ont disparu. Certains éléments sont toutefois accessibles grâce à des archives personnelles, comme celles de Pierre Charnier, chef d'atelier tisseur et conseiller prud'homal durant la première moitié du XIX^e siècle⁷. Un aperçu de ces jugements est également donné par les comptes-rendus des Conseils de Prud'hommes publiés dans la presse à partir des années 1840. Marie Lauricella et moi-même avons mené ce travail de dépouillement des archives de presse à partir du fond de presse lyonnaise de la B.M. de Lyon⁸.

2.1.2 Archives de police et de préfecture

A Lyon, un important corpus de documents a été inventorié à partir de la série M dédiée à l'Administration générale et à l'économie, aux A.D. du Rhône. Il s'agit en particulier des sous-séries dédiées à l'Industrie, à la Police et à la Préfecture⁹.

Ces fonds participent à la documentation du monde ouvrier au XIX^e. Ils contiennent de nombreux rapports et enquêtes sur les fabriques lyonnaises et leur fonctionnement. Dans un contexte d'émergence et de renforcement des associations ouvrières et des syndicats, la police et la préfecture rendent comptes du déroulement des grèves, des tensions entre ouvriers et patrons ou chefs d'ateliers. Les fonds rassemblent également affiches, pétitions et compte-rendus d'assemblées générales qui permettent d'accéder à la parole ouvrière. On trouve également dans ces fonds des données statistiques et des recensements qui distinguent souvent clairement hommes et femmes et fournissent des données précieuses pour articuler les rémunérations et le travail des hommes et des femmes.

2.1.3 Archives privées d'entreprise

A Lille, Paris et Lyon, des archives privées d'ateliers ou de fabriques de textile ont été inventoriées. Elles proviennent de fonds publics administratifs, notamment ceux des tribunaux de commerce, rassemblées à l'occasion de faillites. Il s'agit souvent de livrets

5. Lyon, A. M. de Lyon, HH, Commerce et Industries - en particulier 53 registres recensant les contraventions à la police des arts et métiers entre 1667 et 1781.

6. Lyon, A.M. de Lyon, respectivement, HH214, HH260, HH265 et HH267.

7. Lyon, B. M. de Lyon, fonds Fernand Rude, archives de Pierre Charnier.

8. Lyon, B.M. de Lyon, fonds Presse Lyonnaise.

9. Lyon, A.D. du Rhône, respectivement, 9M, 4M et 10M.

de comptabilité.

Dans les archives de Paris, dans la sous-série dédiée aux justices de paix du 5e arrondissement nouveau (1800-1958), et au tribunal d'instance du même arrondissement¹⁰, on trouve une section de papiers déposés, parmi lesquels un dossier sur la faillite du filateur Dupuis-Drouet (1821-1822)¹¹, et un registre comptable du même atelier pour la période allant de 1817 à 1821¹². A Lyon, ce sont les archives du magasin Papleux-Labaty, contenus dans le fond Labaty-Penelon¹³, qui permettent de retrouver, outre les livrets de comptes, des actes de société, des correspondances ainsi qu'un ensemble de documents de gestion courante tels que des inventaires, des factures, des reçus, etc.

Aux A.D. du Rhône, les archives du Tribunal de Commerce de Lyon¹⁴ contiennent plusieurs dossiers dédiés aux faillites de fabriques ou de marchands de textiles. On y trouve des bilans de créance. Outre ces dossiers de faillite, ce fonds contient également des livrets d'ouvriers¹⁵, ou encore les archives du marchand et fabricant de soie Cochet¹⁶, à savoir plusieurs livrets de compte et un livre de commission.

2.1.4 Tarifs

Les tarifs sont utilisés comme des documents de référence pour le calcul des rémunérations, en particulier pour les travaux payés à la pièce. Ces grilles de prix sont le fruit de négociations opérées entre les représentants des ouvriers, les chefs d'ateliers, et les négo-ciants, qui achètent les pièces. Ils peuvent être placardés ou publiés dans la presse. C'est d'ailleurs le cas, par exemple, des tarifs de 1832, applicables à partir du 1^{er} novembre de la même année, qui sont intégralement reproduits dans le premier numéro du journal *L'Écho de la Fabrique*¹⁷, accompagné du récit des négociations qui y ont conduit.

Les tarifs sont à l'origine de nombreux conflits lorsqu'ils ne sont pas appliqués ou lorsqu'ils sont jugés trop bas. Le « tarif de 1885 » est ainsi à l'origine d'une série de grèves à Lyon en décembre 1894. On en trouve de nombreuses traces dans un dossier de la sous-sous-série 9M5 des A.D. du Rhône dédié aux grèves de 1894-1895¹⁸. Dans une minute adressée au Procureur de la République, rédigée le 19 décembre 1894, un certain M. Giudicelli raconte ainsi :

« L'origine du conflit est la revendication de la part des ouvriers tisseurs à bras de l'application du tarif accepté par les patrons en 1885, tarif dont les

-
- 10. Paris, A.P., D12U1, Justices de paix de Paris. Tribunal de simple de police de Paris.
 - 11. Paris, A.P., D12U1 - 375.
 - 12. Paris, A.P., D12U1 - 376.
 - 13. Lyon, A.M. de Lyon, 26 II, Fonds Labaty-Penelon.
 - 14. Lyon, A.D. du Rhône, 6U, Tribunal de Commerce.
 - 15. Lyon, A.D. du Rhône, 6Up 1 - 3673.
 - 16. Lyon, A.D. du Rhône, 6Up 3664-3667, COCHET fabricant et marchand de soieries. an 10-an 12.
 - 17. *L'Écho de la Fabrique*, n°1, 30 octobre 1831 ; cf. Annexes A.1.4.
 - 18. Lyon, A.D. du Rhône, 9M5, premier dossier, Crise de l'industrie du tissage lyonnais, grèves, enquêtes administrative sur la situation (novembre 1894-juillet 1895).

prix, de 30% plus élevés que ceux payés réellement, n'ont jamais été appliqués mais que les ouvriers n'exigeaient pas à cause de la pénurie des affaires jusqu'à ces derniers temps.

Les nombreuses réunions des ouvriers et chefs d'atelier de divers syndicats, suivies d'entrevues avec les patrons, n'ayant pas abouti à une entente, la grève partielle a été votée à une grande majorité de préférence à la grève générale qui a été repoussée à la presque unanimité ; »

Les tarifs sont des sources précieuses pour les historiens car ils permettent d'accéder à des données sérielles relativement complètes, bien que leur application ne soit pas garantie.

2.1.5 Presse ouvrière

La presse ouvrière qui se développe au XIX^e siècle, constitue une archive de la parole ouvrière, quoiqu'elle fut souvent celle des chefs d'atelier et non celle des ouvrières les moins qualifiés. Outre les comptes-rendus des Conseils de Prud'hommes qui y sont publiés, cette presse rassemble un certain nombre d'articles qui décrivent les conditions de vie et de travail des ouvrières du secteur, et qui les budgétisent. S'y trouvent également des tribunes et des sections qui donnent un aperçu de l'organisation des Conseils de Prud'hommes au fil de leurs réformes, ainsi que de l'organisation progressive de la représentation politique des ouvrières. Enfin, on y trouve des témoignages sur les différents objets de tensions politiques et syndicales qui secouent de temps à autre la classe ouvrière, y compris en ce qui concerne la liberté de cette presse ouvrière.

Les numéros collectés dans le cadre du projet Time Us proviennent principalement du fond de presse régionale de la B.M. de Lyon. Les titres de presse qu'il rassemble sont en grande partie numérisés et accessibles depuis la bibliothèque numérique de l'institution : Numelyo¹⁹. Pour compléter les titres manquants ou incomplets, la bibliothèque numérique de la Bibliothèque nationale de France, Gallica²⁰, a parfois été utilisée.²¹

2.1.6 Enquêtes sociologiques

Les monographies de Le Play constituent un corpus important, traité à part dans le cadre du projet Time Us. Frédéric Le Play (1806-1882), haut fonctionnaire français, est considéré comme l'un des fondateurs de la sociologie²². Il a créé la Société d'Économie Sociale²³ en 1856. C'est par elle que sont publiées 164 enquêtes sociologiques réalisées

19. *Presse lyonnaise de 1790 à 1944*, Numelyo, URL : <http://collections.bm-lyon.fr/PER003> (visité le 17/05/2018).

20. *Presse locale et régionale du Rhône*, Gallica, URL : [/html/und/presse-et-revues/rhone](http://html/und/presse-et-revues/rhone) (visité le 17/05/2018).

21. Cf. Annexes A.1.5.

22. *FRÉDÉRIC LE PLAY*, Encyclopædia Universalis, URL : <http://www.universalis.fr/encyclopedie/frederic-le-play/> (visité le 19/07/2018).

23. Aussi appelée Société Internationale des Études Pratiques d'Économie Sociale.

par Le Play lui-même ou d'après son modèle. Elles paraissent entre 1855 et 1930²⁴ dans deux ensembles : d'une part *Ouvriers des deux mondes* (1857-1930), d'autre part *Ouvriers européens* (1855 pour la première édition, 1877-1879 pour la seconde). C'est le premier ensemble qui intéresse le projet de recherche et sur lequel travaille l'équipe basée au CMH. *Ouvriers des deux mondes* a été publié sous forme de séries, divisées en plusieurs volumes. Nous traitons en particulier des trois premières séries, publiées entre 1857 et 1908. Elles rassemblent 108 de ces enquêtes.

Chaque enquête est une monographie, c'est-à-dire une « étude scientifique, minutieuse et détaillée, portant sur un objet ou un phénomène circonscrit »²⁵, concentrée sur une famille caractérisée par l'activité professionnelle de son « chef ». La première enquête de cet ensemble s'intitule ainsi « Charpentier de Paris (Seine.- France) de la corporation des compagnons du devoir ». En suivant un plan rigoureux, chaque enquête expose les différents aspects qui caractérisent la vie de ces familles : ses membres et leurs occupations, son histoire, ses moeurs, ses habitudes de vie, ses possessions mobilières et immobilières, etc. Ces descriptions sont en outre accompagnées de budgets très détaillés qui comptabilisent toutes les recettes et dépenses d'une famille pendant une année.

Bien que seules dix enquêtes s'attachent à décrire des familles travaillant dans l'industrie du textile²⁶ parmi les 108 de l'ensemble, toutes sont intéressantes car y sont budgétisés et décrits les usages que font ces foyers des matières textiles, notamment les linge et les vêtements. De nombreuses tâches qui s'y rapportent sont de l'ordre du travail domestique non rémunéré : c'est le cas de la fabrication, de la réparation ou du blanchiment du linge et des vêtements. A ce titre, ils intéressent le projet Time Us.

2.2 Collecte et numérisation des sources

Le 1^{er} juin 2018, à l'occasion de la rédaction d'un rapport de mi-parcours pour l'ANR, les représentants des équipes du projet Time Us se sont réunis à Inria pour dresser un bilan du corpus qui avait d'ores et déjà été identifié et des pistes d'accroissement possibles. A Lyon, bien que de très nombreux fonds aient déjà été dépouillés, les archives de la Chambre de Commerce attendent encore de l'être : elles sont inaccessibles en raison de la campagne de reclassement dont elles font l'objet jusqu'à la fin de l'année 2018. Du côté de l'IRHIS, le travail réalisé a permis d'évaluer l'intérêt des justices de paix de Roubaix, dont les archives des années 1847 et 1875 ont déjà été dépouillées, et celui des jugements du tribunal de

24. Louis Hincker, « Les monographies de famille de l'École de Le Play. Les Études sociales, n° 131-132, 1^{er} et 2^e semestres 2000. » *Revue d'histoire du XIXe siècle. Société d'histoire de la révolution de 1848 et des révolutions du XIXe siècle*-23 (1^{er} déc. 2001), p. 274-276, URL : <http://journals.openedition.org/rh19/334> (visité le 19/07/2018).

25. Antoine Savoye, « La monographie sociologique : jalons pour son histoire (1855-1974) », *Les Études Sociales : monographies de familles de l'École de Le Play*-131 (2000).

26. Voir en particulier les enquêtes n°7, 13, 20, 24, 36, 55, 67, 79, 83, 97, 104, 106, 109 et 111 ; cf. Annexes A.1.3.

simple police de la même ville, pour lesquels des sondages seront effectués à partir de la rentrée de septembre 2018. A Paris, enfin, des archives du Conseil de Prud'hommes doivent faire l'objet d'un deuxième dépouillement.

A ce stade du projet, la phase de sondage et d'inventaire est encore relativement loin d'être terminée. Cependant, l'achèvement de cette phase de travail ne conditionne pas le lancement des étapes suivantes. Le travail de numérisation des sources a d'ailleurs été réalisé en même temps qu'une partie du travail de dépouillement. Il convient de revenir sur les conditions pratiques et méthodologique de cette numérisation.

2.2.1 Résultats des premières campagnes de numérisation

La numérisation des documents est envisagée comme une phase intermédiaire, dont la vocation première est de produire des images à partir desquelles réaliser une transcription des textes. Puisque cette transcription devait initialement être faite à la main, la copie numérique des documents permettait de ne pas dépendre des contraintes géographiques et temporelles de leur centres de conservation et de celles des transcripteurs·rices. Toutes les sources inventoriées n'ont pas été numérisées. Les équipes ont opéré une sélection au sein de leurs corpus en se basant sur leurs propres objectifs de recherche, et ceux, généraux, du projet Time Us, mais aussi en veillant à rendre possible la comparaison entre leurs données et celles des autres équipes.

Les images obtenues dans le cadre de cette numérisation sont de qualité réduite. En effet, compte tenu des enjeux associés à cette campagne, il n'était pas nécessaire qu'elle soit réalisée par des professionnels, ni avec un matériel de pointe. Ce sont les personnes missionnées pour consulter les archives, c'est-à-dire les stagiaires, les vacataires ou les chercheurs, qui ont réalisé ces copies numériques, à l'aide d'appareils photo numériques « compact » ou bien grâce aux appareils photo équipant leurs téléphones portables. L'analyse des métadonnées des photographies permet d'identifier au moins neuf appareils différents²⁷. En outre, l'espace de prise de vue est celui de la consultation : les preneur·ses de vues sont toujours dépendant·es des horaires d'ouverture des centres d'archives, de la luminosité ambiante, et de l'espace et du recul permis par le lieu. Sans obligation de résultat autre qu'une photo suffisamment nette pour qu'elle soit transcrive par un humain, et avec l'objectif de photographier un maximum de documents le plus vite possible, la qualité générale de ces photos est relativement basse si l'on compare avec les standards d'une campagne de numérisation réalisée dans un cadre institutionnel. La qualité du cadrage dépend largement de la configuration du document et de la volonté du ou de la preneur·se de vue de réaliser un cadrage de la meilleure qualité possible.

Les fichiers numériques sont toujours au format JPEG²⁸. En fonction des appareils

27. cf. Annexes A.2.2.

28. Pour *Joint Photographic Experts Group*, il s'agit d'une norme d'enregistrement compressé de fichiers images numériques développé par le groupe du même nom depuis 1992. Pour des campagnes de

utilisés, la résolution des images peut varier de 72 dpi (cas le plus fréquent) à 300 dpi²⁹. La taille des images varie énormément, les plus grands peuvent aller jusqu'à 6000 pixels de hauteur pour 4000 pixels de largeur, et les plus petits, 1600 pixels de hauteur pour 1064 pixels de largeur. Il en va de même pour les fichiers qui varient de quelques centaines de kilo-octets à plusieurs méga-octets. Les images numériques n'ont pas fait l'objet d'un traitement ultérieur qui aurait visé à corriger les cadrages en ôtant les éléments parasites tels que les mains, les doigts et les divers objets présents dans le champ de l'appareil photo. Il peut aussi s'agir des pages adjacentes et de leurs contenu. Enfin, ce traitement peut aussi viser à diminuer l'effet déformant de l'angle de prise de vue pour rétablir la rectitude des documents et surtout des lignes. Cela est souvent nécessaire lorsque les éléments photographiés sont des pages d'un registre.³⁰

En juillet 2018, l'équipe ALMAnaCH a acquis une « *ScanTent* » par l'intermédiaire de READ. Il s'agit d'un dispositif prenant la forme d'une plaque souple d'environ 80 x 50 cm surmontée d'une petite tente à trois parois permettant de créer un espace de luminosité contrôlé pour la photographie de documents. Coiffée d'une petite plate-forme percée d'un trou, la tente permet en outre de stabiliser l'appareil de prise de vue et de le maintenir dans une position perpendiculaire au document. Cette plate-forme est prévue spécifiquement pour un *smartphone* car le dispositif fonctionne de paire avec l'application mobile *DocScan*, qui permet de réaliser des photographies de document déjà cadrées. L'application déclenche elle-même la prise de vue en détectant le changement de page, ce qui permet à son utilisateur·rice de tenir le document à deux mains si nécessaire. L'ensemble du dispositif³¹ a été développé par READ, en coopération avec Computer Vision Lab, TU Wien et l'Université d'Innsbruck. Il a été présenté le 8 juin 2018 à l'occasion d'un « scanathon »³² organisé simultanément aux Archives d'État de Zurich, en Suisse, aux Archives Nationale de Finlande à Helsinki et aux Archives Nationales du Royaume-Uni à Londres. Si elle est utilisée par les membres du projet, la *ScanTent* devrait donc permettre d'obtenir des images mieux cadrées et de réduire le temps de prise de vue pour les prochaines campagnes de numérisations. En revanche, la qualité des fichiers numériques restera limitée, étant donné qu'il implique d'utiliser un *smartphone*.

2.2.2 Collecte de documents en ligne

Deux ensembles du corpus ont été collectés en ligne. Il s'agit en premier lieu de la presse ouvrière lyonnaise, numérisée par la B.M. de Lyon et disponible sur sa bibliothèque numérisation d'archives dans un cadre institutionnel, on évite normalement les formats compressés.

29. Le dpi, pour « *dot per inch* », aussi appelé ppp, pour « point par pixel », définit la résolution d'une image : plus ce nombre est élevé, plus l'est la qualité.

30. Cf. Annexes A.1.2.

31. *ScanTent* et *DocScan* sont présentés sur la page : *The ScanTent*, URL : <https://scantent.cvl.tuwien.ac.at/en/> (visité le 12/08/2018).

32. *Join us at the 2018 Scanathon in London, Zurich and Helsinki!*, READ Project, URL : <https://read.transkribus.eu/2018/05/04/join-us-at-the-2018-scanathon/> (visité le 12/08/2018).

numérique Numelyo. S'il est possible de consulter en ligne une copie numérique en couleur et de haute résolution, la version proposée au téléchargement est un fichier au format PDF composé d'images en noir et blanc, probablement numérisées depuis des microfilms. Le texte des images est aligné avec une transcription issue d'une OCRisation dont la qualité est entièrement relative à celle de l'image.

Le second ensemble est constitué des enquêtes sociologiques leplaysiennes. Plusieurs copies numérisées sont disponibles sur Gallica, sur *Google Books* ou encore sur *Internet Archives*. Cette collecte a été réalisée conjointement par l'équipe du CMH et celle d'ALMAnaCH. *Internet Archives* propose au téléchargement des copies du corpus sous de nombreux formats :

- **JP2**, ou JPEG 2000, une norme de compression pour l'enregistrement d'images sans perte, développé par le *Joint Photographic Experts Group* ;
- **DjVu**, un format de fichier pour l'enregistrement compressé et l'archivage des documents numériques, développé par AT&T depuis 1996 ;
- **PNG**, pour *Portable Network Graphics*, un format ouvert de compression sans perte pour les images numériques, développé par le W3C depuis 1996 ;
- **GIF**, pour *Graphics Interchange Format*, un format de compression sans perte pour les images numériques, initialement propriétaire et devenu ouvert, développé par CompuServe depuis 1987 ;
- **PDF**, pour *Portable Document Format*, un format compressé de fichiers numériques permettant de préserver la mise en forme d'un document (police d'écriture, images, etc), développé par Adobe Systems ;
- **EPUB**, pour *Electronic PUBLication*, un format ouvert et standardisé pour les livres numériques, fondé sur XML, développé par l'International Digital Publishing Forum depuis 2007 ;
- **KINDLE**, qui fait en réalité référence à plusieurs formats propriétaires pour l'enregistrement des livres numériques, développés par Amazon ;
- **TXT**, un format de fichier de texte brut ;
- **XML-ABBYY**, un standard XML développé pour les logiciels de l'entreprise ABBYY, notamment FineReader.

Dans un premier temps, les formats JP2 et DjVu ont été favorisés. Pour cet ensemble également la qualité des rendus textuels est très variable.

2.2.3 Stockage des documents numériques

En mai 2018, sur l'initiative de l'équipe ALMAnaCH, le projet Time Us s'est doté d'un espace de stockage grâce au service *ShareDocs* développé et mis à disposition par la

TGIR Huma-Num³³.

La création d'un tel espace de stockage a permis la centralisation de l'ensemble des images et fichiers numériques collectés. Cette centralisation s'est avérée nécessaire pour deux raisons. Premièrement, en l'absence d'une chaîne de traitement pré-établie des fichiers numériques, les différents corpus soit étaient versés directement sur la plate-forme Transkribus, soit étaient envoyés aux autres membres du projets mais sans qu'un·e destinataire précise·e et constant·e de ces fichiers n'ait été désigné·e, soit n'étaient pas transmis aux autres équipes. L'absence d'un espace de stockage commun pour ces fichiers les exposait au risque d'être perdus ou détruits. Deuxièmement, à une étape du projet où il devenait nécessaire de dresser un bilan de la campagne de numérisation, la centralisation des fichiers permettait d'obtenir un aperçu clair de l'état d'avancement de la collecte. En juillet 2018, Time Us avait collecté ou créé plus de 18 500 fichiers images ou PDF dans le cadre de sa collecte de documents d'archives pour l'étude des rémunérations et des budgets-temps des ouvrièr·es du textile. S'y ajoutent les près de 8 000 fichiers DJVU ou JP2 que constituent l'ensemble des numérisations des enquêtes sociologiques de Le Play réalisées entre 1857 et 1908.

ShareDocs n'est pas la seule solution de stockage de fichiers partagé. Il en existe d'autres, plus connus, comme DropBox³⁴ ou *Google Drive*³⁵, mais ce sont des services propriétaires dont l'utilisation dans le cadre du projet induirait des frais³⁶. Le 6 juillet 2018, le dossier partagé du projet sur ShareDocs cumulait déjà 86 Go de fichiers. Il existe des équivalents libres et *open source*, comme Pydio³⁷ et Syncanny³⁸, cependant, ils nécessitent une maintenance que Time Us n'est pas en mesure d'assurer.

Le service de stockage proposé par Huma-Num présentent plusieurs avantages. L'infrastructure est connue de l'équipe ALMAnaCH, car ils sont partenaires dans le cadre du projet Parthenos. S'agissant d'une infrastructure nationale et ses serveurs étant localisés sur le territoire français, le respect du droit européen sur les données est garanti. Le service est gratuit et l'espace est alloué sur demande et sur mesure : Time Us a donc obtenu d'emblée 100Go d'espace de stockage. Enfin, Huma-Num est une infrastructure engagée dans la communauté des humanités numériques, elle propose d'autres outils utiles au projet, parfois intégrés à la plate-forme de stockage. Nous aurons l'occasion de revenir sur ce point.

Afin d'organiser la centralisation des fichiers sur l'espace de stockage, il a été néces-

33. Très Grande Infrastructure de Recherche pour les Sciences Humaines et Sociales, sous la tutelle du Ministère en charge de l'Enseignement Supérieur et de la Recherche et créée en 2013.

34. Développé par DropBox Inc.

35. Développé par Google Inc.

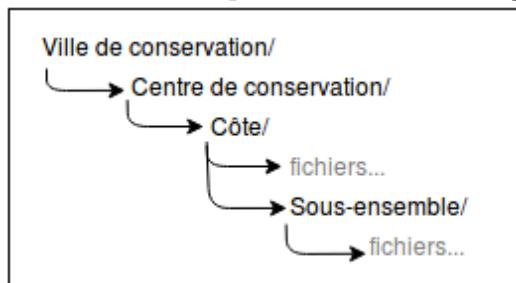
36. 10€/mois pour le forfait « standard » qui permet 2 To de stockage au lieu des 2 Go du forfait gratuit « basic », 2€/mois chez Google Drive pour 100Go au lieu des 15 Go gratuits, 10€/mois pour 2 To (tarifs au 31/07/2018).

37. Développé par Pydio.

38. Développé, entre autres, par Philipp C. Heckel.

saire d'établir un protocole de nommage des fichiers³⁹. Nous avons fait le choix d'organiser cet espace en fonction des lieux de conservation des documents, en reproduisant autant que possible l'état des fonds collectés. Un premier niveau de dossier correspond au centre urbain ciblé (Lille, Paris, Lyon, Marseille). Dans chacun de ces dossiers, un niveau dossier intermédiaire identifie l'institution de conservation. Un troisième niveau de dossier permet de reconstituer les ensembles en fonction des fonds dont ils sont issus : chaque dossier correspond à une côte. Éventuellement, des niveaux supplémentaires peuvent être créés à l'intérieur de chacun de ces dossiers. Cette organisation permet d'échapper à une campagne de renommage des fichiers qui aurait autrement été nécessaire, au risque de perdre les correspondances déjà établies par les chercheurs entre leurs documents de travail et le fichier numérique.

FIGURE 2.1 – Schématisation du protocole de nommage dans ShareDocs.



Outre les dossiers dédiés aux centres urbains, un dossier reçoit le corpus des *Ouvriers des deux mondes*. L'espace de stockage est également utilisé comme un espace de partage de documents de travail. A l'occasion de la réunion organisée le 1^{er} juin pour le bilan de mi-parcours, les comptes-rendus et documents de présentation ont ainsi pu être mis en commun.

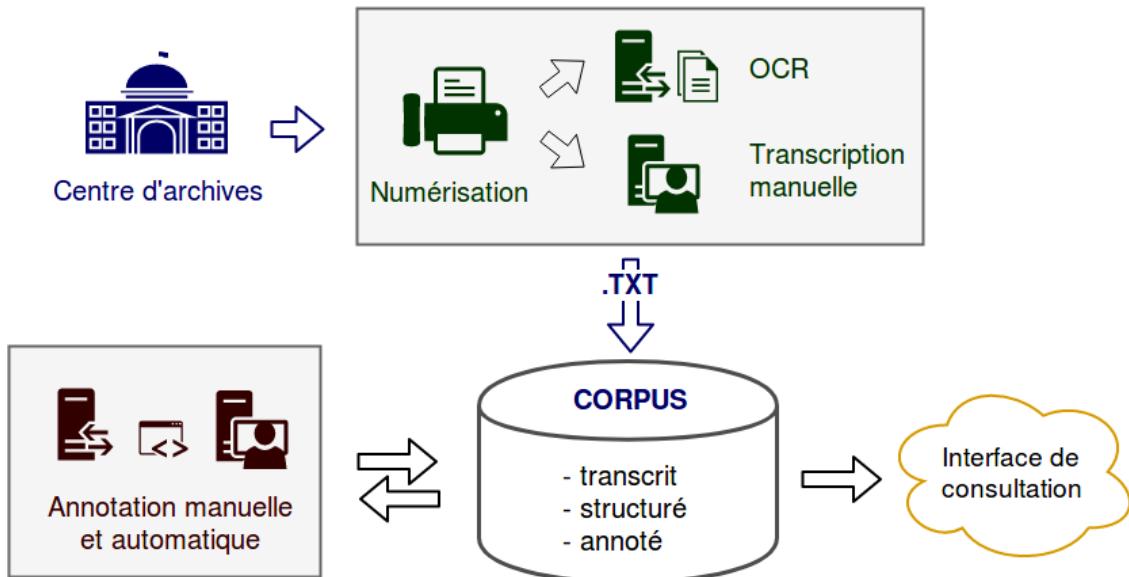
39. Cf. figure 2.1.

Deuxième partie

Des images aux fichiers TEI

Les missions du stage qui a donné lieu à la rédaction de ce mémoire se sont concentrées sur les étapes de traitement des données textuelles extraites des fichiers numérisés. Il s'est agi, d'une part de comprendre le cadre logiciel et méthodologique dans lequel est réalisée la transcription des documents, et d'autre part, de créer un cadre pour organiser l'annotation des textes afin d'en extraire les données. Cette extraction des données doit être réalisée de manière à permettre des traitements supplémentaires sur les fichiers, à rendre possible leur analyse par les historiens, et à les rendre consultables sur une interface dédiée.

FIGURE 2.2 – Schématisation de la chaîne de traitement des documents.



En juillet 2018, la question du traitement des fichiers transcrits, structurés et annotés est encore en pleine exploration. Mes missions se sont avant tout concentrées sur la question de l'annotation et de l'extraction des données, dans la mesure où un cadre relativement stable pour la transcription avait d'ores et déjà été mis en place.

Chapitre 3

Transcrire

Lors de la soumission du projet Time Us à l'ANR, la question du passage des documents d'archives à leur transcription dans des fichiers de texte nativement numériques n'avait pas fait l'objet d'une expertise particulière. Il était donc prévu, comme nous avons pu l'évoquer, que la transcription soit réalisée manuellement en ayant recours à des « petites mains ». Le choix d'intégrer un outil comme Transkribus dans la chaîne de traitement modifie la manière dont la tâche de transcription est envisagée. Comme nous l'avons évoqué, Transkribus permet d'implémenter des outils de reconnaissance automatique de caractères. Il convient de s'attarder sur ce point pour comprendre quel est l'avantage de recourir à une transcription automatique plutôt qu'à une transcription exclusivement manuelle, et pour comprendre ce que cela peut apporter à un projet comme Time Us.

3.1 La reconnaissance automatique de caractères.

3.1.1 Définition

La reconnaissance optique de caractères, ou *optical character recognition* (OCR), désigne le processus informatique qui vise à transposer des éléments textuels présents sur une image numérique vers un fichier de texte de manière automatique. Il s'agit, en d'autres termes, de faire réaliser par l'ordinateur la tâche de copie du texte.

Le document numérique est obtenu à partir de simples photographies, de scans ou de captures d'écran. Il peut contenir du texte imprimé, dactylographié ou manuscrit. Les images font souvent l'objet d'une phase de pré-traitement afin de faciliter la reconnaissance du texte. Il peut s'agir d'ajuster les contrastes ou encore de découper l'image pour ôter les éléments inutiles ou la réorienter.

Plusieurs étapes de traitement sont nécessaires pour mettre en place un OCR. La première consiste à segmenter l'image en zones. On repère ainsi la zone de texte, les paragraphes, les figures, les lignes voire les mots, les lettres, ou encore les parties de lettres. Cette segmentation peut se faire manuellement et/ou automatiquement. En deuxième

lieu vient la reconnaissance des caractères, c'est-à-dire la phase réelle d'extraction du texte. Chaque segment est identifié sur la base de ses caractéristiques typographiques ou grâce à sa comparaison avec une base de caractères¹. Des hypothèses d'interprétation sont formulées et classées à partir d'un modèle statistique (chaîne de Markov)². Une phase de post-traitement est nécessaire afin de corriger les erreurs d'interprétation et d'évaluer les performances du système. Depuis le début des années 2010, la classification des interprétations se fait par le biais de méthodes d'apprentissage profond (ou *Deep Learning*)³.

3.1.2 Le cas des textes manuscrits

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

Ainsi, si la segmentation des textes imprimés est considérée comme une tâche facile à réaliser, ce n'est pas le cas des documents manuscrits⁴. En effet, l'espacement des lignes, des mots ou des lettres n'est pas régulier, et certaines lettres peuvent recouper plusieurs lignes. Les documents endommagés sont difficiles à segmenter et une même page peut combiner plusieurs sens d'écriture⁵. S'y ajoutent les difficultés posées par les nombreuses variations de formes des lettres au sein d'un même document pour une même main d'écriture. Actuellement, les taux d'erreur de reconnaissance de caractères pour une page manuscrite varie de 5%⁶ à 50%⁷, pour une page de mauvaise qualité, ce qui implique de passer du temps en phase de post-traitement pour vérifier et corriger les transcriptions obtenues automatiquement.

L'une des questions posées par le projet Time Us consiste à interroger l'utilité de recourir à des traitements automatiques pour la transcription des sources par rapport au gain de temps que cela représenterait.

Aux nombreux problèmes techniques encore posés par l'HTR, s'ajoute le fait que la

1. Romain Karpinski et Abdel Belaïd, *Rapport Evaluation des OCR*, Research Report, LORIA - Université de Lorraine, 2016, URL : <https://hal.inria.fr/hal-01356824> (visité le 04/06/2018), p. 1.

2. Christopher Kermorvant, *La reconnaissance des écritures manuscrites*, non publié, Reconnaissance par ordinateur des écritures anciennes : le projet HIMANIS, 29 mai 2018.

3. *Ibid.*

4. Aurélie Lemaitre, Jean Camillerapp et Bertrand Coüasnon, « Handwritten text segmentation using blurred image », *DRR - Document Recognition and Retrieval XXI*, DRR - Document Recognition and Retrieval XXI San Francisco, États-Unis (janv. 2014), URL : <https://hal.inria.fr/hal-01087210> (visité le 04/06/2018), p. 1.

5. A. Lemaitre et J. Camillerapp, « Text Line Extraction in Handwritten Document with Kalman Filter Applied on Low Resolution Image », *Second International Conference on Document Image Analysis for Libraries (DIAL)*, Document Image Analysis for Libraries, 2006. DIAL '06. Second International Conference on Lyon, France (avr. 2006), DOI : [10.1109/DIAL.2006.41](https://doi.org/10.1109/DIAL.2006.41).

6. Louise Seaward et Elaine Charwat, *If you teach a computer to READ...* CILIP Update, 2017.

7. C. Kermorvant, *La reconnaissance des écritures manuscrites...*

mise en place de tels traitements requiert au préalable une campagne de numérisation de ces sources ainsi qu'un pré-traitement des copies numériques. En définitive, recourir à la reconnaissance automatique d'écriture implique d'augmenter le temps dédié à la numérisation, au pré-traitement des sources et à la vérification et correction des transcriptions, afin de réduire le temps dédié à la transcription elle-même. Par ailleurs, la tâche de transcription n'est pas entièrement automatisée car il est nécessaire de produire une quantité significative de données d'entraînement avant de pouvoir automatiser le processus. L'ensemble de ces tâches est également multiplié car les meilleurs résultats sont obtenus à l'issu d'un cycle itératif, où les transcriptions corrigées servent de nouvelles données d'entraînement pour relancer l'OCR et obtenir de meilleures transcriptions.

Dans ces conditions, et si l'on considère qu'un·e transcripteur·rice expert·e n'aura pas besoin de plusieurs passages sur un même document pour obtenir le même taux d'erreur que la machine, voire réaliser une transcription sans erreur, la question semble en effet pertinente. Durant mon stage, j'ai à plusieurs reprises eu l'occasion de voir que cette question n'est pas tranchée pour tous les membres du projet, dont certain·es ne sont pas convaincu·es de l'intérêt de l'HTR.

La transcription d'un minimum d'entre 30 et 100 pages d'un corpus homogène est nécessaire avant de pouvoir entraîner convenablement un modèle de reconnaissance d'écriture manuscrite⁸. Pour un projet portant sur la transcription de moins de 300 pages, avoir recours à la transcription automatique par l'HTR présente finalement peu d'intérêt.

La situation est tout autre pour des projets d'ampleur. Le projet *Transcribe Bentham*, par exemple, vise à terme à transcrire près de 45 000 documents manuscrits rédigés par le philosophe Jeremy Bentham et ses secrétaires. Basé sur un système de transcription participative, le projet affichait sur son site, en juin 2018, qu'un peu plus de 45% de ces documents avaient été transcrits.

En janvier 2018, Tim Causer a détaillé les calculs réalisés par l'équipe pour démontrer l'intérêt de mettre en place un système de transcription participative par rapport à un projet de transcription « traditionnel ». Il explique ainsi que si la transcription des documents était menée de la même manière qu'elle l'avait été entre 1984 et 2010, à savoir une équipe de transcripteur·rices travaillant sur de simples éditeurs de texte, on pouvait estimer qu'elle serait achevée en 2081 au plus tôt, soit un minimum de 63 ans pour une transcription manuelle⁹. En revanche, la transcription participative¹⁰ permettait l'achèvement du projet d'ici 31 à 7 ans, en fonction des différents niveaux d'efficacité mesurés¹¹.

8. L. Seaward et E. Charwat, *If you teach a computer to READ.....; Transkribus*, URL : <https://transkribus.eu/Transkribus/> (visité le 04/06/2018).

9. Tim Causer, Kris Grint, Anna-Maria Sichani et Melissa Terras, « "Making such bargain" : Transcribe Bentham and the quality and cost-effectiveness of crowdsourced transcription », *Digital Scholarship in the Humanities* (, 15 janv. 2018), DOI : [10.1093/linc/fqx064](https://doi.org/10.1093/linc/fqx064), p. 16.

10. Un projet participatif repose sur la contribution d'individus extérieurs au projet, sans niveau de qualification minimum et non rémunérés.

11. *Ibid.*

Le recours à la transcription participative avec une plate-forme adaptée a donc permis d'abaisser significativement le temps consacré à la transcription, mais ces délais restent considérables. Si le *Bentham Project* optait pour une transcription automatique plutôt que manuelle, celle-ci ne prendrait pas plus de quelques jours¹².

Dès 2013, le projet *Transcribe Bentham* a participé à des expérimentations avec l'HTR. Les documents transcrits de manière participative ont ainsi servi à produire des données appelées « vérité terrain » (*ground truth*) nécessaires pour entraîner des modèles de reconnaissance d'écriture. Ces expérimentations ont été menées en particulier avec le projet TranScriptorium¹³, devenu ensuite Transkribus. Les résultats de ces tests sont présentés sur le blog de *Transcribe Bentham*¹⁴. Louise Seaward, coordinatrice du projet, a annoncé obtenir des taux d'erreur très convenables, compris entre 5% et 10%, sur des pages lisibles, mais allant jusqu'à 26% pour des pages difficiles à lire et rédigées par Jeremy Bentham lui-même. Un taux d'erreur jugé pour le moment trop élevé pour que la technologie soit intégrée au projet¹⁵. A l'occasion d'une correspondance personnelle, Louise Seaward m'a indiqué que l'objectif du projet Bentham n'est pas que l'HTR remplace la transcription participative, car celle-ci fait partie intégrante de l'identité du projet. Il s'agit plutôt de mettre en place deux stratégies : la première maintiendrait le principe de la transcription manuelle, en ajoutant simplement une assistance sous la forme de suggestions de transcription pour des mots ou des passages difficiles, des suggestions générées grâce à l'HTR. La seconde en revanche consisterait à déplacer le rôle de transcripteur·rice vers des tâches de vérification et de correction des transcriptions automatiques, comme on pourrait s'y attendre.

En définitive, la reconnaissance automatique d'écriture manuscrite représente un énorme gain de temps pour des projets portant sur des corpus importants – plus le corpus est grand, plus le gain de temps est important – en particulier, pour le moment, lorsque ces corpus sont relativement faciles à lire. Et cela, en dépit des tâches supplémentaires de pré- et post-traitement.

3.2 Méthodologie

Dans le projet Time Us, l'adoption de Transkribus comme outil de transcription collaborative est récente, elle date du début de l'année 2018. La plate-forme a fait l'objet d'une demi-journée de formation proposée par l'équipe ALMAnaCH, animée par Marie

12. C. Kermorvant, *La reconnaissance des écritures manuscrites...*

13. *transScriptorium*, URL : <http://transcriptorium.eu/> (visité le 04/06/2018)

14. L. Seaward, *Project Update – teaching a computer to READ Bentham*, UCL Transcribe Bentham, 9 juin 2017, URL : <http://blogs.ucl.ac.uk/transcribe-bentham/2017/06/09/project-update-teaching-a-computer-to-read-bentham/> (visité le 04/06/2018).

15. Id., *Project Update – Bentham vs the computer*, UCL Transcribe Bentham, 23 févr. 2018, URL : <http://blogs.ucl.ac.uk/transcribe-bentham/2018/02/23/project-update-bentham-vs-computer/> (visité le 04/06/2018).

Puren et Charles Riondet, le 15 février 2018, et à laquelle j'ai également participé. Lors de cette formation, les manipulations basiques de la plate-forme, ainsi que son outil de reconnaissance d'écriture, ont été présenté. C'est grâce à cette formation qu'un protocole minimal a été mis en place, faisant office de méthodologie.

3.2.1 Importer les fichiers dans Transkribus

3.2.1.1 Un guide pour les orienter tou·tes

Une collection Transkribus intitulée « timeUs » a été créée par Charles Rionder en février 2018. C'est dans cette collection que tous les versements d'images doivent être réalisés. Pour la plupart, ils constituent des ensembles logiques, correspondant à des parties ou la totalité de fonds d'archives. Le versement « AD Sein Prud'hommes 1847-1849 (D1U10 379) » contient les images réalisées d'après le fond D1U10 379 des A.D. de la Seine contenant les registres des prud'hommes pour les années 1847 à 1849. Nombre des versements listés dans la collection ont été opérés le 15 février, à l'occasion d'exercices de manipulation de Transkribus.

Dans la continuité de cette formation, et afin de fournir aux membres du projet, un guide simple pour verser des images sur la plate-forme, j'ai rédigé un « guide pour charger des fichiers dans Transkribus ». Il est publié sur le wiki de Time Us¹⁶. Ce guide donne des instructions factuelles mais ne fournit pas à ce stade de consignes méthodologiques.

Il répond aussi à un problème d'orientation des images après leur versement dans Transkribus. En effet, la plate-forme se base sur les métadonnées EXIF¹⁷ uniquement, ce qui provoque des erreurs. Nous renvoyons vers un tutoriel externe proposant d'utiliser le logiciel XNView¹⁸ pour corriger ces métadonnées.

La correcte orientation des images est fondamentale pour que puisse être réalisée la segmentation des zones de texte et la reconnaissance automatique d'écriture. Transkribus est doté d'un outil de rotation des images, mais celle-ci n'a d'impact que pour l'affichage de l'image dans l'interface graphique. En d'autres termes : une image correctement réorienté en mode « portrait » dans Transkribus après son versement en utilisant l'outil de rotation sera toujours considérée en mode « paysage » par son outil de segmentation. En outre, la rotation porte sur la totalité du dossier et pas uniquement celui de la page courante.

16. *Guide pour charger des fichiers dans Transkribus*, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_pour_charger_des_fichiers_dans_Transkribus (visité le 11/08/2018).

17. Pour *Exchangeable Image file Format*, une spécification développée depuis 1995 par la Japan Electronic Industry Development Association. Elle permet d'associer des métadonnées aux images créées par les appareils photo numériques, notamment les réglages de l'appareil (dont l'orientation), des données géographiques ou de date, ou encore les droits d'auteurs associés à une image.

18. Il s'agit d'une visionneuse d'images qui permet de réaliser des opérations de conversion de formats et d'édition des images et leurs métadonnées. Le logiciel est développé par XnSoft : *Xnview Software*, URL : <https://www.xnview.com/fr/> (visité le 12/08/2018).

3.2.1.2 Adapter les niveaux de corpus

D'un point de vue méthodologique, il est important d'expliciter l'articulation opérée par les équipes Time Us entre les différents niveaux de corpus permis par Transkribus. La plate-forme fonctionne par groupement généraux appelés « collections ». Chaque versement dans la collection crée un nouveau « document », selon la terminologie employée par Transkribus. Chacun de ces documents pouvant être composé de plusieurs images, celui-ci est divisé en « pages ».

Cette organisation « collection – document – page » n'est pas pleinement satisfaisante pour le projet. Le passage immédiat du niveau « collection » au niveau « document » n'est pas correct et il aurait été intéressant d'avoir un niveau intermédiaire selon le modèle « collection – sous-collection – document – page ». Quoique plus lourde, une autre manière d'envisager l'organisation des versements dans Transkribus aurait pu être la suivante : une collection pour chaque fond documentaire, au sein de laquelle les versements auraient été organisés de manière à reconstituer les unités documentaires. Cela aurait permis de constituer des métadonnées attachées à un niveau fin de description, et de recomposer les unités documentaires dès leur export¹⁹. Toutefois, une telle organisation des versements n'est pas envisageable étant donné l'ampleur du projet Time Us et la quantité de fonds concernés : il aurait ainsi fallu créer pas moins de 37 collections, rien que pour les fonds des Archives municipales de Lyon.

A défaut, et dans le but de ne conserver qu'une seule collection pour tout le projet, nous perdons le niveau « document ». Les fichiers dans Transkribus correspondent à une structure « collection – sous-collection – page ». Dans la phase de transcription des textes, cela ne pose pas de problème. En revanche, cela limite la portée de certains outils prévus par Transkribus, en particulier ceux qui permettent la gestion des métadonnées.

3.2.1.3 Un futur protocole de nommage des sous-collections

A l'heure actuelle, il n'existe pour le projet Time Us ni protocole de constitution des versements, ni protocole de nommage de ces versements. Si Transkribus permet le renommage des ensembles documentaires, on ne peut pas davantage les modifier : à l'issue de la phase d'expérimentation de Transkribus, il sera donc important que des règles soit définies pour garantir la complétude et la pertinence des sous-collections.

Ces règles assurent la constitution de véritables unités documentaires correspondant aux numérisations disponibles pour une côte d'archives comme cela tend à être le cas. Pour les corpus importants, des sous-ensembles peuvent être créés, en suivant des périodes chronologiques par exemple. Cela permettrait d'établir un modèle de nommage tel que le suivant :

19. Lorsque des pages d'une collection sont exportées, on ne distingue pas actuellement les unités documentaires en son sein.

Lieu de conversation ; Côte ; Titre ; Date

Le moment venu, ces éléments méthodologiques pourront être intégrés au « guide pour le versement des fichiers dans Transkribus ».

3.2.2 Créer un modèle de reconnaissance automatique

3.2.2.1 Sur l'importance des pratiques de transcription

La segmentation des zones de texte peut être réalisée manuellement ou automatiquement dans Transkribus. La documentation sur la plate-forme propose plusieurs guides²⁰ détaillant la manipulation, il n'est donc pas nécessaire de la revoir en détail ici. Dans la mesure où les images n'ont pas fait l'objet d'un post-traitement pour éliminer les données textuelles parasites, contrôler la qualité de cette segmentation peut s'avérer nécessaire. Cela permet d'éliminer les données textuelles non signifiantes que l'outil aurait pu prendre en compte.

La délimitation des zones de texte et en particulier des lignes de texte permet de faire correspondre l'image à sa transcription de manière précise. Elle permet en outre de créer des données d'entraînement. C'est la correspondance entre les segments d'images et les lignes de texte transcrit qui constitue la vérité-terrain, essentielle pour l'HTR. Lorsque l'utilisateur·rice déplace son curseur dans l'éditeur de texte, la zone de l'image correspondante est mise en sur-brillance. Cela permet de s'assurer de la bonne coordination des deux éléments.

La correcte adéquation entre la transcription et la zone de texte est fondamentale pour générer des données d'entraînement afin d'obtenir un modèle de transcription automatique efficace. Comme nous l'avons évoqué, le modèle identifie des formes au sein du segment qu'il tâche d'identifier. Si la transcription et l'image ne se correspondent pas, cela crée des données parasites. Pour la même raison, la fidélité de la transcription au texte original est importante : le ou la transcripteur·rice doit préférer laisser une ligne vide plutôt que d'y laisser des blancs, car ces omissions perturbent l'entraînement du modèle.

3.2.2.2 Entraîner et contrôler

La documentation de Transkribus recommande d'entraîner un modèle sur une seule et même main d'écriture : si les formes possibles pour une même lettre sont trop nombreuses, le modèle ne peut pas obtenir de bons résultats. Dans le cas du projet Time Us,

20. Voir en particulier le guide « How to transcribe documents with Transkribus », dont une section entière est dédiée à la segmentation : *How to transcribe : basic instruction*, URL : https://transkribus.eu/wiki/images/5/50/How_To_Transcribe_Documents_with_Transkribus.pdf.

il est difficile de se plier à cette consigne dans la mesure où nos documents de travail sont des documents d'archives produits dans des contextes très variés et par plusieurs personnes. C'est moins le cas des registres de Conseils de Prud'hommes, dans lesquels on retrouve des écritures relativement homogènes sur de longues périodes. En revanche, pour des dossiers comme ceux tirés du fonds 9M5 des A.D. du Rhône, les mains sont très nombreuses. On peut en identifier cinq nettement différentes rien que pour le 25 premières pages du dossier.

Lorsqu'un nombre suffisant de transcriptions manuelles a été réalisé, les membres du projet prennent contact avec l'équipe de développement de Transkribus pour demander l'entraînement d'un modèle. Celui-ci peut n'être basé que sur des données créées par la transcription, mais il peut aussi être alimenté par les données d'entraînement de projets similaires ou pour une langue et une période similaire. Lorsque le modèle est prêt, il est ajouté à la liste des modèles disponibles dans la section « Text Recognition » de l'onglet « Tools » de l'interface de Transkribus. Ici encore, Transkribus propose un guide pour la démarche, il n'est donc pas nécessaire de la détailler²¹.

C'est le taux d'erreurs de reconnaissance de caractères (CER, pour *Character Error Rate*) qui définit l'efficacité du modèle. Un autre taux est disponible, le taux d'erreurs de reconnaissance de mots (WER, pour *Word Error Rate*), mais celui-ci est moins significatif dans la mesure où il est directement corrélé aux erreurs de reconnaissance de caractères. Comme nous l'avons évoqué, pour une reconnaissance automatique de bonne qualité, on vise généralement d'atteindre un CER inférieur à 5%. Dans son guide, Transkribus nuance ce chiffre en indiquant qu'un CER inférieur à 10% est un taux satisfaisant, et qu'un taux compris entre 20 et 30% permet de travailler avec des technologies de type « Word spotting »²².

Entraîné et testé sur les mêmes fichiers, les premiers CER ne sont pas tout à fait fiables. Il est nécessaire de produire de nouvelles données de transcription, une nouvelle « vérité terrain », qui correspond en quelque sorte au corrigé des tests fournis au modèle : elles servent à tester les erreurs de reconnaissance encore présentes.

Deux modèles ont déjà été entraînés pour Time Us. Le premier, « Rhone_XIX_M1 », est basé sur la transcription des pages 1 à 32 de la sous-collection AD69 9M5. La description du modèle fournie par Transkribus indique un CER de 5,75%. Nous avons testé le modèle sur une page de la même sous-collection prise au hasard, la page 292, dont nous avons réalisé une transcription manuelle servant de corrigé au test et permettant de calculer les taux d'erreur²³. Le CER obtenu après l'application du modèle sur la page reste

21. Voir en particulier le guide « How to train a handwritten text recognition model in Transkribus » : *How to train a HTR model in Transkribus*, URL : https://transkribus.eu/wiki/images/3/34/HowToTranscribe_Train_A_Model.pdf.

22. Il s'agit d'identifier, pour un mot donné, toutes ses occurrences dans un document, sans pour autant les avoir reconnues lettre par lettre.

23. Cf. Figure 3.1.

FIGURE 3.1 – Comparaison des transcriptions manuelles et automatiques de la page 292.

fiel dealoe b gx dbeutieund dne facilement à obtenir les revendications demandées, il serit néccenaue serait nécessaire que tsutes toutes les ouvrières et ouvrurs abardonment ouvriers abandonnent leurs usenes-usines pendant nne une matinée, et se portent en masa devreant masse devant l'sme-l'usine de M-gerndre. Mr Gindre. Le né Chovzet comlat Thozet combat cette proposition, proposition ; il'dit q'en il dit qu'en raison du leu-bon vouloir des patrons qui avient auigmente avaient augmenté les salaires-salaires, il valien-y a lieu de se montrer plus pacifique ; que du reste cette mani-maniastion festation n'aurait à autre d'autre résultat que de se faire mette eu mettre en état d'arrestation-d'arrestation. Un ididu 'clay-individu étranger à la corporation qu'on a appélé ele-tuouciard « le Mouchard de lu-grande-la grande usine d-» a été invité à se rétier, retirer ; ce qu'il a fait ummedia-immédia-tement, tens te ment ; tous les assistants l'ont hée-ernant hué criant à "à la porté, a porte ! à la poité. porte !" Un des ouveriers-ouvriers de l'usne-l'usine Bergier viet veut éclaie déclarer que si le personnel de cette maison s'était mis engrêvé en grève le 13 eournant, courant, C'était par Peuite-suite d'un malentendu être-entre la délégation-délégration t-et le patron it-et qu'après plus ample xplication-explications Mo-Bergev Mr Bergier leur avait donne satisfaction donné satisfasaction en aécordant accordant une augmentation de salaire variant de 12 à 1e7, 14 % ; il etme estime que l'index l'index de cette maion-maison doit être lyve. levé. Cette proposition mise aux voix est adte adoptée. Thiozet Thozet dit qu'il feaut emerir faut remercier le pounal journal "Le Peuple" de veupple" de son préciny-précieux concours Oleuieu erns Plusieurs cris de Vite-Vive le Deeple-Peuple se fous enturdre font entendre. Les Le né Simon redacteur rédacteur à ejourviait ce journal

élevé : 17,64%. Il est donc nécessaire d'améliorer encore ce modèle à partir des dizaines de pages supplémentaires qui ont été transcrives.

Le deuxième modèle, intitulé « Prud-Homme_1858 », a été entraîné à partir des cent premières pages transcrives de la sous-collection « AD Seine Prud'homme 1858 (D1U10 386) ». Il en existe trois versions, qui correspondent à des entraînements successifs. La version la plus récente, « Prud-Homme_1858_M3 », est associée à un CER de 10%, ce qui reste élevé étant donné les résultats du précédent test.

Pour améliorer les modèles, il n'existe qu'une seule solution : produire davantage de données d'entraînement, c'est-à-dire, transcrire davantage de texte manuellement. L'ajout d'un dictionnaire de mot lors de l'application de la reconnaissance automatique peut conduire à un meilleur classement des hypothèses de reconnaissance de mots et de lettres.

3.2.2.3 Les outils pour réaliser l'OCR

Si les technologies d'HTR requiert beaucoup de travail préalable, ce n'est pas autant le cas des technologies de reconnaissance automatique des caractères imprimés.

Pour les textes imprimés, Transkribus propose un modèle d'OCR disponible directement au sein de la plate-forme. Cela permet d'utiliser toutes les fonctionnalités de

Transkribus pour des corpus manuscrits et imprimés. La plate-forme gère finement l'application des différents modèles de reconnaissance au niveau de la page. C'est utile dans le cas de corpus mélangeant documents manuscrits et documents imprimés. C'est le cas par exemple de la sous-collection formée des documents tirés du fond 9 M 5 des A.D. du Rhône, qui est composé, en plus des rapports manuscrits, d'articles de presse et de rapports dactylographiés.

Transkribus n'est cependant pas le seul outil d'OCR disponible pour le projet Time Us. En effet, en nous dotant d'un espace de stockage sur ShareDocs, nous bénéficions également des services de traitement des documents proposés par HumaNum. Ceux-ci incluent l'implémentation du logiciel de reconnaissance d'écriture FineReader d'ABBYY directement sur la plate-forme de stockage. Le dépôt d'un fichier à transcrire dans le dossier adéquat permet de déclencher le processus, aboutissant à la création d'un second fichier au format souhaité dans le même dossier²⁴. L'utilisation de cet outil se veut intuitive. La plate-forme de stockage est composé en premier niveau de trois dossiers au minimum : en premier lieu le dossier nommé d'après le projet propriétaire de l'espace²⁵, puis un dossier intitulé « hnTools_Software » et enfin un dossier intitulé « hnTools_watchFolder ». C'est dans ce dernier dossier que se trouve l'outil d'OCR. Il contient plusieurs sous-dossiers : « Audio », « OCR », « PDF » et « Video ». Dans le dossier « OCR », plusieurs dossiers permettent de spécifier le mode de transformation souhaité, par exemple : « toXMLAlto », « toPDF », « toXML », « toEPUB », etc. Enfin, à l'intérieur de chacun de ces dossiers, un sous-dossier correspond à la langue du document à OCRiser.

Cet outil nous permettrait de renouveler l'OCR des deux corpus imprimés dont nous disposons : la presse lyonnaise et les monographies de Le Play.

3.3 La structure logique des fichiers numériques

Pour réaliser une transcription de qualité, les corpus obtenus doivent être structurés de manière à organiser tous les niveaux. Il existe deux cas de figure pour le projet Time Us.

3.3.1 Le modèle Transkribus

Transkribus gère déjà la création d'une structure logique pour les documents créés par son biais. Celle-ci découle directement des outils appliqués aux documents, et elle dépend du format d'export sélectionné. Nous nous concentrerons sur la structure produite pour les documents exportés au format TEI, car c'est le format de sortie choisi pour le corpus numérique du projet Time Us.

24. Cf. Annexes C.1.2.

25. Dans notre cas le dossier s'intitule « ANR ».

Le schéma des fichiers TEI générés par le biais de Transkribus est relativement simple²⁶. Un export étant toujours limité à un document (ou sous-collection) en cours, c'est là le niveau hiérarchique le plus élevé. La transcription est contenue dans un élément `<body>`, lui-même contenu dans une élément `<text>`. Chaque début de page composant le document est rendu par le biais d'un élément vide `<pb>` numéroté grâce à l'attribut `@n`. Chaque région de texte est matérialisée par un élément `<p>` contenant le rendu des lignes. Celui-ci peut être obtenu de deux manières : soit par le biais de balises `<l>` contenant les noeuds de texte et les éventuels balises d'annotation, soit par le biais d'éléments vides `<lb>` signalant le début d'une nouvelle ligne avant les noeuds de texte et les balises d'annotation.

Lorsque le fichier TEI contient un ensemble d'éléments `<facsimile>`, les éléments `<pb>`, `<p>` et `<lb>` sont associés à un attribut `@facs` renvoyant vers l'élément correspondant.

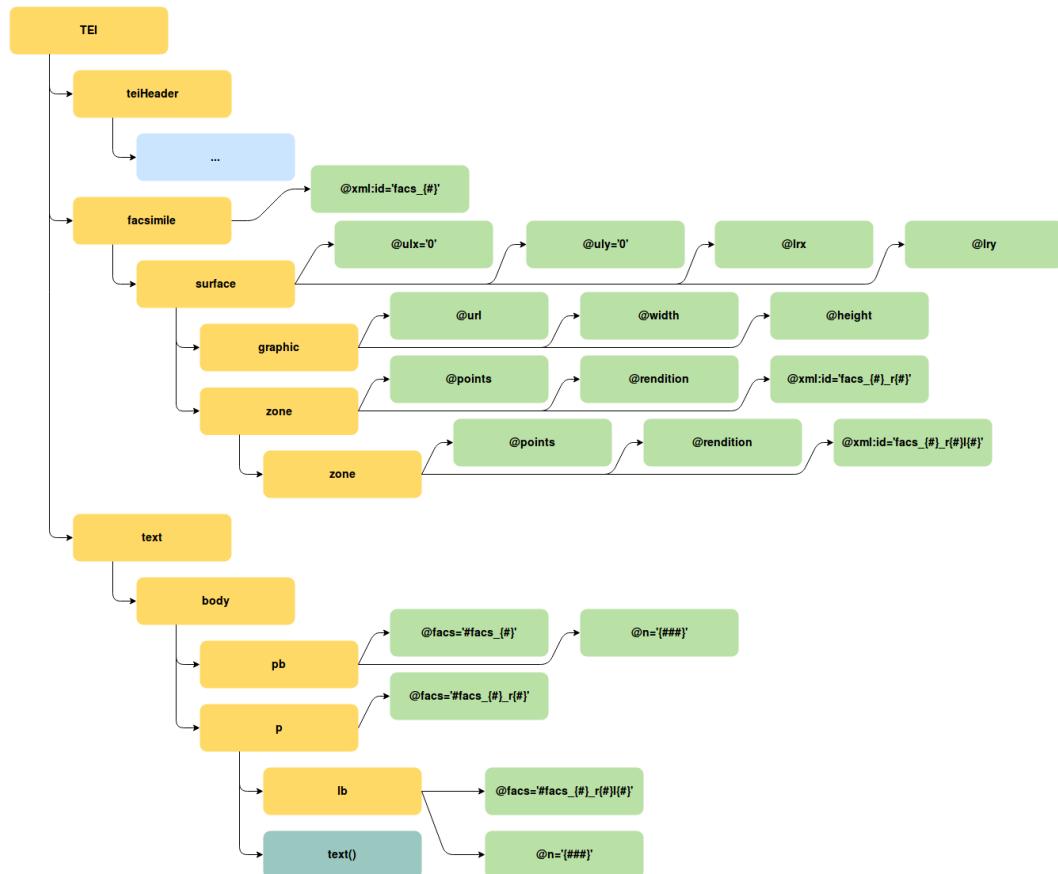


FIGURE 3.2 – Schématisation de la structure logique du fichier TEI, sans annotation.

Pour le projet Time Us, l'option `<lb>` est préférée : elle permet une plus grande liberté pour l'annotation, en évitant la contrainte de non-chevauchement des balises XML. En outre, les exports contiennent généralement un système de renvoi vers les fac-similés : même si nos images sont de faible qualité, dans la mesure où Transkribus permet facilement de réaliser cet alignement, nous avons jugé intéressant de conserver la relation entre

26. Cf. Figure 3.2.

la transcription et l'image par ce biais à ce stade du projet.

3.3.2 Un modèle sur mesure pour Le Play

3.3.2.1 Objectifs

L'édition des monographies de Le Play constitue un projet à part pour deux raisons. En premier lieu, il s'agit d'un corpus connu et exploité par les historiens et les sociologues : si des données importantes pour le projet Time Us peuvent en être extraites, une édition numérique particulière d'enquêtes dédiées à des ouvrière·es du textile fait également partie des objectifs annoncés. En second lieu, il s'agit d'un corpus pris en charge par l'équipe CMH en dehors de ses activités pour le projet Time Us : le CMH envisage en effet de créer une plate-forme de consultation pour la totalité des enquêtes, permettant un accès renouvelé à ses contenus pour la communauté des chercheurs intéressés par le corpus.

Afin de concilier ces deux projets, l'équipe ALMAAnaCH est chargée d'élaborer une modélisation de la structure des ouvrages et des enquêtes pour les séries 1 à 3 . Cette modélisation doit à terme aboutir à l'élaboration d'un schéma pour XML-TEI et pour LaTeX. L'objectif d'une telle structure est double. Il s'agit, d'une part, de repérer plus facilement les contenus susceptibles d'intéresser le projet Time Us, permettant au passage d'améliorer la qualité des OCR. D'autre part, cela doit ouvrir la voie à l'édition numérique de certaines enquêtes, en particulier celle dédiée au tisseur en châle²⁷, qui doit faire l'objet d'une édition critique.

Une fois la structure et les schéma établis, deux approches sont possibles pour la réalisation de l'édition numérique : soit une approche manuelle, où une personne est chargée d'encoder les texte selon en suivant les schéma ; soit une approche automatisée, où un programme prend en charge cet encodage.

Dans le cas des monographies de Le Play, le gain de temps n'est pas un argument de poids car la tâche est répétitive et le CMH juge qu'elle ne nécessite pas plus d'un mois de travail. En revanche, d'un point de vue méthodologique et scientifique, il apparaît pertinent d'interroger la faisabilité de l'approche automatisée. En effet, le projet Time Us vise à explorer les pratiques numériques mis au service de la recherche en Sciences Humaines. Une telle expérimentation entre pleinement dans les enjeux du projet, quel qu'en soit le résultat.

En juillet 2018, j'ai été chargée d'établir un premier bilan sur la structure des quatre premiers volumes de la première série. A partir du premier modèle théorique identifié, il s'est ensuite agi de le confronter aux volumes suivants pour repérer les variations et les limites du modèle. Ce travail a donné lieu à une note interne²⁸ destinée à l'équipe du

27. Il s'agit de l'enquête n°7, publiée dans le premier tome de la première série des *Ouvriers des deux mondes*, intitulée « Tisseur en châles de la Fabrique urbaine collective de Paris » et réalisée en 1857.

28. Cf. Annexes A.2.1.

CMH, à ALMAnaCH et au LARHRA. Le bilan a permis de valider l'existence d'une structure relativement constante les volumes et des enquêtes, ce qui rend théoriquement possible l'application d'un traitement automatique pour l'édition du corpus. Cette phase d'expérimentation pourra débuter à partir de septembre 2018, avec notamment la création des schémas LaTeX et XML-TEI.

3.3.2.2 Bilan de l'analyse

L'analyse de la structure des quatre volumes a porté sur deux niveaux : celui des volumes et celui des enquêtes.

La structure des enquêtes est rendue en premier lieu par le plan suivi par les enquêteurs, qui apparaît au travers des titres de sections et sous-sections. Nous avons identifié un plan type, que nous avons confronté à l'ensemble des enquêtes de la première série pour le valider. Un ensemble de marqueurs typographiques (les tirets, les symboles, les sauts de pages, les espaces, etc) aident à signaler l'articulation des différents niveaux hiérarchiques. En fonction de notre position dans la structure hiérarchique, les en-têtes sont amenées à changer. Si la position des numéros de page reste constante, le texte qui l'accompagne varie : ainsi, on a tantôt le titre court de l'enquête, tantôt le titre de la section en cours. Dans la partie dédiée au budget, le titre court de l'enquête est inscrit en haut de page, que celle-ci soit paire ou impaire. Ce genre d'observation doit permettre de calculer ce qui compose l'en-tête, afin de l'isoler du reste du texte obtenu par reconnaissance automatique des caractères et afin de recomposer les paragraphes interrompus par les en-têtes.

Les volumes ne sont pas uniquement composés des enquêtes : ils contiennent également un ensemble d'articles et de rubriques qui les accompagnent. Nous avons identifié un modèle global pour la composition des volumes. Celle-ci varie beaucoup mais certains éléments sont constants. Par exemple, un volume se termine toujours par une table des matières, précédée d'un index et éventuellement d'un errata. En outre, les enquêtes se suivent et constituent un ensemble ininterrompu : les articles sont toujours placés de part et d'autre du corpus d'enquêtes. Pour finir, la comparaison de ce modèle avec les tables des matières a permis d'évaluer la fiabilité de cet élément d'information. L'outil de traitement automatique pourra donc utiliser les informations contenues dans la table des matières pour mieux appréhender la structure du volume en cours.

Chapitre 4

Annoter

4.1 Quelle annotation pour le projet Time Us ?

4.1.1 Définition

Selon la définition qu'en donnent Lou Burnard¹ et Karën Fort², l'annotation désigne la pratique qui consiste à ajouter à un corpus des informations de type interprétative. Dans son mémoire d'HDR³, Iris Eshkol-Taravella distingue trois applications distinctes de l'annotation : en premier lieu, celle qui consiste à ajouter des remarques ou des commentaires en marge d'un texte ; en deuxième lieu, celle qui correspond à l'ajout de « métadonnées caractérisant et décrivant le document numérique » ; et enfin en dernier lieu, celle, d'ordre linguistique, qui correspond à l'étiquetage morphosyntaxique ou à l'annotation sémantique du corpus⁴. C'est la dernière application qui nous intéresse, en particulier l'annotation sémantique du corpus.

L'annotation est un acte réfléchi qui induit une modélisation préalable du système d'annotation. Cette modélisation est toujours étroitement liée aux usages prévus du corpus annotés et aux théories, en particulier linguistiques, qui entourent son élaboration. Un modèle est matérialisé par des règles d'annotation et par l'élaboration d'une ontologie qui identifie des classes que l'on retrouve dans le corpus. L'annotation est une couche ajoutée au texte : celui-ci doit pouvoir en être détaché, et il peut simultanément exister plusieurs couches d'annotations basées sur des modèles différents. Le corpus annoté peut être exporté sous différents formats, parmi lesquels le format XML. Les couches d'annotations

1. Lou Burnard, *Qu'est-ce que l'annotation et pourquoi en parle-t-on de manière si inquiétante ?*, sept. 2011, URL : <http://www.lattice.cnrs.fr/IMG/pdf/burnard-annotation.pdf> (visité le 24/07/2018).

2. Karën Fort, *Annotation collaborative de corpus : motivations et définitions*, 10 janv. 2018, URL : <http://www.schplaf.org/kf/pdf/AnnotationEtTAL.zip> (visité le 24/07/2018).

3. Habilitation à Diriger des Recherches

4. Iris Eshkol-Taravella, *La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral*, Mémoire d'HDR, Paris, Université d'Orléans, 2015, URL : <https://hal.archives-ouvertes.fr/tel-01250650/document> (visité le 24/07/2018), p. 10.

peuvent être séparées du fichier de texte brut grâce à l'établissement d'un système de pointeurs. Il existe de nombreux standards pour encadrer la réalisation des annotations, mais, comme le rappelle Iris Eshkol-Taravella,

« il n'est pas toujours possible d'être conforme à ces normes s'il s'agit d'un phénomène qui n'a pas été pris en compte dans les conventions proposées. Se pose alors la question d'adapter les étiquettes à celles normalisées ou de développer un nouveau jeu d'étiquettes qui permettra de mieux représenter le phénomène en question. »⁵

Le projet Time Us met en place une approche expérimentale de l'annotation des sources. Pour cette raison, l'une de mes missions a consisté à élaborer un modèle d'annotation basé sur le standard XML-TEI. Ce modèle a ensuite été formalisé sous la forme d'un guide d'annotation pour encadrer l'annotation manuelle et préparer l'annotation automatique.

4.1.2 Annotation manuelle et annotation automatique

Les outils de traitement automatique des langues permettent de mettre en place des campagnes efficaces d'annotation de corpus. Toutefois, de manière similaire aux outils de reconnaissance automatique d'écriture, l'annotation automatique nécessite l'apprentissage d'un modèle par la machine et son entraînement à partir de données préalablement préparées. Ces données d'entraînement sont produites dans le cadre d'une campagne d'annotation manuelle.

L'élaboration du modèle ne doit pas être envisagée comme un enchaînement linéaire d'étapes mais comme un processus itératif qui permet d'améliorer le modèle théorique d'annotation, le modèle applicatif d'annotation automatique et la qualité des corpus annotés obtenus⁶. En amont, le modèle est élaboré et testé à partir d'un échantillon du corpus. Idéalement, cette annotation manuelle doit être réalisée par plusieurs annotateur·rices de manière simultanée. La comparaison des résultats obtenus permet d'identifier les faiblesses du modèle et d'établir un corpus annoté de référence⁷. En outre, le contrôle de la qualité des annotations réalisées automatiquement permet d'améliorer le système d'annotation automatique.

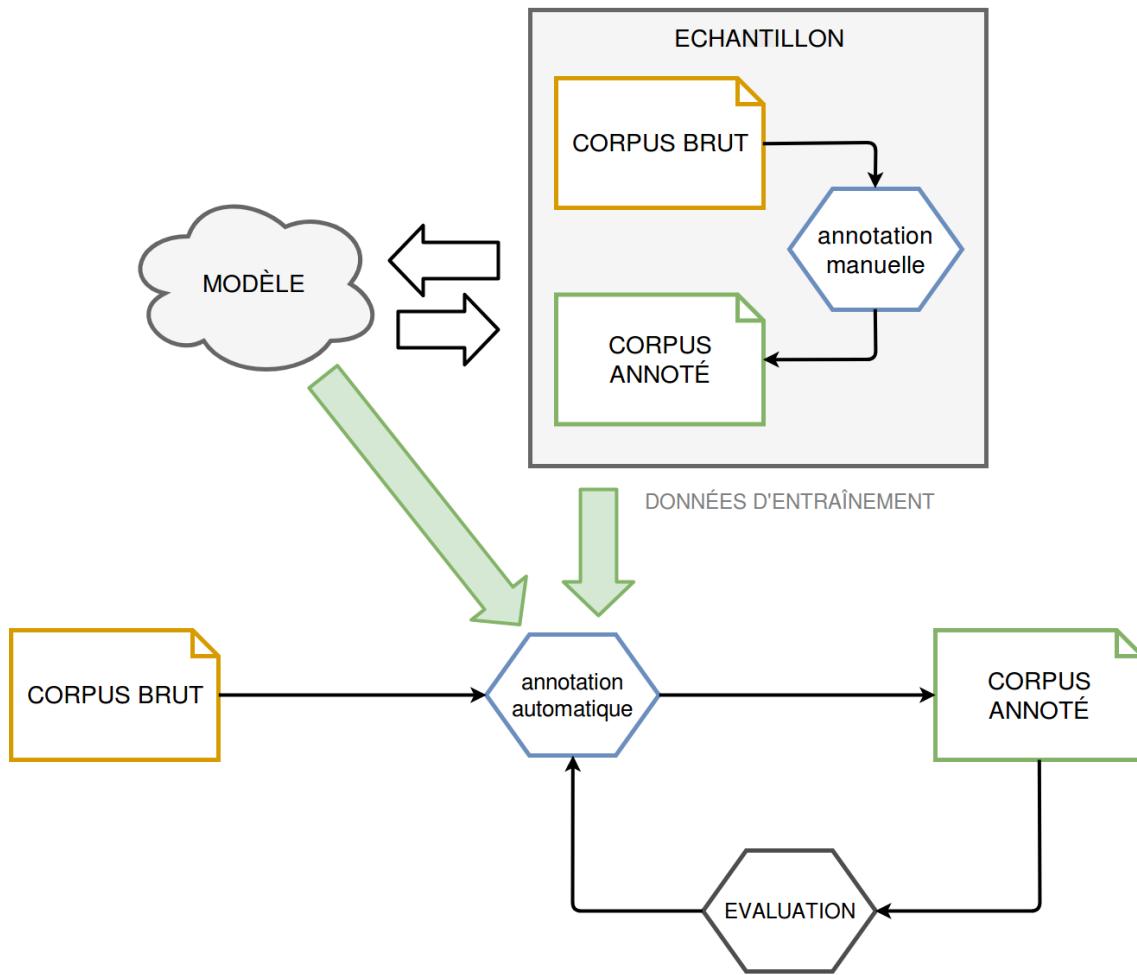
Dans le cas du projet Time Us, l'annotation automatique vise à alléger la tâche d'annotation manuelle en identifiant les éléments redondants comme les noms de personnes ou les noms de métiers. L'objectif est également de réaliser l'annotation d'une partie du corpus avec une intervention humaine minimale. Enfin, elle est également prévue pour aller plus loin en établissant une couche supplémentaire d'annotation qui rend compte

5. *Ibid.*, p. 11.

6. Cf. Figure 4.1.

7. K. Fort, *Annotation collaborative de corpus : motivations et définitions...*

FIGURE 4.1 – Schématisation du processus d'annotation manuelle et automatique.



des relations grammaticales entre les différents éléments annotés, afin d'obtenir des informations contextualisées les plus précises possibles. Les moyens techniques et humains du projet ne permettent pas de recourir à plusieurs annotateur·rices sur les mêmes portions de texte.

4.2 Description du processus de travail

Afin d'encadrer l'annotation manuelle, il est nécessaire de réaliser un guide d'annotation clair et précis qui élimine le plus grand nombre d'interprétations de la règle. Comme le précise Iris Eshkol-Tavarella :

« L'annotation est une façon de s'approprier le corpus. Les différents annotateurs humains interprètent et perçoivent différemment les données. Les résultats d'annotation peuvent dépendre non seulement de leurs connaissances du domaine annoté et de leur orientation théorique mais aussi de variables

sociologiques. Le guide d'annotation se doit donc d'être le plus clair, le plus exhaustif et le moins ambigu possible. »⁸

L'extraction des données en contexte est l'objectif visé par l'annotation du corpus du projet Time Us. Cette annotation est donc directement influencée par les données auxquelles les chercheur·ses souhaitent pouvoir accéder, mais elle prend aussi en compte les besoins techniques exprimés par les spécialistes du traitement automatique des langues.

4.2.1 Présentation des données ciblées

L'annotation du corpus Time Us se concentre sur l'identification d'un certain nombre d'éléments textuels constituant des informations utiles pour la recherche ou comportant des données que l'équipe souhaite rassembler. La démarche mélange reconnaissance d'entités nommées, étiquetage terminologique et simple balisage de segments porteurs de sens dans le cadre du projet.

Nouha Omrane, Adeline Nazarenko et Sylvie Sulzman, qui tentent de définir la notion d'entités nommées rappellent que celle-ci est difficile à cerner car elle implique la définition de concepts ontologiques qui font débat⁹. On peut toutefois tenter une définition avec Maud Ehrman qui indique dans sa thèse : « Étant donnés un modèle applicatif et un corpus, on appelle entité nommée toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus. »¹⁰. Ces expressions linguistiques peuvent relever de différentes catégories linguistiques (noms propres, pronoms, description définies...)¹¹ mais sont généralement à minima des constituants nominaux.

L'objectif de l'annotation précède toujours l'élaboration du modèle conceptuel utilisé pour identifier les éléments à annoter et établir les règles d'annotation. Dans le cas du projet Tine Us, il ne s'agit pas uniquement d'identifier des entités nommées ou des termes, car certaines informations échappent à ces catégories. Certaines expressions sont porteuses de sens et leur encodage permet de les rendre visibles et extractibles pour un traitement ultérieur.

En juillet 2018, nous avions ainsi identifié 17 catégories d'information que les historien·nes ou les TAL-istes souhaitaient annoter/extraire.

Les **personnes**, les **lieux**, les **organisations**, les **dates** et les **adresses** sont des

8. I. Eshkol-Taravella, *La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral...*, p. 12.

9. Nouha Omrane, Adeline Nazarenko et Sylvie Szulman, « Les entités nommées : éléments pour la conceptualisation », *21es Journées francophones d'Ingénierie des Connaissances*, Nîmes, France (juin 2010), http://www.ic2010.mines-ales.fr/index.php?option=com_content&view=article&id=50&Itemid=44, URL : <https://hal.archives-ouvertes.fr/hal-00525530> (visité le 24/07/2018), p. 2.

10. Maud Ehrmann, *Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguïsation*, Thèse de doctorat, Université Paris Diderot, 2008, URL : <https://hal.archives-ouvertes.fr/tel-01639190> (visité le 27/04/2018).

11. N. Omrane, A. Nazarenko et S. Szulman, « Les entités nommées... », p. 2.

types d’entités nommées que l’on retrouve souvent dans les campagnes d’annotation. Dans le cas du projet Time Us, il s’agit de rassembler des éléments de contextualisation sous la forme de listes, de noms par exemple dans le but de modéliser les relations sociales, spatiales et temporelles, et de dresser des notices prosopographiques. De plus, l’annotation des **heures** doit permettre de rassembler des indications sur la manière dont la journée d’un·e ouvrière est rythmée.

Le projet vise ensuite à créer des listes de vocabulaire d’une part pour donner des repères au modèle applicatif pour comprendre le sens du texte, et d’autre part pour créer des index permettant un accès transversal au corpus. En outre, il s’agit également d’identifier des éléments permettant de mieux comprendre la nature des activités ouvrières. Dans cette finalité, les éléments de terminologie comme les **noms de métier ou d’occupation**¹², les **produits** et les **tâches** sont annotés. L’identification des expressions employées pour qualifier les **statuts matrimoniaux** et les **statuts sociaux de travailleur·ses** vise également à l’établissement de listes de vocabulaire pour qualifier les personnes et les métiers auxquels ces statuts sont associés.

Enfin, le projet Time Us visant à reconstituer les budgets temps des travailleur·ses du textile, les expressions contenant des informations sur les **durées ou rythmes de travail**, les **montants**, les **quantités** et les **types de rémunérations** sont d’un intérêt particulier. S’y ajoute, bien sûr, l’expression des **rémunérations**.

Finalement, est jugée intéressante l’identification des **sources bibliographiques** soit servant de référentiels pour contextualiser une information (quand on se réfère à un règlement pour évoquer le montant d’une rémunération par exemple), soit utiles pour éventuellement découvrir de nouvelles sources à traiter.¹³

4.2.2 Une annotation qui pose des difficultés

Plusieurs difficultés émanant des documents ou du contexte d’annotation ont dû être prises en compte pour l’établissement du modèle et du guide d’annotation.

4.2.2.1 Un corpus hétérogène

Le corpus est constitué de documents courant sur plus de deux siècles, durant lesquels la pratique de la langue française a évolué. Ces variations sont doublées des spécificités langagières régionales. De plus, en fonction des contextes de rédaction, l’expression et l’attention portée à telle ou telle information n’est pas la même. Étant donné l’hétérogénéité du corpus, établir un modèle d’annotation valide à tout moment s’est révélé difficile : en particulier pour définir les règles d’annotation et créer des catégories d’informations

12. Dans la mesure où le projet vise également les activités domestiques et/ou non-rémunérées, on ne peut pas parler de « métier » seulement.

13. Cf. Annexes D.1.2.

suffisamment larges pour être valides dans autant de documents que possible, sans pour autant devenir trop vagues. En dépit du manque d'homogénéité du corpus, les objectifs du projet restent les mêmes. S'en tenir à un seul modèle d'annotation est nécessaire pour ne pas compliquer la phase de traitement automatique, mais cela représente un défi.

4.2.2.2 Des données difficiles à définir

Étant donné la variété des documents traités, il est difficile de prévoir la forme prise par certaines informations. C'est en particulier le cas des rémunérations. Pour cette raison, l'annotation des rémunérations est prise dans un sens extrêmement large. Il s'agit de signaler les ensembles textuels constituant l'expression d'une rémunération : ces ensembles pourront servir de base pour correctement paramétriser l'outil d'analyse syntaxique afin qu'il parvienne à reconstituer les informations sur les rémunérations. A une échelle moindre, le même problème est posé par les statuts matrimoniaux, en particulier pour la période moderne ou le début du XIX^e siècle. Les indices sur le statut matrimonial d'une personne ne sont pas toujours accolés à la mention de la personne, et elles sont parfois évoquées plus loin dans la même phrase ou dans le texte. Les durées et rythmes de travail n'y échappent pas non plus.

4.2.2.3 Des annotateur·rices aux compétences variables

Afin de prévenir les erreurs humaines, prendre en compte le niveau de compétence des annotateur·rices est crucial. Dans le cas du projet Time Us, une partie de l'annotation manuelle doit être réalisée par des « petites mains ». Ni spécialistes de l'annotation sémantique, ni informé·es de tous les enjeux de cette annotation au-delà de leur domaine de compétence, ces annotateur·rices doivent pouvoir utiliser des outils qui empêchent autant que possible les erreurs prévisibles. Ne pas se baser uniquement sur le travail d'annotateur·rices expérimenté·es permet de prendre en compte cette difficulté.

4.3 Outils et stratégies pour l'annotation

4.3.1 Oxygen XML Editor pour un encodage manuel

Initialement, comme pour la transcription, l'annotation des textes devait être réalisée manuellement. Elle devait alors prendre la forme d'une tâche d'encodage des textes bruts dans un éditeur de texte, en particulier dans *Oxygen XML Editor*¹⁴. A cette fin, deux

14. Oxygen XML Editor est une plate-forme rassemblant un éditeur de texte avancé pour la rédaction et l'édition de fichiers XML, et un outil de débugage pour XSLT et xQuery. La plate-forme est développée par SyncRo Soft Ltd. Elle est actuellement dans sa version 20.0. Elle est distribuée pour tous les systèmes d'exploitation, sous licence propriétaire.

sessions de formation à la TEI et à l'utilisation d'Oxygen ont été organisées. La première s'est tenue le 11 juillet 2017 à Paris, la deuxième le 19 octobre 2017 à Aix-en-Provence.

Une telle stratégie d'annotation présente plusieurs inconvénients. D'une part, l'encodage manuel, pour un corpus tel que celui rassemblé pour le projet Time Us, est une tâche lourde et répétitive. La qualité des annotations est difficile à contrôler, alors même que les annotateur·rices ont plus de chance d'oublier des annotations si celle-ci devient trop lourde sur une même portion de texte. Ils et elles risquent alors davantage de mal réaliser la segmentation des portions de texte annotées, de faire des fautes de frappes ou de se tromper dans le choix des balises. D'autre part, l'utilisation d'un éditeur de texte et l'encodage de celui-ci confrontent les membres du projet à des objets informatiques dont ils et elles n'ont pas nécessairement l'habitude, voire face auxquels ils et elles sont parfois réticents. Une personne mal à l'aise sur son outil de travail est moins efficace pour réaliser une pareille tâche. Enfin, *Oxygen XML Editor* étant un logiciel propriétaire, son utilisation nécessite l'achat de licences, ce qui induit un coût supplémentaire pour le projet.

La découverte de la plate-forme Transkribus et sa mise en oeuvre dans le cadre du projet Time Us a conduit à abandonner l'utilisation d'*Oxygen XML Editor* pour l'annotation. C'est une meilleure solution, à plusieurs titres.

4.3.2 Transkribus : annoter dans une interface graphique

La plate-forme Transkribus propose un outil d'annotation des transcriptions. Les modalités d'utilisation de cet outil ont fait l'objet d'un *How to guide*¹⁵, il n'est donc pas nécessaire de revenir en détail sur son fonctionnement en dehors des points qui nous intéressent. L'utilisation de Transkribus pour réaliser l'annotation dans le cadre du projet présente trois avantages : une interface graphique ergonomique qui rend la tâche plus agréable, une application des annotations par le biais de « tags » qui limite les risques de fautes dans les noms de balises, et enfin l'utilisation d'un seul et même outil pour réaliser la transcription et l'annotation du corpus, avec la possibilité d'en faire un export au format TEI. Dans ces conditions, l'adoption de l'outil Transkribus pour réaliser la phase d'annotation a paru évidente.

L'outil d'annotation est accessible depuis l'onglet « Metadata » et sa section « Textra ». Dès son installation, la plate-forme est dotée d'un certain nombre de *tags* qui visent avant toute chose à permettre une annotation imitative du texte et la création d'index des personnes, des lieux et des organisations. Ces *tags* ne peuvent pas être supprimés. A titre d'exemple, « abbrev » permet de signaler les abréviations, « blackening » les segments de texte qui ne doivent pas être publiés à l'export (comme les informations personnelles), « sic » les erreurs contenues dans le texte d'origine. S'y ajoutent des *tags*

15. *How to enrich transcribed documents with mark-up*, URL : https://transkribus.eu/wiki/images/e/e8/How_to_enrich_transcribed_documents_with_mark-up.pdf.

sémantiques comme « person », « place » et « organization » qui permettent d'annoter les entités nommées destinées à créer des index.

L'utilisation de Transkribus est relativement simple tant qu'elle reste dans les modes d'application prévus par l'équipe de développement. Nous aurons l'occasion de revenir sur les difficultés posées par cet outil dans le cadre du projet Time Us et sur les moyens mis en oeuvre dans le cadre de mon stage pour les contourner.

4.4 Le guide d'annotation

Une première réflexion sur l'utilisation des *tags* dans Transkribus et sur les éléments à annoter avait été mise en place avant le mois d'avril 2018. Menée en particulier par Charles Riondet et Marie Puren, elle avait conduit à l'établissement d'une liste de 8 *tags*¹⁶ et à la création d'un guide interne pour leur création et leur utilisation sur la plate-forme.

personnes
organisations
lieux
tâches
dates
montant
occupation
type de rémunération

TABLE 4.1 – *Tags* définis avant le mois d'avril.

Pour mener à bien ma mission, deux phases ont été nécessaire : il a fallu d'une part comprendre les spécificités et les difficultés posées par le corpus à annoter, et d'autre part identifier les besoins et les souhaits des historien·nes impliqué·es dans le projet. Plusieurs rendez-vous physiques ou téléphoniques ont permis de dresser un premier état des lieux avant la fin du mois de mai. En parallèle, cet état des lieux a été alimenté par l'analyse d'un travail de transcription et d'annotation réalisé sur une partie de la sous-collection intitulée « AD69 9M5 », correspondant au fond 9M5 des A.D. du Rhône entre février et mars 2018. Cette analyse a permis une entrée en matière très instructive pour la suite sur les enjeux du guide d'annotation.

4.4.1 Analyse de l'annotation de « AD69 9M5 »

L'analyse méthodique de l'annotation de la sous-collection a été réalisée sur les pages 3 à 17 en repérant et en notant dans un tableau de suivi¹⁷ l'ensemble des questions susci-

16. Cf. Table 4.1.

17. Cf. Annexe D.1.1.

tées par les choix d'annotation ainsi que les éventuelles corrections nécessaires. Il est vite apparu que ces observations pouvaient être réparties selon 5 catégories dont l'une, dédiée aux remarques sur la transcription, ne concerne pas directement l'annotation. Outre les questions méthodologiques rattachées à des cas précis, le tableau rendait compte d'erreurs telles que 1) l'usage d'un mauvais *tag* ou le mauvais usage d'un *tag*, 2) des oubli d'annotation, 3) des erreurs de délimitation des segments annotés.

Cette dernière catégorie d'erreurs se traduisait de deux manières : dans un cas il s'agit de segments qui comprenaient les espaces ou les éléments de ponctuations autour de l'ensemble ciblé, dans l'autre, il s'agissait d'une erreur de délimitation du segment qui s'étendait au-delà de l'expression ou de l'entité à annoter, ou au contraire qui ne la recouvrait pas totalement. Un exemple typique de ce dernier point : l'application d'un *tag* pour identifier un métier sur le mot « ouvrier » au lieu de l'étendre également sur l'expression « du tissage mécanique » qui le suit. En cause dans les deux cas, l'absence de consignes méthodologiques. En outre, on peut supposer dans le deuxième cas une mauvaise compréhension de l'expression à annoter, une erreur que le guide d'annotation doit permettre d'éliminer ou d'éviter au maximum.

Les oubli d'annotation ont généralement deux origines : soit une simple erreur humaine qui rend nécessaire un passage de contrôle sur les annotations, par le ou la même annotateur·rice ou par un·e autre ; soit un manque méthodologique sur la nécessité d'annoter *systématiquement* les informations, qu'elles aient ou non déjà été données dans le document.

Enfin, le mauvais usage d'un *tag* ou l'emploi du mauvais *tag* sont directement liés à l'absence d'un guide d'annotation, dont le rôle est justement de délimiter clairement le champ d'action de telle ou telle catégorie d'annotation. On peut mettre en cause un autre élément : au lieu d'utiliser le *tag* « occupation » pour signaler les noms de métier, un *tag* qui devait être créé manuellement, c'est le *tag* « work » qui a été utilisé. Sa traduction peut effectivement être comprise comme métier, mais dans le cas de Transkribus, « work » sert à signaler les références bibliographiques (*work* devant alors être traduit par « oeuvre »). J'ai tiré de cette erreur la conclusion qu'il était nécessaire d'employer des noms de balises non ambigus. Finalement, le fait que le *tag* prévu pour annoter les noms de métier n'ait pas été créé en temps voulu est révélateur des éventuelles difficultés rencontrées lors de la création des *tags* dans Transkribus.

Les questions soulevées à l'occasion de la lecture attentive des documents ont rendu claire la nécessité de discuter directement avec les historien·nes du projet des manières dont ils et elles jugeaient utiles d'y répondre.

4.4.2 Conclusion de l'enquête auprès des historien·nes

Plusieurs discussions téléphoniques avec Marie Lauricella, une rencontre avec Anaïs Albert, le 17 avril à Paris, et avec Anne Montenach, le 9 mai à Aix-en-Provence, m'ont permis de mieux comprendre les différents types de documents sur lesquels allait porter l'annotation, et d'identifier des stratégies d'annotation utiles pour répondre aux besoins alors exprimés.

Elles ont confirmé le besoin d'identifier les informations telles que les noms de personnes et de lieux, les rémunérations/paiements, les tâches et les métiers, tout en précisant certains cas particuliers propre à leur corpus et à leur démarches. Anaïs Albert a ainsi rappelé l'intérêt d'inclure certains verbes d'action exprimant des tâches précises dans les documents. Une annotation plus poussée aurait pu intégrer un traitement spécifique de tous les verbes et des modaux employés afin de rendre compte des relations de subordination telles qu'elles apparaissent dans l'énonciation. Chacun des deux corpus présentait des particularité dans le cas de l'expression des rémunérations. Les compte-rendus du Conseil de Prud'hommes de Paris sur lesquels travaille Anaïs Albert décrivent par exemple les montants des indemnités pour « jour perdu », un montant intéressant car il correspond au coût alors estimé d'une journée perdue pour un·e ouvrièr·e mobilisé·e par un jugement. Pour le corpus des procès de litiges de travailleurs lyonnais¹⁸, étudiés par Anne Montenach et Hugue Serveau, il a été souligné l'intérêt de distinguer paiement exigé et paiement réellement obtenu, voire la nécessité par ailleurs de repérer les modalités d'échelonnement des paiements, qui reflètent souvent la conjoncture économique du procès. Ces deux observations concernant la question des rémunérations ou des paiements ont permis de mettre en évidence le besoin d'inclure une typologie des rémunérations, que celle-ci soit ou non explicite dans le texte annoté.

Outre ces éléments qui avaient déjà été identifiés avant le mois d'avril, ces discussions ont fait émerger le besoin d'annoter de nouvelles informations, notamment :

- les **produits** réalisés dans le cadre de l'activité d'un·e ouvrièr·e ;
- les **sommes** engagées dans le cadre d'une rémunération ou d'un paiement ;
- le **statut social** des travailleur·ses, en particulier pour distinguer les apprentis des autres ouvriers ;
- le **genre** des individus et plus particulièrement le statut matrimonial des femmes ;
- les **quantités**, qu'il s'agisse de produits ou d'effectifs ;
- les **informations de temporalité** liées au travail, soit par l'expression des échéances, soit par celle de la durée de réalisation d'une tâche ou d'un produit, soit par l'expression d'un rythme associé à une tâche¹⁹.

18. Il s'agit des différentes sous-séries de la série HH des A.M. de Lyon.

19. Pour l'interprétation par l'historien·ne, une durée ou une échéance n'a en effet que peu de sens si elle n'est pas associée à un rythme de travail.

Anaïs Albert et Anne Montenach ont toutes deux souligné la structuration logique de leurs sources, pour lesquelles il serait intéressant de réaliser un travail similaire à celui mis en place pour les monographies de Le Play. Dans un cas, il s'agit de distinguer la présentation du contexte juridique et le récit du conflit réalisé à l'occasion du jugement prud'homal, le contexte juridique étant identifié comme la partie la plus riche en informations pour le projet. Dans l'autre cas, il s'agirait de pouvoir recréer la structure des différents compte-rendu de jugements car plusieurs affaires sont renvoyées faute de parution. En définissant une structure pour ces registres, on pourrait mettre en place un système d'identifiants et de renvois pour accéder directement au jugement final, d'une affaire donnée.

Pour finir, plusieurs discussions avec Éric de la Clergerie ont mené à la création de deux mécanismes d'annotation supplémentaires : l'**échappement** et les **commentaires**. L'échappement consiste en la possibilité laissée à l'annotateur·rice de ne pas trancher dans le cas d'une annotation problématique. Cela est nécessaire pour garantir la qualité de l'annotation : en l'absence d'un tel mécanisme, l'annotateur·rice pourrait omettre l'annotation problématique (au risque de perdre une information), ou trancher pour une solution qu'il ou elle sait insatisfaisante (donnant l'illusion d'une annotation non problématique, et au risque de perdre l'information mal annotée). L'outil de création de commentaires existe dans Transkribus. L'échappement, en revanche, a pris la forme d'un *tag* spécifique, éventuellement associé à un commentaire afin que l'annotateur·rice puisse apporter des précisions sur l'origine du doute.

4.4.3 Élaboration des règles et des *tags*

4.4.3.1 La TEI au cœur du modèle d'annotation

Étant donné le format final souhaité pour la sortie des documents annotés, la consultation des *guidelines* de la TEI²⁰ a constitué un élément clef pour établir le modèle d'annotation. Les solutions d'encodage pour un type d'information donné peuvent être très nombreuses. Souvent, il a fallu trancher et choisir la solution la plus appropriée à la forme prise par les noeuds de texte pouvant être contenus dans les éléments choisis. Souvent, c'est aussi l'usage de l'élément TEI qui a déterminé une partie des règles d'annotation.

Par exemple, une personne peut être balisée de plusieurs manières en TEI :

- <rs type="person">
- <name type="person">
- <person>
- <persName>

²⁰ TEI Consortium, *TEI P5 : Guidelines for Electronic Text Encoding and Interchange*. Version 3.4.0. Dernière modification : 23/07/2018. TEI Consortium. <http://www.tei-c.org/Guidelines/P5/> (consulté le 31/07/2018).

Chacune de ces modalités d'encodage induit des contraintes plus ou moins lourdes. Il a fallu s'interroger sur ce que le projet souhaitait identifier dans le texte. Alors que `<persName>` s'attache à la mention d'une personne par son nom, `<person>`, au contraire, peut être appliqué chaque fois que le concept d'une personne est évoqué, quelque soit le référent. Pour le projet Time Us, étant donné qu'il s'agit de dresser une liste de noms de personnes, l'élément `<persName>` est le plus approprié. Il en va de même pour les noms de lieux et les noms d'organisations.

Certaines informations comme les adresses, les références bibliographiques, et les expressions de date et de temps possèdent des éléments spécifiques en TEI dont l'usage est suffisamment libre pour qu'ils puissent être contenu dans les éléments composant la structure du document.

Il est nécessaire que les éléments signalés comme problématiques puissent être identifiés dans la sortie TEI, bien qu'ils aient vocation à disparaître à terme. Nous avons choisi d'utiliser l'élément `<certainty>` associé à attribut `@cert` dont la valeur de sortie serait fixée à « low ». En fonction du traitement apporté à ces cas, la valeur de l'attribut pourrait ensuite être modifiée.

Enfin, nous avons vu qu'il était difficile de prédire à quoi ressemble l'expression de certaines informations dans le corpus : leurs formes varient trop. Pour cette raison, il est nécessaire d'opter pour une balise peu contraignante. Le choix a donc été fait de rendre le reste des informations dans des éléments `<rs>`, associés à un attribut `@type` permettant de distinguer différentes catégories.

4.4.3.2 Nommage et propriétés des *tags*

Comme nous l'avons évoqué, il était nécessaire d'éviter au maximum les ambiguïtés induites par les noms des *tags* créés dans Transkribus. C'est la raison pour laquelle, il n'était pas envisageable de créer ces *tags* en les nommant d'après les éléments TEI qui les manifestent dans les textes de sortie. A l'exception de ceux empruntés au jeu de *tags* prédéfini par Transkribus, la totalité des *tags* créés pour le projet Time Us a été nommée en français et en évitant autant que possible les formulations abrégées. Transkribus gérant mal les caractère accentués, ceux-ci ne sont pas présents dans les tags. On obtient donc « quantite » au lieu de « quantité », ce qui ne pose pas de problème majeur de compréhension. En outre, afin de distinguer les *tags* réalisés sur mesure pour le projet Time Us, ils sont systématiquement précédés du préfixe « TU_ ».

Les *tags* Transkribus peuvent posséder des propriétés. Celles-ci permettent d'ajouter des éléments de spécification semblables à des attributs. Leurs valeurs peuvent être fixes ou entrées manuellement par l'annotateur·rice. Nous avons associé des propriétés « type » ou « subtype » aux *tags* « TU_montant », « TU_document » et « TU_typeRemuneration ». Elles visent à catégoriser ces informations car comme nous l'avons vu dans les cas des rémunérations, celles-ci peuvent être de plusieurs nature. Il en va de même pour les documents

FIGURE 4.2 – *Tags* et équivalents TEI en juillet 2018.

Information	Tag	TEI
Échappement	TU_incertitude	<certainty cert="low">
commentaire		<!-- -->
Adresse	TU_adresse	<address>
Nom de personne	TU_personne	<persName>
Nom d'organisation	organization	<orgName>
Nom de lieux	place	<placeName>
Date	date	<date>
Heure	TU_heure	<time>
Référence bibliographique	TU_document + type	<bibl type="{}">
Somme d'argent	TU_montant + subtype	<measure type="sum" subtype="{}">
Quantité	TU_quantite + unit	<measure type="count" unit="{}">
Métier ou occupation	TU_occurrence + normal	<choice><orig><rs type="occupation">[occupation]</rs></orig><reg><rs type="occupation">{normal}</rs></reg></choice>
Tâche	TU_tache	<rs type="task">
Produit	TU_produit	<rs type="product">
Statut de travailleur-se	TU_statut	<rs type="workerStatus">
Information de durée ou rythme	TU_duree	<rs type="duration">
Expression d'une rémunération ou d'un paiement	TU_remuneration	<s type="revenue">
Expression d'un type de rémunération	TU_typeRemuneration + subtype	<rs type="revenue-type" subtype="{}">
Statut matrimonial	TU_statutMatrimonial	<rs type="matStatus">

auxquels les textes font parfois référence (il est important de distinguer les documents juridiques et les articles de presses, par exemple, car leur exploitation ne sera pas la même à terme). Enfin, les montants sont parfois exprimés de manière relative plutôt qu'absolue, ce qui peut poser un problème pour leur interprétation par la machine. C'est en prévision de cela qu'a été créée la propriété « subtype » qui s'y rattache. La propriété « unit » que nous avons associée au tag « TU_quantite » a pour but de préciser l'unité de mesure d'une quantité. Enfin, afin de reconstruire une terminologie des noms de métiers, il a semblé important de prévoir un mécanisme permettant de normaliser la formulation d'un métier lorsque celle-ci est modifiée par la syntaxe de la phrase. C'est le cas dans une expression comme « les ouvriers du tissage mécanique et ceux du tissage à bras ». La propriété « normal » est prévu pour recevoir la version normalisée de l'expression dans de tels cas de figure.

4.4.4 La rédaction du guide d'annotation

4.4.4.1 Un modèle pour le guide

En raison de sa simplicité de lecture et d'usage, c'est le *Guide d'annotation des entités nommées* du projet d'annotation du corpus ISTEX par le Laboratoire d'Informatique (LI) de l'Université François Rabelais de Tours et l'INIST-CNRS de Vandoeuvre-lès-Nancy²¹ qui a servi de modèle pour la formulation de nos règles d'annotation. Nous nous sommes inspirés de sa structure, le guide étant divisé en deux parties : la première dédiée à des considérations méthodologiques et scientifiques, la seconde aux règles d'annotation illustrées d'exemples concrets. Plusieurs entités identifiées pour le projet ISTEX recoupent celles du projet Time Us. Il s'agit notamment des personnes, des lieux administratifs (par opposition aux lieux naturels), des dates, des organisations et des références bibliographiques. En outre, le projet ISTEX s'est basé sur les *guidelines* de la TEI pour élaborer son système de balisage et ses règles d'annotation, en ce point, il est donc très semblable au nôtre.

Le guide d'annotation est publié sur le wiki du projet Time Us²² sur deux pages web. La première²³ comporte des considérations générales, des remarques méthodologiques sur les bonnes pratiques d'annotation, et des petits guides sur l'utilisation de l'annotation dans Transkribus. La deuxième page²⁴ correspond à la liste des *tags* prévus pour l'annotation du corpus Time Us. Les *tags* sont groupés par catégories logiques, le but étant d'aider l'annotateur·ice à circuler dans la liste :

1. Problèmes d'annotation
2. Personnes et organisations
3. Localisations
4. Temps
5. Rémunérations et activités

21. ISTEX (Initiative d'excellence de l'Information Scientifique et Technique) est un projet d'acquisition de ressources scientifiques et de création de bibliothèque numérique mis en place dans le cadre d'un financement du Ministère en charge de l'Enseignement Supérieur, de la Recherche et de l'Innovation. Il est porté par le CNRS, l'Agence Bibliographique de l'Enseignement Supérieur (ABES), le Consortium Universitaire de Publications Numériques (Couperin) et l'Université de Lorraine. Le projet ambitionne d'implémenter les meilleurs standards sur les corpus rassemblés et mis à disposition dans le cadre de ce projet, et d'offrir la plus grande variété de modalités de consultation des collections possible. *ISTEX - Socle de la bibliothèque scientifique numérique nationale*, URL : <https://www.istex.fr/> (visité le 12/08/2018).

22. *Guide d'annotation*, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_d'annotation (visité le 12/08/2018).

23. *Guide d'annotation : remarques générales*, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_d%27annotation:_remarques_g%C3%A9n%C3%A9rales (visité le 11/08/2018).

24. *Guide d'annotation : tags du projet Time Us*, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_d%27annotation:_tags_du_projet_Time_Us (visité le 11/08/2018).

6. Références bibliographies

Au sein de chaque catégorie, les *tags* sont suivis d'une part l'énumération de leurs règles d'application, d'autre part de différents exemples illustrant les cas typiques ou précisant l'annotation à adopter dans les cas problématiques identifiés. Chaque exemple est tiré du corpus.

4.4.4.2 Afficher l'annotation sur le wiki

Comme son nom l'indique, l'extension TEITags permet de prendre en charge des éléments TEI dans les corps de texte publiés sur des sites réalisés avec le CMS MediaWiki. En réalité, l'extension ne se limite pas uniquement au standard TEI : elle prend en charge n'importe quelle balise XML, dès lors que celle-ci est déclarée. Appliquer un style aux éléments XML insérés dans le wikicode permet de rendre plus visible les segments de texte annoté. De plus, lorsque l'utilisateur·rice survole le segment annoté, une info-bulle affiche le nom de la balise²⁵. C'est pour profiter de cette fonctionnalité que nous avons créé des balises nommées d'après les *tag* Transkribus, et non d'après les éléments TEI de sortie.

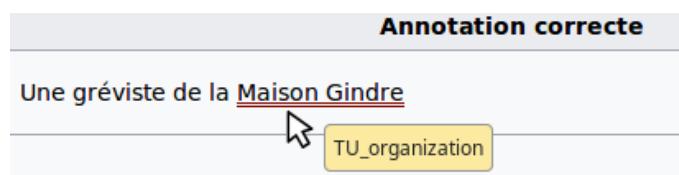


FIGURE 4.3 – Capture d'écran : affichage des info-bulles par TEITags.

Des lignes de style et des commandes ont dû être ajoutées respectivement dans les *scripts* CSS et PHP. Ces deux fichiers se trouvent sur le serveur hébergeant le site internet, dans le dossier de l'extension TEITags. Il s'agit des fichiers `TEITags.body.php` et `ext.teitags.css` présents dans les dossiers `TEITags/` et `TEITags/css/` situés, sur le serveur, à l'emplacement suivant : `/var/www/html/mediawiki/extensions/`.

Le fichier `.php` crée une classe nommée « `TEITagsHooks` » dans laquelle sont définies plusieurs fonctions publiques²⁶.

La première, `ParserFirstCallInit()`, permet de paramétriser la manière dont une balise va être « parsée » en pointant vers une autre fonction nommée d'après le nom de la balise :

```
1 $parser->setHook('{tagname}', array($this, 'Render{tagname}'))
```

Chaque balise est associée à une fonction publique, également déclarée au sein de la classe « `TEITagsHooks` » :

25. Cf. Figure 4.3.

26. Cf. Figure 4.4.

```

35 class TEITagsHooks {
36
37     public function ParserFirstCallInit ( Parser $parser ){
38         global $wgOut;
39         // ...
40
41         $parser->setHook( 'TU_adresse' , array( $this , 'RenderTU_Adresse' ) );
42         // ...
43
44         $wgOut->addModules( 'ext.TEITags' );
45
46         return true;
47
48     }
49     // ...
50
51     public function RenderTU_Adresse(){
52         $HookArgs = func_get_args();
53         return $this->TEITagsRenderer( 'TU_adresse' , $HookArgs );
54     }
55     // ...

```

FIGURE 4.4 – Extraits du fichier TEITags.body.php.

```

1 public function Render{tagname}(){
2     $HookArgs = func_get_args();
3     return $this->TEITagsRenderer('{$tagname}', $HookArgs);
4 }

```

La seule compréhension de cette articulation entre les déclarations dans le *script* php suffisait pour l'adapter à nos besoins. Pour chaque nom de *tag* Transkribus, on reprend la structure identifiée et on remplace *{tagname}* par le nom du *tag*.

TEITags crée des éléments dans le code HTML, dont la classe correspond au modèle `tei-{tagname}`. Dans `ext.teitags.css`, il a suffit d'ajouter les propriétés CSS souhaitées à chacun des éléments dont la classe correspond à un *tag* Transkribus déclaré dans `TEITags.body.php`.

```

44 span.tei-TU_adresse {text-decoration: underline;
                      text-decoration-style: double;
                      text-decoration-color:#34D334; }

```

FIGURE 4.5 – Extrait du fichier ext.teitags.css.

4.4.5 Garantir l'intégrité du jeu de *tags* sur tous les postes

Si Transkribus permet le travail de transcription et d'annotation de manière collaborative par le biais d'un serveur géré par la plate-forme, la configuration des *tags*, elle, se fait localement. Au moment de la création des *tags* pour Time Us, l'annotateur·rice risque d'oublier un *tag* ou de commettre une faute de frappe. De plus, c'est une manipulation longue, puisqu'il s'agit de créer les 15 *tags* personnalisés du projet Time Us.

Il est d'autant plus important que les annotateur·rices débutent leur travail avec un jeu de *tags* fixe que la modification ultérieure des *tags* n'est pas reportée dans le document annoté. Par exemple, si l'on modifie le *tag* « TU_occupation » en « TU_occurrences », toutes les annotations créées avant cette modification seront toujours associées au *tag* « TU_occupation », quand bien même il n'existerait plus. Il en va de même pour les propriétés des tags.

Nous voulons un corpus annoté homogène. Dès le mois d'avril, il a paru nécessaire de trouver une solution permettant de contrôler la création des *tags* dans Transkribus pour les futur·es annotateur·rices. Cela s'est fait par le biais du fichier `config.properties`²⁷ présent dans le dossier de l'application, quel que soit le système d'exploitation. Deux variables déclarées dans ce fichier sont importantes pour cette prise de contrôle :

1. `tagNames`
2. `tagSpecs`

La première permet d'ajouter des *tags* personnalisés dans l'application dès son lancement, sans passer par l'interface graphique. `tagNames` prend comme valeur une chaîne de caractères constituée de la liste des *tags* séparés par de simples espaces. Les *tags* sont formulés selon le modèle suivant : `nom-de-tag{codeHEX, propriété(s)}`. Les propriétés sont séparées par des virgules et, comme pour la création des *tags* dans l'interface graphique, leur dénomination ne doit pas contenir d'accent. Le code hexadécimal, `codeHEX`, n'est pas obligatoire mais il permet d'attribuer au *tag* une couleur fixe. Nous avons choisi de paramétriser la couleur des *tags* pour suggérer des ensembles de *tags* par le biais de camaïeux. Par exemple, « TU_adresse » et « place » sont deux nuances différentes de vert.

La seconde ligne permet de créer une liste personnalisée de *tags* qui seront affichés dans l'interface, en mode « Tag Specifications »²⁸. Le but de cette manipulation est de supprimer les concurrents que constituent les *tags* par défaut de Transkribus (en particulier « work » et « person »). Les *tags* sont donnés comme valeur de la liste `tagSpecs`, séparés, comme il se doit, par des virgules, et selon le modèle suivant : `{\"customTag\": \"nom-de-tag {}\"}`, où « `nom-de-tag` » doit être remplacé par le nom du *tag*.

Le « Guide pour l'installation de la liste des tags Time Us »²⁹ du wiki, rédigé dans le cadre de mon stage, détaille la marche à suivre pour créer les *tags* en passant par le fichier `config.properties`. Afin de ménager les annotateur·rices et limiter les risques de mauvaise manipulation, nous invitons l'utilisateur·rice à remplacer le fichier

27. Cf. Annexes D.3.1.

28. Il existe deux modes d'affichage des *tags* : soit « Tag Specifications », pour afficher une liste personnalisée, soit « All Tags », pour afficher tous les *tags* disponibles.

29. *Guide pour l'installation de la liste des tags Time Us*, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_pour_l%27installation_de_la_liste_des_tags_Time_Us (visité le 11/08/2018)

`config.properties` par un fichier préparé à l'avance par nos soins. Cela permet de simplifier la manipulation au maximum. Cependant, étant donné que certain·es annotateur·rices peuvent être amené·es à travailler avec Transkribus pour d'autres projets Time Us, nous avons jugé important de donner également des instructions pour intervenir dans le fichier `config.properties` et ajouter manuellement les *tags* par ce biais si nécessaire. Cela évite de risquer de supprimer des paramétrages réalisés dans un autre contexte par l'utilisateur·rice. En effet, écraser le fichier `config.properties` est solution radicale : il contient toutes les informations de paramétrage réalisées localement par l'utilisateur·rice.

4.4.6 Un guide toujours en phase d'élaboration

En juillet 2018, le guide d'annotation est toujours en cours d'élaboration car il fait l'objet de tests réalisés par plusieurs partenaires du projet et sur les différents corpus afin d'identifier les incohérences du guide et du modèle, les difficultés d'application des règles et les éventuels manques du modèle.

La première version du guide a été élaborée à partir de la sous-collection « AD69 9M5 ». Il a ensuite été testé par une personne tierce sur les pages 11 à 20 de la même sous-collection durant le mois de mai. Ce premier test a abouti à l'ajout d'un *tag* pour identifier les quantités. Il a conduit à s'interroger sur la manière de distinguer la dénomination d'un métier de celle d'un statut de travailleur : « patron », « chef d'atelier » ou encore « ouvrier » indique d'avantage des statuts vis-à-vis des outils de travail et au sein de la hiérarchie des travailleur·ses que des missions ou tâches précises dans la chaîne de production. Le test a permis de confirmer la plupart des règles déjà établies pour l'annotation, tout en mettant en évidence quelques ambiguïtés que l'ajout d'exemples d'annotation a pu régler.

Une fois le modèle et le guide d'annotation mis à jour, celui-ci a été testé par un partenaire de l'équipe TELEMMÉ en deux sessions sur deux sous-collections créées pour l'exercice. Les documents annotés étaient tirés de sous-séries de la série HH des A.M. de Lyon. La première session n'a pas permis de tester le guide d'annotation lui-même car il n'a pas été appliqué par l'annotateur, qui a tenté de réaliser l'annotation en se basant uniquement sur les noms de balises. Cette expérience a cependant été utile pour se rendre compte de la réticence des annotateur·rices à lire le guide d'annotation et à s'y conformer. C'est une réaction à laquelle on doit s'attendre et pour laquelle il convient de trouver des moyens de rendre la lecture du guide plus agréable, voire ludique. A l'occasion de la deuxième session d'annotation, les retours du testeur ont confirmé le besoin de distinguer statut de travailleur et métier, notamment pour différencier, dans le cas de ces documents, des statuts comme ceux d'apprentis, de maître ou encore de compagnon. Par ailleurs, là aussi, ses retours ont permis d'identifier des cas d'annotation qui nécessitaient l'ajout de règles ou d'exemples pour clarifier les ambiguïtés.

Le guide d'annotation doit encore être testé, notamment sur le corpus parisien,

à nouveau par un·e annotateur·rice expérimenté·e connaissant à la fois les enjeux d'une telle annotation de corpus, les objectifs du projet et les documents à annoter. Cela permet d'identifier les besoins pour le projet Time Us réellement, et pas uniquement les besoins en terme d'encadrement de l'annotateur·rice.

Les deux expériences de test du guide d'annotation, et les différents échanges qui ont conduits à son élaboration, ont mis en évidence la difficulté d'établir un modèle satisfaisant et compréhensible et de mettre en place une communication efficace autour de ces outils. En outre, obtenir des retours sur le guide et le modèle prend du temps. Dans le cadre du projet Time Us, la mise en place d'une session d'annotation collective et intensive peut être imaginée dès le mois de septembre. Cet « annotathon » permettrait de rassembler l'ensemble des partenaires du projet amenés à réaliser l'annotation des documents pour discuter des solutions à adopter afin de régler les questions encore non résolues.

Chapitre 5

Traiter les données

5.1 Un export TEI insatisfaisant

Lorsque nous avons présenté la plate-forme Transkribus, nous avons évoqué la manière dont fonctionne l'outil d'export des fichiers. L'existence d'une telle possibilité est un point extrêmement positif à première vue. Toutefois, cet outil pose de nombreux problèmes d'utilisation et la qualité insatisfaisante des exports nous a un temps conduit à nous interroger sur la nécessité d'avoir finalement recours à un autre outil que Transkribus pour réaliser l'annotation, comme par exemple TXM, un logiciel de textométrie multi-plate-forme développé par l'ANR Textométrie depuis 2007.

5.1.1 Un traitement insuffisant des *tags* personnalisés

La qualité des fichiers TEI obtenus n'est pas sans poser problème, en particulier lorsque l'annotation prévoit de recourir à des *tags* créés sur mesure pour le projet.

Transkribus ne permet pas de paramétriser le comportement du parseur de *tags* au moment de l'export TEI. Seuls ceux prévus par défaut pour la plate-forme font l'objet d'un traitement spécifique qui les transforme en éléments TEI, mais ces paramètres sont intégrés directement dans les classes java qui les définissent dans le cœur du programme. Tous les *tags* créés par l'utilisateur·rice sont transposés tels quels dans le fichier exporté. Le *tag* intitulé « TU_adresse » donne donc lieu à la création d'une balise « <TU_adresse> » dans le fichier XML de sortie. S'il possède une propriété nommée « type », celle-ci est également transposée telle quelle dans la balise correspondante, sous la forme d'un attribut.

Dans ces conditions, à moins de modifier le cœur du code de la plate-forme, l'utilisateur·rice n'a pas d'autre choix que d'avoir recours à une feuille de transformation XSLT¹

1. Pour *eXtensible Stylesheet Language Transformation*, il s'agit d'un langage de transformation exprimé en XML, permettant de créer, à partir de fichiers XML ou HTML, de fichiers dans un format et/ou selon un schéma différent. Pour appliquer une feuille de style XSLT, on donne à un parseur java, tel que Saxon, un fichier d'entrée XML à modifier et une feuille de style en XSLT, et on obtient le ou les fichier(s) de sortie souhaité(s).

pour opérer un post-traitement sur le fichier XML de sortie, afin d'obtenir un document réellement conforme au standard de la TEI. C'est ce que nous avons fait dans un premier temps avec le fichier `modifTag.xsl`².

Cette transformation consistait simplement à copier la totalité du fichier en modifiant l'apparence des balises et les attributs créés à partir des *tags* personnalisés, afin d'obtenir des éléments conformes aux recommandations de la TEI.

FIGURE 5.1 – Extraits de modifTag.xsl.

```
<?xml version="1.0" encoding="utf-8"?>
<xsl:stylesheet version="2.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
  xmlns="http://www.tei-c.org/ns/1.0" xpath-default-namespace="http://www.tei-c.org/ns/1.0">
  <xsl:output method="xml" encoding="utf-8" indent="yes"/>

  <xsl:template match="node() | @*">
    <xsl:copy>
      <xsl:apply-templates select="node() | @*"/>
    </xsl:copy>
  </xsl:template>

  <!-- ... -->
  <xsl:template match="TU_adresse">
    <address><xsl:value-of select=". "/></address>
  </xsl:template>

  <!-- ... -->
</xsl:stylesheet>
```

5.1.2 Une mauvaise gestion des *tags* étendues sur plusieurs lignes

Il doit être possible d'extraire les éléments annotés pour créer, comme nous l'avons évoqué, des listes pour la terminologie ou des index. Cette opération est compliquée par deux aspects de la transformation réalisée par Transkribus vers la TEI.

En premier lieu, chaque annotation est circonscrite à la ligne sur laquelle elle porte. Lorsqu'une annotation s'étend sur plusieurs lignes, elle est segmentée en autant de morceaux qu'il y a de lignes.

FIGURE 5.2 – Exemple d'annotation segmentée.

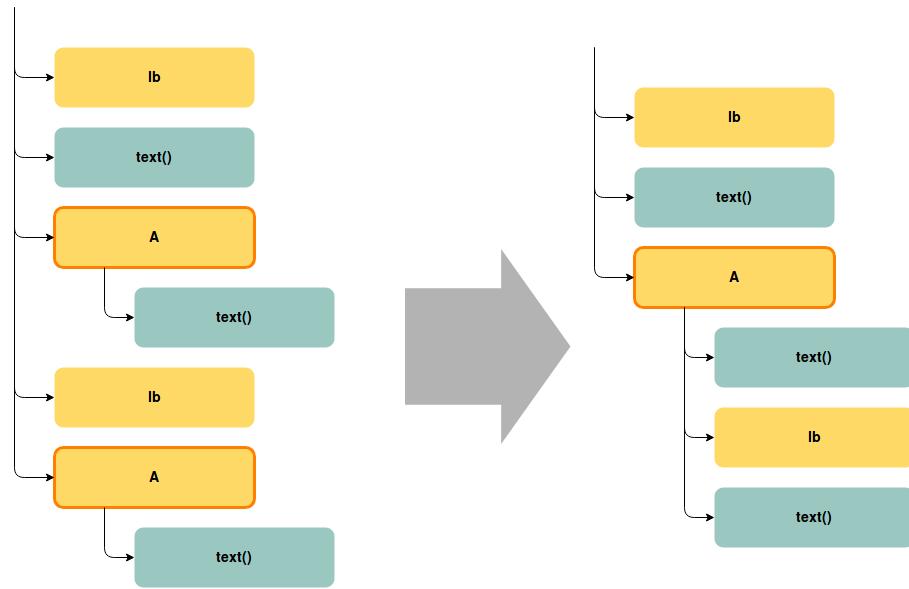
```
<lb n='N007'/>à faire connaître ces conditions à leurs collègues <TU_occupation>chefs</TU_occupation>
<lb n='N008'/><TU_occupation>d&apos;ateliers</TU_occupation> dans la grande réunion qui aura lieu ce <TU_heure>soir</TU_heure>
```

Cela signifie qu'il est nécessaire de reconstruire les ensembles avant de pouvoir les exploiter. Quoique laborieuse, la tâche n'est pas impossible, car on peut envisager de se baser sur la structure de l'arbre TEI pour identifier les éléments identiques situés directement de part et d'autre d'une balise `<lb>` et que cette balise interrompt³.

2. Cf. Figure 5.1 ; cf. Annexes F.1.

3. Cf. Figure 5.3.

FIGURE 5.3 – Schématisation de la reconstruction des annotations segmentées.



Néanmoins, ce processus est encore compliqué par le fait que les imbrications d'éléments ne sont pas nécessairement réalisées dans le même ordre d'une ligne à l'autre. Si Transkribus permet d'appliquer plusieurs *tags* sur une même portion de texte, lorsque ces annotations s'étendent sur plusieurs lignes, leur ordre d'imbrication n'est pas calculé de la même manière. Sur la première ligne où ils apparaissent, l'ordre dépend du premier élément ouvert ; sur la dernière ligne, il dépend du dernier élément fermé. Si les *tags* sont ouverts et/ou fermés au même emplacement dans la ligne de texte, si ces *tags* s'étendent sur plus de deux lignes, pour la ou les ligne(s) intermédiaire(s), l'ordre des éléments dépendra d'un ordre alphabétique basé sur le nom des balises.

FIGURE 5.4 – Exemple d'annotations segmentées se chevauchant.

```

<lb n='N008'>grévistes <TU_remuneration>pouvaient toucher <TU_duree>tous les
soirs</TU_duree> la somme </TU_remuneration>
<lb n='N009'><TU_remuneration>de <TU_montant>1f50</TU_montant></TU_remuneration>
au <placeName>siège du <orgName>syndicat des ouvriers tisseurs</orgName></placeName>
<lb n='N010'><orgName><placeName>et similaires</placeName></orgName>,
<TU_adresse>rue d&apos;Austerlitz 8</TU_adresse>, et que lui, s&apos;il

```

Cette situation n'est pas sans solution, mais elle nécessite alors une transformation lourde et qui ne garantit pas que les éléments recomposés l'auront été correctement.

5.1.3 Des métadonnées pauvres

Les métadonnées contenues dans l'élément `<teiHeader>` des fichiers TEI exportés sont peu détaillées. L'interface graphique de Transkribus permet à l'utilisateur·rice de remplir des champs définis qui sont ensuite pour partie reportés dans les fichiers exportés. Ces champs n'obéissent cependant pas à un modèle standard de métadonnées et

leur pertinence dans le cadre du projet Time Us reste relativement limitée. Les champs « auteur (*author*) », « genre » et « écrivain (*writer*) » ne sont pas utiles pour des documents d'archives. Seuls les champs « titre (*title*) », « date d'écriture (*date of writing*) », « langue (*language*) » et « description » sont intéressants. Par ailleurs, Transkribus permet de créer des déclarations sur les choix éditoriaux ayant guidé la transcription, telles qu'elles peuvent être formulées dans l'élément <*editorialDecl*> de la TEI.

Même lorsque ces champs sont remplis, leur export vers la TEI n'est pas toujours pris en charge. C'est par exemple le cas du champ « langue ». Certaines métadonnées par défaut sont caduques : le contenu de l'élément <*publisher*>, créé par Transkribus lors de la création du fichier exporté, indique « Transcriptorium », soit l'ancien nom du projet Transkribus.

FIGURE 5.5 – Exemple de métadonnées obtenues à l'export depuis Transkribus.

```
<?xml version='1.0' encoding='UTF-8'?>
<TEI xmlns='http://www.tei-c.org/ns/1.0'>
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title type='main'>AD69 9M5</title>
      </titleStmt>
      <publicationStmt>
        <publisher>transcriotorium</publisher>
      </publicationStmt>
      <sourceDesc>
        <bibl><publisher>TRP document creator: charles.riondet_</publisher></bibl>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
```

La construction du <*teiHeader*> telle qu'elle est réalisée par Transkribus nécessite d'être revue. Soit par le biais d'une intervention manuelle destinée à en corriger le contenu et à ajouter des informations (par exemple les noms des personnes ayant participé à l'élaboration du fichier). Soit sous la forme d'une transformation XSLT qui permettrait d'ajouter des éléments de description communs à tous les documents créés dans le cadre du projet Time Us.

5.1.4 Une manipulation lente

L'utilisation de l'outil d'export dans l'interface graphique de Transkribus est un processus lent et répétitif. L'export des fichiers est limité à chaque document, qu'il faut charger un à un, ce qui prend souvent plus de dix secondes (en fonction du nombre de pages et de données qu'il contient). La manipulation pour paramétriser l'export doit être répétée à chaque fois, car l'export est par défaut dirigé vers l'option « Transkribus Document ». Il faut donc décocher ces cases, indiquer le format d'export TEI, puis cocher le bon rendu des retours à la ligne (« <1b/> ») car celui activé par défaut est le rendu « <1>...</1> ». De plus, il peut être nécessaire de modifier à chaque export le chemin de dossier vers lequel

celui-ci est réalisé. A ces inconvénients s'ajoute la lenteur de l'export lui-même, qui peut prendre jusqu'à 12 minutes pour un fichier de 450 pages, pourtant transcrit et annoté à moins de 50%. Dans ces conditions il est nécessaire de trouver une autre méthode pour réaliser un export de manière automatisée.

5.2 L'API Transkribus pour contourner ces problèmes

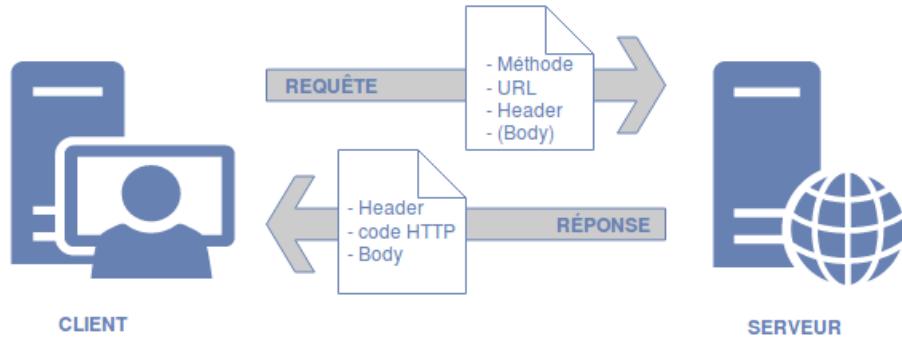
La plate-forme Transkribus est construite sur la base d'une interaction client-serveur. Une telle structure est supportée par l'existence d'une API⁴ de type REST⁵. Cette interface a constitué un élément important pour mieux contrôler l'export des fichiers depuis Transkribus et trouver des solutions alternatives aux problèmes que nous avons soulevés.

5.2.1 Prendre en main les requêtes HTTP

5.2.1.1 Définition des requêtes HTTP

Les API REST fonctionnent grâce à des requêtes HTTP. HTTP, pour *HyperText Transfer Protocol*, est un protocole de communication entre client et serveur pour le *web*. Il permet le transfert des données entre l'un et l'autre de la manière suivante :

FIGURE 5.6 – Schématisation de l'interaction Client-Serveur.



1. Le client envoie une requête HTTP typée par une méthode et composé d'une adresse de requête sous la forme d'une URL (pour *Uniform Resource Locator*), d'un *header* et éventuellement d'un corps, appelé *body*. La méthode permet de spécifier le cadre de l'interaction entre le client et le serveur : une méthode comme GET indique qu'il s'agit d'une simple consultation des ressources du serveur, tandis que des méthodes comme POST ou DELETE modifient les ressources sur le serveur (en les créant, modifiant ou supprimant). Le *header* contient des informations supplémentaires sur la requête qui permettent de la contextualiser ou de la spécifier. Le *body* agit de

4. Application Programming Interface ; un ensemble de requêtes HTTP permettant d'interagir avec un serveur et ses données sans nécessairement passer par une interface graphique.

5. Representational State Transfer ; il s'agit d'un modèle de formulation de requêtes HTTP.

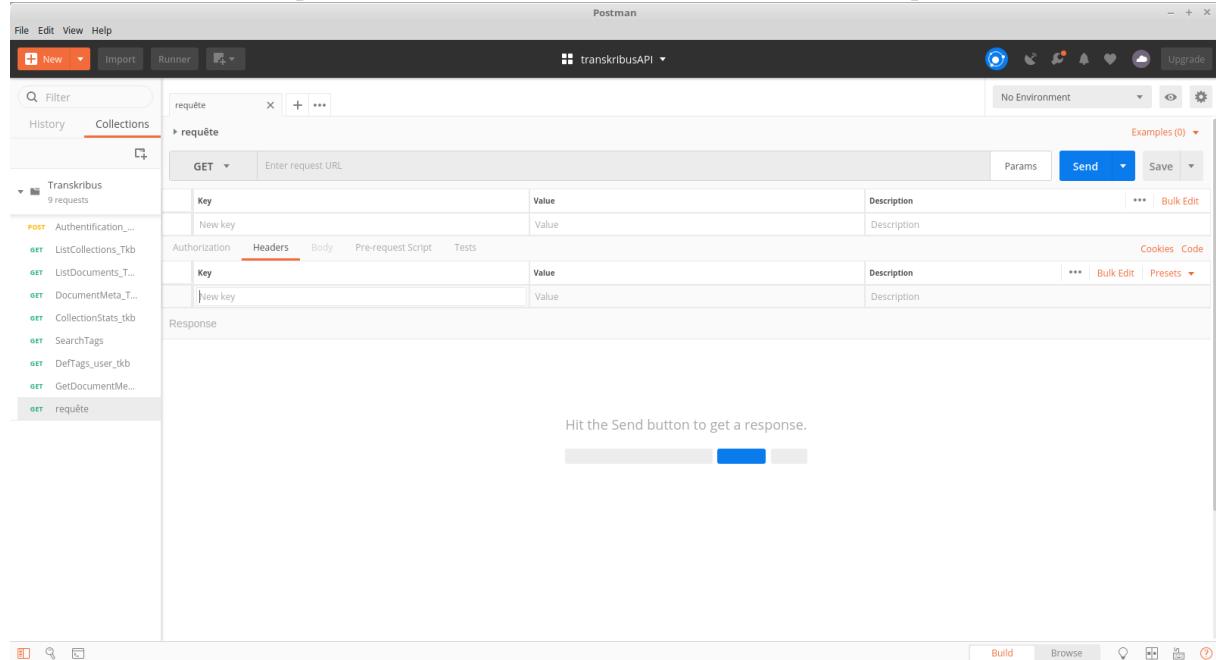
façon similaire en communiquant les informations supplémentaires nécessaires à la formulation de certaines requêtes.

- Le serveur retourne une réponse au client. Son statut est indiqué par un code HTTP, qui permet d'indiquer les erreurs de traitement rencontrées par le serveur (404, pour un erreur de ressource, 500 ou 503, pour des erreurs de serveur, par exemple) ou au contraire son bon fonctionnement (200 le plus souvent). La réponse du serveur est composée d'un *header*, qui donne des informations sur la réponse, et d'un *body*, qui contient le corps de la réponse, sous des formes diverses, telles que JSON⁶, XML, HTML, ou encore du texte brut.

5.2.1.2 Postman : l'interface graphique pour gérer les requêtes

Nous avons utilisé la plate-forme Postman⁷, qui propose un environnement de développement pour les API, afin de tester le fonctionnement de ces requêtes. Postman est composé de deux éléments : une application installée localement, où l'utilisateur·rice peut manipuler ses requêtes⁸, et une interface en ligne où l'utilisateur·rice ne peut que consulter ses collections et les rendre publiques⁹.

FIGURE 5.7 – Capture d'écran de l'interface de création de requêtes de Postman.



Chaque utilisateur enregistré sur Postman peut accéder à un tableau de bord dans lequel sont créés des espaces de travail. Chaque espace de travail est constitué d'une ou

6. *JavaScript Object Notation*

7. *Postman*, URL : <https://www.getpostman.com> (visité le 03/08/2018).

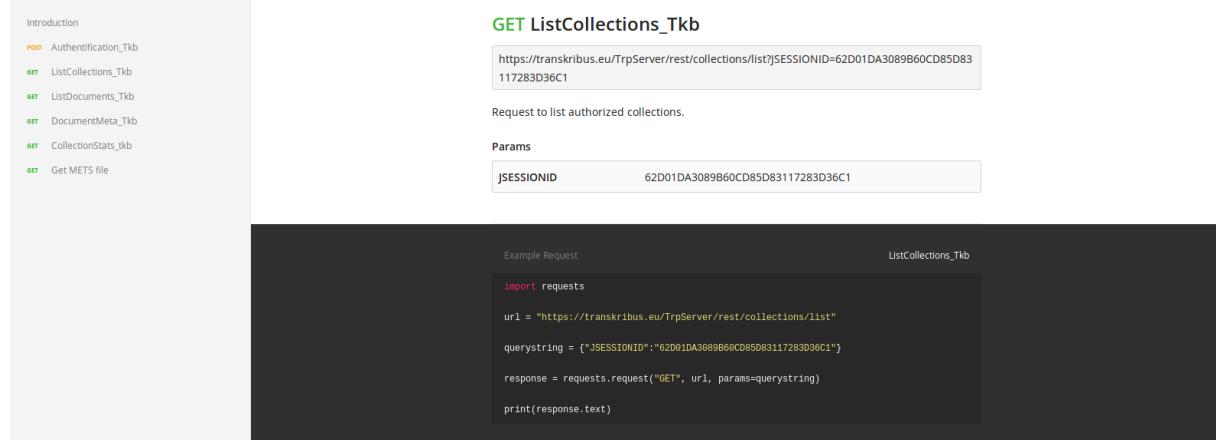
8. Cf. Figure 5.7.

9. Cf. Figure 5.8.

plusieurs collection(s) dans lesquelles sont rassemblées des requêtes HTTP. L'application Postman propose une interface relativement complète pour créer les requêtes. L'utilisateur·rice spécifie la méthode, l'URL, les éventuels arguments (*params*) et le *header* de sa requête. Celle-ci peut alors être envoyée et/ou enregistrée. Postman permet donc de tester facilement la validité des requêtes pour les enregistrer pour un usage ultérieur. Lorsqu'une requête est valide, la réponse HTTP est affichée dans l'interface.

Le système de collections de requêtes dans Postman permet de créer de la documentation personnalisée sur une API. Celle-ci est consultable depuis l'interface en ligne. Chaque requête est affichée sous la forme d'une url, accompagnée d'une documentation rédigée par l'utilisateur·rice et du détail des éléments composant la requête. En parallèle de cette partie descriptive, s'affiche un exemple de formulation de la requête dans un langage choisi parmi ceux proposés¹⁰. C'est ce dernier aspect de l'interface en ligne que nous avons utilisé.

FIGURE 5.8 – Capture d'écran de l'interface de consultation en ligne de Postman.



Postman a été un outil précieux pour tester l'interface REST de Transkribus et formuler les requêtes dont nous avions besoin, tout en ayant facilement accès à des exemples de réponse pour paramétriser notre outil.

5.2.2 Les requêtes de l'API de Transkribus

L'ensemble des requêtes disponibles dans le cadre de l'API de Transkribus est documenté grâce à un fichier au format WADL¹¹. En complément, le fonctionnement d'un certain nombre de ces requêtes est décrit sur le wiki de Transkribus¹². Plusieurs de ces requêtes nous intéressent en particulier.

10. cURL, jQuery, Ruby, Node, Go, PHP ou encore Python requests.

11. WADL pour *Web Application Description Language*. - Transkribus, *Transkribus REST Interface Description*, URL : <https://transkribus.eu/TrpServer/Swadl/wadl.html> (visité le 03/08/2018).

12. Id., *REST Interface - Transkribus Wiki*, URL : https://transkribus.eu/wiki/index.php/REST_Interface (visité le 03/08/2018).

5.2.2.1 Authentification

Toute interaction avec le serveur de Transkribus doit passer en tout premier lieu par une authentification :

```
https://transkribus.eu/TrpServer/rest/auth/login
```

Envoyée en méthode POST, l'URL prend deux paramètres : *user* et *pw*, dont les valeurs sont respectivement l'adresse mail d'identification de l'utilisateur·rice sur Transkribus et son mot de passe. Pour que la requête fonctionne, il est nécessaire qu'un *content-type* soit précisé dans le *header* de la requête : il s'agit du type MIME¹³ « application/x-www-form-urlencoded ».

Cette requête renvoie un fichier au format XML contenant diverses informations sur l'utilisateur·rice, ainsi qu'un élément <sessionId> dont la valeur permet de garantir l'authentification de l'utilisateur·rice dans les requêtes suivantes¹⁴.

5.2.2.2 Récupération des données

Trois requêtes en mode GET permettent d'accéder aux contenus disponibles pour un·e utilisateur·rice identifié·e. Chacune de ces requêtes prend l'argument *JSESSIONID* dont la valeur correspond à celle de l'élément <sessionId> récupéré après authentification.

Il est intéressant de vérifier que l'utilisateur·rice a accès à la collection avec laquelle on souhaite travailler. Deux requêtes permettent de faire cela. L'une renvoie une réponse au format XML, l'autre au format JSON¹⁵ :

```
https://transkribus.eu/TrpServer/rest/collections/list
```

```
https://transkribus.eu/TrpServer/rest/collections/list.xml
```

Quel que soit le format de sortie, la liste constituant la réponse est organisée de la même manière : pour chaque collection sont données des métadonnées, dont le nom de la collection (*colName*) et son identifiant (*colId*). L'identifiant de la collection est un identifiant unique et stable.

Pour lister les documents contenus dans une collection, deux requêtes sont possibles, là encore en fonction du format de sortie souhaité. L'URL contient l'identifiant de la collection dont on souhaite connaître le contenu :

```
https://transkribus.eu/TrpServer/rest/collections/{collection-ID}/list
```

13. *Multipurpose Internet Mail Extension* ; un standard pour décrire les formats de données textuelles ou multimédia et assurer leur correct encodage/décodage dans le cadre des communications sur les *web*.

14. Cf. « Authentication.xml », dans Annexes B.2.1.

15. Cf. « ListCollection.* », dans Annexes B.2.1.

FIGURE 5.9 – Extrait de la réponse à la requête « collections/list ».

```
{
  "type": "trpCollection",
  "colId": 8097,
  "colName": "timeUS",
  "description": "created by charles.riondet[REDACTED]",
  "crowdsourcing": false,
  "elearning": false,
  "pageId": 1181082,
  "url": "https://dbis-thure.uibk.ac.at/f/Get?id=SVIOUMFEMMCHGWVODERFMU&fileType=view",
  "thumbUrl": "https://dbis-thure.uibk.ac.at/f/Get?id=SVIOUMFEMMCHGWVODERFMU&fileType=thumb",
  "nrOfDocuments": 26,
  "role": "Owner"
},
```

<https://transkribus.eu/TrpServer/rest/collections/{collection-ID}/list.xml>

Chaque document dans la collection est listé avec un certain nombre de métadonnées, en particulier son titre et un identifiant unique¹⁶.

FIGURE 5.10 – Extrait de la réponse à la requête « collections/8097/list ».

```
{
  "type": "trpDocMetadata",
  "docId": 41459,
  "title": "AD69 9MS",
  "uploadTimestamp": 1521133660775,
  "scriptType": "HANDWRITTEN",
  "uploader": "charles.riondet[REDACTED]",
  "uploaderId": 4197,
  "nrOfPages": 451,
  "pageId": 1735916,
  "url": "https://dbis-thure.uibk.ac.at/f/Get?id=WJNLBXYXGOKCJZNNZPIGX&fileType=view",
  "thumbUrl": "https://dbis-thure.uibk.ac.at/f/Get?id=WJNLBXYXGOKCJZNNZPIGX&fileType=thumb",
  "status": 0,
  "fimgStoreColl": "TrpDoc DEA_41459",
  "createdFromTimestamp": 0,
  "createdToTimestamp": 0,
  "collectionList": {
    "collist": [
      {
        "colId": 8097,
        "colName": "timeUS",
        "description": "created by charles.riondet[REDACTED]",
        "crowdsourcing": false,
        "elearning": false,
        "nrOfDocuments": 0
      }
    ]
  }
},
```

Finalement, on peut accéder à l'ensemble des transcriptions disponibles pour un document donné grâce aux requêtes suivantes :

<https://transkribus.eu/TrpServer/rest/collections/{collection-ID}/>

16. Cf. « ListDocument.* », dans Annexes B.2.1.

```
{document-ID}/fulldoc

https://transkribus.eu/TrpServer/rest/collections/{collection-ID}/
{document-ID}/fulldoc.xml
```

La réponse obtenue contient les métadonnées de la collection et du document, ainsi que la liste des pages composant le document. Pour chaque page, une liste des différentes transcriptions enregistrées est donnée. Chaque élément dans cette liste contient des métadonnées sur la transcription, comme le nombre de régions de texte, de lignes, de mots, etc, mais également le statut associé à la transcription ainsi qu'une URL de requête de type GET permettant de récupérer un fichier XML contenant la transcription¹⁷.

FIGURE 5.11 – Extrait de la réponse à la requête « collections/8097/41459/fulldoc ».

```
"pageList": {
  "pages": [
    {
      "pageId": 1735916,
      "docId": 41459,
      "pageNr": 1,
      "key": "WJNLBXYGOKKCJZNNZPPIGX",
      "imageId": 1122408,
      "url": "https://dbis-thure.uibk.ac.at/f/Get?id=WJNLBXYGOKKCJZNNZPPIGX&fileType=view",
      "thumbUrl": "https://dbis-thure.uibk.ac.at/f/Get?id=WJNLBXYGOKKCJZNNZPPIGX&fileType=thumb",
      "imgFileName": "IMG_0119.JPG",
      "tsList": {
        "transcripts": [
          {
            "tsId": 4663221,
            "parentTsId": 2854015,
            "key": "OEIAHPZIVRLSSMAUENLIFKIB",
            "pageId": 1735916,
            "docId": 41459,
            "pageNr": 1,
            "url": "https://dbis-thure.uibk.ac.at/f/Get?id=OEIAHPZIVRLSSMAUENLIFKIB",
            "status": "DONE",
            "userName": "charles.riondet[REDACTED]",
            "userId": 4197,
            "timestamp": 1532332880979,
            "md5Sum": "",
            "nrOfRegions": 1,
            "nrOfTranscribedRegions": 1,
            "nrOfWordsInRegions": 6,
            "nrOfLines": 3,
            "nrOfTranscribedLines": 3,
            "nrOfWordsInLines": 6,
            "nrOfWords": 0,
            "nrOfTranscribedWords": 0
          }
        ]
      }
    }
  ]
}
```

```
https://dbis-thure.uibk.ac.at/f/Get?id={transcription-ID}
```

La transcription obtenue grâce à cette requête est donnée dans un fichier XML conforme au standard PAGE, qui gère l'alignement du texte et des segments d'images correspondants. C'est le seul format de transcription disponible par l'intermédiaire de l'interface REST de Transkribus¹⁸.

17. Cf. « fullDoc.* », dans Annexes B.2.1.

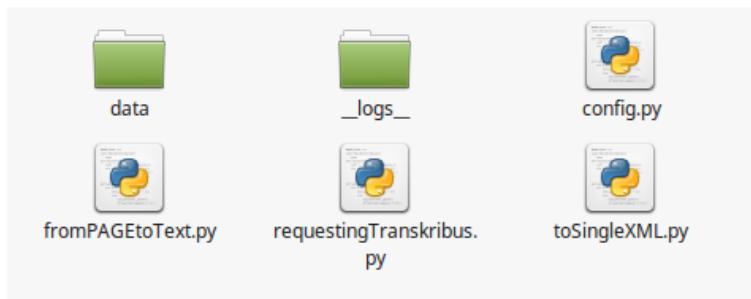
18. Cf. « fichierPage.xml », dans Annexes B.2.1.

5.2.3 Un *script* Python pour extraire les fichiers

Afin d'améliorer les conditions d'extraction des transcriptions, et la qualité des fichiers obtenus, j'ai réalisé un ensemble de *scripts* publiés sur la plate-forme Github sous le nom *UsingTranskribusAPI*¹⁹. Outre le *script* principal (`requestingTranskribus.py`), deux *scripts* permettent des transformations supplémentaires : `toSingleXML.py` et `fromPAGEToText.py`.

Ces trois fichiers doivent être exécutés dans une configuration minimale composée des dossiers `__logs__` et `data` et du fichier `config.py`²¹. Le dossier `__logs__` reçoit les fichiers de rapport créés lors de l'exécution de chacun des *scripts*, le dossier `data` reçoit les fichiers créés lors de l'exécution des *scripts*. Le fichier `config.py` permet à l'utilisateur·rice de paramétriser les *scripts*, en précisant en particulier le nom d'utilisateur Transkribus et son mot de passe, le nom des collections ciblées et les statuts de transcriptions visés.

FIGURE 5.12 – Configuration minimale du dossier pour exécuter les *scripts*.



L'exécution des *scripts* nécessite l'installation d'un environnement virtuel basé sur **python 3** et dans lequel les *packages* suivants doivent être installés :

- **requests** (version 2.19.1) : qui permet la création simplifiée de requêtes HTTP en Python.²²
- **lxml** (version 4.2.1) : parseur pour les fichiers XML et HTML en Python.²³
- **beautifulsoup4** (version 4.4) : qui permet une interaction simplifiée en Python avec les fichiers XML et HTML parsés.²⁴

Tous ces *packages* sont listés dans le fichier `requirements.txt` afin de faciliter leur installation par le biais du gestionnaire de paquets « pip ». Les instructions d'installations sont données dans le wiki du *repository* Github.

19. A. Chagué, *Github / UsingTranskribusAPI*, original-date : 2018-06-20T14:59:14Z, 30 juil. 2018, URL : <https://github.com/alix-tz/UsingTranskribusAPI> (visité le 03/08/2018).

20. Cf. Annexes B.2.2.

21. Cf. Figure 5.12.

22. *Requests 2.19.1 : documentation*, URL : <http://docs.python-requests.org/en/master/> (visité le 12/08/2018).

23. *lxml - Processing XML and HTML with Python*, URL : <https://lxml.de/> (visité le 12/08/2018).

24. *Beautiful Soup 4.4.0 : documentation*, URL : <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> (visité le 12/08/2018).

5.2.3.1 Le *script* principal : requestingTranskribus.py

Le fichier `requestingTranskribus.py` permet de télécharger des transcriptions au format XML-PAGE à partir de Transkribus en partant des données de connexion d'un·e utilisateur·rice, de noms de collections Transkribus et de statuts de transcription. Ces données sont fournies par l'utilisateur·rice par le biais du fichier `config.py`²⁵.

FIGURE 5.13 – Extrait du fichier config.py (l.1 à 19).

```

1 # -----
2 # for requestingTranskribus.py
3 #
4 # Transkribus username / nom d'utilisateur Transkribus
5 # ex : username = 'username@mail.fr'
6 username = ''
7
8 # Transkribus password / mot de passe Transkribus
9 # ex : password = 'mypassword'
10 password = ''
11
12 # Targeted collection name(s)
13 # ex : collectionnames = ['collectionname', 'anothercollectionname', 'yetanotherone']
14 collectionnames = []
15
16 # Targeted document status
17 # values can only be : 'NEW', 'IN PROGRESS', 'DONE' or 'FINAL'
18 # ex : status = ['DONE', 'IN PROGRESS'] or status = ['FINAL']
19 status = []

```

Outre les *packages* **requests** et **BeautifulSoup**, plusieurs module python sont importés en début d'exécution :

- **json** : qui permet de traiter les fichiers JSON²⁶
- **os** : qui permet l'interaction avec le système d'exploitation de l'utilisateur·rice²⁷
- **datetime** : qui permet de manipuler des objets de type date et heure²⁸

Le *script* exécute des tâches de vérification, de requêtes et de modification ou de création de fichiers. Il suit les étapes suivantes :

1. Vérification de la validité des statuts de transcription fournis par l'utilisateur·rice : seuls 4 statuts sont possibles (« NEW », « IN PROGRESS », « DONE » et « FINAL ») ; lorsqu'ils ne sont pas valides, les statuts sont ignorés ; si aucun statut n'est valide, le *script* est interrompu.

25. Cf. Figure 5.13.

26. *JavaScript Object Notation, Python 3.7.0 documentation : json*, URL : <https://docs.python.org/3/library/json.html> (visité le 12/08/2018).

27. *Python 3.7.0 documentation : os*, URL : <https://docs.python.org/3/library/os.html> (visité le 12/08/2018).

28. *Python 3.7.0 documentation : datetime*, URL : <https://docs.python.org/3/library/datetime.html> (visité le 12/08/2018).

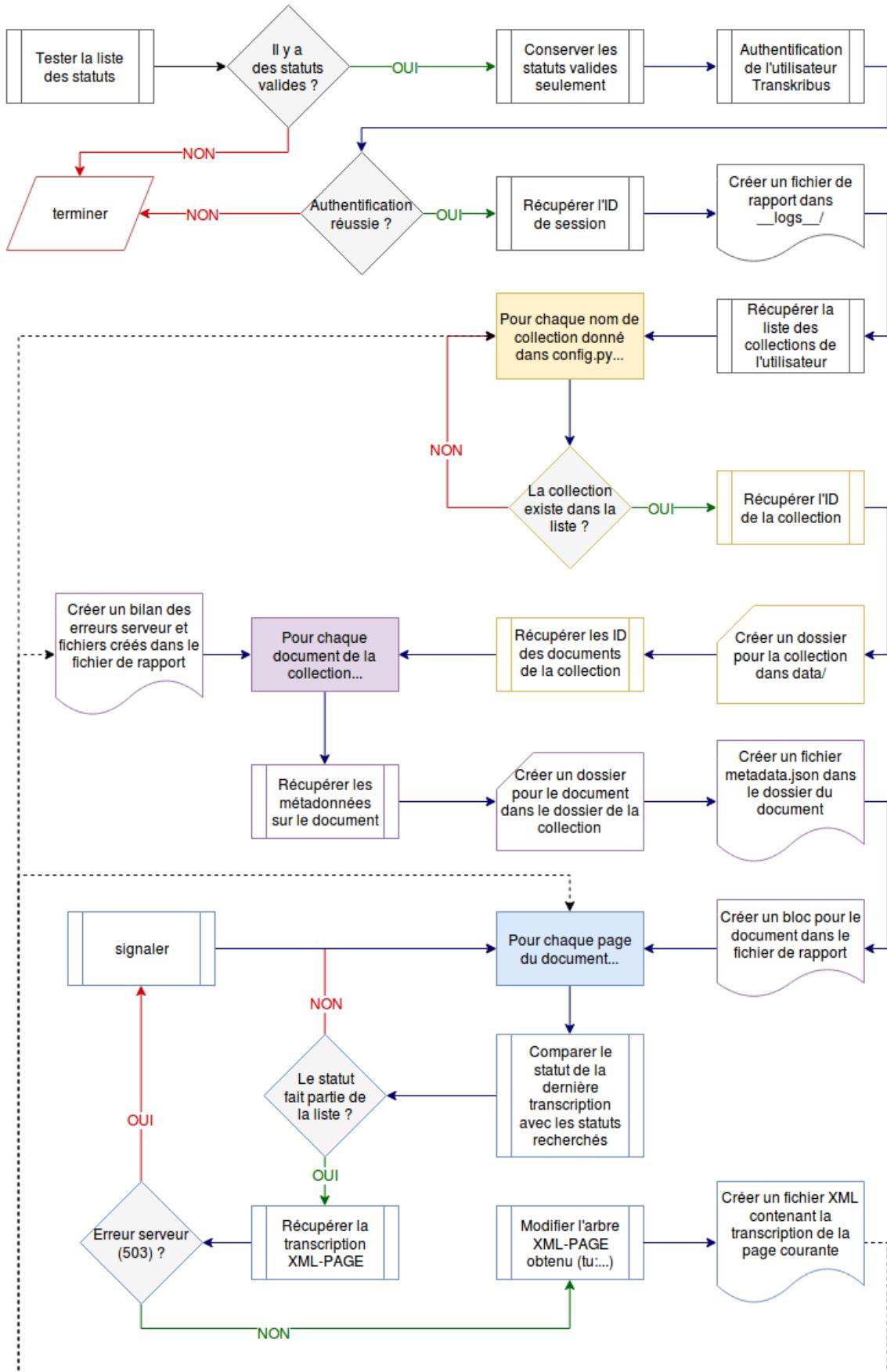
2. Authentification de l'utilisateur·rice : si elle échoue, l'utilisateur·rice en est averti·e et le *script* est interrompu, si elle réussit, l'identifiant de session est utilisé pour la suite des requêtes.
3. Pour chaque collection ciblée, récupération de l'identifiant de la collection et des identifiants de chaque document dans la collection.
4. Récupération des transcriptions correspondant au(x) statut(s) ciblé(s) : pour chaque document, un fichier `metadata.json` est créé pour récupérer les métadonnées de la collection, et pour chaque page dont la transcription possède le statut voulu, un fichier XML est créé, nommé d'après le numéro de la page, après avoir été modifié pour correspondre aux besoins du projet Time Us :
 - ajout de la déclaration d'un espace de nom « Time Us ».
 - ajout de métadonnées dans l'élément `<Metadata>` du fichier PAGE, sous l'espace de nom Time Us : il s'agit des éléments `<tu:pagenumber>`, `<tu:title>`, `<tu:desc>`, et `<tu:language>`.
 - ajout des attributs `@url` et `@id` dans l'espace de nom Time Us contenant, respectivement, l'URL de téléchargement de l'image du fac-similé, le numéro de la page transcrise.

Pour chaque collection, un dossier est créé dans le dossier `data`, nommé d'après le nom de la collection. Pour chaque document dans la collection, un dossier est créé dans le dossier correspondant à la collection, nommé d'après l'identifiant et le nom du document. Chaque transcription et chaque fichier `metadata.json` sont créés dans le dossier du document correspondant.

En outre, un fichier de rapport est créé dans le dossier `__logs__` ; il donne la description de chacun des documents composant la collection en listant le nombre de transcriptions disponibles pour chaque statut, et les numéros de pages correspondants au(x) statut(s) recherché(s). Plusieurs éléments de rapport sont également fournis dans le terminal afin de suivre l'exécution du *script*, qui peut prendre plusieurs minutes, en fonction de la vitesse de traitement du serveur de Transkribus.

La création de ce *script* répond donc à plusieurs problèmes soulevés par l'utilisation de l'interface graphique de Transkribus. Il est désormais possible de récupérer l'ensemble des transcriptions disponibles pour toute la collection en une seule manipulation. Celle-ci donne lieu à la création de fichiers XML-PAGE contenant des métadonnées plus complètes, qui pourront être reportées dans les fichiers XML-TEI créés à partir de ces fichiers XML-PAGE.

FIGURE 5.14 – Modélisation du script requestingTranskribus.py.



5.2.3.2 Deux *scripts* de post-traitement : `tosingleXML.py` et `fromPAGEtoText.py`

Les *scripts* `tosingleXML.py` et `fromPAGEtoText.py` n'interagissent qu'avec les fichiers XML-PAGE téléchargés dans le cadre de l'exécution de `requestingTranskribus.py`. Ils les modifient pour créer deux types de fichiers :

- soit un fichier `.txt` rassemblant l'ensemble des transcriptions disponibles pour un document sous forme de texte brut accompagné de marqueurs de pages et de zones de texte ;
- soit un fichier `.xml` reprenant le standard PAGE et rassemblant l'ensemble des éléments `<Page>` contenus dans les fichiers XML originaux au sein d'un élément `<tu:PageGrp>` créé pour les besoins du projet Time Us mais non conforme au standard PAGE.

Les deux *scripts* fonctionnent de manière similaire. Il utilisent les modules **BeautifulSoup**, **os** et **datetime** que nous avons déjà évoqués. La liste des noms de collection nécessaire à l'exécution des fichiers est donnée dans le fichier `config.py`²⁹. Pour chaque collection, les *scripts* parcourront le contenu du dossier de collection précédemment créé dans le dossier `data`. Un dossier `__TextExports__` dans le cas de la transformation vers un fichier `.txt`, ou `__AllInOne__` dans le cas de la transformation vers un fichier `.xml` est créé. C'est ce dossier qui reçoit le résultat de la transformation. Chaque fichier de synthèse est nommé d'après l'identifiant et le nom du document dont il rassemble les transcriptions.

FIGURE 5.15 – Extrait du fichier `config.py` (l.22 à 35).

```

22 # -----
23 # for fromPAGEtoText.py
24 # -----
25 # Targeted collection name(s). Collection must have been downloaded with requestingTranskribus.py first.
26 # ex : textcollectionnames = ['collectionname'] or textcollectionnames = ['firstcollection', 'secondcollection']
27 textcollectionnames = []
28
29
30 # -----
31 # for toSingleXML.py
32 # -----
33 # Targeted collection name(s). Collection must have been downloaded with requestingTranskribus.py first.
34 # ex : singlecollectionnames = ['collectionname'] or signlecollectionnames = ['firstcollection', 'secondcollection']
35 singlecollectionnames = []

```

Pour chaque dossier de document contenu dans le dossier de collection, le *script* rassemble, ordonne, transforme et fusionne les contenus des fichiers XML-PAGE. Les dossiers ne contenant aucun fichier XML-PAGE sont ignorés. L'exécution des *scripts* donne également lieu à la création d'un fichier de rapport dans le dossier `__logs__`. Y sont listés

29. Cf. Figure 5.15.

les noms des dossiers de document traités et le nombre de fichiers fusionnés pour chaque document.

La création des fichiers `.txt` permet d'obtenir rapidement les transcriptions en texte brut, tout en ne perdant pas la trace des différentes zones de texte, ce qui permet de reconstruire l'enchaînement logique des phrases et des paragraphes étendus sur plusieurs pages. Ces fichiers sont utiles aux spécialistes du traitement automatique de la langue, pour mettre en place des premiers tests sur leurs outils.

Les fichiers `.xml` sont des fichiers de transition à deux titres. D'une part, il s'agissait d'obtenir un fichier commun pour l'ensemble des transcriptions d'un document, à partir duquel réaliser la transformation vers le format XML-TEI. Dans un deuxième temps, cette transition n'apparaît finalement pas nécessaire car les futures améliorations du `script`, que nous évoquerons plus tard, devraient permettre de créer un fichier XML-TEI pour tout le document en partant directement de toutes les transcriptions disponibles pour un document donné, sans passer par la création de ce fichier intermédiaire.

5.2.4 Une feuille de style pour obtenir la TEI : `page2tei.xsl`

En prenant contact avec l'équipe de développement de Transkribus pour leur signaler les problèmes que nous rencontrions avec les exports TEI proposés par la plate-forme, nous avons été redirigés vers le travail de Dario Kampkaspar. Développeur à l'*Austrian Centre for Digital Humanities* à Vienne, il participe au projet Transkribus, notamment en concevant plusieurs feuilles de transformation, dont certaines sont d'ores et déjà implémentées dans la plate-forme. Parmi elles, un projet de transformation du standard PAGE vers le standard TEI, nommé « `page2tei` »³⁰.

L'ensemble de ces feuilles de transformation est publié sur le compte Github de Dario Kampkaspar (`dariok`)³¹, dans un *repository* intitulé « `page2tei` ». Nous avons copié certains fichiers de ce dossier dans un nouveau *repository*³² afin d'adapter la feuille de transformation aux besoins du projet Time Us. Cette feuille est rédigée en XSLT 3.0 et importe les fonctionnalités du fichier XSLT `string-pack.xsl` situé dans le même dossier. Nous avons donc conservé les fichiers `page2tei-0.xsl` et `string-pack.xsl`. Nous avons commencé ce travail à la fin du mois de juin, bien qu'il s'agit, du côté de Dario Kampkaspar, d'un travail toujours en cours. Plusieurs modifications mineures ont d'ailleurs été apportées au fichier `page2tei-0.xsl` depuis que nous avons début notre travail.

30. Dario Kampkaspar, *Github / page2tei*, original-date : 2018-05-02T16 :32 :49Z, 9 août 2018, URL : <https://github.com/dariok/page2tei> (visité le 12/08/2018).

31. Id., *Github / dariok*, GitHub, URL : <https://github.com/dariok> (visité le 12/08/2018).

32. A. Chagué, *Github / page2tei_TimeUS*, original-date : 2018-07-13T15 :04 :18Z, 23 juil. 2018, URL : https://github.com/alix-tz/page2tei_TimeUS (visité le 03/08/2018).

5.2.4.1 Adaptation du projet page2tei pour Time Us

La feuille de style telle qu'elle a été développée par Dario Kampkaspar prévoit d'être appliquée à un fichier XML-METS, tel qu'il peut être exporté depuis l'interface graphique de Transkribus, en choisissant le format de sortie « Transkribus Document ». Le fichier METS est utilisé pour établir un lien entre d'une part les fichiers XML au standard PAGE, qui contiennent le texte issu de la transcription et les coordonnées des zones de texte, et d'autre part les fichiers .png ou .jpg correspondant aux images à partir desquelles sont réalisées les transcriptions.

Le fichier `page2tei-0.xsl` de Dario Kampkaspar crée un fichier XML-TEI où les éléments `<facsimile>` renvoient vers des fichiers .png ou .jpg théoriquement situés dans le même environnement que les fichiers XML-PAGE. En outre, la valeur de l'attribut `@n` de chaque élément TEI signalant une page (`<pb>`) et celle des identifiants des éléments TEI `<facsimile>` sont calculées en prenant en compte le numéro de la page courante.

FIGURE 5.16 – Comparaison d'extraits des fichiers page2tei, original (l.32 à 35) et adapté (l.30 à 33).

```

32  <xd:doc>
33    <xd:desc>Entry point: start at the top of METS.xml</xd:desc>
34  </xd:doc>
35  <xsl:template match="/mets:mets">



```

Dans la mesure où l'API de Transkribus nous permet d'accéder directement aux fichiers XML-PAGE, sans obtenir de fichiers METS, ni de fichiers image, il s'est avéré nécessaire d'adapter le point d'entrée de la transformation. Alors que la feuille originale prenait comme point d'entrée l'élément racine du fichier XML-METS, `<mets>`, celui de notre feuille de style est l'élément racine du fichier XML-PAGE, `<PcGts>`³³.

Deux informations nécessaires à la création de fichiers TEI complets ne sont pas présentes dans le fichier XML-PAGE tel qu'il est fourni par Transkribus. Il s'agit de l'adresse (locale ou non) de l'image correspondant au fac-similé de la transcription, et du numéro de la page courante. C'est pour remédier à cela que nous avons créé les attributs `@tu:url` et `@tu:id` dans les fichiers XML-PAGE créés par le script `requestingTranskribus.py`³⁴.

33. Cf. Figure 5.16.

34. Cf. Figure 5.17.

FIGURE 5.17 – Comparaison d’extraits des fichiers PAGE, original et adapté.

```

1  <?xml version="1.0" encoding="utf-8"?>
2   <PcGts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15 http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15/pagecontent.xsd">
3     <Metadata>
4       <Creator>prov=University of Rostock/Institute of Mathematics/CITlab/Tobias Gruening/tobias.gruening@uni-rostock.de:name=/net_tf/LA73_249_0mod360.pb:de.uro.citlab.module.la.core.CITlab_LA_ML:v=?0.1 TRP</Creator>
5       <Created>2018-03-15T17:44:50.479+01:00</Created>
6       <LastChange>2018-05-07T11:28:53.843+02:00</LastChange>
7     </Metadata>
8     <Page imageFilename="IMG_0121.JPG" imageHeight="5184" imageWidth="3888">

```



```

1  <?xml version="1.0" encoding="utf-8"?>
2  <PcGts xmlns="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15" xmlns:tu="timeUs" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15 http://schema.primaresearch.org/PAGE/gts/pagecontent/2013-07-15/pagecontent.xsd">
3    <Metadata>
4      <Creator>prov=University of Rostock/Institute of Mathematics/CITlab/Tobias Gruening/tobias.gruening@uni-rostock.de:name=/net_tf/LA73_249_0mod360.pb:de.uro.citlab.module.la.core.CITlab_LA_ML:v=?0.1 TRP</Creator>
5      <Created>2018-03-15T17:44:50.479+01:00</Created>
6      <LastChange>2018-05-07T11:28:53.843+02:00</LastChange>
7      <tu:title>AD69 9M5</tu:title>
8      <tu:desc>No description.</tu:desc>
9      <tu:pagenumber>3</tu:pagenumber>
10    </Metadata>
11    <Page imageFilename="IMG_0121.JPG" imageHeight="5184" imageWidth="3888" tu:id="3" tu:url="https://dbis-thure.uibk.ac.at/f/Get?id=JTMXNSQVMNVHWGKLVRHPQCZK&fileType=view">

```

Le calcul de la valeur de la variable `numCurr`, utilisée dans le feuille de transformation pour construire l’identifiant des différents segments de texte, a ainsi été simplifié : c’est la valeur de l’attribut `@tu :id` du fichier PAGE. `numCurr` permet créer la valeur des attributs `@n` (ex : « `n="8"` »), et `@xml :id` et `@facs` (ex : « `xml:id="facs_8"` » et « `facs="#facs_8"` »).

Nous avons fait le choix de donner à l’élément `<graphic>`, qui permet de renvoyer vers le fichier image du fac-similé, un attribut `@url` dont la valeur correspond à celle de l’attribut `@tu :url` du fichier PAGE. C’est l’URL de la requête permettant de télécharger l’image depuis les serveurs de Transkribus. Cela permet à l’utilisateur·rice de contrôler à quel moment le téléchargement de l’image est nécessaire. En effet, celui-ci prend du temps pour une utilité très limitée. Conserver l’URL de téléchargement, une URL constante, permet de garder le lien avec le fichier image en évitant son téléchargement.

La feuille de transformation a été adaptée pour pouvoir prendre comme fichier source un fichier XML-PAGE créé par le *script requestingTranskribus.py* ou par le *script toSingleXML.py* : le *template* vise d’abord l’élément `<PageGrp>` s’il existe. Autrement, il traite directement l’élément `<Page>`, sans passer par `<PageGrp>`³⁵.

35. Voir en particulier les lignes 71 à 124 de `page2tei_TU.xsl`; cf. Figure 5.18.

FIGURE 5.18 – Extrait du fichier page2tei_TU.xsl (l.71 à 89).

```

71      <xsl:choose>
72          <xsl:when test="p:Page">
73              <xsl:apply-templates select="p:Page" mode="facsimile"/>
74          </xsl:when>
75          <xsl:when test="tu:PageGrp">
76              <xsl:apply-templates select="tu:PageGrp" mode="facsimile"/>
77          </xsl:when>
78      </xsl:choose>
79  </xsl:if>
80  <text>
81      <body>
82          <xsl:choose>
83              <xsl:when test="p:Page">
84                  <xsl:apply-templates select="p:Page" mode="text"/>
85              </xsl:when>
86              <xsl:when test="tu:PageGrp">
87                  <xsl:apply-templates select="tu:PageGrp" mode="text"/>
88              </xsl:when>
89          </xsl:choose>

```

5.2.4.2 Améliorer les métadonnées

De même que celles des fichiers TEI créés via l'outil d'export vers TEI de Transkribus, les métadonnées calculées à l'occasion de la transformation de PAGE vers TEI proposé par la feuille de Dario Kampkaspar n'étaient pas satisfaisantes. Nous avons complété les informations en créant des éléments constants, valides pour tous les fichiers TEI créés dans le cadre du projet Time Us. Il s'agit par exemple du contenu des éléments `<encodingDesc>` (qui décrit les conditions d'encodage du fichier original vers le fichier nativement numérique) et `<publicationStmt>` (qui précise l'autorité à l'origine de la création du fichier numérique).

Grâce à l'ajout des éléments `<tu:title>`, `<tu:pagenumber>`, `<tu:language>` et `<tu:desc>`, nous avons modifié le calcul du titre du fichier (il est basé sur le titre du document et le numéro de la page transcrrite quand le fichier XML-PAGE correspond à une seule page), et ajouté des métadonnées : en particulier, la ou les langue(s) du fichier transcrit (dans `<profileDesc>`) et la description rédigée par les transcripteur·rices et annotateur·rices (dans `<sourceDesc>`), lorsque ces champs ont été remplis. Nous avons par ailleurs conservé la plupart des métadonnées créées par la feuille de transformation originale, à l'exception de la mention « TranScriptorium » dans `<publisher>`. A ce stade, nous ne sommes pas encore parvenu à récupérer l'identité de la personne désignée comme créateur·rice du document dans Transkribus, une information qu'affiche toutefois le `<teiHeader>` du fichier TEI produit par Transkribus³⁶.

36. Voir en particulier les lignes 35 à 69 de `page2tei_TU.xsl`; cf. Figure 5.19.

FIGURE 5.19 – Comparaison des métadonnées créées par Transkribus et par page2tei_TU.



5.2.4.3 Modifier la gestion des tags

La feuille XSLT originale prévoit la transformation des mentions de *tags* vers des éléments TEI. Nous avons ajouter plusieurs scénarios de transformation pour que les *tags* personnalisés créés pour le projet Time Us soient directement transformés en éléments TEI également.

Dans le fichier XML-PAGE, tous les *tags* appliqués à une ligne de texte sont contenus dans l'attribut **@custom** de l'élément **<TextLine>**, qui précise le contenu d'un ligne de texte³⁷. Les informations contenues dans l'attribut **@custom** sont exprimés comme des classes Java³⁸, où les éléments donnés entre accolades à la suite du nom de la classe correspondent à sa définition.

La première partie de la valeur de cet attribut correspond toujours à l'indexation de la ligne dans la zone de texte, soit la valeur « **readingOrder {index:#;}** », où # est un nombre. La feuille de transformation originale ignore systématiquement cette partie

37. Cf. Figure 5.20.

38. Java est un langage de programmation orienté objet créé en 1995 et désormais maintenu par Oracle Corporation depuis son rachat en 2009.

FIGURE 5.20 – Extrait d'un fichier XML-PAGE exporté de Transkribus.

```

<TextLine custom="readingOrder {index:7;} TU_personne·{offset:29;·length:6;sex:h;}·TU_personne·
{offset:37;·length:8;}" id="r118">
<Coords points="1024,1526 1135,1531 1246,1536 1358,1542 1469,1544 1581,1548 1692,1551 1803,1553
1915,1555 2026,1556 2138,1558 2249,1558 2360,1558 2472,1558 2583,1558 2695,1556 2806,1556 2917,1554
3029,1554 3140,1552 3252,1550 3252,1453 3140,1455 3029,1457 2917,1457 2806,1459 2695,1459 2583,1461
2472,1461 2360,1461 2249,1461 2138,1461 2026,1459 1915,1458 1803,1456 1692,1454 1581,1451 1469,1447
1358,1445 1246,1439 1135,1434 1024,1429"/>
<Baseline points="1024,1494 1135,1499 1246,1504 1358,1510 1469,1512 1581,1516 1692,1519 1803,1521
1915,1523 2026,1524 2138,1526 2249,1526 2360,1526 2472,1526 2583,1526 2695,1524 2806,1524 2917,1522
3029,1522 3140,1520 3252,1518"/>
<TextEquiv>
<Unicode>Plusieurs fois, les citoyens Naudot, Besacier </Unicode>
</TextEquiv>
</TextLine>

```

au moment de la création des attributs.

Qu'il soit personnalisé ou non, chaque *tag* est ensuite donné selon le modèle suivant : « nomDuTag {offset:#; length:#} », où # est un nombre. Une structure prise en charge par la feuille de transformation grâce au modèle donné en exemple dans la figure 5.21. Sur le même modèle, nous avons ajouté la prise en charge des *tags* personnalisés pour le projet Time Us³⁹.

FIGURE 5.21 – Extrait du fichier page2tei-0.xsl de Dario Kampkaspar (l.434 à 441).

```

434   <xsl:when test="@type = 'organization'">
435     <!-- TODO use of tei:rs would be more appropriate here; change after dicussion -->
436     <orgName>
437       <xsl:call-template name="elem">
438         <xsl:with-param name="elem" select="$elem" />
439       </xsl:call-template>
440     </orgName>
441   </xsl:when>

```

FIGURE 5.22 – Extrait du fichier page2tei_TU.xsl (l.512 à 523).

```

512   <xsl:when test="@type = 'TU_montant'">
513     <measure>
514       <xsl:attribute name="type">sum</xsl:attribute>
515       <xsl:call-template name="attr">
516         <xsl:with-param name="attr" select="map:keys($custom)"/>
517         <xsl:with-param name="custom" select="$custom"/>
518       </xsl:call-template>
519       <xsl:call-template name="elem">
520         <xsl:with-param name="elem" select="$elem"/>
521       </xsl:call-template>
522     </measure>
523   </xsl:when>

```

Les *tags* qui s'étendant sur plusieurs lignes sont signalés, dans le fichier XML-PAGE,

39. Voir en particulier les lignes 443 à 634 de page2tei_TU.xsl ; cf. Figure 5.22.

par la déclaration « `continued:true` »⁴⁰. Afin de créer la possibilité d'un traitement spécifique de ces cas de figure, alors que cette information est ignorée dans la transformation proposée par Dario Kampkaspar, nous avons souhaité qu'elle apparaisse dans les fichiers TEI sous la forme du couple attribut-valeur « `rend="multiline"` ». En outre, étant donné que les *tags* personnalisés peuvent avoir des propriétés qui sont également donnés dans la déclaration du *tag*⁴¹ et puisque nous voulons qu'elles soient exportées comme des attributs, nous avons créé un *template XSLT* spécifique à la création des attributs⁴². Il crée un attribut doté d'une valeur pour chaque propriété associée au *tag* déclaré dans l'attribut `@custom`. Lorsque cette propriété est « `continued:true` », elle est transformée pour prendre la forme que nous avons définie. Ce *template* est appliqué à la transformation de tous les *tags* utilisés pour le projet Time Us, personnalisés ou non.

FIGURE 5.23 – Extrait du fichier page2tei_TU.xsl (l.650 à 672).

```

650  <xd:doc>
651      <xd:desc>TIME US - Create attributes</xd:desc>
652      <xd:param name="attr"/>
653      <xd:param name="custom"/>
654  </xd:doc>
655  <xsl:template name="attr">
656      <xsl:param name="custom"/>
657      <xsl:param name="attr"/>
658      <xsl:for-each select="$attr">
659          <xsl:if test=". != 'length' and . != ''">
660              <xsl:choose>
661                  <xsl:when test=". = 'continued'">
662                      <xsl:attribute name="rend">multiline</xsl:attribute>
663                  </xsl:when>
664                  <xsl:otherwise>
665                      <xsl:attribute name="{.}">
666                          <xsl:value-of select="replace(map:get($custom, .), '\\u0020', ' ')"/>
667                      </xsl:attribute>
668                  </xsl:otherwise>
669              </xsl:choose>
670          </xsl:if>
671      </xsl:for-each>
672  </xsl:template>
```

5.2.5 Un développement à poursuivre

Grâce à la modification de la feuille de transformation de Dario Kampkaspar, nous parvenons à créer des fichiers TEI conformes. Cependant, plusieurs éléments peuvent être améliorés à partir de cette première étape de travail.

40. Par exemple : « `TU_occupation {offset:51; length:5; continued:true;}` ».

41. Par exemple : « `TU_document {offset:36; length:5; type:tarif;}` ».

42. Voir en particulier les lignes 650 à 672 de `page2tei_TU.xsl`; cf. Figure 5.23.

5.2.5.1 Corriger et améliorer les *scripts*

Nous avons évoqué le fonctionnement de `toSingleXML.py` et `fromPAGEToText.py` : à terme la création d'un fichier d'export des textes bruts devrait être intégrée à l'exécution du *script* `requestingTranskribus.py`. Par ailleurs, le *script* `toSingleXML.py` ne devrait plus être nécessaire. Il s'agit en effet de faire en sorte que la fusion des transcriptions téléchargées pour un même document soit réalisée au moment de la transformation vers la TEI.

Le schéma du fichier TEI de sortie doit également être revu afin qu'un document donne lieu à un fichier TEI dont la racine serait un élément `<teiCorpus>`, et où chaque page constituerait un élément `<TEI>` en son sein. Actuellement, le schéma des fichiers TEI, lorsqu'ils sont créés à partir de fichier XML-PAGE transformés par le *script* `toSingleXML.py`, rassemblent l'ensemble des pages transcrites au sein du même élément `<body>`, et chaque page est signalée par un élément `<pb/>`. Ce nouveau schéma permettrait notamment de personnaliser les métadonnées de chaque page.

Lorsqu'un utilisateur cible plusieurs statuts de transcription, ces statuts n'apparaissent pas dans les fichiers téléchargés. Créer des dossiers par statuts, ou intégrer le statut dans le nom du fichier permettrait de ne pas perdre cette information.

Enfin, il est actuellement de la responsabilité de l'utilisateur·rice de sauvegarder les données obtenues grâce à l'exécution des *scripts*. Une nouvelle stratégie de nommage des fichiers et des dossiers devrait permettant d'éviter que les données précédemment créées ne soient écrasées. De la même manière que les fichiers de rapport sont datés pour qu'ils puissent être distingués, il pourrait exister un dossier nommé en fonction de sa date de création au sein de chaque dossier de collection.

5.2.5.2 Ajouter des fonctionnalités

Comme nous l'avons vu, l'un des objectifs du projet Time Us est la création de listes de terminologies ou de noms pour créer des index. Nous avons manqué de temps pour développer cette fonctionnalité dans le cadre du stage. Pour qu'elle soit mis en place, deux traitements doivent être développés :

- d'une part, la gestion du symbole « ↗ », qui sert à signaler, dans Transkribus, les césures de mots en fin de ligne.
- d'autre part, l'utilisation du signalement des annotations étendues sur plusieurs lignes pour recomposer les informations.

5.2.5.3 Améliorer les pratiques sur la plate-forme Transkribus

Une fois la transmission des métadonnées vers les fichiers TEI mieux contrôlée, il est possible d'envisager de nouvelles règles pour les utilisateur·rices de Transkribus.

Celles-ci porteraient en premier lieu sur un encadrement plus fin du nommage des documents/sous-collections dans Transkribus et sur la manière de préparer les ensembles documentaires les composant. Avec des titres homogènes, tels que nous les avons envisagés, il serait possible d'extraire des informations importantes, comme le lieu de conservation et la côte des dossiers documentaires transcrits. En outre, il est nécessaire de mieux encadrer la constitution des ensembles, en exigeant le traitement des métadonnées EXIF avant le versement, et imposant que ces ensembles soient complets.

Il s'agit par ailleurs de donner des consignes pour remplir le champ « description » du panneau de métadonnées dans Transkribus, afin de récupérer des informations comme l'identité des transcripteur·rices, annotateur·rices et correcteur·rices, ou encore comme la description de l'intérêt du fond d'archives dans le cadre du projet Time Us ou de son contenu. Un modèle pour remplir ce champ doit encore être établi.

Ces consignes pourront faire l'objet d'un guide spécifique publié sur le wiki du projet Time Us.

5.3 Une interface de consultation ?

L'intention initiale du projet Time Us était d'utiliser le CMS MediaWiki pour proposer une interface visualisation des fichiers nativement numériques produits à partir de la transcription et de l'annotation des fichiers. Une telle interface permettrait de voir simultanément l'image du document d'origine et sa transcription annotée.

Cette interface de consultation sur le wiki du projet Time Us n'a, pour le moment, pas été mise en place, mais nous avons testé la possibilité d'utiliser la solution de publication en ligne **TEI Boiler Plate**, développée par John Walsh, Grant Simpson et Saeed Moaddeli, à l'Université d'Indiana⁴³, sur la recommandation de Charles Riondet. En effet, familier de cet outil, il a déjà produit un guide pour personnaliser l'affichage des fichiers TEI grâce à Boiler Plate de manière à obtenir un effet similaire à celui de Wikisource⁴⁴ : un affichage simultanée de l'image et du texte.

TEI Boiler Plate permet d'afficher un fichier XML-TEI, avec une mise en page graphique, sans passer par la création d'un fichier HTML⁴⁵ intermédiaire. TEI Boiler Plate repose cependant sur une feuille de transformation XSLT qui sert d'interpréteur pour le navigateur. Toutes les étapes de prise en main de la solution sont détaillées par les développeurs du projet⁴⁶.

43. Grant Simpson, John Walsh et Saeed Moaddeli, *Github / TEI-Boilerplate*, original-date : 2012-09-24T15:33:46Z, 26 mai 2018, URL : <https://github.com/GrantLS/TEI-Boilerplate> (visité le 20/07/2018).

44. C. Riondet, *TEI Boilerplate : Displaying a facsimile beside a transcription*, Bag of tags, URL : <https://tags.hypotheses.org/60> (visité le 20/07/2018).

45. Pour *HyperText Markup Language*, un langage de balisage permettant de représenter des pages web.

46. *TEI Boilerplate / Index*, URL : <http://dcl.ils.indiana.edu/teibp/index.html> (visité le

Notre test a été réalisé sur les vingt premières pages du document « AD69 - 9M5 » car elles sont correctement transcrites et annotées et correspondent à ce titre au résultat que l'on envisage d'obtenir à l'issu des étapes de traitement que nous avons évoquées jusqu'à présent. S'agissant d'une expérimentation, nous avons cependant modifié certains éléments du fichier XML-TEI pour qu'il soit plus proche de la structure attendue par TEI Boiler Plate⁴⁷. Nous avons supprimé les éléments <facsimile>, placés entre l'élément <teiHeader> et l'élément <body>, et nous avons assigné à chaque élément <pb/> un attribut @facs dont la valeur correspond au chemin local vers le fac-similé au format .jpg⁴⁸.

Afin de personnaliser le fonctionnement de TEI Boiler Plate, nous avons réalisé un *fork*⁴⁹ du projet sur Github et travaillé sur une branche intitulée « TimeUs »⁵⁰. Conformément aux instructions, nous avons placé un fichier XML-TEI dans le dossier dist/content et les images dans le dossier dist/images.

Deux fichiers sont mis à disposition des utilisateur·rices pour personnaliser la mise en page des fichiers TEI, dans le dossier dist : css/custom.css et content/custom.xsl. Nous n'avons pas eu, à ce stade, besoin d'interagir avec le fichier content/custom.xsl, mais uniquement avec css/custom.css⁵¹.

La mise en page par défaut proposée par TEI Boiler Plate⁵² prévoit, mise à part l'en-tête qui rassemble des informations issues du header, l'affichage du texte en continu, interrompu uniquement par les marqueurs des pages, qui sont associés aux images sous la forme de miniatures. En cliquant sur ces miniatures, on peut afficher l'ensemble des images convoquées dans le fichier XML-TEI, en taille agrandie. Une boîte à outil (*toolbox*) permet de cacher les marqueurs de page et/ou de modifier le thème de l'affichage. Il en existe trois : le thème par défaut (défini dans teibp.css), Terminal (défini dans terminal.css) et Sleepy Time (défini dans sleepy.css). Notons que nous avons uniquement travaillé à partir du thème par défaut.

Dans son guide, Charles Riondet propose d'ajouter plusieurs règles CSS dans le fichier custom.css afin de cacher la boîte à outil et d'agrandir la taille des images de manière à ce que le texte des fac-similés soit lisible sans qu'il soit nécessaire de cliquer dessus pour agrandir l'image. Leur hauteur passe ainsi à 1000 pixels. C'est ce qui permet d'obtenir un rendu similaire à celui de Wikisource. Ce paramétrage conduit à l'affichage des images de manière flottante : le texte est affiché en continu sur la gauche de l'écran,

20/07/2018).

47. Cf. « 41459 - AD69 9M5 (short-jpg).xml », dans Annexes G.2.

48. Par exemple : « facs="..../images/AD699M5_1.jpg" ».

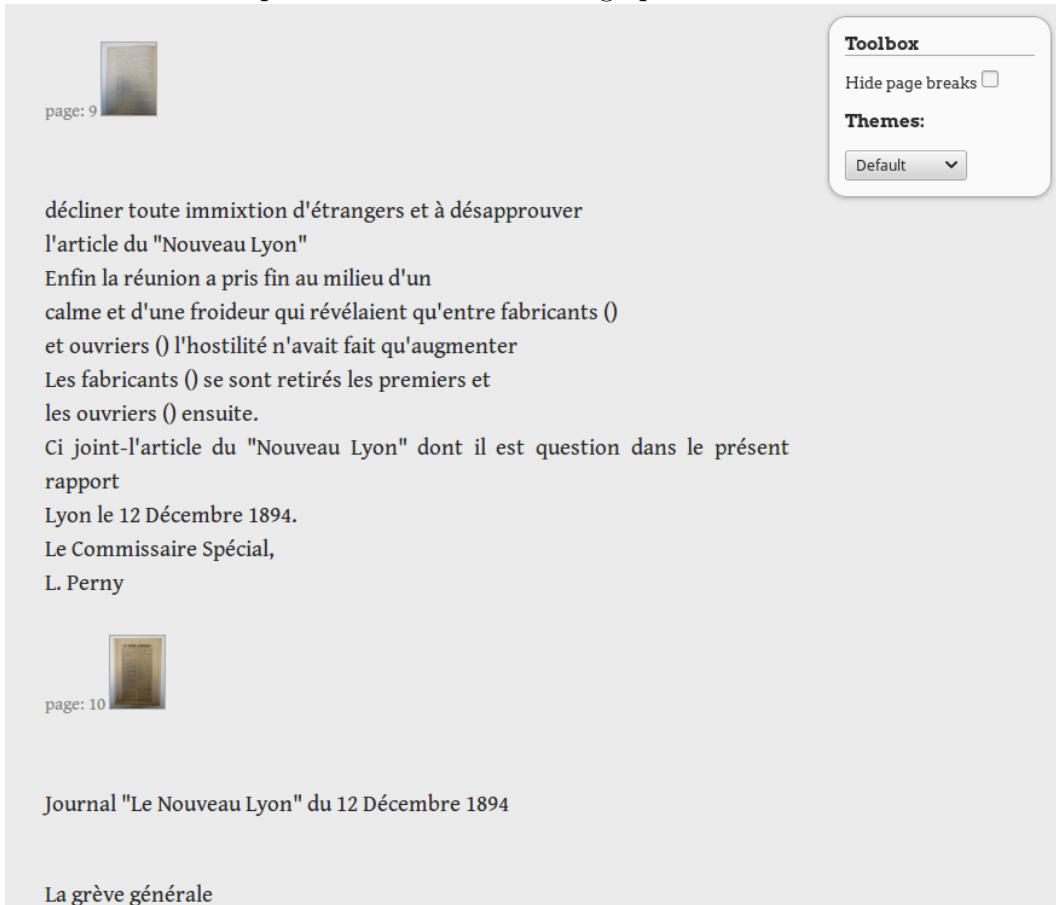
49. Copie d'un projet informatique en vue d'un développement différent, un *fork* permet de conserver la parentalité du code source et les informations de licence s'y rattachant.

50. A. Chagué, *Github / TEI-Boilerplate*, original-date : 2018-07-20T14:54:35Z, 20 juil. 2018, URL : <https://github.com/alix-tz/TEI-Boilerplate/tree/TimeUs> (visité le 03/08/2018).

51. Cf. Annexe G.1.

52. Cf. Figure 5.24.

FIGURE 5.24 – Capture d'écran de l'affichage par défaut de TEI Boiler Plate.



les images le sont sur la droite. Il en résulte que les images et leurs transcriptions ne sont pas alignées.

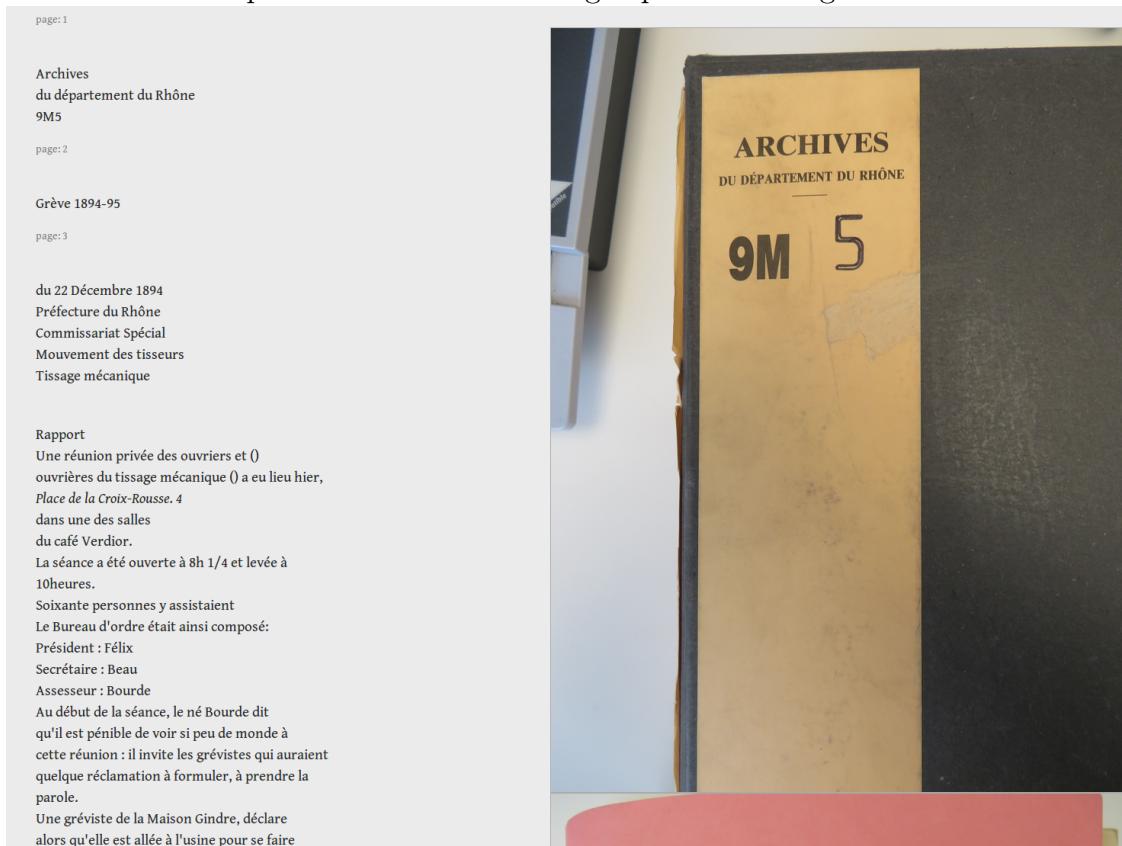
Nous avons ajouté deux éléments de prise en main de l'affichage des fichiers.

D'une part, nous avons ajouté des règles CSS pour colorer ou décorer les segments de texte annotés. Contrairement à ce que permet TEITags sur MediaWiki, ici le survol de segments de texte annotés ne conduit pas à l'affichage d'une info-bulle identifiant l'élément contenant le segment de texte. Cette fonctionnalité est pourtant utile car elle permet à l'utilisateur·rice d'accéder facilement à l'information et de comprendre l'annotation. Deux solutions sont possibles pour remédier à cela. Soit l'affichage d'une légende, de manière similaire à la boîte à outils de TEI Boiler Plate, en prenant soin qu'elle puisse être cachée par l'utilisateur·rice. Soit la modification de la feuille de transformation de TEI Boiler Plate pour créer des éléments permettant d'activer une pseudo-classe CSS comme :hover, qui crée des info-bulles.

D'autre part, des règles CSS supplémentaires ont été mises en place pour corriger le décalage des images et du texte. La structure HTML simulée par TEI Boiler Plate⁵³ est

53. Lorsque l'on affiche le code source de la page prise en charge par TEI Boiler Plate, on accède au fichier XML-TEI d'origine. En revanche, si on affiche les éléments de la page dans la console de débogage du navigateur, c'est la structure HTML que l'on peut consulter.

FIGURE 5.25 – Capture d'écran de l'affichage après les consignes de Charles Riondet.



construite comme suit : dans un élément `<body>`, chaque page donne lieu à la création de deux types d'éléments, `` et `<p>`. Ce premier élément `` contient deux éléments ``, l'un correspondant au marqueur de la page (sa classe est « `-teibp-pageNum` »), l'autre au fac-similé (sa classe est « `-teibp-pbFacs` »). Chaque `<p>` contient le texte d'une zone de texte identifiée sur une page. Lorsque plusieurs zones de texte ont été délimitées sur une page donnée, celle-ci donne lieu à la création de plusieurs éléments `<p>`, donnés à la suite⁵⁴. Tous ces éléments qui composent une page ne sont pas rassemblés dans un élément parent qui permettrait de signaler leur unité⁵⁵.

La solution que nous avons trouvée pour obtenir un alignement plus correct des transcriptions et des images est restée cantonnée à la prise en charge du fichier HTML par le CSS. Nous n'avons pas modifié la feuille de transformation XSLT de TEI Boiler Plate pour créer cet élément parent qui semblerait nécessaire. Nous avons simplement ajouté une règle aux éléments `<p>` afin que leur taille soit toujours de 1000 pixels de hauteur, comme pour les images. En outre, nous avons fait en sorte que l'élément `` qui marque le début d'une nouvelle page se comporte comme un « bloc ». Cela permet de s'assurer que l'affichage d'un nouveau fac-similé ne débute pas tant que l'affichage de la transcription de la page précédente n'est pas terminée. Cette solution n'est pas encore

54. Cf. Figure 5.26.

55. Un élément `<div>`, par exemple.

FIGURE 5.26 – Extrait de la structure HTML interprétée par le navigateur.

```

<html> event
  > <head> ...
  > </head>
  > <body>
    > <div id="teibpToolbox"> ...
    > </div>
    > <div id="tei_wrapper">
      > <TEI id="id0xfffffffffffffedc400">
        > <teiHeader id="id0xfffffffffffffedc680"> ...
        > </teiHeader>
        > <text id="id0xffffffffffff4bd80">
          > <body lang="">
            > <span id="id0xffffffffffff4c080" class="-teibp-pb" style="display: block;" lang="en">
              > <span class="-teibp-pageNum"> ...
              > <span class="-teibp-pbFacs"> ...
              > </span>
            > <p id="id0xffffffffffff4c180"> ...
            > <span id="id0xffffffffffff4c700" class="-teibp-pb" style="display: block;" lang="en"> ...
            > </span>
            > <p id="id0xffffffffffff2900"> ...
            > <span id="id0xffffffffffff2b80" class="-teibp-pb" style="display: block;" lang="en"> ...
            > </span>
            > <p id="id0xffffffffffff2c80"> ...
            > </p>
            > <p id="id0xffffffffffff3600"> ...
            > </p>
            > <span id="id0xffffffffffff2800" class="-teibp-pb" style="display: block;" lang="en"> ...
            > </span>
            > <p id="id0x1b7100"> ...
            > </p>

```

pleinement satisfaisante car lorsque plusieurs zones de texte sont associées à une image, chaque `<p>` fait une taille de 1000 pixels : la transcription n'est donc pas toujours en face d'une image, celle-ci étant alors affichée plus haut. Enfin, pour les éléments `<p>` dont le contenu dépassait 1000 pixels de hauteur, nous avons ajouté une règle pour gérer les dépassements (*overflow*). Ceux-ci donnent lieu, le cas échéant, à la création d'un ascenseur de défilement⁵⁶.

Nous nous sommes contenté à ce stade d'intervenir uniquement au niveau de la CSS, car nous manquions de temps pour manipuler la feuille de transformation XSLT. Une adaptation future pourrait permettre de mieux gérer l'alignement des images et des texte en créant pour chaque ensemble texte-image, un élément `<div>` parent.

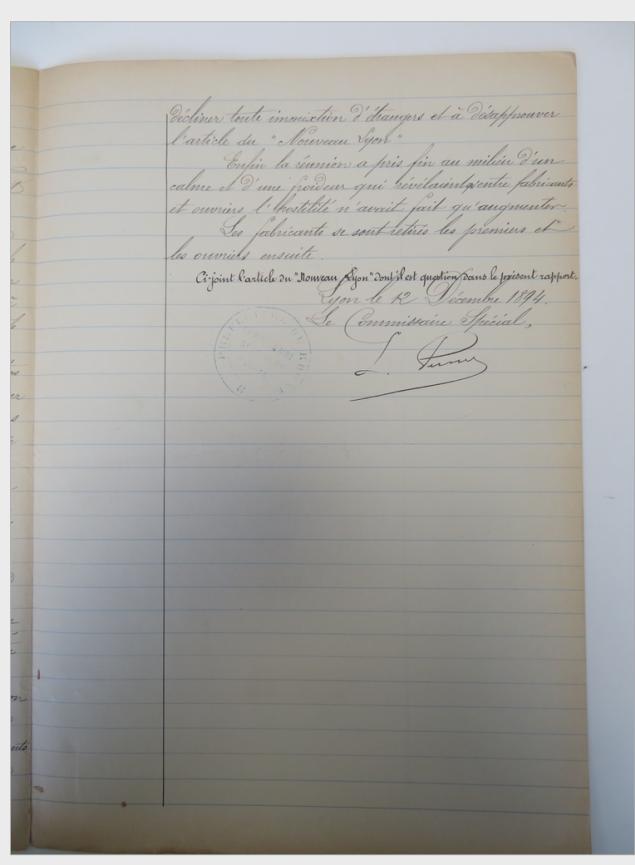
Les résultats de cette expérimentation sont cependant encourageants et pourraient donner lieu à un développement supplémentaire. TEI Boiler Plate paraît donc une alternative au site MediaWiki pour créer une interface de consultation des fichiers TEI créés dans le cadre du projet Time Us. Sa manipulation est relativement simple, pour un résultat de bonne qualité.

56. Cf. Figure 5.27.

FIGURE 5.27 – Capture d'écran de l'affichage après nos modifications.

page: 9

décliner toute immixtion d'étrangers et à désapprouver
l'article du "Nouveau Lyon"
Enfin la réunion a pris fin au milieu d'un
calme et d'une froideur qui révélaient qu'entre **fabricants** ()
et **ouvriers** () l'hostilité n'avait fait qu'augmenter
Les **fabricants** () se sont retirés les premiers et
les **ouvriers** () ensuite.
Ci joint l'article du "Nouveau Lyon" dont il est question dans le présent rapport
Lyon le 12 Décembre 1894.
Le Commissaire Spécial,
L. Perny



Conclusion

Le présent mémoire s'est attaché à retracer les différentes opérations nécessaires pour aboutir à des fichiers XML conformes et standardisés en partant de documents d'archives papiers, en vue de rendre possibles des traitements quantitatifs sur des sources qualitatives. Les missions qui m'ont été confiées m'ont permis de prendre part à la réflexion menée par l'équipe ALMAnaCH, en collaboration avec les autres équipes participant au projet de recherche, sur la question des outils informatiques et des bonnes pratiques à mettre en oeuvre pour permettre au projet de réaliser ses objectifs. S'agissant d'un projet de recherche expérimentant un certain nombre d'outils informatiques, tels que les outils de traitement automatique des langues, il convenait en effet de trouver les meilleures solutions possibles aux obstacles rencontrés tout en prenant en compte les contraintes logistiques, financières et de compétences d'un projet débuté depuis un an et demi.

Nous avons reconstitué une partie de la chaîne de traitement au fil de ce mémoire en analysant les difficultés posées par chacune d'entre elles. En premier lieu, nous avons vu que la collecte et la numérisation des documents d'archives avaient fait abstraction de questions qui sont réapparues plus tard dans le processus. Les objectifs associés à une tâche ont un impact sur l'importance accordée à certains aspects de sa réalisation. Ainsi, dans la mesure où la campagne de numérisation visait à créer des copies numériques des sources en vue de faciliter leur transcription manuelle, aucune consigne sur l'équipement et les pratiques de prise de vue n'a été donnée. Ces numérisations n'ont par ailleurs pas fait l'objet d'une campagne de post-traitement. Toutefois, ces images servant finalement de support pour une annotation automatique, et dans la mesure où leur existence permet d'envisager qu'elles accompagnent la visualisation des transcriptions, nous nous sommes aperçus qu'il aurait été utile de mettre en place, à minima, une étape de recadrage et de correction des images avant leur versement dans l'outil de transcription automatique. En outre, l'accumulation d'images numériques venant des différentes équipes et correspondant à plusieurs fonds d'archives a rapidement rendu nécessaire de trouver une solution de stockage adéquate pour garantir la cohérence du corpus d'images et conserver les informations sur leurs origines.

En deuxième lieu, le recours à une plate-forme de transcription automatique telle que Transkribus, et à des outils de reconnaissance automatique de caractères, a modifié l'organisation de la tâche de transcription telle qu'elle avait été envisagée en amont. Mon arrivée au sein du projet s'est déroulée après l'implémentation de cette plate-forme dans la chaîne de traitement. Il a donc fallu explorer les fonctionnalités de l'outil pour en comprendre les avantages et les limites, et il s'est aussi agi d'accompagner les membres du projet amenés à utiliser Transkribus. La réflexion menée sur le schéma TEI utilisé pour représenter les documents transcrits a fortement été influencée par le schéma TEI prévu par Transkribus, bien que nous avons apporté des modifications ultérieures au fur et à mesure que nous développions des outils complémentaires.

Le cœur de mes missions était associé à la phase d'annotation du corpus, pour

laquelle peu de choses avaient été mises en place. L’identification des objectifs de l’annotation tant d’un point de vue scientifique, car il s’agissait de définir les termes et notions à annoter dans le texte, que d’un point de vue technique, ici l’annotation manuelle vise en effet à produire des données d’entraînement pour une annotation automatique, a constitué une étape de travail cruciale. La mise en place d’un modèle d’annotation solide est allée de paire avec l’établissement d’un guide d’annotation dont la confrontation aux sources textuelles doit permettre, encore, l’amélioration. En outre, il était nécessaire de définir les bons outils pour réaliser l’annotation : Transkribus n’était satisfaisant que dans la mesure où nous pouvions davantage contrôler l’environnement local d’annotation et la qualité des fichiers de sortie.

Alors que Transkribus permet d’exporter les transcriptions au format TEI via son interface graphique, il s’est avéré nécessaire de développer un ensemble de moyens pour extraire les données de Transkribus par un autre biais. Cela nous a conduit à concevoir un *script* Python d’interaction avec l’API de Transkribus, et à élaborer notre propre feuille de transformation XSLT pour améliorer la qualité des fichiers XML-TEI de sortie. Cette meilleure prise en main nous a également permis d’initier une réflexion plus poussée sur la manière d’améliorer la qualité des métadonnées dans ces fichiers. La stratégie de création, d’enregistrement et de mise en forme des métadonnées dépendait de l’efficacité de notre outil d’extraction depuis Transkribus. C’est la raison pour laquelle la réflexion sur cette question a débuté tardivement. Elle doit être poursuivie et discutée avec les autres membres du projet.

Deux aspects en particulier doivent encore être développés dans le cadre du projet Time Us. Il s’agit tout d’abord de la mise en place des outils de transcription et d’annotation automatique, sur lesquels les spécialistes du traitement automatique des langues peuvent désormais se pencher, les opérations précédant cette étape étant désormais cadrees et en partie enclenchés. Le deuxième aspect concerne la visualisation du corpus annoté dans le cadre d’une plate-forme de consultation. Dans cette optique, deux outils ont pour le moment été testés : d’une part le CMS MediaWiki, équipé des modules adéquats, et d’autre part l’outil TEI Boiler Plate, avec lequel nous avons obtenu des résultats encourageants.

L’ensemble des *work packages* identifiés en 2016 lors de la soumission du projet à l’ANR est nécessairement remis en question par la réalité du terrain. Participer à la rationalisation de la chaîne de traitement des documents et à la mise en place d’outils solides pour obtenir des données standardisées de qualité m’a permis de mobiliser des connaissances acquises durant ma formation et de développer ainsi de réelles compétences pour la gestion de projets en humanités numériques. Outre les enjeux techniques liés à l’utilisation d’outils informatiques, tel que Transkribus, et de standards, comme la TEI, la réalisation de ce stage impliquait également de prendre en compte la réalité humaine d’un tel projet. Celui-ci existe en effet en dehors du cadre d’une institution unique ; il

rassemble des personnalités, des compétences et des intérêts différents. Il s’agissait bien de mettre les outils informatiques au service de la recherche historique telle qu’elle est pensée pour Time Us, en visant la meilleure qualité possible, sans pour autant contraindre les ambitions scientifiques du projet.

L’état de mes connaissances en PHP m’a conduit, en accord avec mes encadrant·es, à ne pas approfondir la question du développement de l’interface de consultation sur le CMS MediaWiki. Il reste par ailleurs de nombreuses pistes d’amélioration des dispositifs `usingTranskribusAPI` et `page2tei_TimeUs` : ils fonctionnent correctement et peuvent être adaptés à un projet autre que Time Us, mais ils gagneraient à fonctionner comme un ensemble unique. Mon stage ayant pris place à mi-parcours du projet de recherche, les solutions que j’ai été amenée à proposer devaient prendre en compte l’existant (qui ferait difficilement l’objet d’une nouvelle réalisation) et le fait que ces solutions devaient être suffisamment souples pour pouvoir encore être modifiables et adaptables d’ici la fin du projet, en fonction des nouvelles problématiques techniques rencontrées ou de la modification de certains objectifs. Documenter mes choix et les outils développés durant ce stage était d’autant plus crucial pour la poursuite du projet.

En terme d’expérience professionnelle, à titre personnel, j’ai apprécié le cadre institutionnel dans lequel s’est déroulé mon stage. Travailler dans les locaux d’Inria m’a en effet permis de développer ma connaissance du monde de l’entrepreneuriat à travers plusieurs initiatives lancées par l’établissement. S’y ajoutent les nombreux séminaires portant sur des sujets étrangers au domaine des humanités numériques auxquels j’ai pu assister (cryptographie, *machine learning*, applications numériques à la biologie, ...). Au sein de l’équipe ALMANaCH, j’ai également beaucoup appris sur le fonctionnement du traitement automatique des langues en me renseignant sur le travail de mes collègues. J’ai pu profiter de leurs connaissances en développement informatique pour soumettre mes codes à leurs critiques et améliorer la qualité de ma rédaction en Python. En outre, échanger avec les ingénieurs de recherche et d’étude de l’institution m’a permis de mieux comprendre le rôle que j’ai eu et aurai à jouer dans des projets semblables ainsi que la dimension collective du travail mené et l’importance de son rayonnement international.

La complexité du contexte institutionnel du projet Time Us a présenté dans un premier temps une difficulté puisque cela ne simplifiait pas l’identification de mes interlocuteur·rices. De manière générale, cela implique également que le temps de réflexion, de communication et de prise de décision est plus long, et je reconnais à ce titre avoir bénéficié du contexte favorable de l’établissement du bilan de mi-parcours pour l’ANR. Ce contexte institutionnel m’a poussée à assurer mon autonomie et m’a également aidé à développer un regard critique sur mon environnement de travail et sur l’organisation générale du projet. Les missions telles qu’elles m’ont été confiées m’ont permis de découvrir des logiciels et des outils, comme Transkribus ou Postman, et de les expérimenter en toute liberté, voire de leur chercher par moi-même des alternatives. J’ai également

été en mesure de mieux appréhender le réseau d'acteurs intervenant dans le monde des humanités numériques, en France et en Europe.

Annexes

Les annexes accompagnant le présent mémoire sont présentées sur un support multimédia « clef USB » et sur un *repository* Github, situé à l'adresse : https://github.com/alix-tz/M2TNAH_memoire-de-stage. Elles sont aussi en partie reproduites ici.

Cette section contient des indications sur la localisation des fichiers dans le dossier d'annexes.

Annexe A

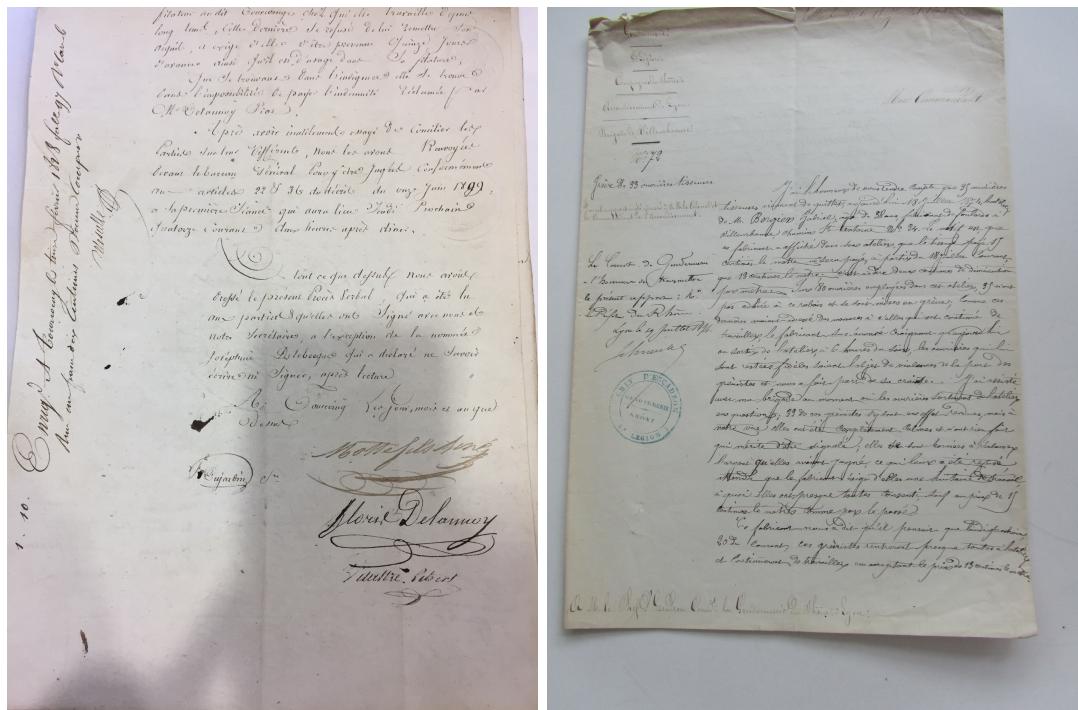
Archives

A.1 Exemples

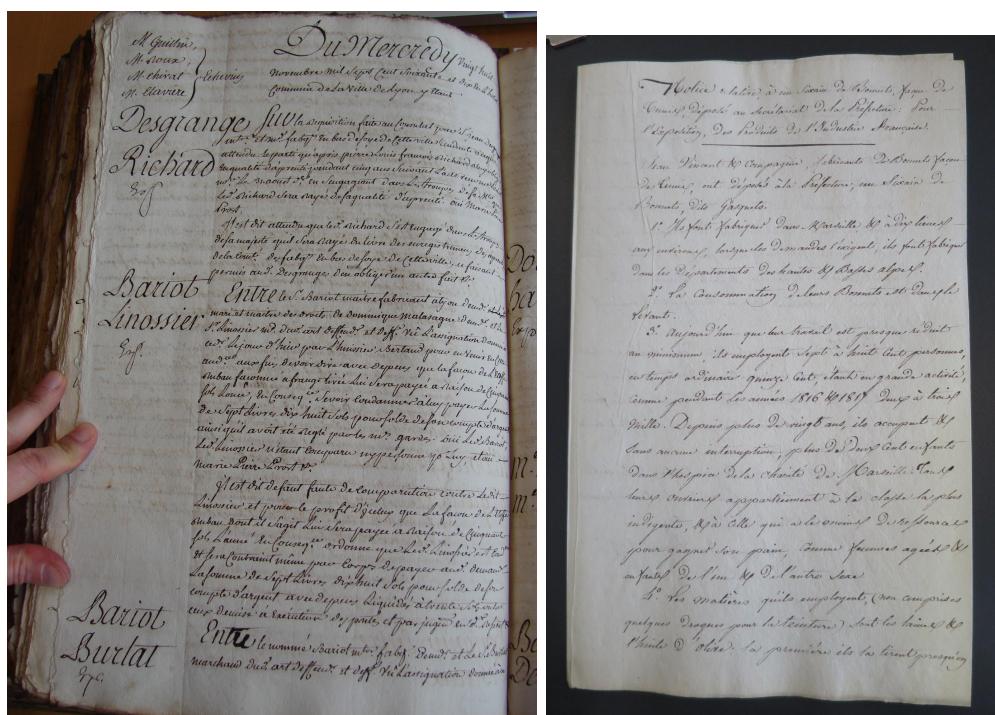
A.1.1 Extraits du corpus d'images

Dans le dossier /A - Archives/A1 - Exemples/images/, voir en particulier :

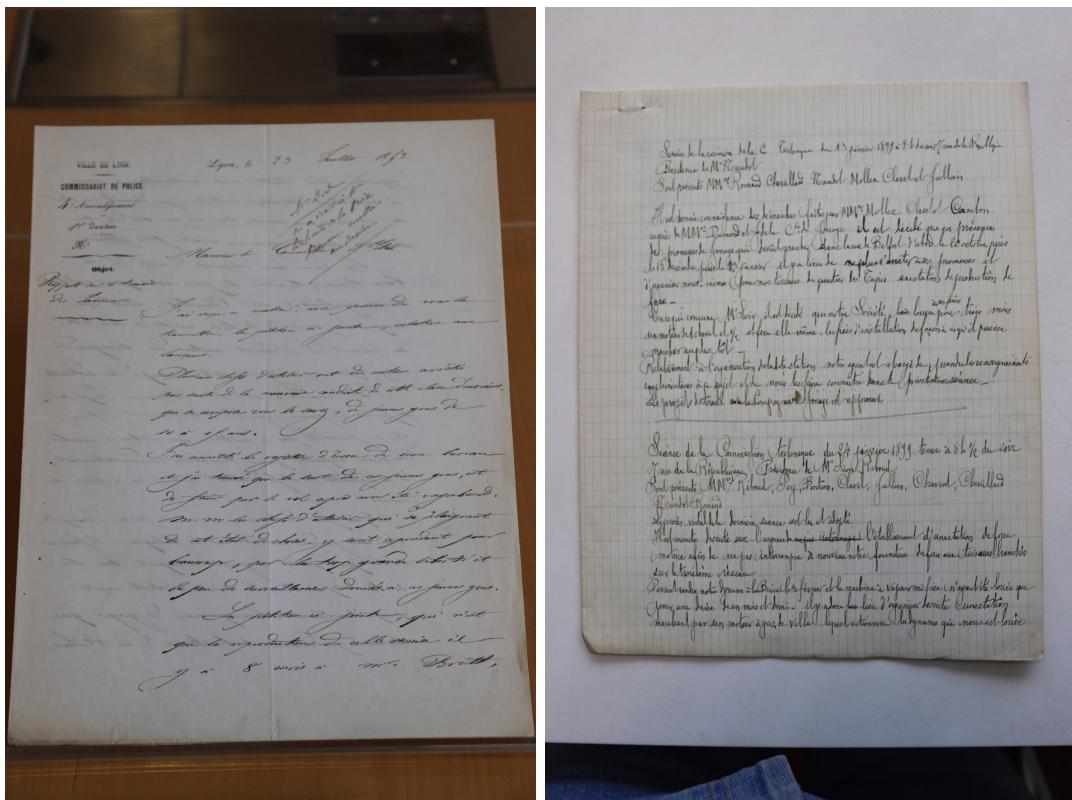
- 0014_image2 18.42.45.JPG
- 1874-10M395.JPG
- DSC08285.JPG
- IMG_0003.JPG
- IMG_2849.JPG
- IMG_6725.JPG
- IMG_9012.JPG
- P1050137.JPG
- PH 1848-8.jpg
- PH 1858-2 d.jpg



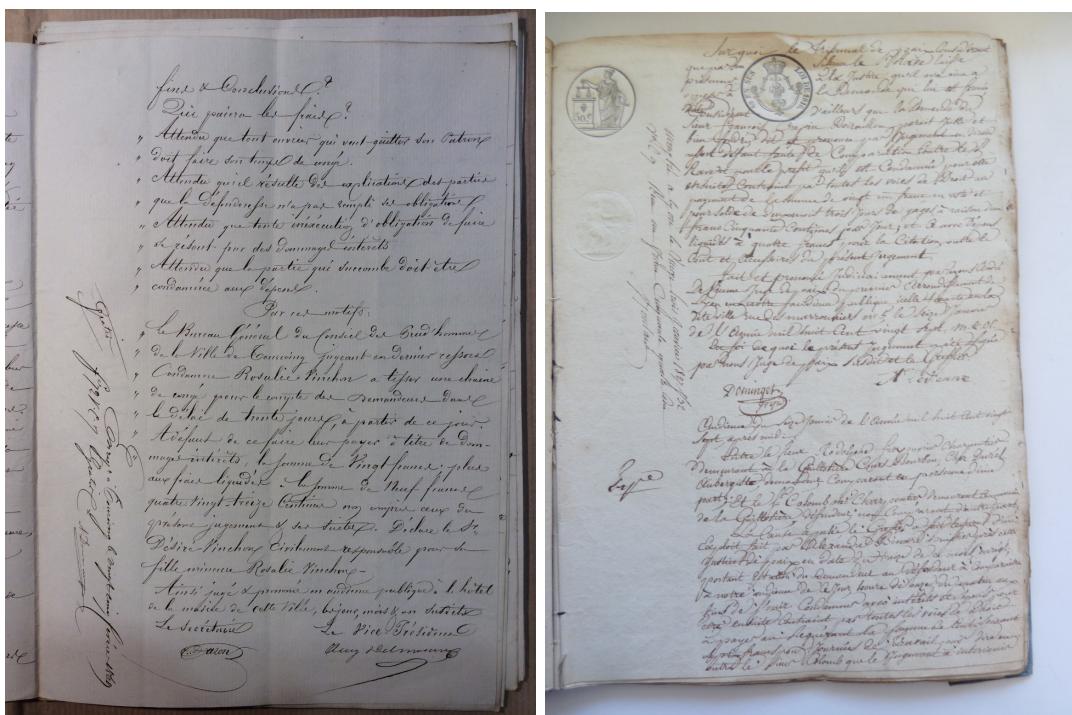
0014_image2 18.42.45.JPG et 1874-10M395.JPG



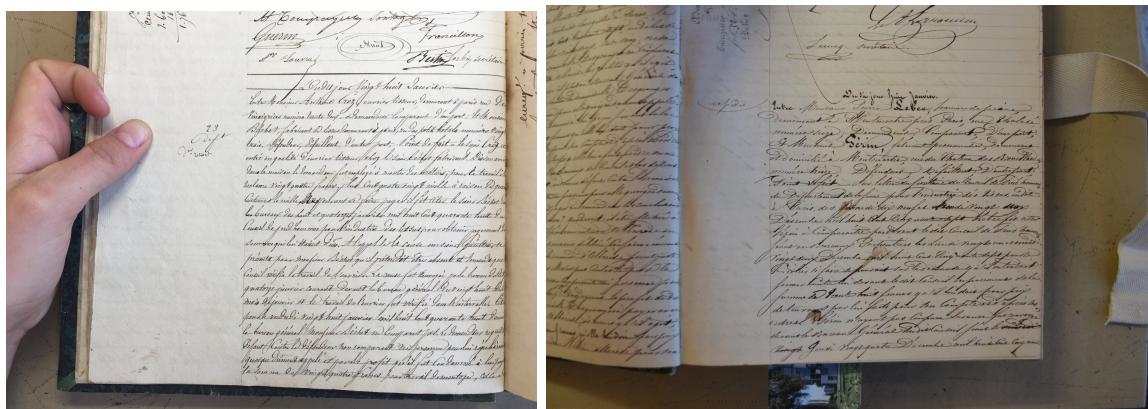
DSC08285.JPG et IMG_0003.JPG



IMG_2849.JPG et IMG_6725.JPG



IMG_9012.JPG et P1050137.JPG



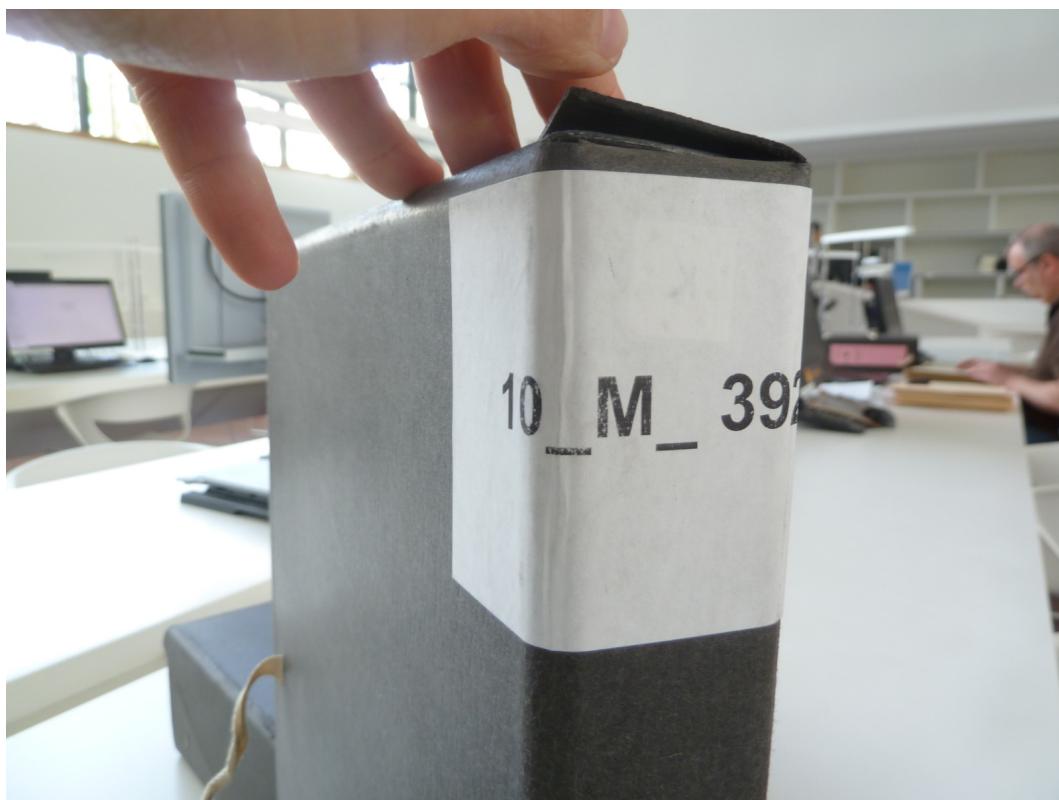
PH 1848-8.jpg et PH 1858-2 d.jpg

A.1.2 Exemples de prises de vue problématiques

Dans le dossier /A - Archives/A1 - Exemples/problematique/, voir en particulier :

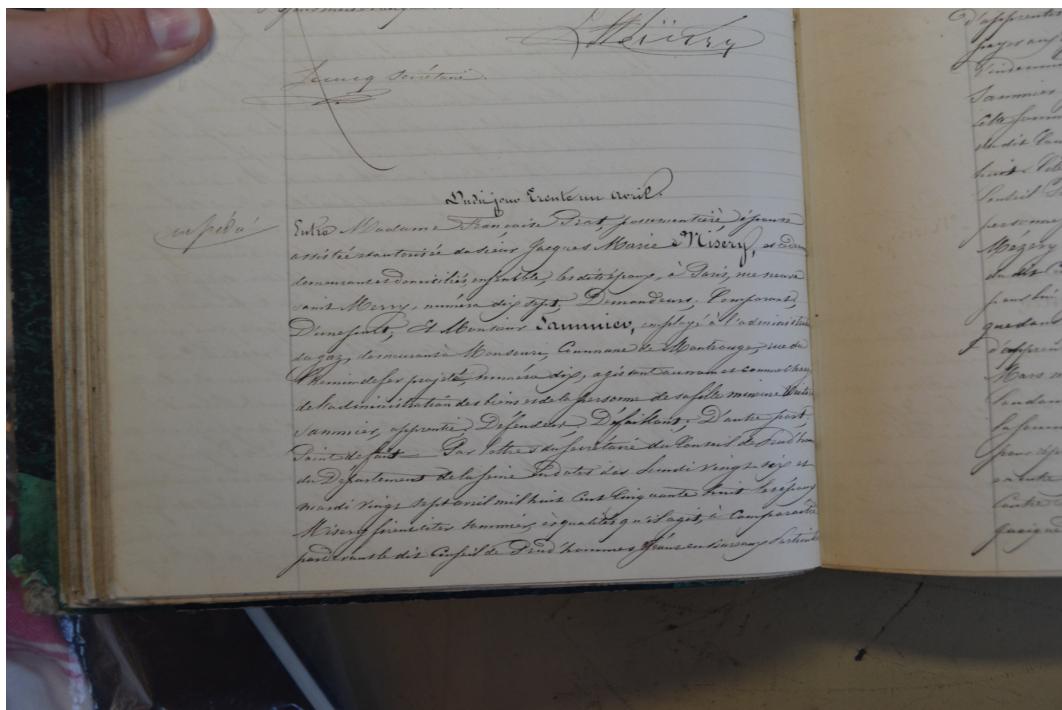
- P1050278.JPG
- PH 1858-428.jpg
- PH 1858-1072.jpg

A.1.2.1 Prise de vue servant au repérage dans les corpus



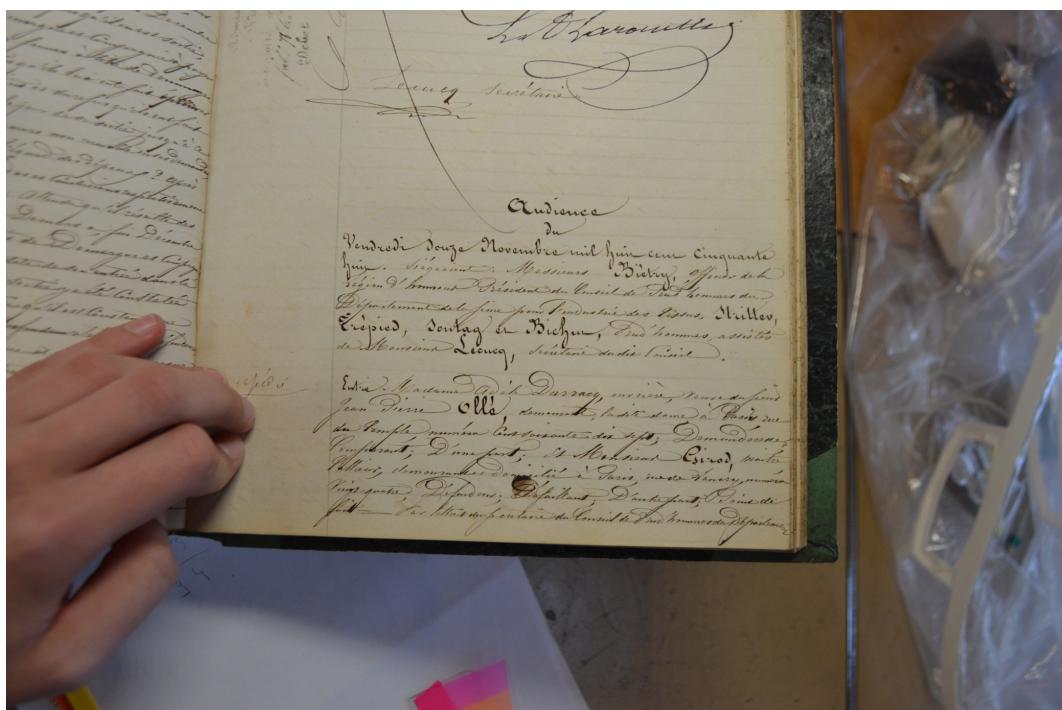
P1050278.JPG

A.1.2.2 Prise de vue nécessitant un recadrage et un redressement



PH 1858-428.jpg

A.1.2.3 Prise de vue nécessitant un recadrage



PH 1858-1072.jpg

A.1.3 Enquêtes sociologiques de Le Play sur les métiers du textile

Quatorze enquêtes sociologique publiées parmi les monographies de Le Play s'attachent aux métiers du textile :

- Série 1, Tome I, n°7 : *Tisseur en châle de Paris*
- Série 1, Tome II, n°13 : *Tailleur d'habit de Paris*
- Série 1, Tome III, n°20 : *Brodeuse des Vosges*
- Série 1, Tome III, n°24 : *Lingère de Lille*
- Série 1, Tome IV, n°36 : *Tisserand des Vosges*
- Série 2, Tome I, n°55 : *Gantier de Grenoble*
- Série 2, Tome III, n°67 : *Tisserand du Gand (Belgique)*
- Série 2, Tome IV, n°79 : *Tisseur de San Leucio (Italie)*
- Série 2, Tome V, n°83 : *Fileur du Val-des-Bois (Marne)*
- Série 3, Tome I, n°97 : *Tisserand d'usine de Gladbach (Prusse-rhénane)*
- Série 3, Tome II, n°104 : *Teinturier de ganterie de Saint-Junien (H^{te}-Vienne*
- Série 3, Tome II, n°106 : *Corsetière du Raincy, banlieue de Paris*
- Série 3, Tome III, n°109 : *Tisseur de Saint-Quentin*
- Série 3, Tome III, n°111 : *Tisserand de Roulers (Belgique)*

Dans le dossier /A - Archives/A1 - Exemples/, voir en particulier :

- Le-Play_7.zip : extrait du tome 1, de la première série, correspondant à l'enquête sur le *tisseur en châle de Paris*.

A.1.4 Exemple de titre de presse ouvrière

Dans le dossier /A - Archives/A1 - Exemples/, voir en particulier :

- Echo_de_la_Fabrique_001.pdf

A.1.5 Liste des titres de presse ouvrière lyonnaise recherchés

- 1831-1835 : *L'Écho de la fabrique*
- 1833-1834 : *L'Écho des travailleurs*
- 1834-1835 : *L'indicateur, journal industriel de Lyon*
- 1834-1835 : *La tribune prolétaire*
- 1835 : *Le nouvel Écho de la fabrique, journal industriel de Lyon*
- 1835 : *L'union des travailleurs*
- 1835-1836 : *Le Vigilant lyonnais*

- 1837-1843 : *Le Lyonnais*
- 1840-1841 : *L'Écho des ouvriers, journal des intérêts de la fabrique et des chefs d'atelier*
- 1841-1845 : *L'Écho de la fabrique de 1841*
- 1845 : *L'Écho de l'industrie*
- 1845-1851 : *La Tribune lyonnaise*
- 1846 : *L'Avenir*
- 1852-1853 : *Le Moniteur de la fabrique, journal spécial des chefs d'atelier, écho des prud'hommes.*

A.2 Éléments d'analyse

A.2.1 Analyse de la structure des monographies de Le Play

Dans le dossier /A - Archives/A2 - Elements d'analyse/, voir en particulier :

- *LePlay_Analyse.pdf* : note à usage interne détaillant la structure des quatre premiers tomes de la première série des monographies de Le Play, ainsi que les marqueurs signalant cette structure.

A.2.2 Tableau comparatif des métadonnées des échantillons du corpus d'images

0014_image2 18.42.45.JPG	Lille, A. D. du Nord, 5U 1 1
1536 x 2048 px	556 KB
iPhone 6 (Apple)	72 dpi
1874-10395.JPG	Lyon, A.D. du Rhône, 10 M 395
3240 x 4320 px	2,9 MB
PowerShot SX210 IS (Canon)	180 dpi
DSC08285.JPG	Lyon, A.M. de Lyon, HH 260
2112 x 2816 px	2,4 MB
DSC-H2 (Sony)	72 dpi
IMG_0003.JPG	Marseille, A.D. des Bouches du Rhône, 6 M 1620
1600 x 1064 px	482 KB
Powershot S120 (Canon)	180 dpi
IMG_2849.JPG	Lyon, A.M. de Lyon, 784 WP 8
6000 x 4000 px	5,5 MB
EOS 200D (Canon)	72 dpi
IMG_6725.JPG	Lyon, A.D. du Rhône, 14 J 99
6000 x 4000 px	6,2 MB
EOS 200D (Canon)	72 dpi
IMG_9012.JPG	Lille, A.D. du Nord, 5U 1 2
1936 x 2592 px	1,2 MB
iPad Air (Apple)	72 dpi
P1050137.JPG	Lyon, A.D. du Rhône, 10 M 392
4000 x 3000 px	3,1 MB
DMC-TZ8 (Panasonic)	180 dpi
PH 1848-8.jpg	Paris, Archives de Paris, D1U10 379
4032 x 3024 px	2,5 MB
iPhone 6s (Apple)	72 dpi
PH 1858-2 d.jpg	Paris, Archives de Paris, D1U10 386
4608 x 3072 px	5,6 MB
D3100 (Nikon)	300 dpi

Annexe B

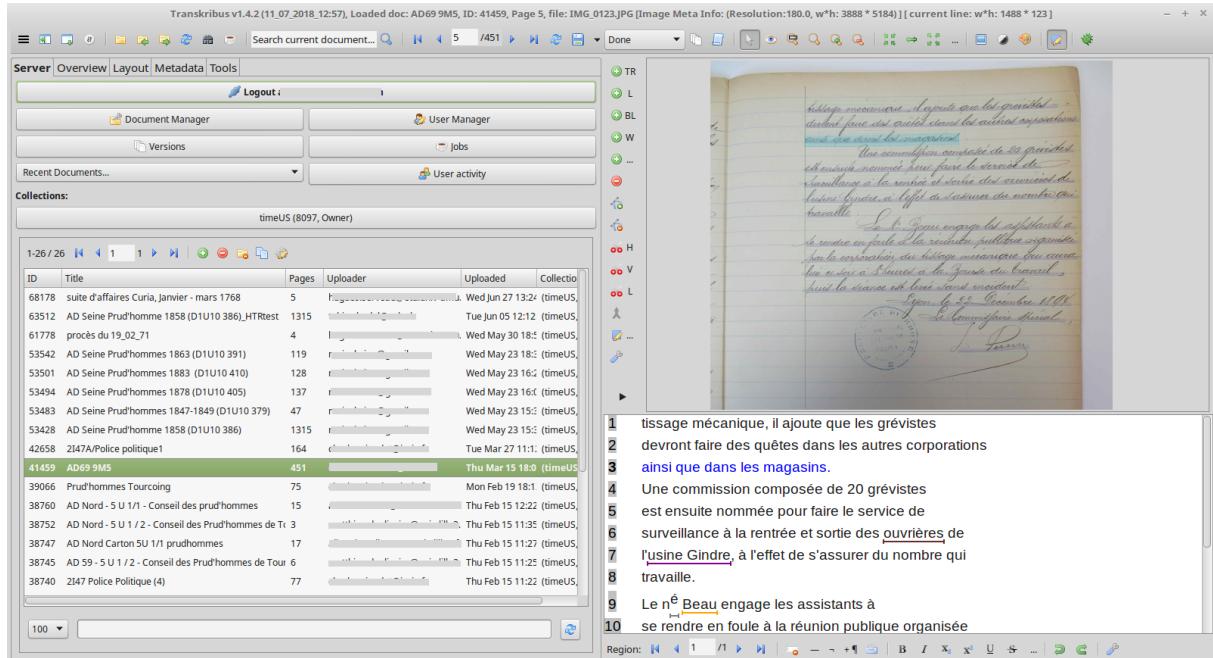
Transkribus

B.1 Interface graphique

Dans le dossier /B - Transkribus/B1 - Interface graphique/, voir en particulier :

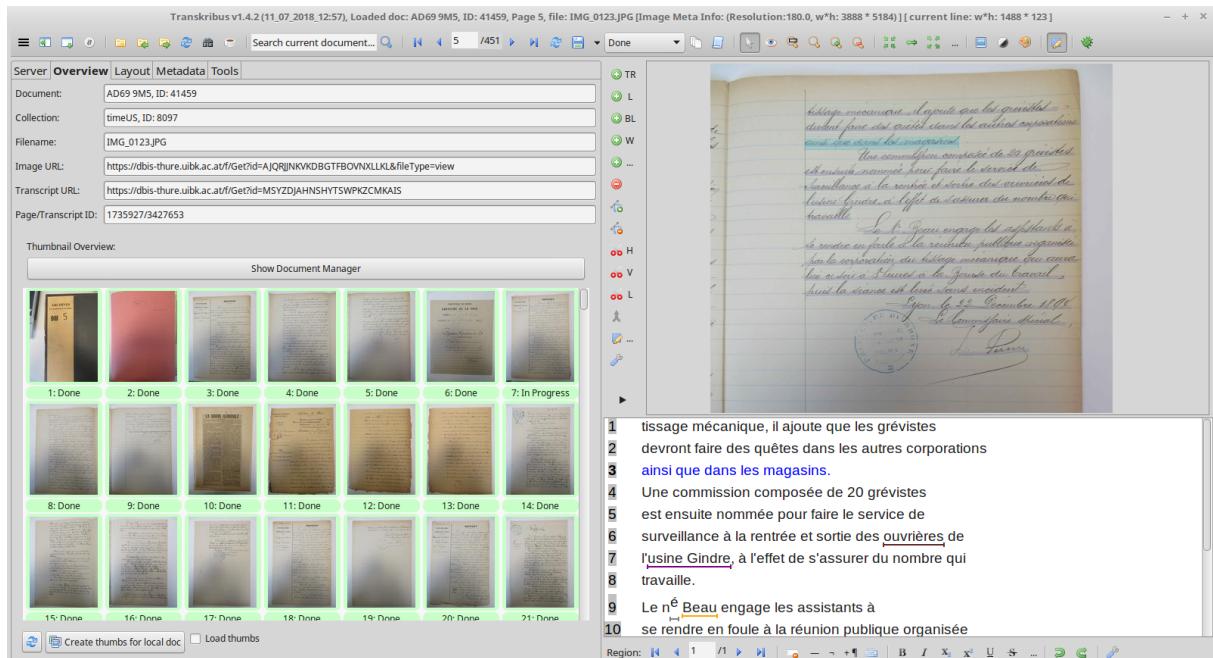
- 01-Tkb_Server.png
- 02-Tkb_Overview.png
- 03-Tkb_Layout.png
- 04-Tkb_Metadata_document.png
- 05-Tkb_Metadata_structural.png
- 06-Tkb_Metadata_textual.png
- 07-Tkb_Metadata_textual_tag-configuration.png
- 08-Tkb_Metadata_comments.png
- 09-Tkb_Tools.png
- 10-Tkb_Tools_HTR-model.png
- 11-Tkb_export_PAGE.png
- 12-Tkb_export_TEI.png

B.1.1 Capture d'écran de la page d'accueil de Transkribus, onglet « Server »



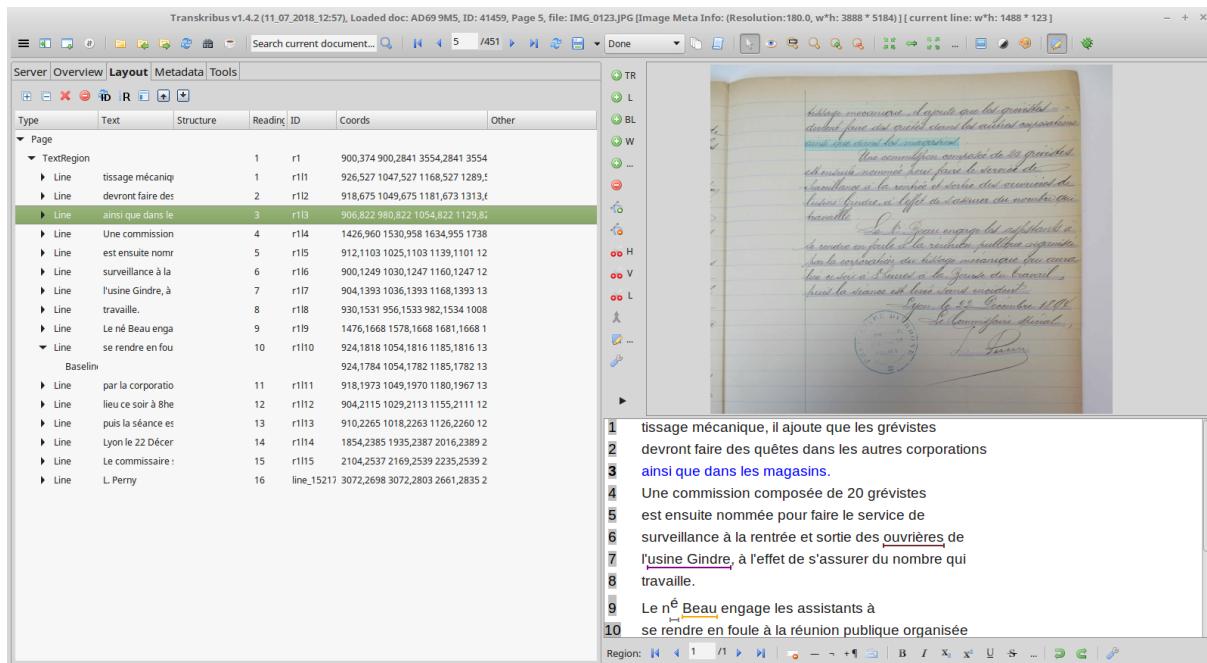
01-Tkb_Server.png

B.1.2 Capture d'écran avec l'onglet « Overview » développé



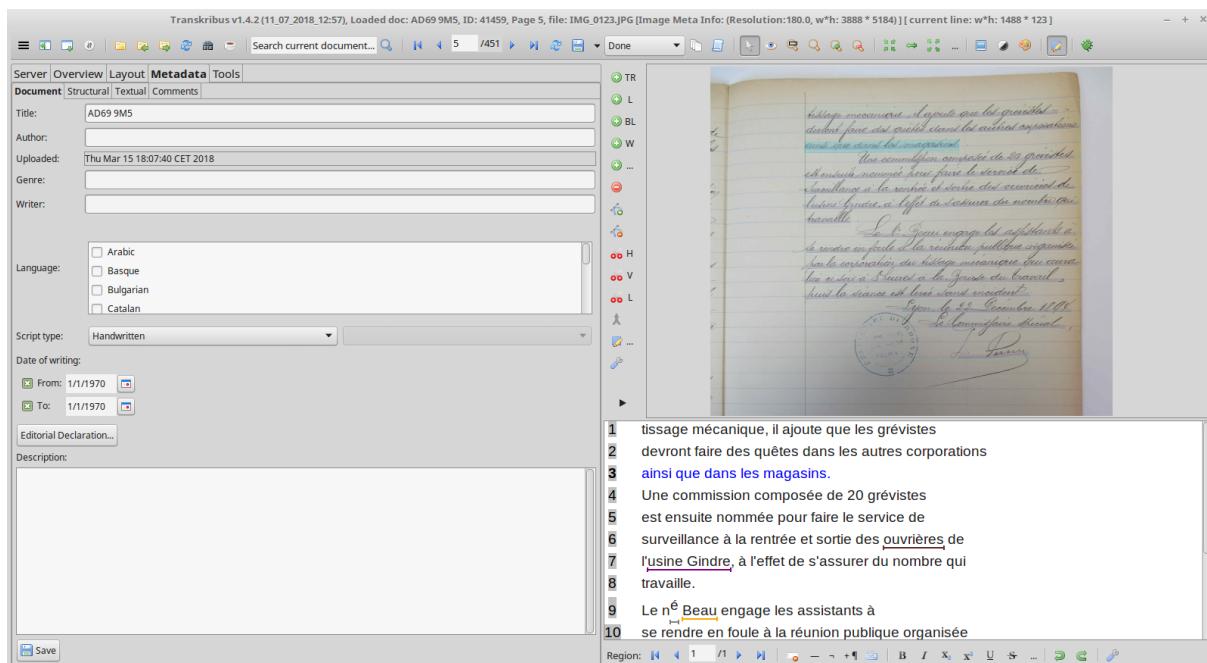
02-Tkb_Overview.png

B.1.3 Capture d'écran avec l'onglet « Layout » développé



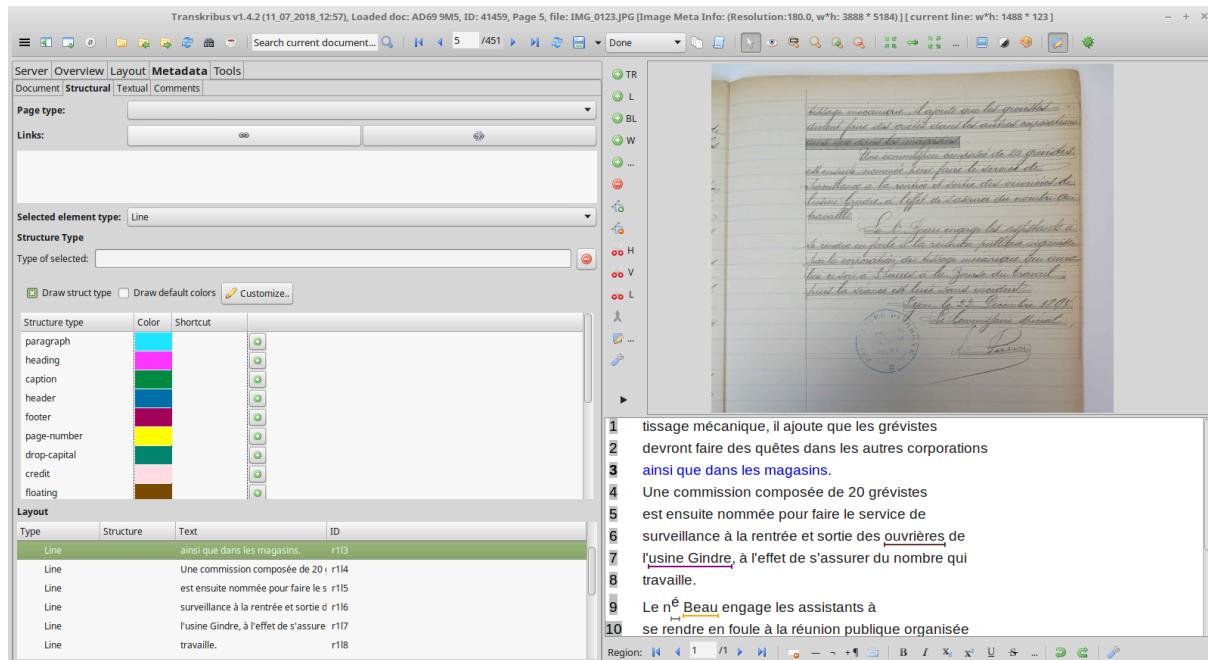
03-Tkb_Layout.png

B.1.4 Capture d'écran avec les onglets « Metadata » et « document » développés



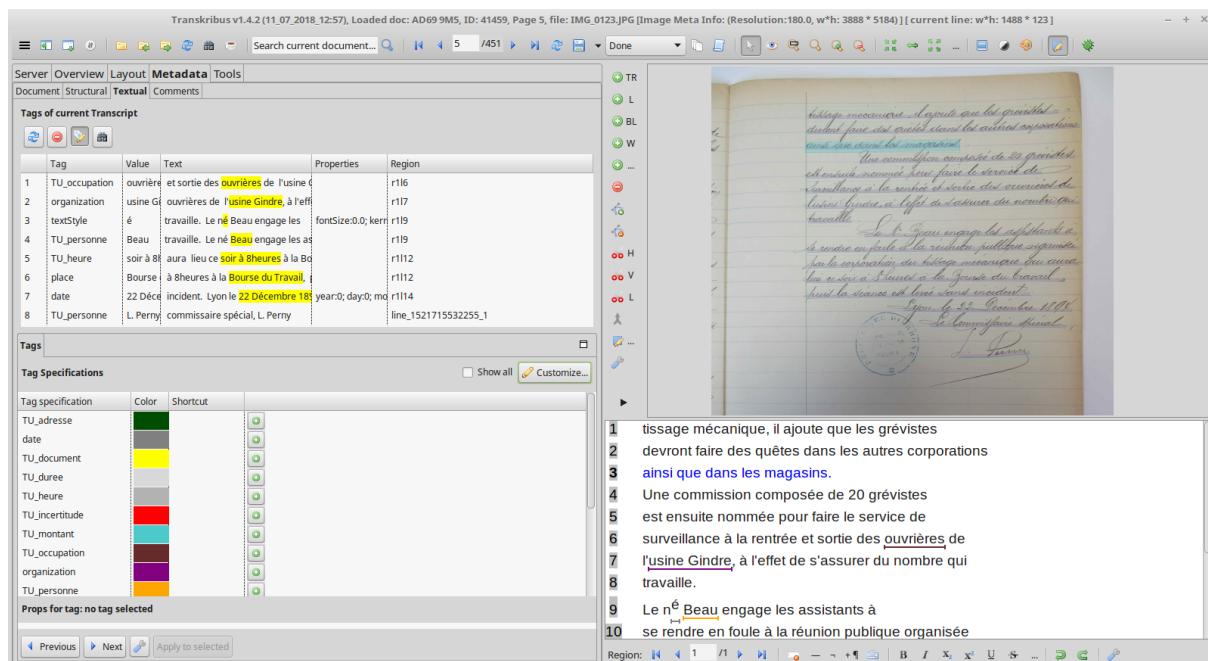
04-Tkb_Metadata_document.png

B.1.5 Capture d'écran avec les onglets « Metadata » et « structural » développés



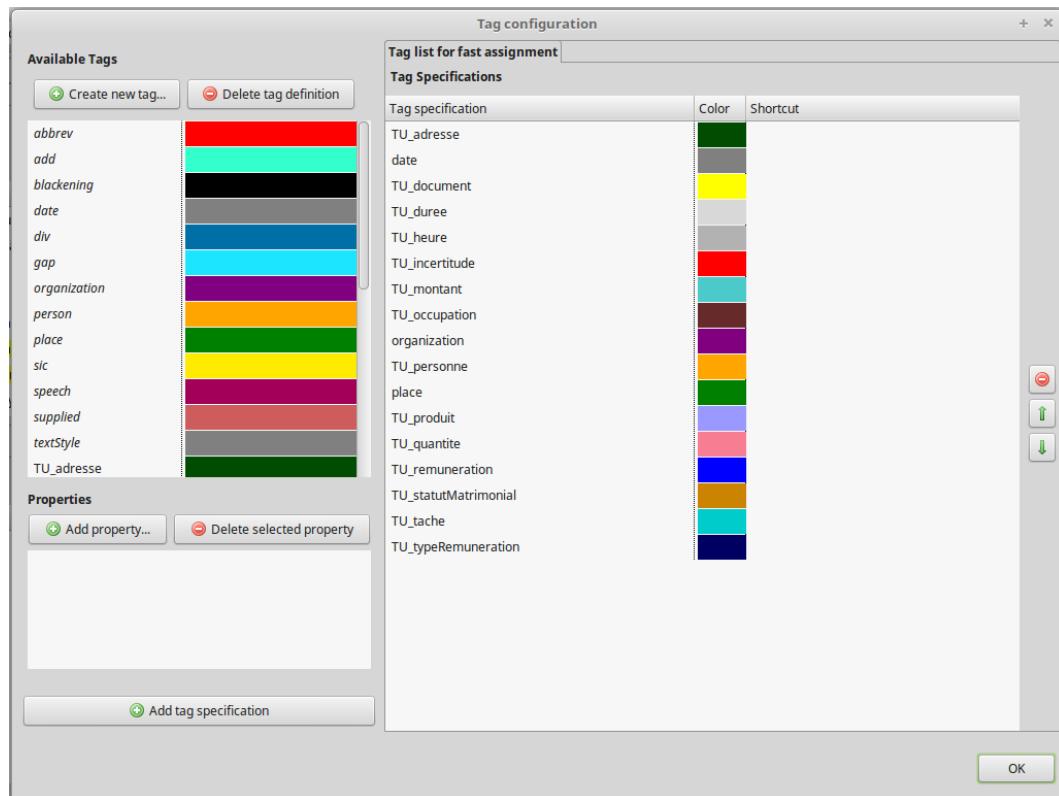
05-Tkb_Metadata_structural.png

B.1.6 Capture d'écran avec les onglets « Metadata » et « textual » développés



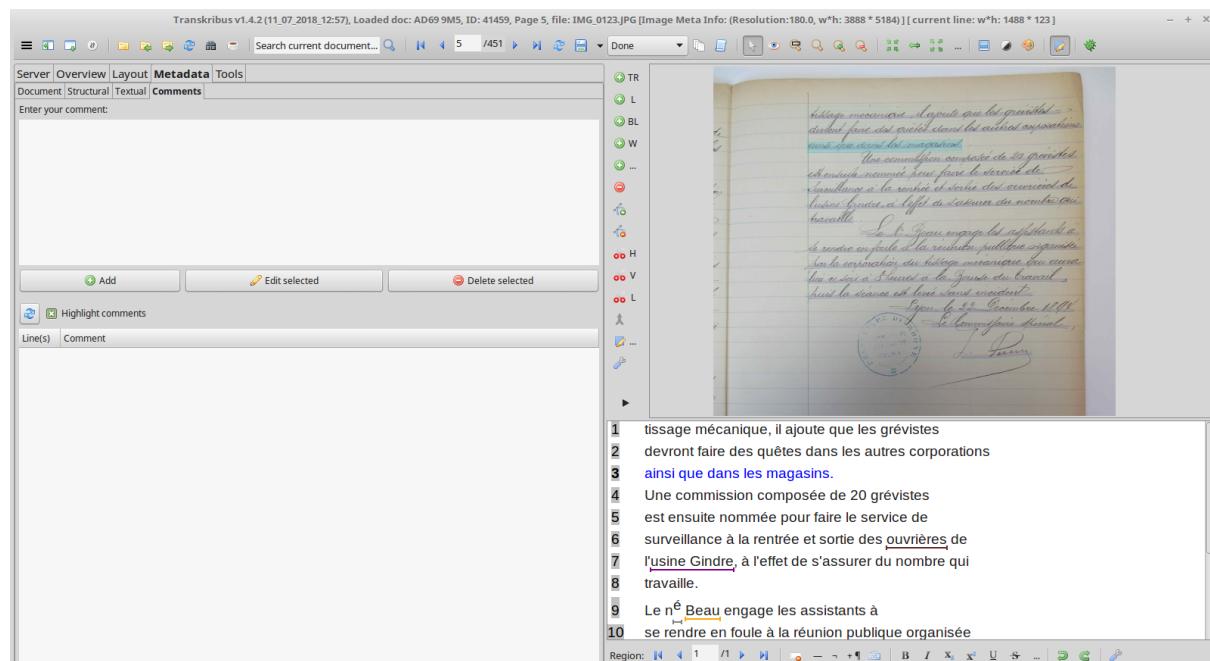
06-Tkb_Metadata_textual.png

B.1.7 Capture d'écran de la fenêtre de personnalisation des *tags*



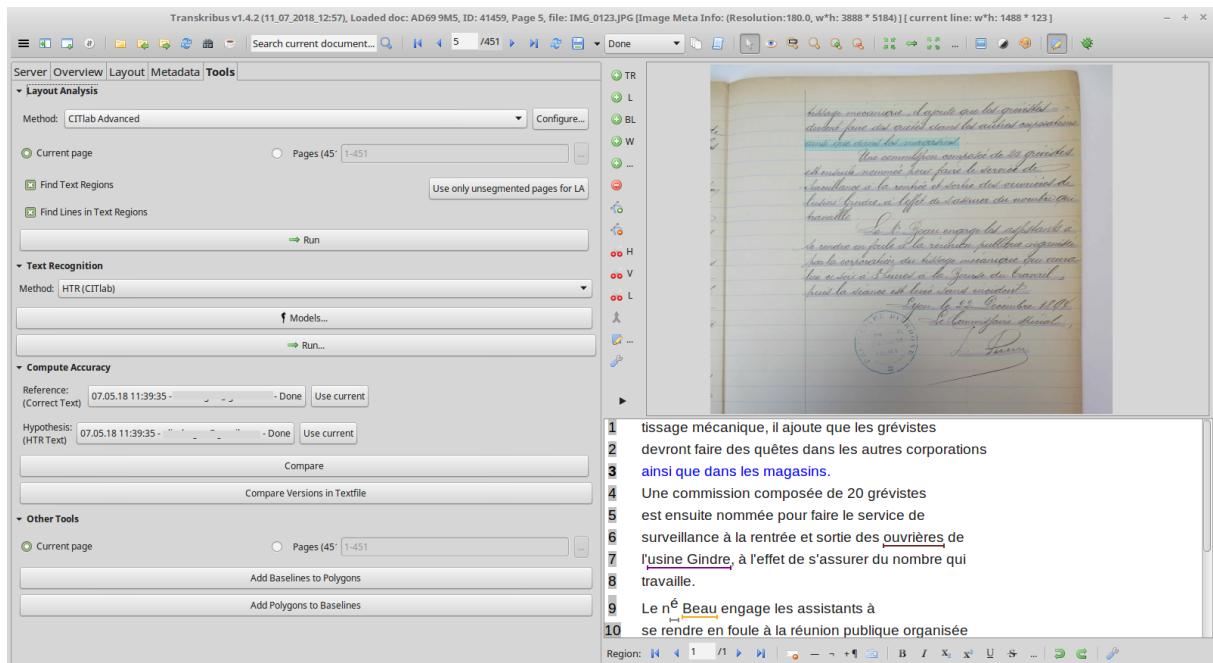
07-Tkb_Metadata_textual_tag-configuration.png

B.1.8 Capture d'écran avec les onglets « Metadata » et « comments » développés



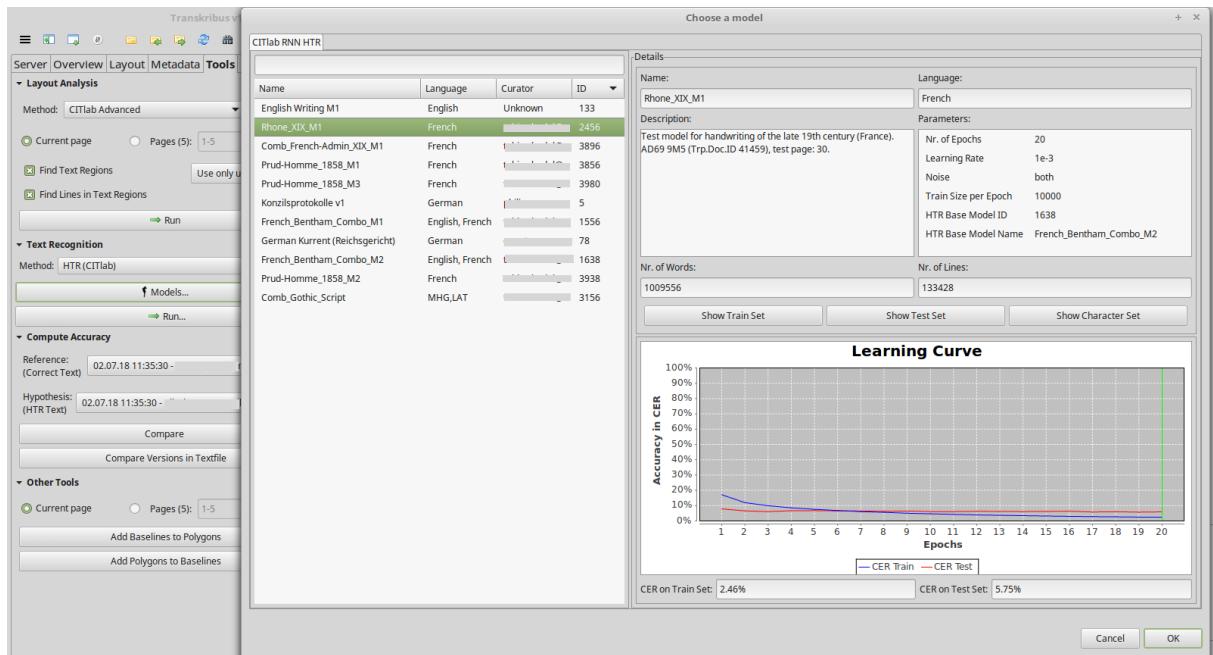
08-Tkb_Metadata_comments.png

B.1.9 Capture d'écran avec l'onglet « Tools » développé



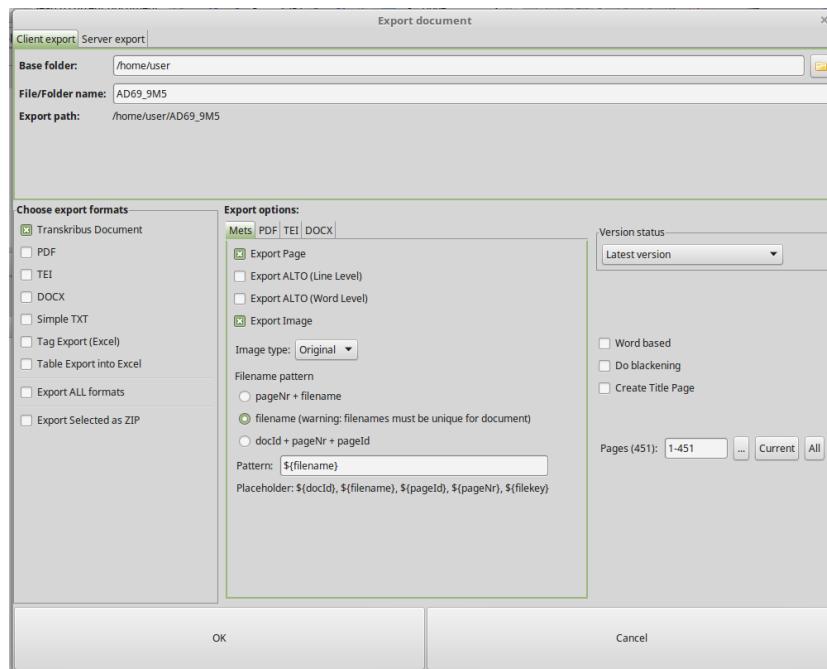
09-Tkb_Tools.png

B.1.10 Capture d'écran de la fenêtre de présentation des modèles d'HTR disponibles



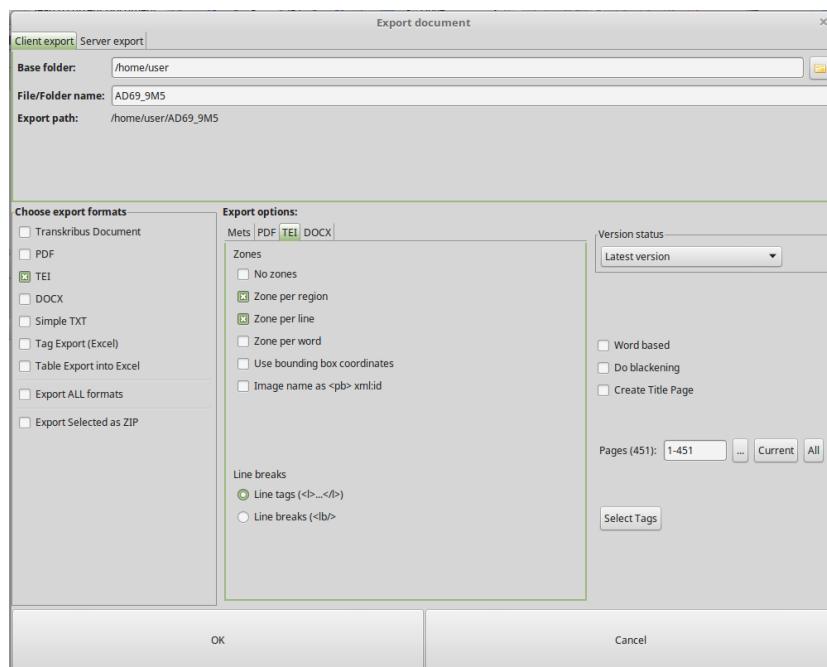
10-Tkb_Tools_HTR-model.png

B.1.11 Capture d'écran de la fenêtre d'export, pour le format « Transkribus Document »



11-Tkb_export_PAGE.png

B.1.12 Capture d'écran de la fenêtre d'export, pour le format TEI



12-Tkb_export_TEI.png

B.2 API

B.2.1 Requêtes

L'API de Transkribus permet d'obtenir divers fichiers au format JSON ou XML contenant des informations relatives aux utilisateurs, aux collections, aux documents et aux fichiers.

Dans le dossier /B - Transkribus/B2 - API/Réponses-Requêtes/, voir en particulier :

- Authentication.xml
- ListCollection.*
- ListDocument.*
- fullDoc.*
- fichierPage.xml

B.2.2 *Script* pour l'API Transkribus

La rédaction de *scripts* Python pour intéragir avec Transkribus via l'API a permis de mettre au point des outils d'export et de transformation des fichiers annotés. Dans le dossier /B - Transkribus/B2 - API/UsingTranskribusAPI/, voir en particulier les fichiers :

- config.py
- requestTranskribus.py
- fromPAGEtoText.py
- toSingleXML.py

Les dossiers data/ et __logs__/ de ce même dossier contiennent le résultat de l'exécution de ces *scripts* pour la collection « Time Us » et le statut « DONE ».

Ce dossier est une copie du *repository* Github disponible à l'adresse suivante : <https://github.com/alix-tz/UsingTranskribusAPI>.

Annexe C

Outils techniques

C.1 Sharedocs

ShareDocs est un service de stockage partagé en ligne fourni par Huma-Num. Il intègre également un outil d'OCR.

Dans le dossier /C - Outils techniques/C1 - Sharedocs/, voir en particulier :

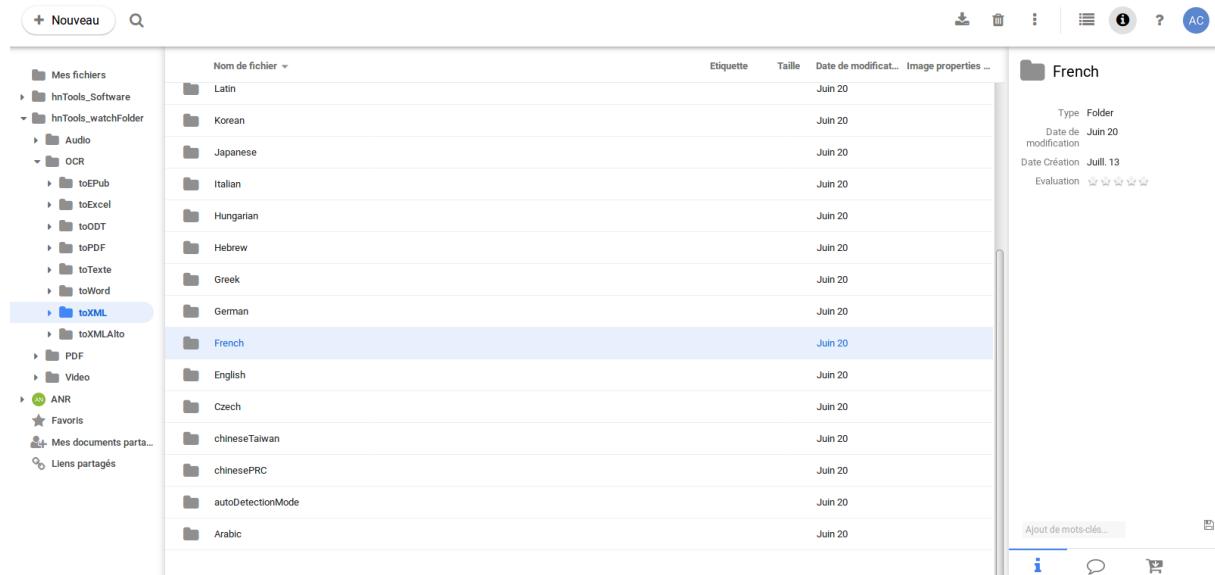
- sharedocs-dossier.png
- sharedocs-OCR.png
- Echo_de_la_Fabrique_001_hnOCR.xml : résultat de l'utilisation de l'OCR de ShareDocs.

C.1.1 Capture d'écran du dossier ShareDocs pour Time Us

The screenshot shows a file management interface for ShareDocs. On the left, there is a sidebar with navigation links like 'Mes fichiers', 'hnTools_Software', 'hnTools_watchFolder', 'ANR', 'Time-Us' (which is expanded to show 'Gallica', 'Lille', and 'Lyon'), 'Marseille', 'Ouvriers des 2 mondes...', 'Paris', 'Réunion consortium ...', 'test OCR', 'Transcriptions', 'Favoris', 'Mes documents partagés', and 'Liens partagés'. The main area displays a list of files in the 'AD du Rhône' folder. The columns are 'Nom de fichier', 'Etiquette', 'Taille', 'Date de modifcat...', and 'Image properties ...'. The files listed are: 4M209, 6M1019, 6UP1_373, 6UP1_1741, 6UP1_1742, 6UP1_1744, 6UP1_1745, 6UP1_1746, 6UP1_1755, 6UP1_1825, 6UP1_1844, 6UP1_2398, 6UP1_3055, 6UP1_3057, 6UP1_3666, and 6UP1_3667. A tooltip at the bottom indicates 'sharedocs-dossier.png'.

sharedocs-dossier.png

C.1.2 Capture d'écran de la configuration de l'outil d'OCR



sharedocs-OCR.png

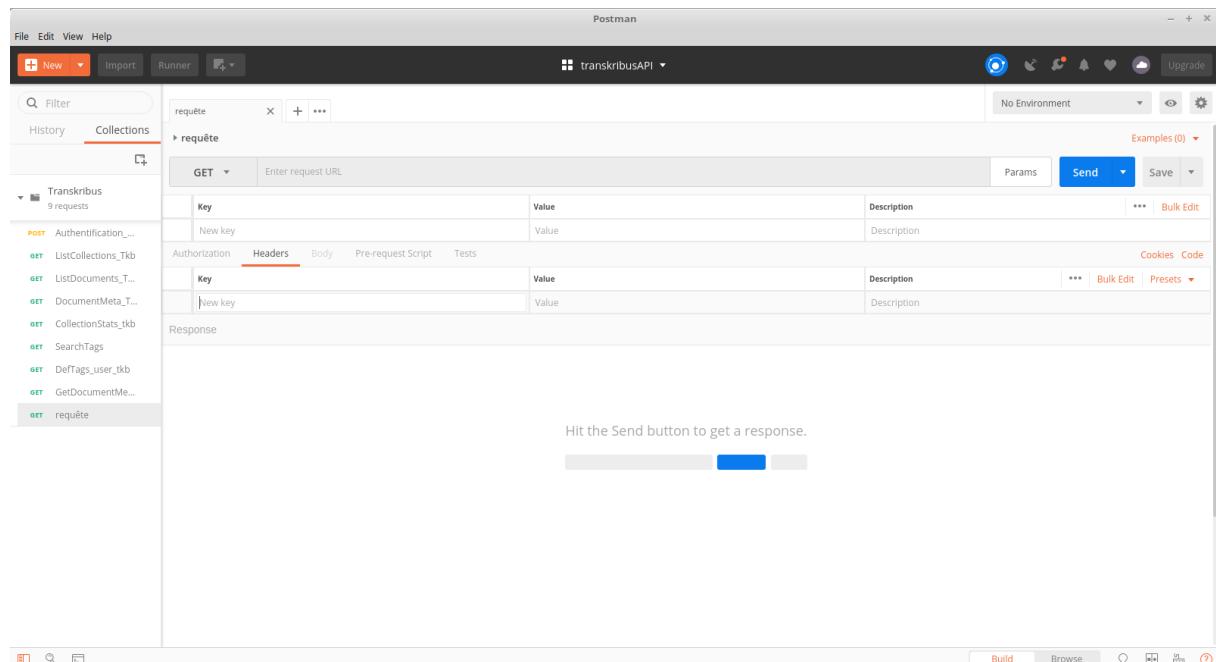
C.2 Postman

Postman est un outil qui permet de réaliser des requêtes HTTP, dans le cadre d'API, par le biais d'une interface graphique. Il permet aussi de réaliser de la documentation sur ces requêtes et de produire leur formulation en différents langages.

Dans le dossier /C - Outils techniques/C1 - Postman/, voir en particulier :

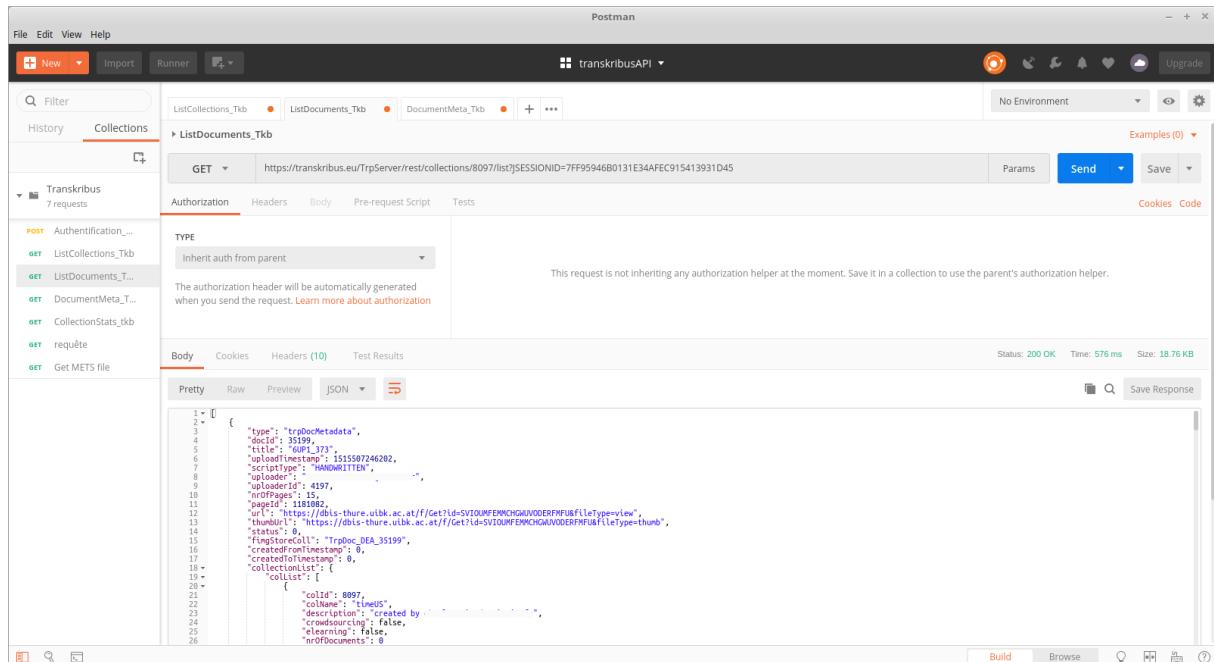
- postman-local-blanc.png
- postman-local-exemple.png
- postman-onligne.png

C.2.1 Capture d'écran de la page de création d'une requête, sur l'interface locale



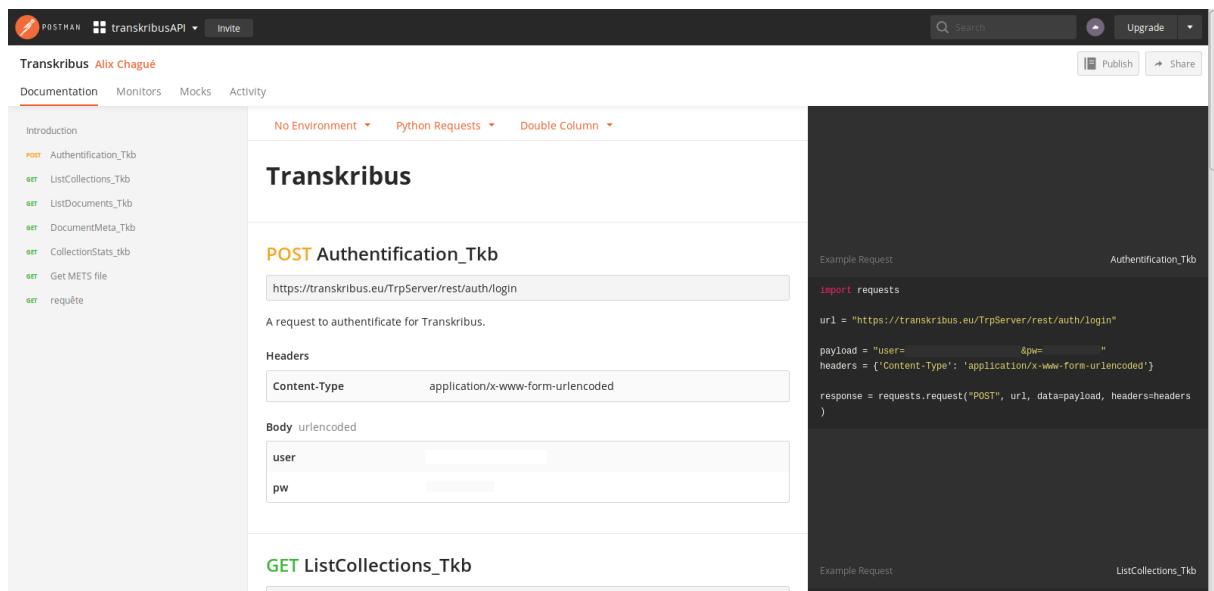
postman-local-blanc.png

C.2.2 Capture d'écran de la page de création d'une requête remplie, sur l'interface locale



postman-local-exemple.png

C.2.3 Capture d'écran de la page de consultation de la documentation d'une requête, sur l'interface en ligne



postman-onligne.png

Annexe D

Annotation

D.1 Éléments d'analyse

La réflexion préalable à l'établissement d'un modèle d'annotation a donné lieu à la création de deux documents internes rassemblant l'ensemble des remarques, interrogations et suggestions de solutions sur la question de l'annotation pour le projet Time Us.

D.1.1 Analyse des pages 3 à 17 du document « AD69 9M5 »

Dans le dossier /D - Annotation/D1 - Éléments d'analyse/, voir en particulier :

- bilanAD699M5avril.xlsx

D.1.2 Note interne sur le fonctionnement des *tags* dans Transkribus et sur l'annotation pour Time Us

Dans le dossier /D - Annotation/D1 - Éléments d'analyse/, voir en particulier :

- analyse-Transkribus-tags.*

D.2 Guides

Plusieurs guides portant sur des éléments d'installation, l'annotation et les bonnes pratiques de travail pour le projet sont publiés sur le wiki du projet. Leur rédaction a constitué l'une des réalisations attendues du stage.

En lignes, voir les adresses suivantes :

- Documentation sur les tags Transkribus : http://timeusage.paris.inria.fr/mediawiki/index.php/Documentation_sur_les_tags_Transkribus
- Guide d'annotation - remarques générales : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_d%27annotation_-_remarques_g%C3%A9n%C3%A9rales

- Guide d'annotation - *tags* du projet Time Us : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_d%27annotation_:_tags_du_projet_Time_Us
- Guide pour charges des fichiers dans Transkribus : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_pour_charger_des_fichiers_dans_Transkribus
- Guide pour l'installation des *tags* Times Us : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_pour_l%27installation_de_la_liste_des_tags_Time_Us

Pour les impressions de ces mêmes pages, dans le dossier /D - Annotation/D2 - Guides/, voir en particulier :

- documentation-sur-les-tags-Transkribus.pdf
- guide-d-annotation-remarques-générales.pdf
- guide-d-annotation-tags-du-projet-time-us.pdf
- guide-pour-charger-des-fichiers-dans-Transkribus.pdf
- guide-pour-l-installation-de-la-liste-des-tags-times-us.pdf

D.3 Fichiers

D.3.1 Configuration locale de Transkribus

Manipuler la configuration locale de Transkribus pour les collaborateurs du projet, quel que soit leur système d'exploitation, permet de simplifier l'installation des *tags* nécessaires pour réaliser l'annotation et pour minimiser les risques d'erreur. Dans /D - Annotation/D3 - Fichiers/, voir en particulier le fichier config.properties, qui nous a permis cela.

D.3.2 TEITags

Paramétriser la coloration syntaxique et la décoration des contenus des balises XML utilisées au sein du wikicode permet de mettre en valeur les éléments annotés. C'est utile pour publier les textes annotés, mais aussi pour dynamiser le guide d'annotation.

Dans le dossier /D - Annotation/D3 - Fichiers/TEITags/, voir en particulier les fichiers :

- ext.teitags.css
- TEITags.body.php

Leurs versions avant modification sont par ailleurs données dans les annexes dans le dossier /D - Annotation.D3 - Fichiers/TEITags/fichiers originaux/.

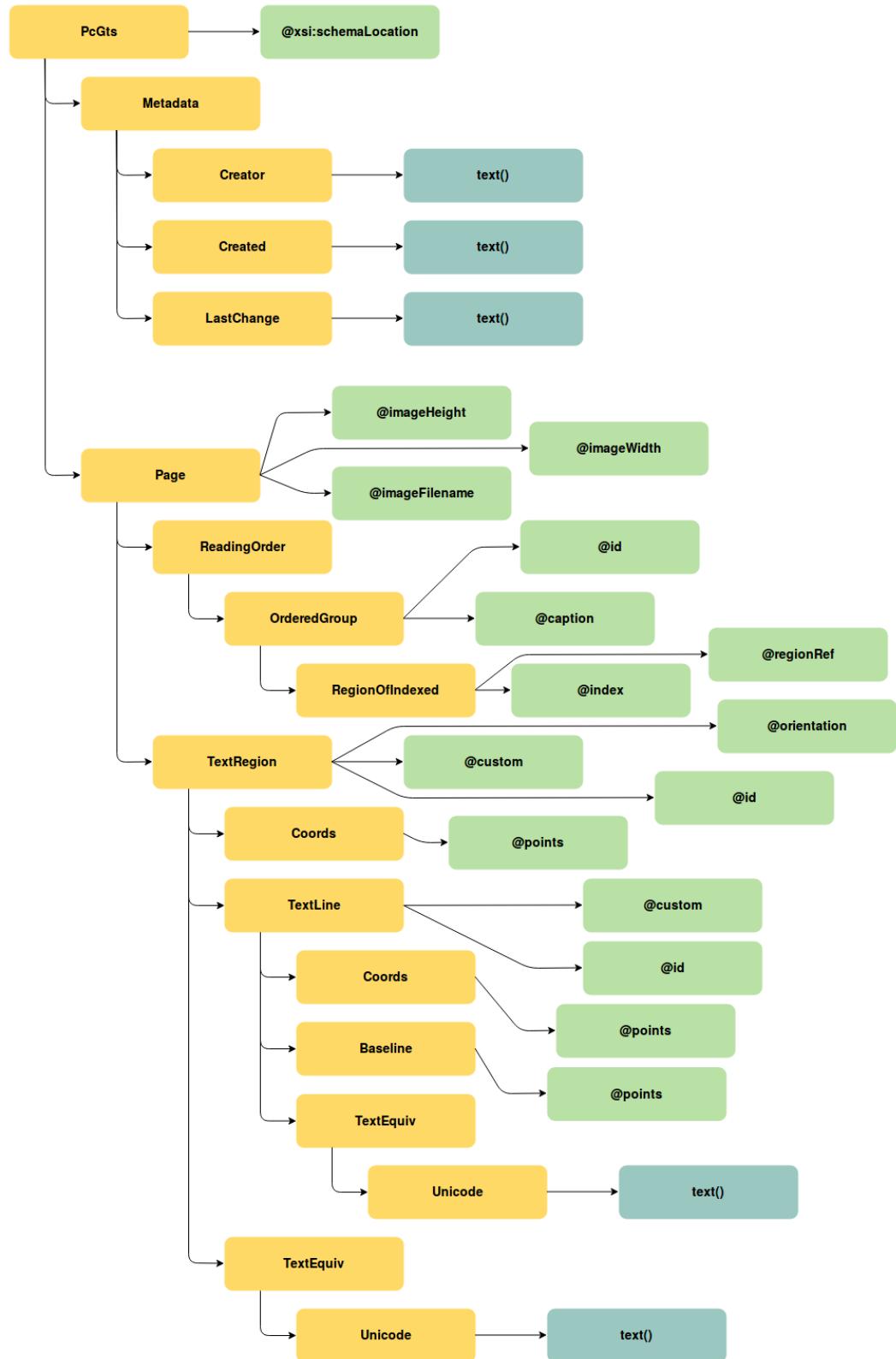
Annexe E

Modélisation des schémas XML

Dans le dossier `/E - Modélisations XML/`, voir les fichiers :

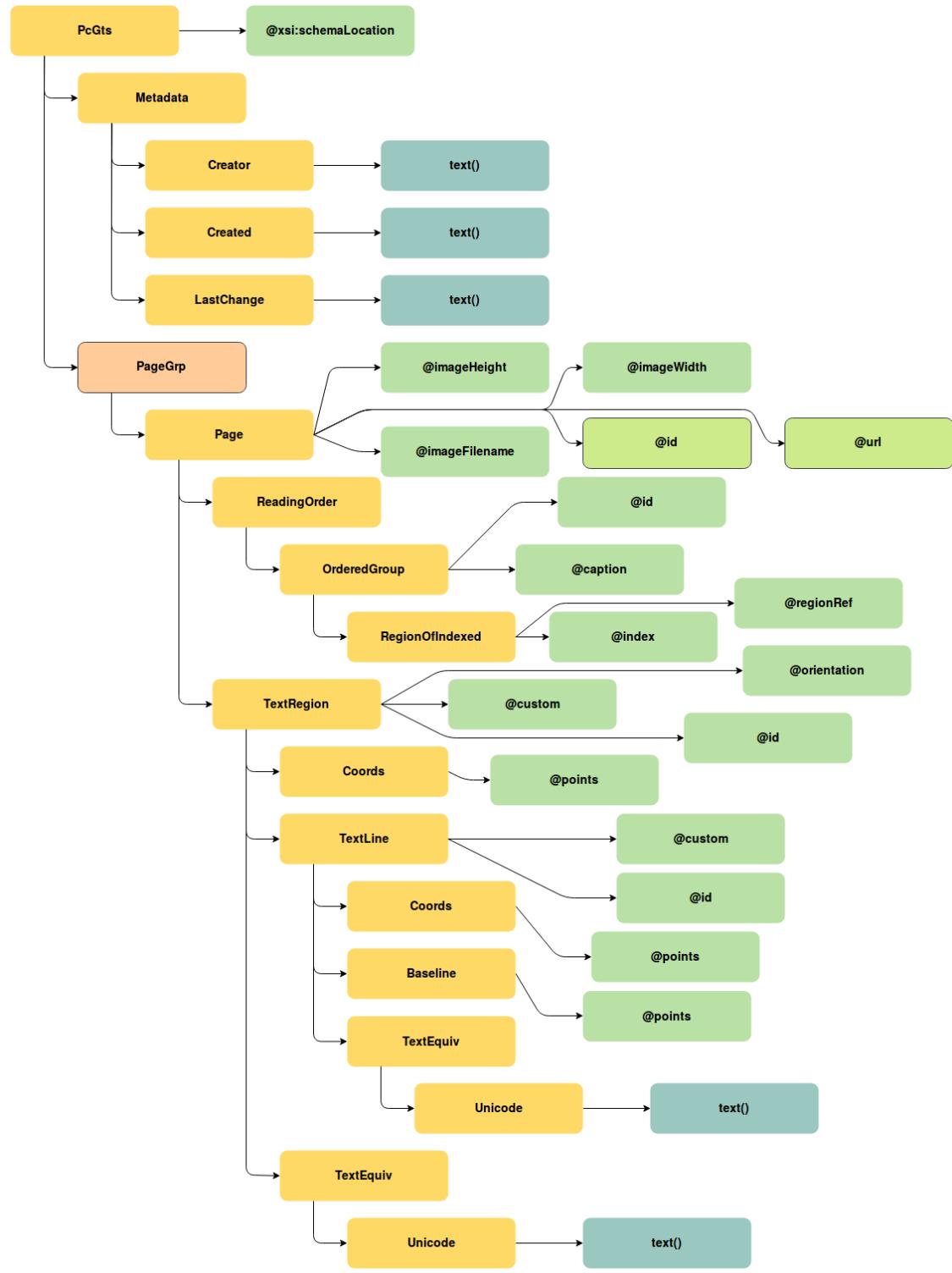
- `model-METS.png` : modélisation du schéma des fichiers XML-METS produits par Transkribus.
- `model-PAGE.png` : modélisation du schéma des fichiers XML-PAGE produits par Transkribus.
- `model-PAGE-TimeUs.png` : modélisation du schéma des fichiers XML-PAGE produits grâce à UsingTranskribusAPI.
- `model-TEI.png` : modélisation du schéma des fichiers XML-TEI produits pour Time Us grâce à Transkribus, sans le détail de l'élément `<teiHeader>`.
- `model-TEI-header.png` : modélisation du contenu de l'élément `<teiHeader>` des fichiers XML-TEI pour le projet Time Us.

E.1 Modélisation du schéma XML-PAGE original



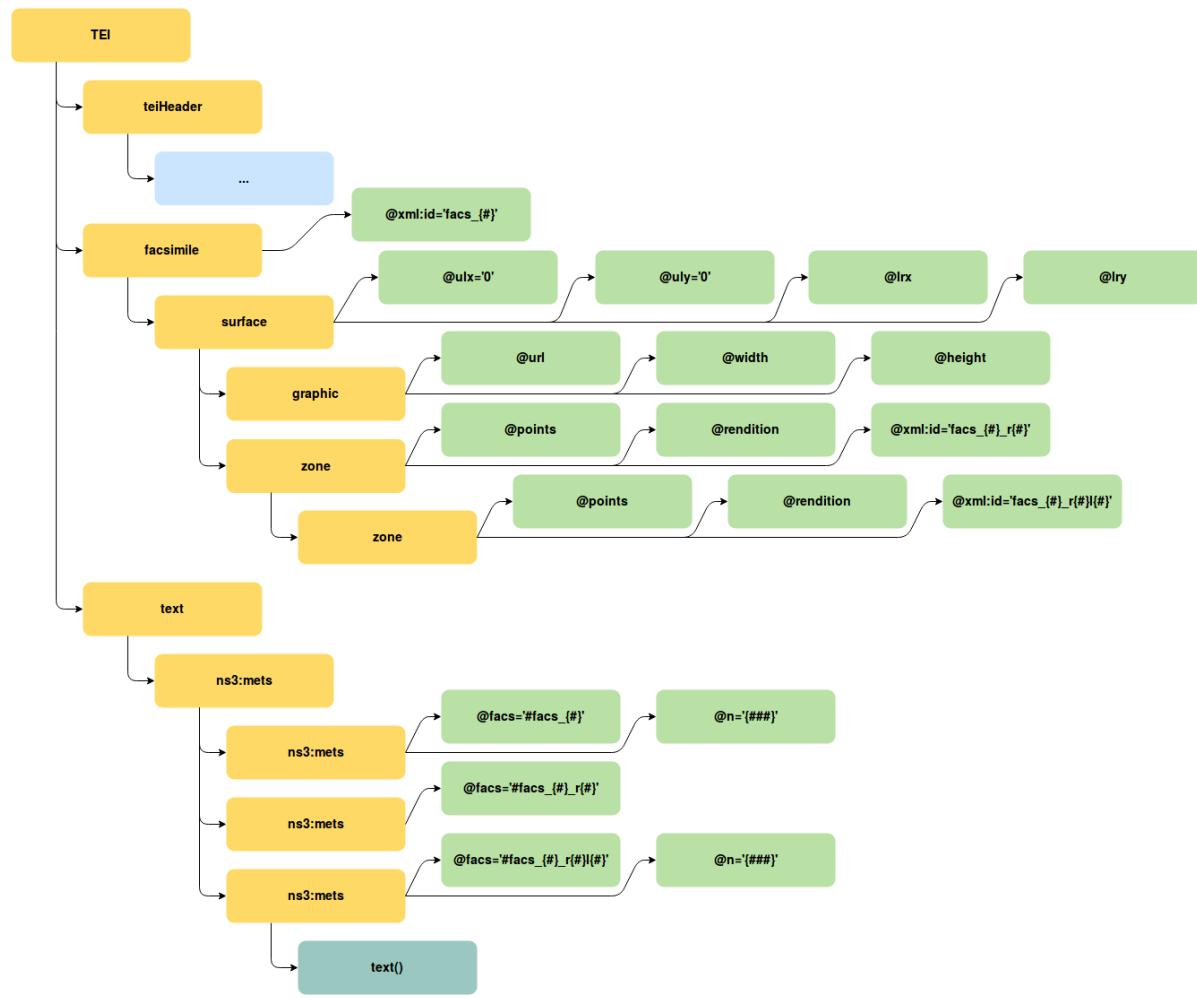
model-PAGE.png

E.2 Modélisation du schéma XML-PAGE adapté pour Time Us



model-PAGE-TimeUs .png

E.3 Modélisation des fichiers XML-TEI du projet Time Us



Annexe F

Transformation XSLT

F.1 Modification des *tags*

Un premier *script* de transformation des fichiers XML-TEI exportés depuis Transkribus a été nécessaire pour rendre les *tags* créés sur mesure conformes aux recommandations de la TEI.

Dans le dossier `/F - Transformation XSLT/`, voir en particulier :

— `modifTag.xsl`

F.2 page2tei

Le travail de Dario Kampkaspar a servi de base pour l’élaboration d’une feuille de transformation du standard PAGE vers le standard TEI répondant aux besoins du projet Time Us.

Voir le contenu du dossier `/F - Transformation XSLT/page2tei/` pour la copie du travail de Dario Kampkaspar, dans la version utilisée pour le projet. Ce travail se trouve sur le *repository* Github du même nom à l’adresse suivante : <https://github.com/dariok/page2tei>.

Voir le contenu du dossier `/F - Transformation XSLT/page2tei_TimeUs/` pour la copie du contenu du *repository* Github du même nom, qui contient la feuille de transformation `page2tei_TU.xsl` réalisée dans le cadre du stage. Ce *repository* se trouve à l’adresse suivante : https://github.com/alix-tz/page2tei_TimeUS.

La feuille de transformation a été appliquée aux fichiers XML-PAGE obtenus à l’issu de l’exécution des *scripts* du projet UsingTranskribusAPI, grâce au parseur Saxon HE (version 9-8-0-12). Voir en particulier le contenu du dossier `/F - Transformation XSLT/Résultat-page2tei_TimeUs/output/`.

Annexe G

TEI Boiler Plate

La réalisation d'un *fork* du projet TEI Boiler Plate nous a permis de tester l'affichage d'un fichier TEI dans un navigateur avec application d'une feuille de style. Ce *fork* se trouve à l'adresse suivante : <https://github.com/alix-tz/TEI-Boilerplate/tree/TimeUs>. Il est reproduit dans le dossier d'annexes.

G.1 Environnement TEI-Boilerplate

Dans le dossier `/G - TEI Boiler Plate/TEI Boiler Plate/`, voir en particulier :

- `dist/css/custom.css`
- `dist/content/41459 - AD69 9M5 (short-jpg).xml` (à ouvrir dans un navigateur, de préférence *Mozilla Firefox*).

G.2 Fichiers préparatoires

Des images et un fichier XML-TEI ont été rassemblés pour préparer ce test. Voir en particulier le contenu du dossier `G - TEI Boiler Plate/AD 69 9M5`.

Annexe H

Documents

H.1 Digital Humanities 2018

Dans le cadre de la valorisation du projet au sein de la communauté des Humanités Numériques, Time Us a été présenté à l'occasion d'un *short paper* durant le congrès international *Digital Humanities* de 2018 qui s'est tenu à Mexico. J'ai élaboré le diaporama et les notes en collaboration avec Marie Puren.

Dans le dossier /H - Documents/H1 - DH2018/, voir en particulier :

- TimeUs-DH2018-Présentation.pdf
- TimeUs-DH2018-Notes.pdf

H.2 Réunion du 1^{er} juin 2018

Dans le cadre de la réunion de bilan de mi-parcours du projet pour l'ANR, qui s'est tenue le 1^{er} juin dans les locaux d'Inria à Paris, j'ai été amenée à présenter mon travail et les enjeux de l'annotation pour le projet. Cette présentation a été réalisée à l'aide d'un diaporama.

Dans le dossier /H - Documents/H2 - ANR-01-06-2018/, voir en particulier :

- presentationANR1juin.pdf
- presentationANR1juin_notes.*

Bibliographie

Archives et centres d'archives

Accueil, [Site web des Archives municipales de Lyon], URL : <http://www.archives-lyon.fr/archives/> (visité le 26/04/2018).

Accueil, [Site web des Archives Départementales et métropolitaines du Rhône], URL : <http://archives.rhone.fr/> (visité le 26/04/2018).

Archives de Lyon, numérisation de la page 20, [Site web des Archives municipales de Lyon], URL : http://www.archives-lyon.fr/static/archives/contenu/serie_I/SII_20.jpg (visité le 26/04/2018).

Archives en ligne - Aide et outils - Aide et outils, Site web des Archives départementales du Nord, URL : http://archivesdepartementales.lenord.fr/?id=436_440 (visité le 26/04/2018).

Bibliothèque, [Site web du Musée des Tissus et Musée des Arts décoratif de Lyon], URL : http://www.mtmad.fr/fr/pages/topnavigation/musees_et_collections/documentation/blibliotheque.aspx (visité le 26/04/2018).

Faire des recherches - Les instruments de recherche - Recherche selon le plan de classement, Site web des Archives départementales du Nord, URL : [http://archivesdepartementales.lenord.fr/?id=recherche_guidee_plan_detail&doc=accounts%2Fmnesys_ad59%2Fdatas%2Fir%2FEtat%20des%20fonds%2FFRAD059_etat_fonds_publiques%2Exml&open=23515&page_ref=9629&unittitle=%205%20U%208%20\(Tourcoing\)%20&unitid=&unitdate=#node23515](http://archivesdepartementales.lenord.fr/?id=recherche_guidee_plan_detail&doc=accounts%2Fmnesys_ad59%2Fdatas%2Fir%2FEtat%20des%20fonds%2FFRAD059_etat_fonds_publiques%2Exml&open=23515&page_ref=9629&unittitle=%205%20U%208%20(Tourcoing)%20&unitid=&unitdate=#node23515) (visité le 26/04/2018).

Faire des recherches - Les instruments de recherche - Recherche selon le plan de classement, Site web des Archives départementales du Nord, URL : [http://archivesdepartementales.lenord.fr/?id=recherche_guidee_plan_detail&open=23427&doc=accounts%2Fmnesys_ad59%2Fdatas%2Fir%2FEtat%20des%20fonds%2FFRAD059_etat_fonds_publiques%2Exml&page_ref=23427&unittitle=Sous%20S%C3%A9rie%205%20U%20\(Conseil%20de%20prud%27homme\)&unitid=&unitdate=#node23427](http://archivesdepartementales.lenord.fr/?id=recherche_guidee_plan_detail&open=23427&doc=accounts%2Fmnesys_ad59%2Fdatas%2Fir%2FEtat%20des%20fonds%2FFRAD059_etat_fonds_publiques%2Exml&page_ref=23427&unittitle=Sous%20S%C3%A9rie%205%20U%20(Conseil%20de%20prud%27homme)&unitid=&unitdate=#node23427) (visité le 26/04/2018).

Fonds Anciens : Série HH : Commerce et Industrie, [Site web des Archives municipales de Lyon], URL : <http://www.archives-lyon.fr/static/archives/contenu/old/fonds/04.htm#44> (visité le 26/04/2018).

LARHRA, MARIN (Anne-Catherine), ROBERT (François) et VERNUS (Pierre), *Bicentenaire du premier conseil des prud'hommes ; Lyon 1806-2006*, avec la coll. de Clé-

mentine Breed, Delphine Digout et Jean-luc Bouville, URL : <http://larhra.ish-lyon.cnrs.fr/cdeprudhomme2/index.htm> (visité le 17/07/2018).

Le répertoire de la série I, [Site web des Archives municipales de Lyon], URL : http://www.archives-lyon.fr/archives/sections/fr/sorienter_fonds/archives_lyon9849/archives_apres1789/reperatoire_serie_i/ (visité le 26/04/2018).

Les archives de la Ville antérieures à 1790, [Site web des Archives municipales de Lyon], URL : http://www.archives-lyon.fr/archives/sections/fr/entete/sorienter_fonds/archives_lyon9849/archives_avant1789/ (visité le 26/04/2018).

Presse locale et régionale du Rhône, Gallica, URL : </html/und/presse-et-revues/rhone> (visité le 17/05/2018).

Presse lyonnaise de 1790 à 1944, Numelyo, URL : <http://collections.bm-lyon.fr/PER003> (visité le 17/05/2018).

Histoire de l'industrie du textile et de la sociologie

ASSIER-ANDRIEU (Louis), « Le Play et la famille-souche des Pyrénées : politique, juridisme et science sociale », *Annales*, 39–3 (1984), p. 495–512, DOI : [10.3406/ahess.1984.283074](https://doi.org/10.3406/ahess.1984.283074).

AUZIAS (Claire), HOUEL (Annik) et PERROT (Michelle), *La grève des ovalistes : Lyon, juin-juillet 1869*, Lyon, France, 2016.

COTTEREAU (Alain), « The fate of manufacture in the industrial world : the silk industries of Lyons and London, 1800-1850 », dans *World of possibilities : flexibility and mass production in western industrialization*, dir. Charles F. Sabel et Jonathan Zeitlin, trad. par Cyprian Blamires, Paris, France, 1997.

FOHLEN (Claude), « La concentration dans l'industrie textile française au milieu du XIXe siècle », *Revue d'Histoire Moderne & Contemporaine*, 2–1 (1955), p. 46–58, DOI : [10.3406/rhmc.1955.2596](https://doi.org/10.3406/rhmc.1955.2596).

FRÉDÉRIC LE PLAY, Encyclopædia Universalis, URL : <http://www.universalis.fr/encyclopedie/frederic-le-play/> (visité le 19/07/2018).

HINCKER (Louis), « Les monographies de famille de l'École de Le Play. Les Études sociales, n 131-132, 1er et 2e semestres 2000. » *Revue d'histoire du XIXe siècle. Société d'histoire de la révolution de 1848 et des révolutions du XIXe siècle*–23 (1er déc. 2001), p. 274–276, URL : <http://journals.openedition.org/rh19/334> (visité le 19/07/2018).

LÉON (Pierre), « Les industries textiles en France au XIXe siècle », *Annales*, 12–2 (1957), p. 326–330, DOI : [10.3406/ahess.1957.2642](https://doi.org/10.3406/ahess.1957.2642).

LEQUIN (Yves), *Aspects économiques des industries lyonnaises de la soie (1870-1900) : la fin de la fabrique*, 2 t., Lyon, France, 1958.

— *Les ouvriers de la région lyonnaise (1848-1914)*. 2, *Les intérêts de classe et la république*, Lyon, France, 1977.

Les Études Sociales : monographies de familles de l'Ecole de Le Play (1855-1930), 2000, URL : <https://gallica.bnf.fr/ark:/12148/bpt6k62254910> (visité le 19/07/2018).

SAVOYE (Antoine), « La monographie sociologique : jalons pour son histoire (1855-1974) », *Les Études Sociales : monographies de familles de l'École de Le Play*–131 (2000).

Fouille de données, TAL et OCR

- EHRMANN (Maud), *Les Entités Nommées, de la linguistique au TAL : Statut théorique et méthodes de désambiguisation*, Thèse de doctorat, Université Paris Diderot, 2008, URL : <https://hal.archives-ouvertes.fr/tel-01639190> (visité le 27/04/2018).
- Linguistique et traitements automatiques des langues*, dir. Catherine Fuchs, OCLC : 246361071, Paris, 1993 (Hachette université langue linguistique communication).
- GENET (Jean-Philippe), « Informatique et Histoire », *Le bulletin de l'EPI*–49 (mars 1988), p. 12.
- GRANET (Adeline), MORIN (Emmanuel), MOUCHÈRE (Harold), QUINIOU (Solen) et VIARD-GAUDIN (Christian), « Décodeur neuronal pour la transcription de documents manuscrits anciens », *25e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Rennes, France (17 mai 2018).
- IBEKWE-SANJUAN (Fidelia) et SANJUAN (Eric), « Ingénierie linguistique et fouille de textes », *Veille stratégique, scientifique et technologique* (, 24 oct. 2004), URL : https://archivesic.ccsd.cnrs.fr/sic_00001396/document (visité le 11/08/2018).
- KARPINSKI (Romain) et BELAID (Abdel), *Rapport Evaluation des OCR*, Research Report, LORIA - Université de Lorraine, 2016, URL : <https://hal.inria.fr/hal-01356824> (visité le 04/06/2018).
- KERMORVANT (Christopher), *La reconnaissance des écritures manuscrites*, non publié, Reconnaissance par ordinateur des écritures anciennes : le projet HIMANIS, 29 mai 2018.
- LEMAITRE (Aurélie) et CAMILLERAPP (Jean), « Text Line Extraction in Handwritten Document with Kalman Filter Applied on Low Resolution Image », *Second International Conference on Document Image Analysis for Libraries (DIAL)*, Document Image Analysis for Libraries, 2006. DIAL '06. Second International Conference on Lyon, France (avr. 2006), DOI : [10.1109/DIAL.2006.41](https://doi.org/10.1109/DIAL.2006.41).
- LEMAITRE (Aurélie), CAMILLERAPP (Jean) et COÜASNON (Bertrand), « Handwritten text segmentation using blurred image », *DRR - Document Recognition and Retrieval XXI*, DRR - Document Recognition and Retrieval XXI San Francisco, États-Unis (janv. 2014), URL : <https://hal.inria.fr/hal-01087210> (visité le 04/06/2018).
- MACMURRAY (Erin), *Discours de presse et veille stratégique d'évènements. Approche textométrique et extraction d'informations pour la fouille de textes*, Thèse de doctorat,

Paris, Université de la Sorbonne nouvelle, 2012, URL : <https://tel.archives-ouvertes.fr/tel-01157562> (visité le 11/08/2018).

MAGALLON (Thibault), BECHET (Frédéric) et FAVRE (Benoît), « Détection d'erreurs dans des transcriptions OCR de documents historiques par réseaux de neurones récurrents multi-niveau », *15e Conférence en Recherche d'Information et Applications (CORIA)*, Rennes, France (17 mai 2018).

MARTINEAU (Claude), TOLONE (Elsa) et VOYATZI (Stavroula), « Les Entités Nommées » (, 2007), p. 25.

OMRANE (Nouha), NAZARENKO (Adeline) et SZULMAN (Sylvie), « Les entités nommées : éléments pour la conceptualisation », *21es Journées francophones d'Ingénierie des Connaissances*, Nîmes, France (juin 2010), http://www.ic2010.mines-ales.fr/index.php?option=com_content&view=article&id=50&Itemid=44, URL : <https://hal.archives-ouvertes.fr/hal-00525530> (visité le 24/07/2018).

Reconnaissance d'entités nommées, dans *Wikipédia*, Page Version ID : 144628341, 2018, URL : https://fr.wikipedia.org/w/index.php?title=Reconnaissance_d%27entit%C3%A9s_nomm%C3%A9es&oldid=144628341 (visité le 27/04/2018).

SEAWARD (Louise) et CHARWAT (Elaine), *If you teach a computer to READ...* CILIP Update, 2017.

Traitemenat automatique du langage naturel, dans *Wikipédia*, Page Version ID : 148328049, 2018, URL : https://fr.wikipedia.org/w/index.php?title=Traitemenat_automatique_du_langage_naturel&oldid=148328049 (visité le 17/05/2018).

TUFFÉRY (Stéphane), *Data mining et statistique décisionnelle - 4ème édition*, 4e édition, Paris, 2012.

Numérisation

ANDRO (Mathieu), « Actualité de la numérisation », *Bulletin des bibliothèques de France*, Supplément 2011 (2011), p. 27–29, URL : <https://hal.archives-ouvertes.fr/hal-01094553> (visité le 04/06/2018).

BELAÏD (Abdel), PIERRON (Laurent), NAJMAN (Laurent) et REYREN (Dominique), « La numérisation de documents : principes et évaluation des performances », dans *Bibliothèques numériques. Cours INRIA, octobre 2000, La Bresse*, dir. Bernard Hidoine et Jean-Claude Le Moal, 2000, 35 p, URL : <https://hal.inria.fr/inria-00099148> (visité le 04/06/2018).

BÜLOW (Anna), AHMON (Jess) et SPENCER (Ross), *Preparing collections for digitization*, avec la coll. de National Archives (Great Britain), OCLC : ocn519248700, London, 2011.

MAUREL (Lionel), « Quel modèle économique pour une numérisation patrimoniale respectueuse du domaine public ? », dans *Communs du savoir et bibliothèques*, 2017, URL : <https://hal-univ-paris10.archives-ouvertes.fr/hal-01528096> (visité le 04/06/2018).

Inria et ALMANaCH

Accueil, ALMANaCH, URL : <https://team.inria.fr/almanach/fr/> (visité le 12/08/2018).

European Holocaust Research Infrastructure, URL : <https://ehri-project.eu/> (visité le 12/08/2018).

Iperion CH, URL : <http://www.iperionch.eu/> (visité le 12/08/2018).

Le modèle « Équipe-projet » Inria, Inria, URL : <https://www.inria.fr/recherches/structures-de-recherche/modele-equipe-projet> (visité le 12/08/2018).

PARTHENOS Project, URL : <http://www.parthenos-project.eu/> (visité le 12/08/2018).

Plan stratégique scientifique, Inria, URL : <https://www.inria.fr/institut/strategie/plan-strategique> (visité le 12/08/2018).

Politique européenne & internationale, Inria, URL : <https://www.inria.fr/europe-international/politique-europeenne-internationale> (visité le 12/08/2018).

Présentation - ALMANACH, Inria, URL : <https://www.inria.fr/equipes/almanach> (visité le 12/08/2018).

Projet PARSITI, ANR, URL : [http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2\[CODE\]=ANR-16-CE33-0021](http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2[CODE]=ANR-16-CE33-0021) (visité le 12/08/2018).

Projet PROFITEROLE, ANR, URL : [http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2\[CODE\]=ANR-16-CE38-0010](http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2[CODE]=ANR-16-CE38-0010) (visité le 12/08/2018).

Projets, ALMANaCH, URL : <https://team.inria.fr/almanach/fr/projects/> (visité le 12/08/2018).

SAGOT (Benoit), *Un projet ANR-NSF sur le développement d'outils informatiques de modélisation de données neurolinguistiques*, ALMANaCH, URL : <https://team.inria.fr/almanach/fr/anr-nsf-project-to-develop-computational-tools-for-modeling-neurolinguistic-data/> (visité le 12/08/2018).

Site web Alpage, URL : <https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=Accueil> (visité le 12/08/2018).

SoSweet, URL : <http://sosweet.inria.fr/> (visité le 12/08/2018).

Syntactic Parsing and Multiword Expressions in French, URL : <http://parsemefr.lif.univ-mrs.fr/doku.php> (visité le 12/08/2018).

Time Us

Accueil, TimeUsage, URL : <http://timeusage.paris.inria.fr/mediawiki/index.php/Accueil> (visité le 03/08/2018).

ALBERT (Anaïs), « Consumption as "hidden transcript" in the 1917 midinettes' strike in Paris », *European Social Science History Conference*, Belfast, Irlande du Nord (4 avr. 2018).

Centre Maurice Halbwachs : Accueil, URL : <https://www.cmh.ens.fr/> (visité le 03/08/2018).

Documentation sur les tags Transkribus, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Documentation_sur_les_tags_Transkribus (visité le 11/08/2018).

Guide d'annotation, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_d'annotation (visité le 12/08/2018).

Guide d'annotation : remarques générales, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_d%27annotation:_remarques_g%C3%A9n%C3%A9rales (visité le 11/08/2018).

Guide d'annotation : tags du projet Time Us, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_d%27annotation:_tags_du_projet_Time_Us (visité le 11/08/2018).

Guide pour charger des fichiers dans Transkribus, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_pour_charger_des_fichiers_dans_Transkribus (visité le 11/08/2018).

Guide pour l'installation de la liste des tags Time Us, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Guide_pour_l%27installation_de_la_liste_des_tags_Time_Us (visité le 11/08/2018).

Inventaire des sources, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Inventaire_des_sources (visité le 11/08/2018).

IRHiS : Accueil, URL : <https://irhis.univ-lille.fr/> (visité le 03/08/2018).

Laboratoire ICT : Accueil, URL : <http://www.ict.univ-paris-diderot.fr/> (visité le 03/08/2018).

LARHRA : Accueil, URL : <http://larhra.ish-lyon.cnrs.fr/> (visité le 03/08/2018).

MARTINI (Manuela), « Gendered division of work and wage conflicts in the Lyon silk trades at the end of the 19th century », *European Social Science History Conference*, Belfast, Irlande du Nord (4 avr. 2018).

MONTENACH (Anne), « "The Bazaar economy of trades" : gender and wage systems in the late seventeenth and eighteenth century Lyon textile industry », *European Social Science History Conference*, Belfast, Irlande du Nord (4 avr. 2018).

Participants au projet, TimeUsage, URL : http://timeusage.paris.inria.fr/mediawiki/index.php/Participants_au_projet (visité le 03/08/2018).

PUREN (Marie), CHAGUÉ (Alix), MARTINI (Manuela), CLERGERIE (Éric de la) et RIONDET (Charles), « Creating gold data to understand the gender gap in the French textile trades (17th - 20th centuries) », *Digital Humanities*, Mexico City, Mexique (28 juin 2018).

RIONDET (Charles), *Gitlab / TimeUs*, URL : <https://gitlab.inria.fr/criondet/TimeUs> (visité le 11/08/2018).

TELEMME : Accueil, URL : <http://telemme.mmsh.univ-aix.fr/> (visité le 03/08/2018).

TIME US, *Bibliographie partagée*, Zotero, URL : <https://www.zotero.org/groups/2174782/time-us> (visité le 03/08/2018).

- *TIME-US / Présentation du programme de recherche TIME-US*, TIME-US, URL : <https://timeus.hypotheses.org/1> (visité le 19/06/2018).
- *TIME-US / Travail, rémunération, textile et foyer (XVIIe-XXe siècle)*, URL : <https://timeus.hypotheses.org/> (visité le 03/08/2018).

Transcribe Bentham

CAUSER (Tim) et WALLACE (Valerie), « Building A Volunteer Community : Results and Findings from Transcribe Bentham », *Digital Humanities Quarterly*, 006–2 (12 oct. 2012).

CAUSER (Tim), GRINT (Kris), SICHANI (Anna-Maria) et TERRAS (Melissa), « "Making such bargain" : Transcribe Bentham and the quality and cost-effectiveness of crowd-sourced transcription », *Digital Scholarship in the Humanities* (, 15 janv. 2018), DOI : [10.1093/llc/fqx064](https://doi.org/10.1093/llc/fqx064).

COHEN (Patricia), « For Bentham and Others, Scholars Enlist Public to Transcribe Papers », *The New York Times* (, 27 déc. 2010), URL : <https://www.nytimes.com/2010/12/28/books/28transcribe.html> (visité le 11/06/2018).

SEWARD (Louise), *Project Update – teaching a computer to READ Bentham*, UCL Transcribe Bentham, 9 juin 2017, URL : <http://blogs.ucl.ac.uk/transcribe-bentham/2017/06/09/project-update-teaching-a-computer-to-read-bentham/> (visité le 04/06/2018).

— *Project Update – Bentham vs the computer*, UCL Transcribe Bentham, 23 févr. 2018, URL : <http://blogs.ucl.ac.uk/transcribe-bentham/2018/02/23/project-update-bentham-vs-computer/> (visité le 04/06/2018).

Transcription Desk, Transcribe Bentham, URL : http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham (visité le 04/06/2018).

Transcription Desk - Technical Requirements, Transcribe Bentham, URL : http://www.transcribe-bentham.da.ulcc.ac.uk/td/Technical_Requirements (visité le 12/08/2018).

UNIVERSITY COLLEGE LONDON, *READ project*, UCL Transcribe Bentham, URL : <http://blogs.ucl.ac.uk/transcribe-bentham/category/read-project/> (visité le 04/06/2018).

— *UCL Transcribe Bentham*, URL : <http://blogs.ucl.ac.uk/transcribe-bentham/> (visité le 04/06/2018).

Media Wiki

DjVuLibre, URL : <http://djvu.sourceforge.net/> (visité le 12/08/2018).

MediaWiki / Extension :TEITags, URL : <https://www.mediawiki.org/wiki/Extension:TEITags> (visité le 12/08/2018).

MediaWiki / Manual :Configuration settings, URL : https://www.mediawiki.org/wiki/Manual:Configuration_settings (visité le 27/04/2018).

MediaWiki / Manual :LocalSettings.php, URL : <https://www.mediawiki.org/wiki/Manual:LocalSettings.php> (visité le 27/04/2018).

MediaWiki / Manual :Short URL, URL : https://www.mediawiki.org/wiki/Manual:Short_URL (visité le 27/04/2018).

MediaWiki / Manual :Short URL/Apache, URL : https://www.mediawiki.org/wiki/Manual:Short_URL/Apache (visité le 27/04/2018).

MediaWiki / Manual :\$wgFavicon, URL : [https://www.mediawiki.org/wiki/Manual:\\$wgFavicon](https://www.mediawiki.org/wiki/Manual:$wgFavicon) (visité le 27/04/2018).

MediaWiki / Manual :\$wgResourceBasePath, URL : [https://www.mediawiki.org/wiki/Manual:\\$wgResourceBasePath](https://www.mediawiki.org/wiki/Manual:$wgResourceBasePath) (visité le 27/04/2018).

Semantic MediaWiki, Semantic MediaWiki, URL : https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki (visité le 12/08/2018).

Version, TimeUsage, URL : <http://timeusage.paris.inria.fr/mediawiki/index.php/Sp%C3%A9cial:Version> (visité le 12/08/2018).

TEI Boiler Plate

RIONDET (Charles), *TEI Boilerplate : Displaying a facsimile beside a transcription*, Bag of tags, URL : <https://tags.hypotheses.org/60> (visité le 20/07/2018).

SIMPSON (Grant), WALSH (John) et MOADDELI (Saeed), *Github / TEI-Boilerplate*, original-date : 2012-09-24T15:33:46Z, 26 mai 2018, URL : <https://github.com/GrantLS/TEI-Boilerplate> (visité le 20/07/2018).

TEI Boilerplate / Index, URL : <http://dcl.ils.indiana.edu/teibp/index.html> (visité le 20/07/2018).

Transformation XSL

KAMPKASPAR (Dario), *Github / page2tei*, original-date : 2018-05-02T16 :32 :49Z, 9 août 2018, URL : <https://github.com/dariok/page2tei> (visité le 12/08/2018).

— *Github / dariok*, GitHub, URL : <https://github.com/dariok> (visité le 12/08/2018).
Saxon Documentation, URL : <https://www.saxonica.com/documentation9.5/using-xsl/commandline.html> (visité le 12/07/2018).

Saxon XSLT and XQuery Processor : Saxon-HE/9.8, SourceForge, URL : <https://sourceforge.net/projects/saxon/files/Saxon-HE/9.8/> (visité le 12/07/2018).

xml - What this stands for in xsl ?, Stack Overflow, URL : https://stackoverflow.com/questions/17210368/what-this-stands-for-in-xsl-match-node?utm_medium=organic&utm_source=google_rich_qa&utm_campaign=google_rich_qa (visité le 22/05/2018).

Transkribus

About, READ Project, URL : <https://read.transkribus.eu/about/> (visité le 06/06/2018).

DÉJEAN (Hervé) et MEUNIER (Jean-Luc), *Github / TranskribusPyClient*, original-date : 2016-11-29T09:06:19Z, 22 mai 2018, URL : <https://github.com/Transkribus/TranskribusPyClient> (visité le 03/08/2018).

Github / Transkribus, URL : <https://github.com/transkribus/> (visité le 03/08/2018).

Home, Transkribus, URL : https://transkribus.eu/wiki/index.php/Main_Page (visité le 12/08/2018).

How to enrich transcribed documents with mark-up, URL : https://transkribus.eu/wiki/images/e/e8/How_to_enrich_transcribed_documents_with_mark-up.pdf.

How to train a HTR model in Transkribus, URL : https://transkribus.eu/wiki/images/3/34/HowToTranscribe_Train_A_Model.pdf.

How to transcribe : basic instruction, URL : https://transkribus.eu/wiki/images/5/50/How_To_Transcribe_Documents_with_Transkribus.pdf.

Join us at the 2018 Scanathon in London, Zurich and Helsinki!, READ Project, URL : <https://read.transkribus.eu/2018/05/04/join-us-at-the-2018-scanathon/> (visité le 12/08/2018).

PLETSCHACHER (Stefan) et ANTONACOPOULOS (Apostolos), « The PAGE (Page Analysis and Ground-truth Elements) format framework », dans, 2010, p. 257–260, DOI : [10.1109/ICPR.2010.72](https://doi.org/10.1109/ICPR.2010.72).

Questions and Answers, Transkribus, URL : https://transkribus.eu/wiki/index.php/Questions_and_Answers (visité le 12/08/2018).

READ Web UI, TranskribusWeb, URL : <https://transkribus.eu/readTest/library/> (visité le 06/06/2018).

SALMONS (Jim), *Transkribus & Magazines : Transkribus' Transcription & Recognition Platform (TRP) as Social Machine...* Medium, 10 juin 2015, URL : <https://medium.com/factminers-musings/transkribus-magazines-5-e75070d5f00f> (visité le 20/07/2018).

The ScanTent, URL : <https://scantent.cvl.tuwien.ac.at/en/> (visité le 12/08/2018).

transScriptorium, URL : <http://transcriptorium.eu/> (visité le 04/06/2018).

- TRANSKRIBUS, *application.wadl*, URL : <https://transkribus.eu/TrpServer/rest/application.wadl> (visité le 03/08/2018).
- REST Interface - Transkribus Wiki, URL : https://transkribus.eu/wiki/index.php/REST_Interface (visité le 03/08/2018).
- Transkribus*, READ Project, URL : <https://read.transkribus.eu/transkribus/> (visité le 06/06/2018).
- Transkribus*, URL : <https://transkribus.eu/Transkribus/> (visité le 04/06/2018).
- TRANSKRIBUS, *Transkribus REST Interface Description*, URL : <https://transkribus.eu/TrpServer/Swadl/wadl.html> (visité le 03/08/2018).
- Transkribus - Interfaces Maps*, URL : https://read.transkribus.eu/wp-content/uploads/2017/07/Interfaces_Map_v4.0.pdf (visité le 06/06/2018).

Webographie générale

Creation of a TEI-based corpus, SSK at Parthenos, URL : <https://ssk-application.parthenos.d4science.org/ssk/#/scenarios/AWIAia1ie-S72mFoMfb6/1> (visité le 25/05/2018).

Digitizing textual material, SSK at Parthenos, URL : <https://ssk-application.parthenos.d4science.org/ssk/#/scenarios/AWIAie9fe-S72mFoMfca/1> (visité le 25/05/2018).

Horizon 2020, URL : <http://www.horizon2020.gouv.fr/> (visité le 12/08/2018).

Hypotheses, URL : <https://hypotheses.org/> (visité le 12/08/2018).

ISTEX - Socle de la bibliothèque scientifique numérique nationale, URL : <https://www.istex.fr/> (visité le 12/08/2018).

Laboratoire de linguistique formelle, URL : <http://www.llf.cnrs.fr/> (visité le 12/08/2018).

LARHRA, Symogih : un système modulaire de gestion de l'information historique, URL : <http://symogih.org/> (visité le 15/05/2018).

Xnview Software, URL : <https://www.xnview.com/fr/> (visité le 12/08/2018).

Utilitaires

CAMPS (Jean-Baptiste), *Github / biblatex-enc*, 23 mai 2017, URL : <https://github.com/Jean-Baptiste-Camps/biblatex-enc> (visité le 11/08/2018).

HOLZNER (Steven), *XSLT fondamental*, OCLC : 422244771, Paris, 2002.

Inclure sa bibliographie : de Zotero à LaTex. - Renault Jonathan (muchos), URL : <http://renoult-jonathan.tilde3.eu/docs/inclure-bibliographie-zotero-latex> (visité le 04/06/2018).

ROUQUETTE (Maïeul), CHABANNES (Brendan) et ROUQUETTE (Enimie), *(Xe)LaTeX appliqué aux sciences humaines*, Tampere, Finlande, 2012.

TEI CONSORTIUM, *P5 : Guidelines for Electronic Text Encoding and Interchange*, 23 juil. 2018, URL : <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/> (visité le 03/08/2018).

Livrables techniques

- CHAGUÉ (Alix), *Github / page2tei_TimeUS*, original-date : 2018-07-13T15:04:18Z, 23 juil. 2018, URL : https://github.com/alix-tz/page2tei_TimeUS (visité le 03/08/2018).
- *Github / TEI-Boilerplate*, original-date : 2018-07-20T14:54:35Z, 20 juil. 2018, URL : <https://github.com/alix-tz/TEI-Boilerplate/tree/TimeUs> (visité le 03/08/2018).
 - *Github / UsingTranskribusAPI*, original-date : 2018-06-20T14:59:14Z, 30 juil. 2018, URL : <https://github.com/alix-tz/UsingTranskribusAPI> (visité le 03/08/2018).

Logiciels et services utilisés

ALPAGE, *FRMG Wiki*, URL : <http://alpage.inria.fr/frmgwiki/> (visité le 03/08/2018).
Github, URL : <https://github.com> (visité le 03/08/2018).
Gitlab, URL : <https://about.gitlab.com/> (visité le 03/08/2018).
JETBRAINS, *PyCharm*, URL : <https://www.jetbrains.com/pycharm/> (visité le 11/08/2018).
JGRAPH LTD, *draw.io*, URL : <https://www.draw.io/> (visité le 03/08/2018).
Postman, URL : <https://www.getpostman.com> (visité le 03/08/2018).
ShareLatex, URL : <https://fr.sharelatex.com/project> (visité le 03/08/2018).
Sublime Text, URL : <https://www.sublimetext.com/> (visité le 11/08/2018).

Packages python utilisés

Beautiful Soup 4.4.0 : documentation, URL : <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> (visité le 12/08/2018).

lxml - Processing XML and HTML with Python, URL : <https://lxml.de/> (visité le 12/08/2018).

Python 3.7.0 documentation : datetime, URL : <https://docs.python.org/3/library/datetime.html> (visité le 12/08/2018).

Python 3.7.0 documentation : json, URL : <https://docs.python.org/3/library/json.html> (visité le 12/08/2018).

Python 3.7.0 documentation : os, URL : <https://docs.python.org/3/library/os.html> (visité le 12/08/2018).

Requests 2.19.1 : documentation, URL : <http://docs.python-requests.org/en/master/#> (visité le 12/08/2018).

Glossaire

- **ALTO** : *Analyzed Layout and Text Object* - Standard XML permettant de stocker des données techniques de description de la structure d'un document ayant fait l'objet d'un OCR, généralement articulé avec un fichier XML-METS. Maintenant par le Bibliothèque nationale de France et la Bibliothèque du Congrès.
- **API** : *Application Programming Interface* - Ensemble de requêtes, souvent HTTP, permettant d'interagir avec un serveur et ses données sans passer par une interface graphique.
- **Bibliothèque numérique** : Un ensemble organisé de documents nativement numériques ou numérisés accessibles à distance par le biais d'Internet.
- **Blog** : Type de site web articulé autour de publications régulières et datées généralement personnelles et/ou informelles sur un sujet donné. Son fonctionnement est similaire à celui des journaux intimes ou carnets de bord.
- **Chaîne de traitement** : Un ensemble organisé d'opérations de transformation en vue d'atteindre un objectif.
- **CMS** : *Content Management System* - Une plate-forme de déploiement de sites web, permettant de gérer les contenus mis en ligne en passant par une interface graphique.
- **Exposition virtuelle** : Une exposition diffusée sur Internet, généralement sous la forme d'un site internet dédié.
- **Fork, fork** : Copie d'un projet informatique en vue d'un développement différent ; permet de conserver la parentalité du code source et les informations de licence s'y rattachant.
- **Format** : Manière normalisée de représenter des données ou des fichiers sous la forme d'informations binaires.
- **Gestionnaire de paquets** : Un outil permettant d'automatiser l'installation, la mise à jour ou la désinstallation de logiciels ou de *packages* dans un environnement informatique.
- **HTTP** : *HyperText Transfer Protocol* - Protocole pour le transfert d'informations sur le web.

- **Interface-graphique** : Souvent par opposition aux interfaces en lignes de commande, un dispositif visuel et symbolique permettant l’interaction entre l’humain et la machine, souvent accompagné d’un pointeur de type souris.
- **JSON** : *JavaScript Object Notation* - Format de données textuelles structurées dérivé de la notation des objets du langage JavaScript.
- **Logiciel libre** : Par opposition aux logiciels propriétaires, un logiciel dont l’utilisation, la copie et la modification sont permises légalement.
- **METS** : *Metadata Encoding Transmission Schema* - Standard XML permettant de conserver les métadonnées et la structure hiérarchique d’objets faisant partie d’une collection numérique, ainsi que les liens vers ces objets. Développé par la *Digital Library Federation*.
- **Module** : En Python, un fichier pouvant contenir des fonctions, des classes et des données, et pouvant être importé dans un script.
- **Open-source** : Un mouvement qui vise à garantir la possibilité de distribuer librement des logiciels, d’accéder à leur code source et de créer des logiciels dérivés de ces codes sources. Ces ouverture est régie par divers types de licences.
- **Package, library** : Un ensemble de modules contenant des outils tels que des fonctions. Pour être utilisé, il doit être importé entièrement ou partiellement, par module.
- **PAGE** : *Page Analysis and Ground truth Elements* - Standard XML permettant de stocker la description et la transcription de fichiers (*ground truth*) de fichiers transcrits. Développé par le laboratoire PRIma (*Pattern Recognition & Image Analysis*) de l’université de Salford à Manchester.
- **Parser** : Processus d’analyse d’un élément textuel le rendant intelligible par la machine, sous la forme d’un encodage numérique.
- **Pixel** : Unité permettant de mesurer la définition d’une image numérique matricielle.
- **Pseudo-classe** : En CSS, un mot-clé ajouté à un sélecteur afin d’indiquer l’état spécifique dans lequel l’élément doit être pour être ciblé par la déclaration.
- **Python** : Langage de programmation informatique à usage général, multi-plateforme et *open-source*.
- **Script** : Un programme ou extrait de programme informatique dont l’exécution conduit à la réalisation d’une ou plusieurs actions définies dans le programme.
- **Standard** : Un texte de référence reconnu, documenté et élaboré par un groupe de travail spécialisé, visant à harmoniser l’activité d’un secteur donné. Pour XML, les standards prennent la forme de schémas et de règles de balisage permettant de créer des documents de structures comparables au sein d’un même standard.

- **TEI** : *Text Encoding Initiative* - Standard de description de documents textuels pour XML. Développé par le TEI Consortium.
- **Repository** : Aussi appelé « dépôt informatique » ; un espace organisé de stockage de fichiers.
- **Résolution** : Mesure, généralement exprimée en « dpi », exprimant la finesse d'une image numérique matricielle à partir de la comparaison de ses dimensions en pixels avec sa taille d'impression (en centimètre, pouces, etc).
- **Wiki** : Application web dont le contenu peut être édité par les visiteurs, ce qui permet la création et la modification des pages de manière collaborative. Il est généralement dédié à un projet ou à une thématique précise.
- **Work package** : Un sous-ensemble cohérent de tâches et d'objectifs au sein d'un projet, dont l'exécution peut être attribuée à un ou plusieurs acteurs désigné(s).
- **XML** : *eXtensible Markup Language* - Un langage de balisage générique permettant de décrire des informations de manière organisée et standardisée.
- **XSLT** : *Extensible Stylesheet Language Transformations* - Un langage basé sur XML permettant de styliser ou transformer des fichiers XML ou HTML.

Table des matières

Résumé	iii
Remerciements	v
Liste des sigles et abréviations	vii
Table des figures	ix
Introduction	3
I Un projet ambitieux	7
1 Un contexte pluri-institutionnel	11
1.1 Inria et Almanach : soutiens du projet	11
1.1.1 Présentation générale d’Inria	11
1.1.2 ALMAcH	12
1.2 ANR Time US : le projet de recherche	14
1.2.1 Les équipes de Time Us	14
1.2.2 Un projet de trois ans en plusieurs étapes	16
1.2.3 Les principaux outils numériques du projet	18
2 Les sources du projet	25
2.1 Typologie des sources	25
2.1.1 Archives des Conseils de Prud’hommes	25
2.1.2 Archives de police et de préfecture	27
2.1.3 Archives privées d’entreprise	27
2.1.4 Tarifs	28
2.1.5 Presse ouvrière	29
2.1.6 Enquêtes sociologiques	29
2.2 Collecte et numérisation des sources	30
2.2.1 Résultats des premières campagnes de numérisation	31

2.2.2 Collecte de documents en ligne	32
2.2.3 Stockage des documents numériques	33
II Des images aux fichiers TEI	37
3 Transcrire	41
3.1 La reconnaissance automatique de caractères.	41
3.1.1 Définition	41
3.1.2 Le cas des textes manuscrits	42
3.2 Méthodologie	44
3.2.1 Importer les fichiers dans Transkribus	45
3.2.2 Créer un modèle de reconnaissance automatique	47
3.3 La structure logique des fichiers numériques	50
3.3.1 Le modèle Transkribus	50
3.3.2 Un modèle sur mesure pour Le Play	52
4 Annoter	55
4.1 Quelle annotation pour le projet Time Us ?	55
4.1.1 Définition	55
4.1.2 Annotation manuelle et annotation automatique	56
4.2 Description du processus de travail	57
4.2.1 Présentation des données ciblées	58
4.2.2 Une annotation qui pose des difficultés	59
4.3 Outils et stratégies pour l'annotation	60
4.3.1 Oxygen XML Editor pour un encodage manuel	60
4.3.2 Transkribus : annoter dans une interface graphique	61
4.4 Le guide d'annotation	62
4.4.1 Analyse de l'annotation de « AD69 9M5 »	62
4.4.2 Conclusion de l'enquête auprès des historien·nes	64
4.4.3 Élaboration des règles et des <i>tags</i>	65
4.4.4 La rédaction du guide d'annotation	68
4.4.5 Garantir l'intégrité du jeu de <i>tags</i> sur tous les postes	70
4.4.6 Un guide toujours en phase d'élaboration	72
5 Traiter les données	75
5.1 Un export TEI insatisfaisant	75
5.1.1 Un traitement insuffisant des <i>tags</i> personnalisés	75
5.1.2 Une mauvaise gestion des <i>tags</i> étendues sur plusieurs lignes	76
5.1.3 Des métadonnées pauvres	77

5.1.4	Une manipulation lente	78
5.2	L'API Transkribus pour contourner ces problèmes	79
5.2.1	Prendre en main les requêtes HTTP	79
5.2.2	Les requêtes de l'API de Transkribus	81
5.2.3	Un <i>script</i> Python pour extraire les fichiers	85
5.2.4	Une feuille de style pour obtenir la TEI : page2tei.xsl	90
5.2.5	Un développement à poursuivre	96
5.3	Une interface de consultation ?	98

Conclusion 107

Annexes 113

A Archives 115

A.1	Exemples	115
A.1.1	Extraits du corpus d'images	115
A.1.2	Exemples de prises de vue problématiques	118
A.1.3	Enquêtes sociologiques de Le Play sur les métiers du textile	120
A.1.4	Exemple de titre de presse ouvrière	120
A.1.5	Liste des titres de presse ouvrière lyonnaise recherchés	120
A.2	Éléments d'analyse	122
A.2.1	Analyse de la structure des monographies de Le Play	122
A.2.2	Tableau comparatif des métadonnées des échantillons du corpus d'images	123

B Transkribus 125

B.1	Interface graphique	125
B.1.1	Capture d'écran de la page d'accueil de Transkribus, onglet « Server »	126
B.1.2	Capture d'écran avec l'onglet « Overview » développé	126
B.1.3	Capture d'écran avec l'onglet « Layout » développé	127
B.1.4	Capture d'écran avec les onglets « Metadata » et « document » développés	127
B.1.5	Capture d'écran avec les onglets « Metadata » et « structural » développés	128
B.1.6	Capture d'écran avec les onglets « Metadata » et « textual » développés	128
B.1.7	Capture d'écran de la fenêtre de personnalisation des <i>tags</i>	129
B.1.8	Capture d'écran avec les onglets « Metadata » et « comments » développés	129

B.1.9	Capture d'écran avec l'onglet « Tools » développé	130
B.1.10	Capture d'écran de la fenêtre de présentation des modèles d'HTR disponibles	130
B.1.11	Capture d'écran de la fenêtre d'export, pour le format « Transkribus Document »	131
B.1.12	Capture d'écran de la fenêtre d'export, pour le format TEI	131
B.2	API	132
B.2.1	Requêtes	132
B.2.2	<i>Script</i> pour l'API Transkribus	132
C	Outils techniques	133
C.1	Sharedocs	133
C.1.1	Capture d'écran du dossier ShareDocs pour Time Us	133
C.1.2	Capture d'écran de la configuration de l'outil d'OCR	134
C.2	Postman	135
C.2.1	Capture d'écran de la page de création d'une requête, sur l'interface locale	135
C.2.2	Capture d'écran de la page de création d'une requête remplie, sur l'interface locale	136
C.2.3	Capture d'écran de la page de consultation de la documentation d'une requête, sur l'interface en ligne	136
D	Annotation	137
D.1	Éléments d'analyse	137
D.1.1	Analyse des pages 3 à 17 du document « AD69 9M5 »	137
D.1.2	Note interne sur le fonctionnement des <i>tags</i> dans Transkribus et sur l'annotation pour Time Us	137
D.2	Guides	137
D.3	Fichiers	138
D.3.1	Configuration locale de Transkribus	138
D.3.2	TEITags	138
E	Modélisation des schémas XML	139
E.1	Modélisation du schéma XML-PAGE original	140
E.2	Modélisation du schéma XML-PAGE adapté pour Time Us	141
E.3	Modélisation des fichiers XML-TEI du projet Time Us	142
F	Transformation XSLT	143
F.1	Modification des <i>tags</i>	143
F.2	page2tei	143

<i>Glossaire</i>	191
G TEI Boiler Plate	145
G.1 Environnement TEI-Boilerplate	145
G.2 Fichiers préparatoires	145
H Documents	147
H.1 Digital Humanities 2018	147
H.2 Réunion du 1 ^{er} juin 2018	147
Bibliographie	151
Glossaire	183
Table des matières	187