

ÉCOLE NATIONALE DES CHARTES
UNIVERSITÉ PARIS, SCIENCES & LETTRES

Maxime HUMEAU

licencié ès lettres

master ès lettres

Ocupación de la Araucanía

Mise en place d'une chaîne de traitement de
la transcription à l'édition numérique
d'archives manuscrites espagnoles du XIX^e
siècle

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

2022

Résumé

Ce présent mémoire rend compte du stage effectué entre avril et juillet 2022 au centre *Archivo Central Andrés Bello - Universidad de Chile* à Santiago du Chili pour le traitement éditorial des sources autour de l'« Occupation de l'Araucanie » (1850-1881). Il a conduit à la conception d'une chaîne de traitement de documents numérisés vers l'encodage TEI p5, décrit par une documentation ODD. Pour se faire, un modèle HTR a été produit afin de reproduire le contenu et la mise en page des documents numérisés *via* le moteur OCR Kraken.

Dans un second temps, le stage a donné lieu au développement d'un modèle de reconnaissance d'entités nommées sur la base du modèle BETO. Cette utilisation des techniques du traitement automatique du langage a eu pour effet de permettre l'indexation des personnes, lieux, organisations et dates présentes au sein des fichiers XML-TEI. Enfin, ces entités ont fait le fruit d'un processus d'enrichissement automatique à partir des bases de données Wikidata.

Mots-clés : Humanités numériques ; Histoire ; Chili ; Araucania ; Archives politiques et militaires ; Histoire contemporaine ; XIX^e siècle ; Data engineering ; XML-TEI ; HTR ; Kraken ; Traitement automatique du langage ; Reconnaissance des entités nommées ; Apprentissage automatique ; Édition numérique ; Universidad de Chile ; Histoire coloniale ; Mapuches ; Web sémantique.

Informations bibliographiques : Maxime Humeau, « *Ocupación de la Araucanía* ». *Mise en place d'une chaîne de traitement de la transcription à l'édition numérique d'archives manuscrites espagnoles du XIX^e siècle*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Thibault Clérice et Ariane Pinche, École nationale des chartes, 2022.

Remerciements

Mes souhaits vont tout d'abord à mon tuteur de stage M. Alessandro Chiaretti et l'ensemble de l'équipe du centre Archivo Central Andrés Bello pour l'ensemble de leurs aides, leurs accompagnements et leur accueil extrêmement chaleureux et bienveillant. Découvrir ce pays et sa culture à leurs côtés fut d'un très grand plaisir.

Je tiens à remercier Thibault Clérice et Ariane Pinche, mes directeurs de mémoire, pour leurs suivies, leurs recommandations précieuses tout au long de stage. Ces remerciements s'étendent à l'ensemble de leur travail sur cette année pour une volonté sans faille pour transmettre leurs savoirs-faires et leurs soutiens infaillibles envers l'ensemble de la promotion.

Je remercie l'ensemble de l'équipe pédagogique du master TNAH et le personnel de l'École nationale des Chartes sans qui cette formation ne pourrait aboutir. Cette année fut celle d'un très grand enrichissement et signé d'une très grande curiosité intellectuelle et technique.

Je tiens à remercier l'ensemble de la promotion pour le partage de cette année sous le sigle d'une aventure numérique, historique et de colonnades.

Je remercie spécialement Antoine Lauer, Laura Catrou, Grégoire Hör et Léo Ludovic pour leurs aides et leurs relectures précieuses.

Je souhaite remercier l'ensemble des personnes qui ont su m'accueillir avec une extrême générosité tout au long de ce périple à l'autre bout du monde, et plus particulièrement à toi Josefa sans qui rien ne serait possible.

Enfin, je remercie ma famille et mes proches pour leurs soutiens et leurs encouragements indéfectibles sur l'ensemble de mon histoire universitaire.

Bibliographie

Études sociopolitiques autour du Chili

- ACEVEDO (Fernández) et J (Fernando), « El Documento Electrónico En El Derecho Civil Chileno : Análisis de La Ley 19.799 », *Ius et Praxis*, 10–2 (2004), p. 137-167, DOI : 10.4067/S0718-00122004000200005.
- ANDRADE (María José), « La lucha por el territorio mapuche en Chile : una cuestión de pobreza y medio ambiente », *L'Ordinaire des Amériques*–225 (225[2019]), DOI : 10.4000/orda.5132.
- ARRUÉ (Michèle), « Les Mapuches du Chili et la question de leur identité », *Amérique Latine Histoire et Mémoire. Les Cahiers ALHIM. Les Cahiers ALHIM*–10 (10[2004]), DOI : 10.4000/alhim.123.
- BARBUT (Michael), « « Qui sont les terroristes ? » Lutte de classement autour de la radicalité mapuche », *Lien social et Politiques*–68 (2012), p. 79-100, DOI : 10.7202/1014806ar.
- BENGOA (José), « Los Mapuches : historia, cultura y conflicto », *Cahiers des Amériques latines*, 2011/3–68 (68[2011]), p. 89-107, DOI : 10.4000/cal.118.
- CARVAJAL-DEL MAR (Zunilda), « La criminalisation du conflit Mapuche : l'application discriminatoire de la loi antiterroriste chilienne », *Archives de politique criminelle*, 36–1 (2014), p. 213-226, DOI : 10.3917/apc.036.0213.
- CHIARETTI (Alessandro) et BILBAO (Claudio Ogass), *Transformar El Archivo En Un Archivo. Retos y Desafíos En El Archivo Central Andrés Bello*, Santiago, Chili, Comité Nacional de la Memoria del Mundo, 2016, URL : https://www.academia.edu/32354445/Transformar_el_Archivo_en_un_Archivo_Retos_y_desaf%C3%ADos_en_el_Archivo_Central_Andr%C3%A9s_Bello (visité le 16/08/2022).
- DELFAU (Antonio) et KAUFMANN SALINAS (Sebastián), « Le Chili, un pays en transformation », *Études*, avril–4 (2014), p. 19-28, DOI : 10.3917/etu.4204.0019.
- FABRY (Pierre), *Archives, Archivistes et Crise Au Chili, Quelques Réflexions*, Ecole des chartes, janv. 2020, URL : https://ecoledeschartes.tumblr.com/post/190494197307/archives-archivistes-et-crise-au-chili-quelques?is_related_post=1 (visité le 13/09/2022).

- GROOPPO (Bruno), « Chapitre v – Les archives des droits humains. Documenter la répression et la résistance au Chili et en Argentine », dans *Documenter les violences : Usages publics du passé dans la justice transitionnelle*, dir. Camille Goirand et Angélica Müller, Paris, 2020 (Travaux et mémoires), p. 131-149, URL : <http://books.openedition.org/iheal/8887> (visité le 13/09/2022).
- HERNÁNDEZ (Salvador Millaleo), « Los pueblos originarios ante el horizonte de una nueva constitución », *Anales de la Universidad de Chile*–13 (13[2017]), p. 241-259, DOI : 10.5354/0717-8883.2017.49005.
- HUENCHO (Verónica Figueroa), « Pueblos indígenas y derechos : una discusión a la luz de las políticas públicas desde el caso chileno », *Anales de la Universidad de Chile*–13 (13[2017]), p. 97-114, DOI : 10.5354/0717-8883.2017.48999.
- LACOSTE (Pablo Segovia), « La construction discursive de l'événement conflit mapuche dans la presse écrite chilienne », *Synergies Chili*–12 (2016), p. 73-87.
- LE BONNIEC (Fabien), « La culture mapuche à la barre : pouvoir et médiation linguistico-culturels des facilitateurs interculturels dans les tribunaux pénaux du sud du Chili », *Autrepart*, 73–1 (2015), p. 55-71, DOI : 10.3917/autr.073.0055.
- N.C., *Estado del arte nacional e internacional en materia de gestión de datos de investigación e información científica y tecnológica y recomendaciones de buenas prácticas*, Santiago, Chili, Comisión Nacional de Investigación Científica y Tecnológica, 2010, p. 108, URL : <http://datoscientificos.cl/files/ufro-2010.pdf>.
- *Manual de Datos Abiertos*, Santiago, Chili, Comisión Nacional de Investigación Científica y Tecnológica, 2014, URL : <http://datoscientificos.cl/files/manual-2014.pdf>.
- OBREGÓN ITURRA (Jimena Paz), « Enjeux conflictuels liés à la terre et au territoire en pays mapuche », *Nuevo Mundo Mundos Nuevos. Nouveaux mondes mondes nouveaux - Novo Mundo Mundos Novos - New world New worlds* (, 8 oct. 2020), DOI : 10.4000/nuevomundo.82647.
- SEPÚLVEDA (Bastien), « Le pays mapuche, un territoire « à géographie variable » », *Espace populations sociétés. Space populations societies*–2012/1 (2012/1[2012]), p. 73-88, DOI : 10.4000/eps.4872.

Histoire et historiographie des relations Chileno-Mapuche

- BENGOA (José), *Historia Del Pueblo Mapuche (Siglo XIX y XX)*, Ediciones Sur, Santiago, Chili, 1987.
- *La memoria olvidada : historia de los pueblos indígenas de Chile*, Cuadernos Bicentenario, Presidencia de la República, Santiago, Chili, 2004.

- BOCCARA (Guillaume), « Organisation sociale, guerre de captation et ethnogenèse chez les Reche-Mapuche à l'époque coloniale », *Homme*, 39–150 (1999), p. 85-117, DOI : 10.3406/hom.1999.453568.
- CANALES TAPIA (Pedro) et PINTO RODRÍGUEZ (Jorge), « Historiografía Mapuche : balances y perspectivas de discusión en el Chile reciente », *Izquierdas*–24 (24[2015]), URL : <https://journals.openedition.org/izquierdas/358?lang=fr> (visité le 15/09/2022).
- CARMEN RAMALLO DÍAZ (DEL) (Cecilia), *Transcripción de Los Documentos de La Serie “Pacificación de La Araucanía” (Colección Manuscritos)*, Santiago, Chili, Archivo Central Andrés Bello, 2014, p. 30.
- GROOPPO (Bruno), « Chapitre v – Les archives des droits humains. Documenter la répression et la résistance au Chili et en Argentine », dans *Documenter les violences : Usages publics du passé dans la justice transitionnelle*, dir. Camille Goirand et Angélica Müller, Paris, 2020 (Travaux et mémoires), p. 131-149, URL : <http://books.openedition.org/iheal/8887> (visité le 13/09/2022).
- LACOSTE (Pablo Segovia), « La construction discursive de l'événement conflit mapuche dans la presse écrite chilienne », *Synergies Chili*–12 (2016), p. 73-87.
- NOUAILLE (Thierry), « L'indépendance du Chili : les conséquences sur le peuple mapuche (1810-2010) », *América. Cahiers du CRICCAL*–42 (42[2012]), p. 131-143, DOI : 10.4000/americam.1111.
- QUEMENADO (Pablo Mariman), « La geoestrategia en el conflicto chileno mapuche : la configuración del Estado Nación (1830-1869) », *Anales de la Universidad de Chile*–13 (13[2017]), p. 39-57, DOI : 10.5354/0717-8883.2017.48995.
- SOLANO ASTA-BURUAGA (Francisco), *Diccionario Geográfico de la República de Chile*, Segunda edición corregida y aumentada, Santiago, Chili, 1899, URL : https://es.wikisource.org/wiki/Diccionario_Geogr%C3%A1fico_de_la_Rep%C3%BAblica_de_Chile (visité le 31/08/2022).

Humanités numériques et science ouverte

- BARDIOT (Clarisso), *Happy APIs : Débridons les APIS pour développer les humanités numériques*, DORRA-DH, URL : <https://dorradh.hypotheses.org/66> (visité le 18/07/2022).
- BOULLIER (Dominique), *Sociologie du numérique*, Paris, 2016.
- BOURGEOIS (Nicolas), PELLET (Aurélien) et PUREN (Marie), « Using Topic Generation Model to Explore the French Parliamentary Debates during the Early Third Republic (1881-1899) », dans *DiPaDA 2022 Digital Parliamentary Data in Action 2022. Proceedings of the Digital Parliamentary Data in Action (DiPaDA 2022) Workshop Co-Located with 6th Digital Humanities in the Nordic and Baltic Countries Conference*

- rence (*DHNB 2022*), dir. Matti La Mela, Fredrik Norén et Eero Hyvönen, 2022 (CEUR Workshop Proceedings), t. 3133, p. 35-51, URL : <https://hal.archives-ouvertes.fr/hal-03526254> (visité le 07/09/2022).
- COMMUNICATION SCIENTIFIQUE DIRECTE (Centre pour la), *Principes FAIR*, CCSD | Centre pour la Communication Scientifique Directe, URL : <https://www.ccsd.cnrs.fr/principes-fair/> (visité le 05/09/2022).
- EDER (Maciej), « Mind Your Corpus : Systematic Errors in Authorship Attribution », *Literary and Linguistic Computing*, 28–4 (1^{er} déc. 2013), p. 603-614, DOI : 10.1093/linc/fqt039.
- FOUCAULT (Michel) et LAGRANGE (Jacques), *Dits et écrits, 1954-1988*, dir. Daniel Defert et François Ewald, Paris, France, impr. 1994.
- GAYOL (Víctor) et MELO FLÓREZ (Jairo Antonio), « Presente y perspectivas de las humanidades digitales en América Latina », *Mélanges de la Casa de Velázquez. Nouvelle série*, 2-47-2 (47[2017]), p. 281-284, DOI : 10.4000/mcv.7907.
- GEFEN (Alexandre), « Les enjeux épistémologiques des humanités numériques », *Socio. La nouvelle revue des sciences sociales*-4 (4[2015]), p. 61-74, DOI : 10.4000/socio.1296.
- JACQUEMIN (Bernard), SCHÖPFEL (Joachim) et FABRE (Renaud), « Libre accès et données de recherche. De l'utopie à l'idéal réaliste », *Études de communication. langages, information, médiations*-52 (52[2019]), p. 11-26, DOI : 10.4000/edc.8468.
- MCGILLIVRAY (Barbara) et POIBEAU (Thierry), « Digital Humanities and Natural Language Processing : “Je t'aime... Moi Non Plus” », dans 2020, t. 14–2, p. 11.
- MONCLA (Ludovic), GAIO (Mauro), EGOROVA (Ekaterina) et CLARAMUNT (Christophe), « An Automatic Extraction Method of Static and Dynamic Spatial Contexts from Texts », dans *Atelier Science Des Données et Humanités Numériques (SDHN)*, Conférence Internationale Francophone Sur l'Extraction et La Gestion de Connaissance (EGC 2018), Paris, France, 2018, URL : <https://hal.archives-ouvertes.fr/hal-01695643> (visité le 02/09/2022).
- MOUNIER (Pierre), *Les humanités numériques : Une histoire critique*, Paris, 2018 (Interventions), URL : <http://books.openedition.org/editionsmsmsh/12006> (visité le 16/08/2022).
- OGILVIE (Denise), « Paradoxes de « l'archive » », *Sociétés & Représentaions*, 43–1 (2017), p. 121-134, DOI : 10.3917/sr.043.0121.
- Open Science in Latin America and the Caribbean : A Strong Tradition with a Long Journey Ahead*, UNESCO, 2020, URL : <https://en.unesco.org/news/open-science-latin-america-and-caribbean-strong-tradition-long-journey-ahead> (visité le 05/09/2022).

- PUREN (Marie), *La Lecture Distante : Introduction et Exemples d'application*, France, nov. 2020, URL : <https://hal.archives-ouvertes.fr/hal-03152747> (visité le 07/09/2022).
- RUSSELL (Isabel Galina), « Geographical and Linguistic Diversity in the Digital Humanities », *Literary and Linguistic Computing*, 3–29 (sept. 2014), p. 307-316, DOI : 10.1093/linc/fqu005.
- SFORZINI (Arianna), *Michel Foucault numérique. Implications philosophiques*, 16 juill. 2021, URL : <https://www.implications-philosophiques.org/michel-foucault-numerique/> (visité le 16/08/2022).
- UNDP (éd.), *The next Frontier : Human Development and the Anthropocene*, New York, Etats-Unis, 2020 (Human Development Report, 2020).
- WILKINSON Mark D., DUMONTIER Michel, AALBERSBERG IJsbrand Jan, APPLETON Gabrielle, AXTON Myles, BAAK Arie, BLOMBERG Niklas, BOITEN Jan-Willem, DA SILVA SANTOS Luiz Bonino, BOURNE Philip E., *et al.*, « The FAIR Guiding Principles for Scientific Data Management and Stewardship », *Scientific Data*, 3–1 (1[2016]), p. 160018, DOI : 10.1038/sdata.2016.18.
- WISSIK Tanja, EDMOND Jennifer, FISCHER Frank, DE JONG Franciska, SCAGLIOLA Stefania, SCHARNHORST Andrea, SCHMEER Hendrik, SCHOLGER Walter et WESSELS Leon, *Teaching Digital Humanities Around the World : An Infrastructural Approach to a Community-Driven DH Course Registry*, mars 2020, URL : <https://hal.archives-ouvertes.fr/hal-02500871> (visité le 17/08/2022).

Éditions numériques

- ÁLVAREZ-MELLADO (Elena), DÍEZ-PLATAS (María Luisa), RUIZ-FABO (Pablo), BERMÚDEZ (Helena), Ros (Salvador) et GONZÁLEZ-BLANCO (Elena), « TEI-friendly Annotation Scheme for Medieval Named Entities : A Case on a Spanish Medieval Corpus », *Language Resources and Evaluation*, 55–2 (1^{er} juin 2021), p. 525-549, DOI : 10.1007/s10579-020-09516-2.
- BISSON (Marie) et GOLOUBKOFF (Anne), « Les notices d'autorité en XML-TEI : un outil pour l'accroissement collaboratif de connaissances et l'indexation d'éditions de sources », *Tabularia. Sources écrites des mondes normands médiévaux* (, 3 mars 2020), DOI : 10.4000/tabularia.4176.
- BURNARD (Lou), *Qu'est-ce que la Text Encoding Initiative ?*, trad. par Marjorie Burghart, Marseille, 2015 (Encyclopédie numérique), URL : <http://books.openedition.org/oep/1237> (visité le 02/09/2022).
- CAMPS (Jean-Baptiste), « La TEI, Une Communauté de Pratiques », dans *Édition Électronique et TEI : Enjeux, Pratiques et Perspectives - _dayClic()*, Le Mans, 2017.

CAMPS (Jean-Baptiste), « Où va la philologie numérique ? », *Fabula-LhT-20* (29 janv. 2018), URL : <https://www.fabula.org:443/lht/20/camps.html> (visité le 20/08/2022).

CASENAVE (Joana), « Le positionnement éditorial dans l'édition critique numérique », *Digital Studies / Le champ numérique*, 9–1 (1[2019]), DOI : 10.16995/dscn.348.

Étienne Cavalié, et al. (éd.), *Expérimenter les humanités numériques. Des outils individuels aux projets collectifs*, Nouvelle édition [en ligne], Montréal, 2018 (Parcours numérique), URL : <http://books.openedition.org/pum/11091> (visité le 02/09/2022).

CHAGUÉ (Alix), FOURNER (Victoria Le) et MARTINI (Manuela), « Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ? », dans *Colloque DHNord 2019 "Corpus et archives numériques"*, MESHS Lille Nord de France, 2019, URL : <https://hal.inria.fr/view/index/identifiant/hal-02448921>.

CHAGUÉ (Alix), SCHEITHAUER (Hugo), TERRIEL (Lucas), CHIFFOLEAU (Floriane) et TADJO-TAKIANPI (Yves), « Take a Sip of TEI and Relax : A Proposition for an End-to-End Workflow to Enrich and Publish Data Created with Automatic Text Recognition », dans *Digital Humanities 2022 : Responding to Asian Diversity*, Tokyo, Japan, 2022, URL : <https://hal.inria.fr/hal-03739767> (visité le 19/08/2022).

CHIFFOLEAU (Floriane), *DAHN Project : Digital Edition of Historical Manuscripts*, 22 juin 2022, URL : <https://github.com/FloChiff/DAHNProject/blob/e19fbc38476305a941ff7f5a6d> Correspondence / Guidelines / Documentation – Correspondance . pdf (visité le 22/08/2022).

CHRISTENSEN (Kelly), *Alto2tei*, Gallic(orpor)a, 2022, URL : <https://github.com/katkel/alto2tei> (visité le 03/09/2022).

CHRISTENSEN (Kelly), PINCHE (Ariane) et GABAY (Simon), « Gallic(Orpor)a : Traitement Des Sources Textuelles En Diachronie Longue de Gallica », dans *DataLab de La BnF*, Paris, France, 2022, URL : <https://hal.archives-ouvertes.fr/hal-03716534> (visité le 21/08/2022).

CLAVAUD (Florence), « Vers l'édition En Ligne Des Testaments de Poilus », dans *Les Éditions Savantes Numériques : Enjeux et Réalisation*, Lille, France, 2019, URL : <https://hal.archives-ouvertes.fr/hal-02469768> (visité le 17/08/2022).

CLÉRICE (Thibault), « Les outils CapiTainS, l'édition numérique et l'exploitation des textes », *Médiévales. Langues, Textes, Histoire*, 73–73 (73[2017]), p. 115-131, DOI : 10.4000/medievales.8211.

CORBIÈRES (Caroline), JOYEUX-PRUNEL (DIR.) (Béatrice) et CLÉRICE (DIR.) (Thibault), *Du Catalogue Au Fichier TEI : Création d'un Workflow Pour Encoder Automatiquement En Xml-Tei Des Catalogues d'exposition*, mémoire pour le diplôme

- de master « Technologies numériques appliquées à l'histoire », dir. Thibault Clérice et Béatrice Joyeux-Prunel, Paris, France, Ecole nationale des chartes, 2020.
- DUVAL (Frédéric), « Pour des éditions numériques critiques. L'exemple des textes français », *Médiévales. Langues, Textes, Histoire*–73 (73[2017]), p. 13-29, DOI : 10 . 4000/medievales.8165.
- GABAY (Simon), CAMPS (Jean-Baptiste), PINCHE (Ariane) et CARBONI (Nicola), *SegmOnto, A Controlled Vocabulary to Describe the Layout of Pages*, version 0.9, Ecole Nationale des Chartes, Université de Genève, 2021, URL : <https://github.com/SegmOnto> (visité le 22/04/2022).
- « SegmOnto : Common Vocabulary and Practices for Analysing the Layout of Manuscripts (and More) », dans *1st International Workshop on Computational Palaeography (IWCP@ICDAR 2021)*, Lausanne, Suisse, 2021, URL : <https://hal.archives-ouvertes.fr/hal-03336528> (visité le 22/08/2022).
- JANES (Juliette), PINCHE (Ariane), JAHAN (Claire) et GABAY (Simon), « Towards Automatic TEI Encoding via Layout Analysis », dans *Fantastic Future 21, 3rd International Conference on Artificial Intelligence for Librairies, Archives and Museums*, Paris, France, 2021, URL : <https://hal.archives-ouvertes.fr/hal-03527287> (visité le 21/08/2022).
- JANÈS (Juliette), *Du catalogue papier au numérique : Une chaîne de traitement ouverte pour l'extraction d'informations issues de documents structurés*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Thibault Clérice et Simon Gabay, Paris, France, École nationale des chartes, 2021, URL : https://github.com/Juliettejns/Memoire_TNAH.
- LE PEVEDIC (Solenn) et MAUREL (Denis), « Retour sur les annotations des entités nommées dans les campagnes d'évaluation françaises et comparaison avec la TEI », *Corela. Cognition, représentation, langage*, 2–14-2 (14[2016]), DOI : 10 . 4000/corela.4644.
- MAYEUR (Ingrid) et PAVEAU (Marie-Anne), « Présentation. Les devenirs du texte numérique natif », *Corela. Cognition, représentation, langage*–HS-33 (HS-33[2020]), DOI : 10 . 4000/corela.11749.
- MUÑOZ (Trevor) et VIGLIANTI (Raffaele), « Texts and Documents : New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive », *Journal of the Text Encoding Initiative*–Issue 8 (8[2014]), DOI : 10 . 4000/jtei.1270.
- Names Sell : Named Entity Recognition in TEI Publisher – e-Editiones*, URL : <https://e-editiones.org/names-sell-named-entity-recognition-in-tei-publisher/> (visité le 06/08/2022).
- PALOQUE-BERGES (Camille), « Les sources nativement numériques pour les sciences humaines et sociales », *Histoire@Politique*, 30–3 (2016), p. 221-244, DOI : 10 . 3917/hp . 030 . 0221.

- PINCHE (Ariane), « Faire Une Édition Nativement Numérique : Un Nouveau Rapport Au Texte ? », dans *Séminaire Du Laboratoire ÉRIC*, Lyon, France, 2018, URL : <https://halshs.archives-ouvertes.fr/halshs-01801316> (visité le 21/08/2022).
- PINCHE (Ariane), GABAY (Simon) et CHRISTENSEN (Kelly), « SegmOnto – A Control-led Vocabulary to Describe Historical Textual Sources », dans *Documents Anciens et Reconnaissance Automatique Des Écritures manuscrites/Documents Anciens et Reconnaissance Automatique Des Écritures Manuscrites*, Paris, France, 2022.
- RONDEAU DU NOYER (Lucie), *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la revue des autographes de la librairie Charavay*, mémoire pour le diplôme de master « Technologies numériques appliquées à l'histoire », dir. Thibault Clérice et Simon Gabay, Paris, France, Ecole nationale des chartes, 2019, URL : https://github.com/lairaines/M2TNAH/blob/429079b2d9723dc85a3e8ba07f84c0b7e18717d1/RondeauduNoyer_M2TNAH.pdf (visité le 01/09/2022).
- TEI Publisher*, version 7.1.0, e-editiones, 3 juill. 2022, URL : <https://github.com/eeditiones/tei-publisher-app> (visité le 02/09/2022).
- TOSELLI (Alejandro H.), WU (Si) et SMITH (David A.), « Digital Editions as Distant Supervision for Layout Analysis of Printed Books », dans *International Conference on Document Analysis and Recognition*. Springer, Lausanne, Suisse, 2021, t. 12822, p. 462-476, DOI : 10.1007/978-3-030-86331-9_30, arXiv : 2112.12703 [cs].
- VERTONGEN (Caroline Blanc Feracci et Marthe), *Qu'est-ce qu'un projet d'édition critique numérique ?*, DLIS, 1^{er} mars 2022, URL : <https://dlis.hypotheses.org/5790> (visité le 20/08/2022).

Généralités autour de l'apprentissage Machine

- GARDNER (M. W) et DORLING (S. R), « Artificial Neural Networks (the Multilayer Perceptron)—a Review of Applications in the Atmospheric Sciences », *Atmospheric Environment*, 32–14 (1^{er} août 1998), p. 2627-2636, DOI : 10.1016/S1352-2310(97)00447-0.
- MCCULLOCH (Warren S.) et PITTS (Walter), « A Logical Calculus of the Ideas Immanent in Nervous Activity », *The bulletin of mathematical biophysics*, 5–4 (1^{er} déc. 1943), p. 115-133, DOI : 10.1007/BF02478259.
- NOBLE (William S), « What Is a Support Vector Machine ? », *Nature Biotechnology*, 24–12 (déc. 2006), p. 1565-1567, DOI : 10.1038/nbt1206-1565.
- RAMCHOUN (Hassan), AMINE (Mohammed), IDRISI (Janati), GHANOU (Youssef) et ETTAOUIL (Mohamed), « Multilayer Perceptron : Architecture Optimization and Training », *International Journal of Interactive Multimedia and Artificial Intelligence*, 4–1 (2016), p. 26, DOI : 10.9781/ijimai.2016.415.

- REFAEILZADEH (Payam), TANG (Lei) et LIU (Huan), « Cross-Validation », dans *Encyclopedia of Database Systems*, dir. LING LIU et M. TAMER ÖZSU, Boston, MA, 2009, p. 532-538, DOI : 10.1007/978-0-387-39940-9_565.
- Scikit-Learn/Scikit-Learn*, version 1.1.2, scikit-learn, 9 sept. 2022, URL : <https://github.com/scikit-learn/scikit-learn> (visité le 09/09/2022).
- TAPPERT (Charles C.), « Who Is the Father of Deep Learning ? », dans *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019, p. 343-348, DOI : 10.1109/CSCI49370.2019.00067.
- TOUGUI (Ilias), JILBAB (Abdelilah) et EL MHAMDI (Jamal), « Impact of the Choice of Cross-Validation Techniques on the Results of Machine Learning-Based Diagnostic Applications », *Healthcare Informatics Research*, 27–3 (juill. 2021), p. 189-199, DOI : 10.4258/hir.2021.27.3.189, pmid : 34384201.
- TURING (Alan), « Computing Machinery and Intelligence », *Mind*, LIX-236 (1^{er} oct. 1950), p. 433-460, DOI : 10.1093/mind/LIX.236.433.

Reconnaissance d'écriture manuscrite

- ALKHALAF (Khalaf), « OCR-Based Electronic Documentation Management System », *International Journal of Innovation, Management and Technology*, 5–6 (2014), DOI : 10.7763/IJIMT.2014.V5.560.
- ANVARI (Zahra) et ATHITSOS (Vassilis), *A Survey on Deep Learning Based Document Image Enhancement*, 3 janv. 2022, arXiv : 2112.02719 [cs], URL : <http://arxiv.org/abs/2112.02719> (visité le 11/05/2022).
- ARADILLAS (José Carlos), MURILLO-FUENTES (Juan José) et OLMOS (Pablo M.), « Boosting Handwriting Text Recognition in Small Databases with Transfer Learning », dans *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2018, p. 429-434, DOI : 10.1109/ICFHR-2018.2018.00081, arXiv : 1804.01527 [cs, stat].
- CAMPS (Jean-Baptiste) et PERREAUX (Nicolas), « Reconnaissance optique des caractères et des écritures manuscrites - Projet E-NDP », dans *Séminaire Notre-Dame de Paris et son cloître*, Paris, France, 2021, URL : https://outils.lamop.fr/lamop/mp3/E-Ndp/JBC-NP_e-NDP_OCR-et-HTR.pdf.
- CARON (Bertrand) et CAVALIÉ (Etienne), *Formats de données pour la préservation à long terme : la politique de la BnF*, 1, Paris, France, Bibliothèque nationale de France, 2021, URL : https://www.bnf.fr/sites/default/files/2021-04/politiqueFormatsDePreservationBNF_20210408.pdf.
- CHAGUÉ (Alix), « Conditions de La Mutualisation : Les Principes FAIR et HTR-United », dans *Humanistica 2022*, Montréal, Canada, 2022, URL : <https://hal.inria.fr/hal-03685731> (visité le 13/09/2022).

- CHAGUÉ (Alix) et CHIFFOLEAU (Floriane), *An Accessible and Transparent Pipeline for Publishing Historical Egodocuments*, mars 2021, URL : <https://hal.archives-ouvertes.fr/hal-03180669> (visité le 19/08/2022).
- CHAGUÉ (Alix) et CLÉRICE (Thibault), *HTR-United - Manu McFrench V1 (Manuscripts of Modern and Contemporaneous French)*, version 1.0.0, Zenodo, 17 juin 2022, URL : <https://zenodo.org/record/6657809> (visité le 27/08/2022).
- CHAGUÉ (Alix), CLÉRICE (Thibault) et ROMARY (Laurent), « HTR-United : Mutualisons La Vérité de Terrain ! », dans *DHNord2021 - Publier, Partager, Réutiliser Les Données de La Recherche : Les Data Papers et Leurs Enjeux*, Lille, France, 2021, URL : <https://hal.archives-ouvertes.fr/hal-03398740> (visité le 17/08/2022).
- CHAGUÉ (Alix) et ROSTAING (Aurélia), « Présentation Du Projet Lectaurep (Lecture Automatique de Répertoires) », dans *Atelier Sur La Transcription Des Écritures Manuscrites - BnF DataLab*, Paris, France, 2021, URL : <https://hal.archives-ouvertes.fr/hal-03122019> (visité le 19/08/2022).
- CHIFFOLEAU (Floriane), *How to Produce a Model for the Segmentation*, Digital Intellectuals, URL : <https://digitalintellectuals.hypotheses.org/3844> (visité le 19/08/2022).
- CHIFFOLEAU (Floriane) et BAILLOT (Anne), *Le Projet DAHN : Une Pipeline Pour l'édition Numérique de Documents d'archives*, avr. 2022, URL : <https://hal.archives-ouvertes.fr/hal-03628094> (visité le 19/08/2022).
- CLÉRICE (Thibault), *You Actually Look Twice At It (YALTAi) : Using an Object Detection Approach Instead of Region Segmentation within the Kraken Engine*, juill. 2022.
- CLÉRICE (Thibault) et CHAGUÉ (Alix), *CREMMA-AN-TestamentDePoilus*, 2022, URL : <https://github.com/HTR-United/CREMMA-AN-TestamentDePoilus>.
- CLÉRICE (Thibault), CHAGUÉ (Alix) et JACSONT (Pauline), *HTR-United/HTRVX : HTRVX : HTR Validation with XSD*, version 0.0.10, HTR-United, mars 2022, URL : <https://github.com/HTR-United/HTRVX> (visité le 18/04/2022).
- DE SOUSA NETO Arthur Flor, BEZERRA Byron Leite Dantas, TOSELLI Alejandro Hector et LIMA Estanislau Baptista, « HTR-Flor : A Deep Learning System for Offline Handwritten Text Recognition », dans *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Recife/Porto de Galinhas, Brazil, 2020, p. 54-61, DOI : [10.1109/SIBGRAPI51738.2020.00016](https://doi.org/10.1109/SIBGRAPI51738.2020.00016).
- DURAND (Marc), ROSTAING (Aurélia) et CHAGUÉ (Alix), *Notaires de Paris - Répertoires, Ground Truth for Various Parisian Registries of Notary Deeds (French 19th and 20th Centuries)*, HTR-United, 2021, URL : <https://github.com/HTR-United/lectaurep-repertoiries> (visité le 23/08/2022).
- ELAGOUNI (Khaoula), GARCIA (Christophe), MAMALET (Franck) et SÉBILLOT (Pascale), « Combining Multi-Scale Character Recognition and Linguistic Knowledge for Natural Scene Text OCR », dans *10th IAPR International Workshop on Document*

- Analysis Systems, DAS*, Gold Coast, Queensland, Australia, 2012, p. 120-124, URL : <https://hal.archives-ouvertes.fr/hal-00753908> (visité le 24/08/2022).
- ESPAÑA-BOQUERA (Salvador) et CASTRO-BLEDA (María José), « A Spanish Dataset for Reproducible Benchmarked Offline Handwriting Recognition », *Language Resources and Evaluation*, 56 (1^{er} sept. 2022), p. 1-14, DOI : 10.1007/s10579-022-09587-3.
- GRANELL (Emilio), CHAMMAS (Edgard), LIKFORMAN-SULEM (Laurence), MARTÍNEZ-HINAREJOS (Carlos-D.), MOKBEL (Chafic) et CÎRSTEÀ (Bogdan-Ionuț), « Transcription of Spanish Historical Handwritten Documents with Deep Neural Networks », *Journal of Imaging*, 4–1 (1[2018]), p. 15, DOI : 10.3390/jimaging4010015.
- GRANET (Adeline), MORIN (Emmanuel), MOUCHÈRE (Harold), QUINIOU (Solen) et VIARD-GAUDIN (Christian), « Transfer Learning for Handwriting Recognition on Historical Documents », dans *7th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, Madeira, Portugal, 2018, URL : <https://hal.archives-ouvertes.fr/hal-01681126> (visité le 27/08/2022).
- GRAVES (Alex) et SCHMIDHUBER (Jürgen), « Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks », dans *Advances in Neural Information Processing Systems*, 2008, t. 21, URL : <https://proceedings.neurips.cc/paper/2008/hash/66368270ffd51418ec58bd793f2d9b1b-Abstract.html> (visité le 25/08/2022).
- GROVER (O.), NARTKER (Thomas A.), RICE (Stephen V.), KANAI (Junichi) et NARTKER (Thomas A.), « An Evaluation of OCR Accuracy » () .
- GUPTA (Maya R.), JACOBSON (Nathaniel P.) et GARCIA (Eric K.), « OCR Binarization and Image Pre-Processing for Searching Historical Documents », *Pattern Recognition*, 40–2 (févr. 2007), p. 389-397, DOI : 10.1016/j.patcog.2006.04.043.
- KASS (Dmitrijs) et VATS (Ekta), *AttentionHTR : Handwritten Text Recognition Based on Attention Encoder-Decoder Networks*, 1^{er} avr. 2022, arXiv : 2201.09390 [cs], URL : <http://arxiv.org/abs/2201.09390> (visité le 06/09/2022).
- KIESSLING (Benjamin), « Kraken - a Universal Text Recognizer for the Humanities », dans Utrecht , Pays-Bas, 8, DOI : 10.34894/Z9G2EX.
- « The Kraken OCR System, version 4.1.2, avr. 2022, URL : <https://kraken.re> (visité le 27/08/2022).
- KIESSLING (Benjamin), STÖKL BEN EZRA (Daniel) et MILLER (Matthew Thomas), *BADAM : A Public Dataset for Baseline Detection in Arabic-script Manuscripts*, juill. 2019, URL : <https://hal.archives-ouvertes.fr/hal-02167164> (visité le 19/08/2022).
- KIESSLING (Benjamin), TISSOT (Robin), STÖKL BEN EZRA (Daniel) et STOKES (Peter Anthony), « eScripta : A New Digital Platform for the Study of Historical Texts and Writing », dans *Digital Humanities 2019*, Utrecht , Pays-Bas, 2019, URL : <https://hal-ephe.archives-ouvertes.fr/hal-02310781> (visité le 19/08/2022).

- KIESSLING (Benjamin), TISSOT (Robin), STOKES (Peter) et STÖKL BEN EZRA (Daniel), « eScriptorium : An Open Source Platform for Historical Document Analysis », dans *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019, t. 2, p. 19-19, DOI : 10.1109/ICDARW.2019.90032.
- MEMON (Jamshed), SAMI (Maira), KHAN (Rizwan Ahmed) et UDDIN (Mueen), « Handwritten Optical Character Recognition (OCR) : A Comprehensive Systematic Literature Review (SLR) », *IEEE Access*, 8 (2020), p. 142642-142668, DOI : 10.1109/ACCESS.2020.3012542.
- MOUFFLET (Jean-François), « 5 ans d’expérimentation de la technologie HTR aux Archives nationales », dans *Futurs Fantastiques*, Paris, France, 2021, p. 39, URL : https://www.bnf.fr/sites/default/files/2022-01/futurs_fantastiques_moufflet.pdf.
- N.C., *Expérimentations – LECTAUREP*, URL : <https://lectaurep.hypotheses.org/category/experimentations> (visité le 20/08/2022).
- NOËMIE, LUCAS et FUR (Doria Le), *OCR / HTR et graphie arabe*, 3, GIS Moyen-Orient et Mondes musulmans, 2022, URL : <http://majlis-remomm.fr/72481> (visité le 19/08/2022).
- PINCHE (Ariane), « CREMMALab Project : Handwritten Text Recognition (HTR) for Medieval Manuscripts », dans *Digital Humanities 2022*, Tokyo, Japan, 2022, URL : <https://hal.archives-ouvertes.fr/hal-03719504> (visité le 21/08/2022).
- *HTR Models and Genericity for Medieval Manuscripts*, juill. 2022, URL : <https://hal.archives-ouvertes.fr/hal-03736532> (visité le 21/08/2022).
- PUIGCERVER (Joan), « Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition ? », dans *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto, Japon, 2017, t. 01, p. 67-72, DOI : 10.1109/ICDAR.2017.20.
- SÁNCHEZ (Joan Andreu), ROMERO (Verónica), TOSELLI (Alejandro H.), VILLEGAS (Mauricio) et VIDAL (Enrique), « A Set of Benchmarks for Handwritten Text Recognition on Historical Documents », *Pattern Recognition*, 94 (oct. 2019), p. 122-134, DOI : 10.1016/j.patcog.2019.05.025.
- SOUZA NETO (Arthur Flor de), LEITE DANTAS BEZERRA (Byron), HECTOR TOSELLI (Alejandro) et BAPTISTA LIMA (Estanislau), « A Robust Handwritten Recognition System for Learning on Different Data Restriction Scenarios », *Pattern Recognition Letters*, 159 (1^{er} juill. 2022), p. 232-238, DOI : 10.1016/j.patrec.2022.04.009.
- STOKES (Peter A.), *eScriptorium : un outil pour la transcription automatique des documents*, EphéNum, URL : <https://ephenum.hypotheses.org/1412> (visité le 22/08/2022).
- STOKES (Peter A.), KIESSLING (Benjamin), STÖKL BEN EZRA (Daniel), TISSOT (Robin) et GARGEM (El Hassene), « The eScriptorium VRE for Manuscript Cultures », *Clas-*

- sics@Journal (), URL : <https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/> (visité le 02/08/2022).
- STOKES (Peter Anthony), « Palaeography and Image-Processing : Some Solutions and Problems », *Digital Medievalist*, 3 (mars 2007), DOI : 10.16995/dm.15.
- STRÖBEL (Phillip Benjamin), CLEMATIDE (Simon) et VOLK (Martin), « How Much Data Do You Need ? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR », dans *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, 2020, p. 3551-3559, URL : <https://aclanthology.org/2020.lrec-1.436> (visité le 21/08/2022).
- STRÖBEL (Phillip Benjamin), CLEMATIDE (Simon), VOLK (Martin), SCHWITTER (Raphael), HODEL (Tobias) et SCHOCH (David), *Evaluation of HTR Models without Ground Truth Material*, 29 avr. 2022, arXiv : 2201.06170 [cs], URL : <http://arxiv.org/abs/2201.06170> (visité le 29/08/2022).
- TERRIEL (Lucas), *Représenter et Évaluer Les Données Issues Du Traitement Automatique d'un Corpus de Documents Historiques. L'exemple de La Reconnaissance Des Écritures Manuscrites Dans Les Répertoires de Notaires Du Projet LectAuRep.* mémoire de master « Technologies numériques appliquées à l'histoire », dir. Alix Chagué et Thibault Clérice, Paris, École nationale des chartes, 2020, URL : https://github.com/Lucaterre/L-TERRIEL_memoireDeStage_M2TNAH_ENC.
- TERRIEL (Lucas) et CHAGUÉ (Alix), *KaMI-lib*, version 0.1.3, 8 août 2022, DOI : 10.5281/zenodo.1234.
- TOMOIAGA (Ciprian), FENG (Paul), SALZMANN (Mathieu) et JAYET (Patrick), *Field Typing for Improved Recognition on Heterogeneous Handwritten Forms*, 22 sept. 2019, arXiv : 1909.10120 [cs], URL : <http://arxiv.org/abs/1909.10120> (visité le 28/08/2022).
- TORRES (Sergio) et JOLIVET (Vincent), « HTR Fine-tuning for Medieval Manuscripts Models : Strategies and Evaluation », dans 2022, URL : <https://dahtr.sciencesconf.org/> (visité le 21/08/2022).
- XAMENA (Eduardo), BARBOSA (Héctor) et OROZCO (Carlos Ismael), « Evaluación de una plataforma completa para Reconocimiento de Textos Manuscritos en Español », *Ciencia y tecnología-21* (2021), p. 6, URL : <https://dialnet.unirioja.es/servlet/articulo?codigo=8148856> (visité le 21/08/2022).
- XAMENA (Eduardo), BARBOZA (Héctor Emanuel) et OROZCO (Carlos Ismael), « End-to-End Platform Evaluation for Spanish Handwritten Text Recognition », *Ciencia y Tecnología* (, 20 déc. 2021), p. 81-95, DOI : 10.18682/cyt.vi21.4327.
- ZARRI (Gian Piero), « Quelques aspects techniques de l'exploitation informatique des documents textuels : saisie des données et problèmes de sortie », *Publications de l'École Française de Rome*, 31-1 (1977), p. 399-413, URL : https://www.persee.fr/doc/efr_0000-0000_1977_act_31_1_2286 (visité le 24/08/2022).

Traitements automatiques du langage

- ÁLVAREZ-MELLADO (Elena), DÍEZ-PLATAS (María Luisa), RUIZ-FABO (Pablo), BERMÚDEZ (Helena), Ros (Salvador) et GONZÁLEZ-BLANCO (Elena), « TEI-friendly Annotation Scheme for Medieval Named Entities : A Case on a Spanish Medieval Corpus », *Language Resources and Evaluation*, 55–2 (1^{er} juin 2021), p. 525-549, DOI : 10.1007/s10579-020-09516-2.
- BARRUS (Tyler), *Pyspellchecker*, version 0.6.3, 29 août 2022, URL : <https://github.com/barrust/pyspellchecker> (visité le 30/08/2022).
- CANETE (José), CHAPERON, FUENTES (Rodrigo), HO (Jou-Hui), KANG (Hojin) et PÉREZ (Jorge), *Spanish Pre-Trained BERT Model and Evaluation Data*, Santiago (Chili), 2020, URL : <https://github.com/dccuchile/beto>.
- CHAABI (Youness) et ATAA ALLAH (Fadoua), « Amazigh Spell Checker Using Damerau-Levenshtein Algorithm and N-gram », *Journal of King Saud University - Computer and Information Sciences*, 34 (8, Part B[2022]), p. 6116-6124, DOI : 10.1016/j.jksuci.2021.07.015.
- CLÉRICE (Thibault), *Détection d'isotopies Par Apprentissage Profond : L'exemple de La Sexualité En Latin Classique et Tardif*, thèse de doctorat en lettres et civilisations antiques, dir. Christian Nicolas, Lyon, Université Lyon 3, 2022, URL : <https://github.com/PonteIneptique/these-redaction/releases/tag/1.0.1>.
- DEVLIN (Jacob), CHANG (Ming-Wei), LEE (Kenton) et TOUTANOVA (Kristina), *BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding*, 24 mai 2019, DOI : 10.48550/arXiv.1810.04805, arXiv : 1810.04805 [cs].
- EHRMANN (Maud), « Les Entités Nommées, de La Linguistique Au TAL : Statut Théorique et Méthodes de Désambiguïsation » (), p. 296.
- GABAY (Simon), CAMPS (Jean-Baptiste) et CLÉRICE (Thibault), *Manuel d'annotation Linguistique Pour Le Français Moderne (XVI^e -XVIII^e Siècles)*, avr. 2022, URL : <https://hal.archives-ouvertes.fr/hal-02571190> (visité le 07/09/2022).
- GAIKWAD (Shital) et BOGIRI (Nagaraju), « Effective and Efficient XML Duplicate Detection Using Levenshtein Distance Algorithm », 4–6 (2013), p. 5.
- GAIKWAD (Shital) et NAGARAJU (Bogiri), « Levenshtein Distance Algorithm for Efficient and Effective XML Duplicate Detection », dans 2015, p. 1-5, DOI : 10.1109/IC4.2015.7375698.
- HALDAR (Rishin) et MUKHOPADHYAY (Debajyoti), « Levenshtein Distance Technique in Dictionary Lookup Methods : An Improved Approach » (), p. 5, DOI : 10.48550.
- KARTHIKEYAN Srinidhi, DE HERRERA Alba G. Seco, DOCTOR Faiyaz et MIRZA Asim, « An OCR Post-Correction Approach Using Deep Learning for Processing Medical Reports », *IEEE Transactions on Circuits and Systems for Video Technology*, 32–5 (mai 2022), p. 2574-2581, DOI : 10.1109/TCSVT.2021.3087641.

- KISSOS (Ido) et DERSHOWITZ (Nachum), « OCR Error Correction Using Character Correction and Feature-Based Word Classification », dans *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, Santorini, Greece, 2016, p. 198-203, DOI : 10.1109/DAS.2016.44.
- LABBÉ (Dominique), « Normalisation et lemmatisation d'une question ouverte » (), p. 21.
- LOPRESTI (Daniel), « Optical Character Recognition Errors and Their Effects on Natural Language Processing », *International Journal on Document Analysis and Recognition (IJDAR)*, 12–3 (1^{er} sept. 2009), p. 141-151, DOI : 10.1007/s10032-009-0094-8.
- MOODY (Chris), *Stop Using Word2vec / Stitch Fix Technology – Multithreaded*, multi-threaded, 18 oct. 2017, URL : <https://multithreaded.stitchfix.com/blog/2017/10/18/stop-using-word2vec/> (visité le 06/09/2022).
- NGUYEN (Thi-Tuyet-Hai), JATOWT (Adam), COUSTATY (Mickael), NGUYEN (Nhu-Van) et DOUCET (Antoine), « Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing », dans *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2019, p. 29-38, DOI : 10.1109/JCDL.2019.00015.
- NORVIG (Peter), *How to Write a Spelling Corrector*, févr. 2007, URL : <https://norvig.com/spell-correct.html> (visité le 31/08/2022).
- PAL (Aditya) et MUSTAFI (Abhijit), *Vartani Spellcheck – Automatic Context-Sensitive Spelling Correction of OCR-generated Hindi Text Using BERT and Levenshtein Distance*, 14 déc. 2020, DOI : 10.48550/arXiv.2012.07652, arXiv : 2012.07652 [cs].
- POIBEAU (Thierry) et NAZARENKO (Adeline), « L'extraction d'information, Une Nouvelle Conception de La Compréhension de Texte ? », *Traitemen Automatique des Langues*, 2–40 (1999), p. 87-115.
- RIGAUD (Christophe), DOUCET (Antoine), COUSTATY (Mickaël) et MOREUX (Jean-Philippe), « ICDAR 2019 Competition on Post-OCR Text Correction », dans *15th International Conference on Document Analysis and Recognition*, Sydney, Australia, 2019, p. 1588-1593, URL : <https://hal.archives-ouvertes.fr/hal-02304334> (visité le 31/08/2022).
- Spacy-Transformers : Use Pretrained Transformers like BERT, XLNet and GPT-2 in spaCy*, version 1.1.8, Explosion, 7 sept. 2022, URL : <https://github.com/explosion/spacy-transformers> (visité le 08/09/2022).
- STRUABELL (Emma), GANESH (Ananya) et MCCALLUM (Andrew), *Energy and Policy Considerations for Deep Learning in NLP*, 5 juin 2019, DOI : 10.48550/arXiv.1906.02243, arXiv : 1906.02243 [cs].
- TAKAHASHI (H.), ITOH (N.), AMANO (T.) et YAMASHITA (A.), « A Spelling Correction Method and Its Application to an OCR System », *Pattern Recognition*, 23–3 (1^{er} janv. 1990), p. 363-377, DOI : 10.1016/0031-3203(90)90023-E.

- TANGUY (Ludovic), *Traitemet Automatique de la Langue Naturelle et interprétation : Contribution à l'élaboration d'un modèle informatique de la Sémantique Interprétabile*, thèse d'informatique, dir. Jean-Pierre Barthélémy et Ioannis Kanellos, Université de Rennes 1, 1997, URL : <https://halshs.archives-ouvertes.fr/tel-01322692> (visité le 05/09/2022).
- TANGUY (Ludovic) et FABRE (Cécile), « Évolutions de la linguistique outillée : méfaits et bienfaits du TAL », *L'information grammaticale*–142 (2014), p. 15, URL : <https://halshs.archives-ouvertes.fr/halshs-01057493> (visité le 05/09/2022).
- TERRIEL (Lucas), « Atelier : Production d'un Modèle Affiné de Reconnaissance d'écriture Manuscrite Avec eScriptorium et Évaluation de Ses Performances. Évaluer Son Modèle HTR/OCR Avec KaMI (Kraken as Model Inspector) », dans *Les Futurs Fantastiques - 3e Conférence Internationale Sur l'Intelligence Artificielle Appliquée Aux Bibliothèques, Archives et Musées*, Paris, France, 2021, URL : <https://hal.archives-ouvertes.fr/hal-03495762> (visité le 29/08/2022).
- *spaCy Fishing*, version 0.1.7, Inria, 24 août 2022, URL : <https://github.com/Lucaterre/spacyfishing> (visité le 10/09/2022).
- TONG (Xiang) et EVANS (David A.), « A Statistical Approach to Automatic OCR Error Correction in Context », dans *Fourth Workshop on Very Large Corpora*, Herstmonceux Castle, Sussex, UK, 1996, URL : <https://aclanthology.org/W96-0108> (visité le 30/08/2022).
- TOUSSAINT (Yannick), « Extraction de connaissances à partir de textes structurés », *Document numérique*, 8–3 (2004), p. 11-34, DOI : [10.3166/dn.8.3.11-34](https://doi.org/10.3166/dn.8.3.11-34).
- VASWANI (Ashish), SHAZER (Noam), PARMAR (Niki), USZKOREIT (Jakob), JONES (Llion), GOMEZ (Aidan N.), KAISER (Lukasz) et POLOSUKHIN (Illia), *Attention Is All You Need*, 5 déc. 2017, DOI : [10.48550/arXiv.1706.03762](https://doi.org/10.48550/arXiv.1706.03762), arXiv : [1706.03762 \[cs\]](https://arxiv.org/abs/1706.03762).
- WHITE LAW (Casey), HUTCHINSON (Ben), CHUNG (Grace Y.) et ELLIS (Gerard), « Using the Web for Language Independent Spellchecking and Autocorrection », dans *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 2 - EMNLP '09*, Singapore, 2009, t. 2, p. 890, DOI : [10.3115/1699571.1699629](https://doi.org/10.3115/1699571.1699629).
- WIDDOWSON (Henry), « J.R. Firth, 1957, Papers in Linguistics 1934–51 », *International Journal of Applied Linguistics*, 17–3 (2007), p. 402-413, DOI : [10.1111/j.1473-4192.2007.00164.x](https://doi.org/10.1111/j.1473-4192.2007.00164.x).
- YUE (Xiang) et ZHOU (Shuang), *PHICON : Improving Generalization of Clinical Text De-identification Models via Data Augmentation*, 10 oct. 2020, arXiv : [2010.05143 \[cs\]](https://arxiv.org/abs/2010.05143), URL : [http://arxiv.org/abs/2010.05143](https://arxiv.org/abs/2010.05143) (visité le 01/07/2022).

Reconnaissance d'entités nommées

- ABADIE (N.), CARLINET (E.), CHAZALON (J.) et DUMÉNIEU (B.), « A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories », dans *Document Analysis Systems. DAS 2022*. Dir. S. Uchida, E. Barney et V. Eglin, La Rochelle, France, 2022 (Document Analysis Systems. DAS 2022. 13237), DOI : 10.1007/978-3-031-06555-2_30.
- ALSHAMMARI (Nasser) et ALANAZI (Saad), « The Impact of Using Different Annotation Schemes on Named Entity Recognition », *Egyptian Informatics Journal*, 22–3 (1^{er} sept. 2021), p. 295-302, DOI : 10.1016/j.eij.2020.10.004.
- BOROS (Emanuela), HAMDI (Ahmed), LINHARES PONTES (Elvys), CABRERA-DIEGO (Luis Adrián), MORENO (Jose G.), SIDERE (Nicolas) et DOUCET (Antoine), « Alleviating Digitization Errors in Named Entity Recognition for Historical Documents », dans *Proceedings of the 24th Conference on Computational Natural Language Learning*, Online, 2020, p. 431-441, DOI : 10.18653/v1/2020.conll-1.35.
- BUNESCU (Razvan) et PAŞCA (Marius), « Using Encyclopedic Knowledge for Named Entity Disambiguation », dans *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 2006, p. 9-16, URL : <https://aclanthology.org/E06-1002> (visité le 09/09/2022).
- CARBONELL (Manuel), FORNÉS (Alicia), VILLEGAS (Mauricio) et LLADÓS (Josep), *A Neural Model for Text Localization, Transcription and Named Entity Recognition in Full Pages*, 4 mai 2020, arXiv : 1912.10016 [cs], URL : <http://arxiv.org/abs/1912.10016> (visité le 26/08/2022).
- CLAVAUD (Florence), ROMARY (Laurent), CHARBONNIER (Pauline), TERRIEL (Lucas), PIRAINO (Gaetano) et VERDESE (Vincent), « NER4Archives (Named Entity Recognition for Archives) : Conception et Réalisation d'un Outil de Détection, de Classification et de Résolution Des Entités Nommées Dans Les Instruments de Recherche Archivistiques Encodés En XML/EAD. » Dans *Atelier Culture-INRIA*, Pierrefitte sur Seine, France, 2022, URL : <https://hal.archives-ouvertes.fr/hal-03625734> (visité le 17/08/2022).
- EHRMANN (Maud), *Les Entités Nommées, de La Linguistique Au TAL : Statut Théorique et Méthodes de Désambiguïsation*, thèses d'informatique et langage, Paris Diderot University, 2008, URL : <https://hal.archives-ouvertes.fr/tel-01639190> (visité le 06/09/2022).
- EHRMANN (Maud), HAMDI (Ahmed), PONTES (Elvys Linhares), ROMANELLO (Matteo) et DOUCET (Antoine), *Named Entity Recognition and Classification on Historical Documents : A Survey*, 23 sept. 2021, arXiv : 2109.11406 [cs], URL : <http://arxiv.org/abs/2109.11406> (visité le 08/09/2022).

- GROUIN (Cyril), « Simplification de Schémas d'annotation : Un Aller sans Retour ? », dans *Actes de TALN*, Rennes, France, 2018, URL : <https://hal.archives-ouvertes.fr/hal-01831221> (visité le 08/09/2022).
- HENGCHEN Simon, VAN HOOLAND Seth, VERBORGH Ruben et DE WILDE Max, « L'extraction d'entités nommées : une opportunité pour le secteur culturel ? », *I2D - Information, données & documents*, 52–2 (2015), p. 70-79, DOI : [10.3917/i2d.152.0070](https://doi.org/10.3917/i2d.152.0070).
- JIANG (Jing) et ZHAI (ChengXiang), « Exploiting Domain Structure for Named Entity Recognition », dans *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* -, New York, New York, 2006, p. 74-81, DOI : [10.3115/1220835.1220845](https://doi.org/10.3115/1220835.1220845).
- LABUSCH (Kai), NEUDECKER (Clemens) et ZELLHOFER (David), « BERT for Named Entity Recognition in Contemporary and Historical German » (), p. 9.
- LASSNER (David), *Standoffconverter*, standoff-nlp, 2021, URL : <https://github.com/standoff-nlp/standoffconverter> (visité le 10/09/2022).
- LEMMENS (Jens), *CoNLL 2022 / CoNLL*, CoNLL, 2022, URL : <https://conll.org/> (visité le 28/08/2022).
- LI (Jing), SUN (Aixin), HAN (Jianglei) et LI (Chenliang), *A Survey on Deep Learning for Named Entity Recognition*, 18 mars 2020, arXiv : [1812.09449 \[cs\]](https://arxiv.org/abs/1812.09449), URL : <https://arxiv.org/abs/1812.09449> (visité le 07/09/2022).
- MORENO José G., BESANCON Romaric, BEAUMONT Romain, D'HONDT Eva, LIGOZAT Anne-Laure, ROSSET Sophie, TANNIER Xavier et GRAU Brigitte, « Apprendre Des Représentations Jointes de Mots et d'entités Pour La Désambiguïsation d'entités », dans *24ème Conférence Sur Le Traitement Automatique Des Langues Naturelles - TALN 2017*, Orléans, France, 2017, URL : <https://hal.archives-ouvertes.fr/hal-01626197> (visité le 09/09/2022).
- NAKAYAMA (Hiroki), KUBO (Takahiro), KAMURA (Junya), TANIGUCH (Yasufumi) et LIANG (Xu), *Doccano*, version 1.8.0, doccano, 8 sept. 2022, URL : <https://github.com/doccano/doccano> (visité le 08/09/2022).
- NOUVEL (Damien), *Reconnaissance Des Entités Nommées Par Exploration de Règles d'annotation : Interpréter Les Marqueurs d'annotation Comme Instructions de Structuration Locale*. These de doctorat, Tours, 2012, URL : <http://www.theses.fr/2012TOUR4011> (visité le 24/08/2022).
- PERSSON (Adam), « The Effect of Excluding Out of Domain Training Data from Supervised Named-Entity Recognition », dans *Proceedings of the 21st Nordic Conference on Computational Linguistics*, Gothenburg, Sweden, 2017, p. 289-292, URL : <https://aclanthology.org/W17-0240> (visité le 09/09/2022).
- RIZZO (Giuseppe) et TRONCY (Raphael), « NERD : A Framework for Unifying Named Entity Recognition and Disambiguation Extraction Tools » (), p. 4.

- RÖDER (Michael), USBECK (Ricardo), HELLMANN (Sebastian), GERBER (Daniel) et BOTH (Andreas), « N3 - A Collection of Datasets for Named Entity Recognition and Disambiguation in the NLP Interchange Format », dans *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014, p. 3529-3533, URL : http://www.lrec-conf.org/proceedings/lrec2014/pdf/856_Paper.pdf (visité le 06/07/2022).
- SANG (Erik F. Tjong Kim) et DE MEULDER (Fien), *Introduction to the CoNLL-2003 Shared Task : Language-Independent Named Entity Recognition*, 12 juin 2003, arXiv : cs/0306050, URL : <http://arxiv.org/abs/cs/0306050> (visité le 09/09/2022).
- SCHEITHAUER (Hugo), *La Reconnaissance d'entités Nommées Appliquées à Des Données Issues de La Transcription Automatique de Documents Manuscrits Patrimoniaux. Expérimentations et Préconisations à Partir Du Projet LECTAUREP*, mémoire de master "Technologies numériques appliquées à l'histoire", dir. Alix Chagué et Thibault Clérice, Paris, École nationale des chartes, 2021, URL : https://github.com/HugoSchtr/memoire_TNAH_M2_HugoScheithauer.
- SOUDANI (Aicha), MEHERZI (Yosra), BOUHAFS (Asma), FRONTINI (Francesca), BRANDO (Carmen), DUPONT (Yoann) et MÉLANIE-BECQUET (Frédérique), « Adaptation et Évaluation de Systèmes de Reconnaissance et de Résolution Des Entités Nommées Pour Le Cas de Textes Littéraires Français Du 19ème Siècle », dans *Atelier Humanités Numériques Spatialisées (HumaNS'2018)*, Montpellier, France, 2018, URL : <https://hal.archives-ouvertes.fr/hal-01925816> (visité le 06/09/2022).
- SUÁREZ (Pedro Javier Ortiz), DUPONT (Yoann), MULLER (Benjamin), ROMARY (Laurent) et SAGOT (Benoît), « Establishing a New State-of-the-Art for French Named Entity Recognition » (), p. 9.
- TORRES AGUILAR (Sergio), « La reconnaissance des entités nommées dans les bases numériques de chartes médiévales en latin : le cas du Corpus Burgundiae Medii Aevi (xe-xiiie siècle) », *Médiévales. Langues, Textes, Histoire*, 73–73 (73[2017]), p. 47-65, DOI : [10.4000/medievales.8182](https://doi.org/10.4000/medievales.8182).

Le web sémantique

- BERNERS-LEE (Tim), HENDLER (James) et LASSILA (Ora), « The Semantic Web », *Scientific American*, 284–5 (2001), p. 34-43, JSTOR : 26059207.
- CIFUENTES-SILVA (Francisco), SIFAQUI (Christian) et LABRA-GAYO (Jose Emilio), « Towards an Architecture and Adoption Process for Linked Data Technologies in Open Government Contexts : A Case Study for the Library of Congress of Chile », dans *Proceedings of the 7th International Conference on Semantic Systems*, New York, NY, USA, 2011 (I-Semantics '11), p. 79-86, DOI : [10.1145/2063518.2063529](https://doi.org/10.1145/2063518.2063529).

- CIOTTI (Fabio) et TOMASI (Francesca), « Formal Ontologies, Linked Data, and TEI Semantics », *Journal of the Text Encoding Initiative*—Issue 9 (Issue 9[2016]), DOI : 10.4000/jtei.1480.
- DAVIS (Edie) et HERAVI (Bahareh), « Linked Data and Cultural Heritage : A Systematic Review of Participation, Collaboration, and Motivation », *Journal on Computing and Cultural Heritage*, 14–2 (10 mai 2021), 21 :1-21 :18, DOI : 10.1145/3429458.
- ERXLEBEN (Fredo), GÜNTHER (Michael), KRÖTZSCH (Markus), MENDEZ (Julian) et VRANDEČIĆ (Denny), « Introducing Wikidata to the Linked Data Web », dans *The Semantic Web – ISWC 2014*, dir. Peter Mika, *et al.*, Cham, 2014 (Lecture Notes in Computer Science), p. 50-65, DOI : 10.1007/978-3-319-11964-9_4.
- FOPPIANO (Luca) et ROMARY (Laurent), « Entity-Fishing : A DARIAH Entity Recognition and Disambiguation Service », *Journal of the Japanese Association for Digital Humanities*, 5–1 (20 nov. 2020), p. 22-60, DOI : 10.17928/jjadh.5.1_22.
- GONZALEZ-ZAPATA (Felipe) et HEEKS (Richard), « The Multiple Meanings of Open Government Data : Understanding Different Stakeholders and Their Perspectives », *Government Information Quarterly*, 32–4 (1^{er} oct. 2015), p. 441-452, DOI : 10.1016/j.giq.2015.09.001.
- MARTINS (Pedro Henrique), MARINHO (Zita) et MARTINS (André F. T.), *Joint Learning of Named Entity Recognition and Entity Linking*, 18 juill. 2019, arXiv : 1907.08243 [cs], URL : <http://arxiv.org/abs/1907.08243> (visité le 10/09/2022).
- PÉREZ (Jorge), ARENAS (Marcelo) et GUTIERREZ (Claudio), « Semantics and Complexity of SPARQL », *ACM Transactions on Database Systems*, 34–3 (3 sept. 2009), 16 :1-16 :45, DOI : 10.1145/1567274.1567278.
- TERRIEL (Lucas), *spaCy Fishing*, version 0.1.7, Inria, 24 août 2022, URL : <https://github.com/Lucaterre/spacyfishing> (visité le 10/09/2022).

Codes et scripts produits durant le stage

- HUMEAU (Maxime), *Enrichment Wikisource*, avec la coll. d’Alessandro Chiaretti, Archivo Central Andres Bello, mai 2022, URL : https://github.com/Proyecto-Ocupacion-Araucania-UChile/data_enrichment.
- *Entraînement et Annotations NER*, avec la coll. d’Alessandro Chiaretti, Archivo Central Andres Bello, juin 2022, URL : https://github.com/Proyecto-Ocupacion-Araucania-UChile/NER_Araucania.
 - *HTR Evaluation*, avec la coll. d’Alessandro Chiaretti, Archivo Central Andres Bello, avr. 2022, URL : https://github.com/Proyecto-Ocupacion-Araucania-UChile/model-HTR/tree/main/test_kami (visité le 28/08/2022).

- *Postprocess HTR*, avec la coll. d'Alessandro Chiaretti, Archivo Central Andres Bello, mai 2022, URL : https://github.com/Proyecto-Ocupacion-Araucania-UChile/postprocess_alto.
- *Preprocessing HTR*, avec la coll. d'Alessandro Chiaretti, Archivo Central Andres Bello, avr. 2022, URL : <https://github.com/Proyecto-Ocupacion-Araucania-UChile/model-HTR/tree/main/Preprocess>.
- *Tei Transformation*, avec la coll. d'Alessandro Chiaretti, Archivo Central Andres Bello, juill. 2022, URL : https://github.com/Proyecto-Ocupacion-Araucania-UChile/TEI_tranformation.

HUMEAU (Maxime) et CHIARETTI (Alessandro), *HTR - Araucania - XIX Manuscript*, Archivo Central Andres Bello, 13 sept. 2022, URL : <https://zenodo.org/record/7075075> (visité le 14/09/2022).

- *HTR - Araucania Manuscript XIXe*, Archivo Central Andres Bello, 2022, URL : https://github.com/Proyecto-Ocupacion-Araucania-UChile/HTR_Araucania_XIX (visité le 13/09/2022).

Introduction

« La masse des choses dites dans une culture, conservées, valorisées, réutilisées, répétées et transformées. Bref, toute cette masse verbale qui a été fabriquée par les hommes, investie dans leurs techniques et leurs institutions, et qui est tissée avec leur existence et leur histoire »

Michel Foucault¹

Une querelle persiste depuis le milieu du XX^e sur la notion « archive » entre l’archéologie foucaldienne et les institutions patrimoniales. La seconde la définit comme le produit ontologique d’une action de stockage, de préservation et de classification de documents divers. Dans une autre optique, Michel Foucault lui apporte une nouvelle dimension, celle d’une reconstitution d’un discours dans sa matérialité historique. C’est « le système général de la formation et de la transformation des énoncés ² ».

Le souci culturel et intellectuel de conserver cette « masse des choses » et l’éclosion des outils numériques au sein des institutions patrimoniales et des sciences humaines ont permis une réconciliation de ces deux dimensions contemporaines de l’archive. La mutation de cette ”masse” en « *data* » pouvant être à la fois être conservée, disséquée et contextualisée, voire même enrichie, permet l’acquisition de nouvelles compétences aux institutions patrimoniales et scientifiques. L’automatisation des processus de traitements documentaires (dans le sens archivistique ou éditorial) a fait l’objet d’un véritable investissement, devenant une priorité d’action de développement au sein de nombreux centres ces dix à vingt dernières années³. Les nouvelles méthodes informatiques donnent la capa-

1. Michel Foucault, « La naissance d’un monde », in Michel Foucault et Jacques Lagrange, *Dits et écrits, 1954-1988*, dir. Daniel Defert et François Ewald, Paris, France, impr. 1994, texte n°68, p. 814-815

2. Denise Ogilvie, « Paradoxes de « l’archive » », *Sociétés & Représentations*, 43-1 (2017), p. 121-134, DOI : 10.3917/sr.043.0121, Michel Foucault, *L’archéologie du savoir*, Paris, Gallimard, 1969, p. 177-179. *via*.

3. Arianna Sforzini, *Michel Foucault numérique. Implications philosophiques*, 16 juill. 2021, URL : <https://www. implications - philosophiques.org/michel-foucault-numerique/> (visité le 16/08/2022).

cité de renouveler les méthodes d'exploration des documents et donner un nouveau sens à l'information. Elles permettent un accroissement de l'offre de valorisation considérable⁴. Le numérique est devenu un outil indispensable à la fois au scientifique humaniste comme à l'archiviste face à l'accroissement exponentiel de bases de données dédiées à l'exploration de corpus documentaire.

Toutefois, cette réalité numérique reste concentrée aux plus grandes puissances économiques. Cette transformation numérique fait encore face à de nombreuses disparités géographiques ; de nombreux pays et institutions sont cantonnés à la marge de cette transformation numérique de par les besoins matériels, financiers et techniques que cela exige⁵. Si l'archive reste le symbole de l'affirmation de l'État et de ses besoins bureaucratiques, la pérennisation des instruments culturels reste bien souvent un domaine sacrifié.

Le Chili, pays bien souvent considéré comme le plus développé d'Amérique du Sud au vu de ses infrastructures politiques, économiques et sanitaires, reste le symbole de ce contraste du déploiement du numérique au sein des institutions culturelles⁶. Un contraste à la fois continental où les projets numériques se concentrent aux pays de la pointe du continent, mais où les initiatives numériques restent encore marginales, mais aussi avec les pays occidentaux. Un groupe de chercheurs pointent justement un système concentrique autour des projets anglo-saxons, puis européens puis le reste du monde⁷.

Néanmoins, les humanités numériques et l'accroissement des projets numériques suscitent de plus en plus d'intérêt au sein des institutions patrimoniales et universitaires chiliennes. Dans ce cadre, le centre Archivo Central Andres Bello (ACAB) a déployé un certain nombre de projets autour de la valorisation numérique de leurs ressources archivistiques, plus particulièrement sur la numérisation, mais aussi dernièrement une volonté de s'inscrire dans le processus de Linked Open Data, (données ouvertes et liées)

4. Les récents projets des archives nationales répondent à ce besoin de transcender la fonction archivistique traditionnelle à travers le numérique à l'image du projet d'édition numérique des testaments de poilus ou du projet NER4Archives. Florence Clavaud, « Vers l'édition En Ligne Des Testaments de Poilus », dans *Les Éditions Savantes Numériques : Enjeux et Réalisation*, Lille, France, 2019, URL : <https://hal.archives-ouvertes.fr/hal-02469768> (visité le 17/08/2022) ; F. Clavaud, Laurent Romary, Pauline Charbonnier, Lucas Terriel, Gaetano Piraino et Vincent Verdese, « NER4Archives (Named Entity Recognition for Archives) : Conception et Réalisation d'un Outil de Détection, de Classification et de Résolution Des Entités Nommées Dans Les Instruments de Recherche Archivistiques Encodés En XML/EAD. » Dans *Atelier Culture-INRIA*, Pierrefitte sur Seine, France, 2022, URL : <https://hal.archives-ouvertes.fr/hal-03625734> (visité le 17/08/2022)

5. Tanja Wissik, Jennifer Edmond, Frank Fischer, Franciska de Jong, Stefania Scagliola, Andrea Scharnhorst, Hendrik Schmeer, Walter Scholger et Leon Wessels, *Teaching Digital Humanities Around the World : An Infrastructural Approach to a Community-Driven DH Course Registry*, mars 2020, URL : <https://hal.archives-ouvertes.fr/hal-02500871> (visité le 17/08/2022).

6. UNDP (éd.), *The next Frontier : Human Development and the Anthropocene*, New York, Etats-Unis, 2020 (Human Development Report, 2020).

7. Isabel Galina Russell, « Geographical and Linguistic Diversity in the Digital Humanities », *Literary and Linguistic Computing*, 3–29 (sept. 2014), p. 307-316, DOI : 10.1093/lit/fqu005.

notamment autour des archives du célèbre poète et député communiste chilien Pablo Neruda⁸.

Le centre ACAB a pris naissance officiellement en 1994, bien que son héritage institutionnel remonte à la première partie du XIX^e siècle comme bibliothèque de l'*Instituto Nacional* (Institut Nationale)⁹. Il appartient à l'Universidad de Chile, situé à Santiago, qui est la plus grande université publique du Chili. La structure est sous la responsabilité du Bureau du vice-recteur pour la vulgarisation et la communication et elle est divisée en trois aires de compétences :

- *Información Bibliográfica y Archivística* (Information bibliographie et archivage) : Le service est en charge de l'inventorisation et la classification des collections tout en assurant l'accès aux documents pour le public.
- *Conservación y Patrimonio* (Conservation et patrimoine) : l'équipe est responsable de la préservation (biochimique) et de la restauration des différents documents abîmés. Il assure aussi l'accès au musée et la numérisation des collections.
- *Investigación Patrimonial* (Recherche patrimoniale) : Il s'agit de la partie scientifique et éditoriale du centre afin d'exploiter les documents possédés. Des programmes éducatifs sont mis en place par l'équipe afin de valoriser les fonds documentaires.

Comme nous venons de le voir, le centre culturel est avant tout une structure pluridisciplinaire alliant des compétences éditoriales, archivistiques, conservations et scientifiques. C'est plus d'une vingtaine de personnes qui travaillent ainsi dans les locaux de la maison centrale de l'Universidad de Chile. Au total, ACAB gère plus de 18 collections documentaires, dont trois sont classés comme « Monument National » par le ministère de l'Éducation.

Suite à plusieurs échanges et du fait de l'intérêt soucieux pour les nouvelles technologies et leurs apports aux compétences patrimoniales, un projet est né autour des archives de l'« *Ocupación de la Araucanía*¹⁰ » (1850-1883). Cette volonté est double puisqu'il s'agit de faire une première prospective de l'intérêt des outils numériques et notamment de l'apprentissage machine comme appui aux politiques patrimoniales. Dans un second temps, il s'agit de mettre en place un processus d'édition nativement numérique de ces archives.

Dans ce cadre, un stage de 4 mois entre avril et juillet 2022 a été mis en place au sein de ACAB et l'Universidad de Chile avec la collaboration et le soutien de l'École nationale des chartes (ENC) ainsi que de la Région Île-de-France afin de développer un

8. Note interne de service.

9. *Archivo Central Andres Bello*, url : <http://archivobello.uchile.cl/acerca-del-archivo/historia>, consulté le 17/02/2022.

10. « Occupation de l'Araucanie ». Suite à de nombreuses contestations sociales, mais aussi scientifiques, le terme a substitué la dénomination de « *Pacificación de la Araucanía* » (Pacification de l'Araucanie).

certain nombre d'outils permettant l'automatisation de l'édition numérique des archives autour de l'Occupation de l'Araucanie.

Première partie

Production d'une *pipeline* de
transcription automatisée

Chapitre 1

Les archives et ses données : un enjeu technique, méthodologique et juridique

La constitution d'un jeu de données, bien souvent appelé par son anglicisme *dataset*, est une étape primordiale dans la construction d'un projet d'édition numérique à la fois dans son aspect technique et juridique. Durant cette préparation, il s'agit de construire les données des vérités terrain (*ground truth*) permettant la construction d'une chaîne de traitement automatisée de sa transcription à son édition numérique.

Cette opération ne doit pas être négligée puisqu'elle va être le socle du projet. Ces premières réflexions doivent ainsi problématiser et identifier les particularités du corpus afin d'en garantir les qualités et ses caractéristiques, tout en essayant de minimiser au maximum l'impact de ses limites.

1.1 *Araucania* : conflit, histoire et archives

Dans un premier temps, il convient d'apporter une brève contextualisation historique autour de la jeune République du Chili et de son processus d'extension territoriale au XIX^e siècle. L'Araucanie est une région emblématique de ce processus colonial de par la sécularisation d'un conflit socio-ethnique autour de la question des Mapuches et plus global.

La question et les enjeux de ces sources s'inscrivent ainsi dans cette mémoire lourde portée par une nation en volonté de rupture avec son passé. Afin de saisir tous les aspects, il nous est donc impératif d'entrevoir ce qu'est cette collection et ce qu'elle représente.

1.1.1 Une brève histoire d'un conflit

Il faut rappeler que l'État-nation chilien est fondé sur un territoire pluriethnique et pluriculturel. Alors qu'en 1820 survient la victoire officielle de la révolution de l'indépendance chilienne face à Madrid, l'appropriation de l'ensemble du territoire revendiqué. Certaines alliances avec les peuples indigènes et le régime républicain ont persisté afin de faire face aux derniers soutiens implantés de la couronne espagnole jusqu'au début des années 1830¹. Face au besoin d'affirmation des jeunes institutions étatiques, celles-ci ont premièrement déterminé une politique d'assimilation des peuples aborigènes dans les régions centrales du Chili. Cette politique construit un imaginaire commun autour des racines indigènes et des frontières, renvoyant la question de l'altérité au-delà des nouvelles frontières².

Toutefois, cette pratique d'assimilation et d'appropriation territoriale trouve rapidement des limites au sein des territoires encore marginaux du Sud du Chili avec une résistance accrue menée par le peuple Mapuche, dont l'ethnogenèse s'est construite autour des précédentes tentatives d'invasions coloniales³. L'arrivée des conquistadors espagnols au Chili au cours de la première moitié du XVI^{ème} siècle s'est soldée par la défaite militaire de la Guerre d'Arauco (1546-1641)⁴.

La multiplication des résistances des peuples autochtones et plus particulièrement Mapuche, conduit à une radicalisation des velléités politiques chiliennes. Les différents gouvernements successifs ont mis en place une politique d'expansion plus agressive, en facilitant à partir des années 1850 une immigration de plus en plus massive de citoyens chiliens vers la récente région administrative de l'Arauco⁵. De plus, il s'agit pour le gouvernement de renforcer son appareil capitaliste en introduisant une concurrence féroce au sein des marchés agricoles⁶. Face à la perception de cette invasion, les crispations et les conflits locaux se multiplièrent avec ce que l'on appelle les « guerras civiles "montistas" » entre 1851-1859. Pour l'État, elle est une source de justification pour multiplier l'implantation militaire dans la région afin d'en protéger ces ressortissants⁷.

Cette répression s'intensifie à partir de 1859, après le soutien explicite de certains chefs mapuches auprès des révolutionnaires libéraux lors de l'insurrection de 1859 contre

1. Pablo Mariman Quemenado, « La geoestrategia en el conflicto chileno mapuche : la configuración del Estado Nación (1830-1869) », *Anales de la Universidad de Chile*–13 (13[2017]), p. 39-57, DOI : 10.5354/0717-8883.2017.48995.

2. José Bengoa, *La memoria olvidada : historia de los pueblos indígenas de Chile*, Cuadernos Bicentenario, Presidencia de la República, Santiago, Chili, 2004, p .19-21.

3. Guillaume Boccara, « Organisation sociale, guerre de captation et ethnogenèse chez les Reches-Mapuche à l'époque coloniale », *Homme*, 39–150 (1999), p. 85-117, DOI : 10.3406/hom.1999.453568.

4. Bastien Sepúlveda, « Le pays mapuche, un territoire « à géographie variable » », *Espace populations sociétés. Space populations societies*–2012/1 (2012/1[2012]), p. 73-88, DOI : 10.4000/eps.4872.

5. J. Bengoa, *La memoria olvidada...*, p.331-332.

6. B. Sepúlveda, « Le pays mapuche, un territoire « à géographie variable » »...

7. J. Bengoa, *La memoria olvidada...*, p .313.

8. Thierry Nouaille, « L'indépendance du Chili : les conséquences sur le peuple mapuche (1810-2010) », *América. Cahiers du CRICCAL*–42 (42[2012]), p. 131-143, DOI : 10.4000/americam.1111

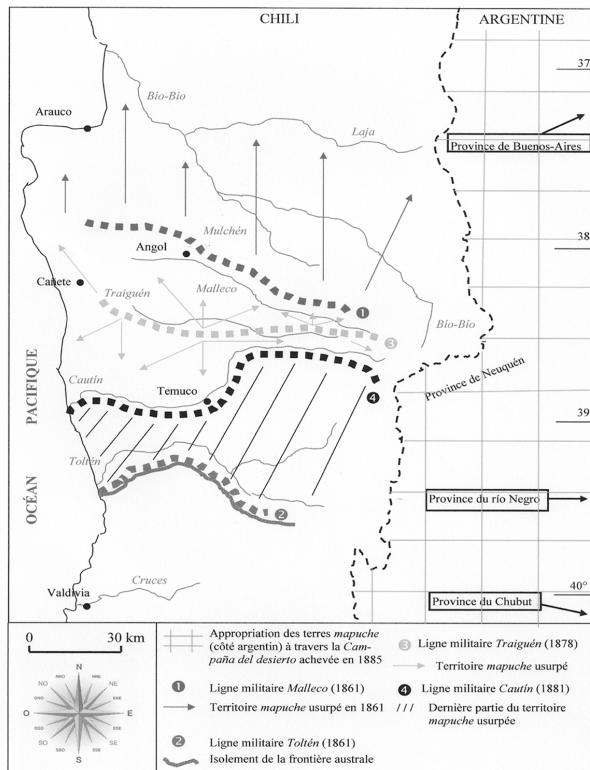


FIGURE 1.1 – Carte de l'usurpation progressive du territoire Mapuche (1810-1885)⁸

le gouvernement. En 1866, les premières lois d'occupation ont été adoptées en modifiant la nature du « territoire indigène » pour « territoire de colonisation ». Elle entérine officiellement le processus colonial entrepris par l'État chilien avec la nomination de Cornelio Saavedra, initiateur de cette expansion, au poste d'intendant d'Arauco⁹. Celui-ci entama la fortification de la région tout en multipliant les lois facilitant l'appropriation des terres jusqu'au bord des rivières Malleco River et Toltén River. Durant ces 20 années, le conflit enferme progressivement le peuple Mapuche aux confins des terres araucaniennes et de la Cordillère des Andes, perdant ainsi 90,7% de son territoire original¹⁰.

Le conflit autour de l'Occupation de l'Araucanie qui dura jusqu'en 1881, n'est pas un épiphénomène au sein d'un processus d'extension territoriale et d'affirmation politique. Il est un tournant progressif majeur en enterrant définitivement une identité fondée sur un État-nation pluriel, comme nous l'évoque l'historien Pablo Mariman Quemenado :

« Il s'est terminé l'idée d'un État qui abhorre la diversité qui le constitue, fondant la relation et la situation coloniales avec des peuples désormais catégorisés comme "communautés indigènes"¹¹ »

9. J. Bengoa, *La memoria olvidada...*, p .314.

10. *Ibid.*, p .350.

11. « Concluida esta, se termina consumando la idea de un Estado que abomina de la diversidad que lo constituye, fundando la relación y situación colonial con pueblos que fueron categorizados en adelante como "comunidades indígenas" » P. M. Quemenado, « La geoestrategia en el conflicto chileno mapuche... »

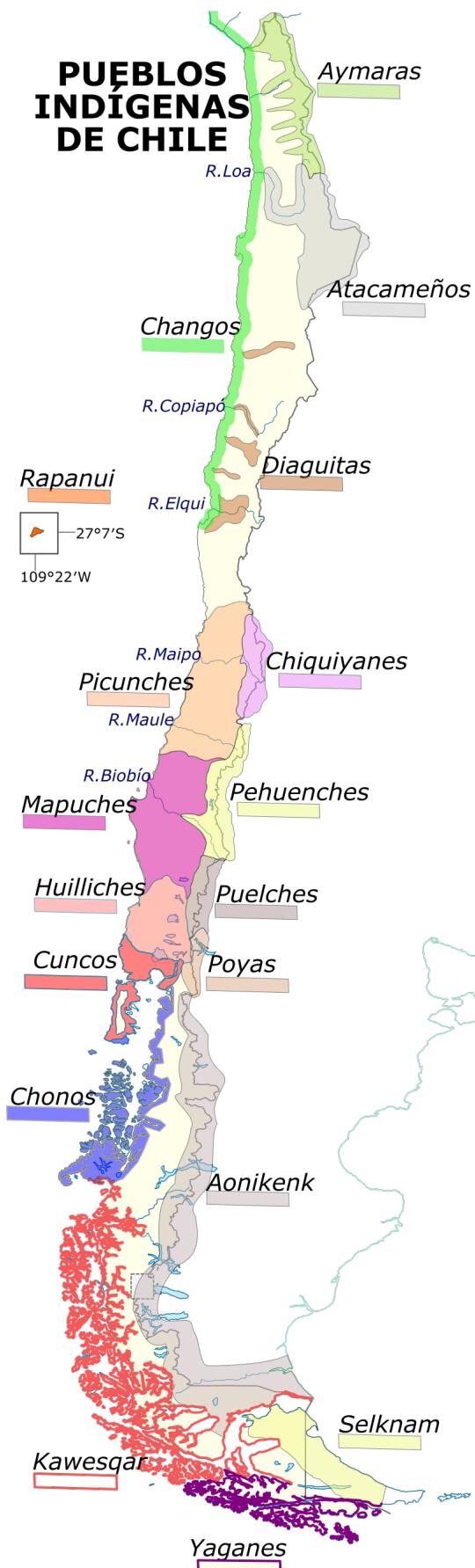


FIGURE 1.2 – Carte schématique des principaux peuples autochtones existants ou ayant existé au sein du territoire chilien actuel - ©Wikipedia

1.1.2 La question Mapuche : symbole des problèmes socio-politiques du Chili contemporain

À l'heure actuelle, la lutte pour la légitimité des droits territoriaux et politiques Mapuche est devenue un symbole de mobilisation ethno-sociale à la fois sur le plan national qu'international. Sa lutte singulière s'est progressivement transformée en une source d'inspiration culturelle inaltérable et presque romantique face à l'étatisme et la capitalisme néo-libéral¹².

Dès 1910, un premier mouvement autochtone Mapuche se forme à travers le collectif *Sociedad Caupolicán de Defensa de la Araucanía*. Au cours du XX^e siècle, de nombreuses organisations politiques voient le jour avant de disparaître selon les tournants de l'histoire politique chilienne. Ce conflit latent accompagne au contraire le débat et les querelles sur fond d'inimitié ou au contraire de solidarité. Cette réclamation territoriale est devenue multiple et s'étend jusqu'aux territoires argentins, là où la culture et l'influence Mapuche demeure : « en renouant avec les connexions transandines de l'époque coloniale, le déploiement des systèmes circulatoires contemporains à l'échelle du cône sud-américain redonne une indéniable matérialité¹³ ».

Si des territoires exclusifs et protégés ont été déterminés afin de reconnaître certains droits aux peuples indigènes présents sur le territoire chilien. Cette rédemption hésitante reste constamment sous le feu des critiques politiques et économiques. Aujourd'hui, si l'appartenance ethnique et les traditions culturelles se perpétuent, cette population est confrontée à de nombreux problèmes socio-économiques représentant environ un million de personnes¹⁴. En effet, la plupart des territoires hors réserves indigènes, restent soumis au monopole d'une grande industrie forestière, agricole et piscicole réfractaire face à la perte de leurs capitaux¹⁵. Cette dépossession économique alimente les conflits locaux, mais aussi nationaux.

Ces crispations sociales se sont exacerbées à partir des années 1990, où l'utilisation de la violence politique comme une pratique de contestation connaît un léger regain suite à la fébrilité de la loi Indigène de 1993. Au sein des sphères politiques et médiatiques se dresse progressivement une criminalisation de la lutte politique mapuche, dépeignant son intensification comme une menace contre l'État¹⁶. Sous fond d'une politique de discrimination raciale, la politique répressive de la contestation va être étendue par l'emploi de la

12. B. Sepúlveda, « Le pays mapuche, un territoire « à géographie variable » »...

13. *Ibid.*

14. J. Bengoa, « Los Mapuches : historia, cultura y conflicto », *Cahiers des Amériques latines*, 2011/3–68 (68[2011]), p. 89-107, DOI : 10.4000/cal.118.

15. María José Andrade, « La lucha por el territorio mapuche en Chile : una cuestión de pobreza y medio ambiente », *L'Ordinaire des Amériques*–225 (225[2019]), DOI : 10.4000/orda.5132.

16. Zunilda Carvajal-del Mar, « La criminalisation du conflit Mapuche : l'application discriminatoire de la loi antiterroriste chilienne », *Archives de politique criminelle*, 36–1 (2014), p. 213-226, DOI : 10.3917/apc.036.0213.

loi antiterroriste promulguée en 1984 sous de la dictature de Pinochet¹⁷.

Depuis le début des années 2000, l'État chilien prétend mettre fin avec cette politique répressive des contestations avec un peuple chilien qui souhaite démontrer une plus large volonté de conciliation¹⁸. Néanmoins, les tensions localisées persistent dans la région de Biobío et de l'Araucanie. Certains territoires de la région de Wallmapu sont encore récemment soumis à un état d'exception juridique et une militarisation de la région. Ces tensions encore très vives démontrent la continuité d'un conflit historique dont les enjeux sociaux et politiques persistent.

1.1.3 La sous-collection de l'Araucania

La "Colección Manuscritos" est l'un des fonds les plus importants détenus par le centre ACAB, avec plus de 2200 documents manuscrits datant de 1642 à 1952. Cette collection s'est progressivement constituée grâce à des dons d'universitaires et de politiciens liés à l'Universidad de Chile, mais aussi grâce à diverses acquisitions au cours du XX^e siècle¹⁹. Ces documents se regroupent autour de différentes grandes thématiques de l'histoire du Chili et dont une grande partie a été produite par des figures éminentes de la vie politique, militaire et intellectuelle (Andrés Bello, Manuel Montt, etc.). De par le prestige de ses producteurs et de l'importance historique de ces documents, ce fonds a été classé comme « Monument historique » par le décret n°295-2009 du ministère de l'Éducation du Chili durant l'année 2009²⁰.

Les archives concernant l'Occupation de l'Araucanie en constituent une part essentielle. Ce sous-fonds spécifique autour des archives de l'Occupation de l'Araucanie a été inventorié partiellement à partir de la fin des années 2000, dans le prolongement de la création de la collection "Colección Manuscritos". Cette première classification et la restauration parcellaire et progressive à permis de mettre à jours des centaines d'archives, suscitant l'intérêt des milieux scientifiques et patrimoniaux.

En 2013, un important travail de numérisation a été mis en place, axée sur la pacification de l'Araucanie, grâce à l'octroi d'un fonds du programme ADAI (Apoyo al Desarrollo de los Archivos Iberoamericanos) qui permettra la numérisation de 249 documents, pour un total de 530 pages. Ce programme a permis de donner suite à une première transcription manuelle les années suivantes. Ce projet d'édition numérique s'inscrit donc

17. *Ibid.*

18. T. Nouaille, « L'indépendance du Chili... ».

19. Voir le registre de l'inventaire de "Colección Manuscritos", *Archivo Central Andrés Bello*, url : http://archivobello.uchile.cl/content/Registro%20Guia/2016/enero/registro_guia_coleccion_manuscritos.pdf, consulté le 02/08/2022.

20. Décret n°295-2009 du ministère de l'Éducation du Chili, *Declara Monumento Nacional en la categoría de Monumento Histórico las colecciones Neruda, Americana y Manuscritos, pertenecientes al Archivo Central Andrés Bello, de la Universidad de Chile*, promulgué le 5 août 2009 et publié le 5 septembre 2009

dans cette continuité en s'appuyant sur les différents travaux précédents.

En observant plus en détail, les archives numérisées du sous-fonds de la "Colección Manuscritos", nous pouvons remarquer que l'essentiel de l'activité de la production documentaire se concentre sur la période 1859-1860. Cette période de transition se situe plus exactement entre la fin des « guerras civiles "montistas" » (1850-1859) dont l'année 1859 est marquée par le soulèvement général de nombreuse tribut mapuche. Cette répression fut suivi d'une campagne de militarisation et pacification de la région dirigée par Nicolas Saavedra. Il mit ensuite en œuvre du plan d'occupation de la région Biobío à partir de 1861²¹. Le recensement de l'activité se concentre surtout la fin de l'année 1859 à février 1860, avec un certain regain pour le mois de juin et octobre 1860 (voir annexe A, figure A.3). Toutefois, on retrouve de nombreux documents s'éparpillant sur l'ensemble de la seconde moitié du XIX^e siècle, bien qu'une part essentielle n'est pas encore pu être numérisé.

1.2 La construction d'un jeu de donnée : vers une utilisation tout terrain

1.2.1 Préparation des données numérisées et leur inventorisation

La constitution d'un jeu de données homogène et structuré a été une étape importante dans la préparation, à la fois pour s'immerger et comprendre les informations et les subtilités qui en résident. Comme nous l'avons vu précédemment, ce corpus documentaire autour des archives de l'« Occupation de l'Araucanie » a déjà été le fruit d'un long travail de révision et d'inventorisation minutieux de la part de l'équipe en charge du processus d'archivage. L'ensemble des informations ont été recueilli sous la forme d'un tableur Excel (Word office) en rassemblant des données sur chaque pièce de ce fonds partiel : date, côtes et identifiant des numérisations, indications géographiques et humaines, type, nombre.

Toutefois, un certain nombre d'éléments ne furent pas adaptés à une exploitation machine, mais davantage à une interprétation humaine. Un travail de réconciliation, d'uniformisation et d'épuration des données a été mis en place en extrayant les données sous le format CSV. Ce format fut ainsi interprétable par le logiciel Dataiku DSS. Cette plateforme de développement intégré est destinée à l'ensemble des besoins dans le cadre traitement et analyse de la donnée. Dataiku a ainsi permis le retypages des données initiales, mais aussi de quantifier les archives numérisées²². Les enjeux de ce traitement se

21. J. Bengoa, *Historia Del Pueblo Mapuche (Siglo XIX y XX)*, Ediciones Sur, Santiago, Chili, 1987, p. 165-171.

22. L'ensemble du flux de travail est disponible à l'adresse suivante : <https://github.com/>

déclinent en trois objectifs :

- Inventaire des fichiers permettant de vérifier si certains fichiers ne sont pas manquants. Il fut ensuite possible de les classer en fonction de la nature et la typologie des documents.
- La conversion des images. À l'origine, le service de numérisation de ACAB a produit les images sous le format Tagged Image File Format (TIFF) (à 1200 dpi) permettant une qualité maximale. C'est un format d'image matricielle permettant de ne pas compresser l'image et ainsi de garder celle-ci la plus authentique possible en limitant la perte des données. Ce standard au sein des institutions patrimoniales reste en revanche un objet extrêmement lourd avec une moyenne autour de 60 mo (mégaoctets). Des projets similaires ont démontré qu'une telle qualité n'était pas nécessaire et qu'une image compressée est amplement suffisante²³. Ce script shell a permis d'automatiser la conversion de l'ensemble des images vers le format Joint Photographic Group (JPG)²⁴.
- Faciliter l'interprétation des données dans un travail d'analyse ou comme support à la production de métadonnées lors de son édition.

Cette première étape permet de mettre en place une stratégie, évolutive néanmoins quant à la constitution du set de données mais aussi sur la mise en place de certains stratagèmes afin de récupérer la donnée le plus efficacement possible avec une perte marginale. Notre jeu de données disponibles est en effet assez restreint au vue de certains autres projets de reconnaissance d'écritures manuscrites ; il est donc impératif de maximiser celle-ci. Il faut pour cela essayer de sélectionner les plus qualitatives, représentatives et homogènes possibles. En observant les différents graphiques, on peut apercevoir la sur-représentation des lettres, et secondairement les notes manuscrites au sein du corpus documentaire (graphique A.2). Nous avons donc pris la décision de centraliser les efforts sur ces deux typologies, avant d'étendre le processus en fonction du résultat. Dans un deuxième temps, la distribution des archives montrent une grande pluralité d'auteurs au sein de ce fonds. En revanche, six personnes se manifestent comme amplement plus prolifiques que la moyenne (graphique A.1).

Proyecto-Ocupacion-Araucania-UChile/Data_preparation.

23. N.C., *Expérimentations – LECTAUREP*, URL : <https://lectaurep.hypotheses.org/category/experimentations> (visité le 20/08/2022) ; Floriane Chiffolleau, *DAHN Project : Digital Edition of Historical Manuscripts*, 22 juin 2022, URL : <https://github.com/FloChiff/DAHNProject/blob/e19fbc38476305a941ff7f5a6db1a32f26a9acf5/Correspondence/Guidelines/Documentation-Correspondance.pdf> (visité le 22/08/2022).

24. voir annexe 4.

1.2.2 Constitution d'un corpus interopérable

À la suite de différentes recherches sur l'existence de *dataset*, il est rapidement apparu que la langue espagnole occupait une position plus marginale dans les domaines de la recherche et de l'ingénierie autour de l'Handwritten Text Recognition (HTR). Si quelques projets ont été mis en place, l'essentiel se tourne vers la recherche fondamentale sur les réseaux neuronaux ou ne sont pas librement accessibles²⁵. De la même manière, des plateformes telles que HTR-United ne recensent que très peu de données disponibles sur cette langue, ou alors inadaptées à la construction d'un modèle adapté ACAB.

Face à ces premiers constats, l'édification d'un modèle dédié paraît donc indispensable, ce qui nécessite de construire au préalable un lot de données d'entraînement. Le premier enjeu de cette automatisation est en réalité triple puisqu'il s'agit d'établir les règles structurelles, épistémologiques et analytiques des transcriptions et de l'interprétation machine. La qualité des données utilisées et leur cohérence sont à la source du succès des prédictions, dans le cadre d'un apprentissage machine supervisée²⁶.

En effet, des erreurs d'acquisition ou de transformation liées à des fautes humaines ou techniques peuvent considérablement ralentir, voire mettre à mal la poursuite du projet. La performance de cette chaîne de traitement, mais aussi la réutilisation à moindre coût énergétique et de temps va dépendre d'une analyse préalable des sources. La dimension d'interopérabilité doit permettre l'utilisation multiple de ce jeu de donnée sur l'ensemble du processus de traitement : segmentation, reconnaissance d'écriture manuscrite et reconnaissance d'entités nommées.

La constitution d'un corpus documentaire dans le but de l'exploiter dans un projet d'édition numérique, nécessite donc un set quantitatif, mais surtout qualitatif afin de procéder à une restitution de l'information la plus complète possible. Si l'aspect quantitatif semble de prime abord plus que négligeable en comparaison d'autres projets, plusieurs études ont démontré l'efficacité de jeu de données à la taille plus modeste tout en permettant un coût marginal²⁷. La principale difficulté réside dans la segmentation et non la reconnaissance de caractères. Par la suite, les chercheurs Phillip Benjamin Strobel, Simon Clematide et Martin Volk ont démontré qu'un jeu de données de plus de 50 pages était suffisant pour avoir des prédictions pouvant être qualifiées de bonnes sur le moteur

25. À cet égard, on peut remarquer que l'*Universitat Politècnica de València* est particulièrement actif à ce sujet comme le démontre les récentes publications : Emilio Granell, Edgard Chammas, Laurence Likforman-Sulem, Carlos-D. Martínez-Hinarejos, Chafic Mokbel et Bogdan-Ionuț Cîrstea, « Transcription of Spanish Historical Handwritten Documents with Deep Neural Networks », *Journal of Imaging*, 4-1 (1[2018]), p. 15, DOI : 10.3390/jimaging4010015 ; Salvador España-Boquera et María José Castro-Bleda, « A Spanish Dataset for Reproducible Benchmarked Offline Handwriting Recognition », *Language Resources and Evaluation*, 56 (1^{er} sept. 2022), p. 1-14, DOI : 10.1007/s10579-022-09587-3

26. Cf. chapitre 4

27. Joan Andreu Sánchez, Verónica Romero, Alejandro H. Toselli, Mauricio Villegas et Enrique Vidal, « A Set of Benchmarks for Handwritten Text Recognition on Historical Documents », *Pattern Recognition*, 94 (oct. 2019), p. 122-134, DOI : 10.1016/j.patcog.2019.05.025.

Kraken²⁸.

Au vu de nos précédentes analyses, il est frappant de voir la forte hétérogénéité de notre corpus ce qui induit la nécessité d'un modèle polyvalent en conséquence d'un fonds extrêmement composite. Ces données préliminaires vont ainsi révéler un ensemble de facteurs et problématiques à prendre en compte afin de tendre vers une automatisation la plus efficiente possible²⁹.

Nom	Pages	Dates extrêmes	Particularités	Données test
Cornelio, Saavedra	43	nov. 1859 - août 1877	style propre, bon état, présence de notes post-scriptum	Non
Avello, Juan	12	nov. 1859 - nov. 1860	style propre, quelques feuilles abimées, présence de notes post-scriptum	Non
Díaz, José Del Carmen	21	déc. 1859 - mars 1860	style propre parfois condensé, bon état	Non
Escala, Manuel Segundo	13	nov. 1859 - nov. 1860	style très propre, bon état	Non
García Videla, Daniel	8	mai 1860 - juin 1860	style très propre, bon état	Non
Pérez Rosales, Vicente	12	déc. 1860 - nov. 1860	style parfois hésitant et condensé, bon état	Non
Sepúlveda, José	8	déc. 1859 - nov. 1860	style intermédiaire, quelques feuilles abimées, présence de notes post-scriptum	Oui
Villalón, Vicente	22	nov. 1859 - nov. 1859	style propre, parfois effacé, présence de notes post-scriptum	Non
Contreras, Juan	16	mars 1860 - juin 1860	style convenable, quelques feuilles légèrement abimées	Non
Echantillon composite	25	nov. 1859 - nov. 1859	16 mains, bon état	Non
TOTAL	180			

TABLE 1.1 – Détails du jeu de données sélectionné

Cet échantillonnage a été réalisé, de manière progressive et empirique, à partir de multiples mains en s'appuyant sur le schéma du projet « LECTAUREP³⁰ » mis en place par l'équipe ALMAAnaCH de l'Institut national de recherche en sciences et technologies du numérique (INRIA). En premier lieu, le corpus s'est construit autour de plusieurs lots alternant entre 10 et 15 numérisations par auteur au nombre 6. Petit à petit, d'autres transcriptions ont été agrégées au lot déjà existant. À partir de là, la stratégie d'apprentissage s'est appuyée sur un lot conséquent d'une main unique (plus de 40 documents), un nombre de mains variables avec une production documentaire modérée voire élevée et un lot s'appuyant sur de très courtes transcriptions venant de nombreuses mains différentes. Ce choix vise à permettre un apprentissage stylistique et graphologique extrêmement varié et ainsi permettre la production d'un modèle le plus polyvalent possible en s'appuyant sur une méthodologie dite générique³¹.

28. Phillip Benjamin Ströbel, Simon Clematide et Martin Volk, « How Much Data Do You Need ? About the Creation of a Ground Truth for Black Letter and the Effectiveness of Neural OCR », dans *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, 2020, p. 3551-3559, URL : <https://aclanthology.org/2020.lrec-1.436> (visité le 21/08/2022).

29. Alix Chagué, Victoria Le Fournier et Manuela Martini, « Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ? », dans *Colloque DHNord 2019 "Corpus et archives numériques"*, MESHS Lille Nord de France, 2019, URL : <https://hal.inria.fr/view/index/identifiant/hal-02448921>.

30. *LECTure Automatique de REpertoire*; pour plus d'informations, voir <https://lectaurep.hypotheses.org/>.

31. Ariane Pinche, « CREMMLAB Project : Handwritten Text Recognition (HTR) for Medieval Manuscripts », dans *Digital Humanities 2022*, Tokyo, Japan, 2022, URL : <https://hal.archives-ouvertes.fr/hal-03719504> (visité le 21/08/2022).

1.3 Disséquer l'image et reconstruire l'information : une étape préliminaire à la transformation éditoriale

La segmentation est un principe général du traitement et de l'analyse optique d'une image afin d'obtenir la reconnaissance des régions et des lignes d'écriture. De nombreux projets d'éditions numériques s'appuient justement sur cette reconnaissance pour transformer une image brute en édition numérique native. La mise en place de ce procédé nécessite au préalable une clarification des enjeux sémantiques et en définissant au préalable un protocole d'annotation et de transcription. L'emploi d'une telle méthode doit ainsi permettre une automatisation de l'extraction de l'information dans son contexte, s'apparentant à une analyse diplomatique, et vérifier la solidité du processus de transformation.

1.3.1 Ontologie d'une procédure de segmentation d'un document

Comme nous pourrons l'observer plus en détail au sein du chapitre 4, l'un des principes novateurs de l'HTR est l'analyse de la mise en page (*layout analysis*) dans le processus de reconnaissance des écritures. Cette étape consiste à détecter les zones d'intérêts dans une image, les différencier et spécifier leur intérêt : ligne de texte, régions, réclames, marges. On déconcentre le processus d'apprentissage³². C'est-à-dire que l'écriture n'est plus l'unique point de fixation, c'est désormais l'ensemble de la page qui est prise en compte pour en attraper l'information substantielle.

Cette technologie dans le domaine de la reconnaissance d'écriture offre ainsi un double avantage : la reconnaissance d'écriture manuscrite complexe et une reconnaissance morphologique de l'image au travers de son processus de segmentation. Ces régions et ses bases de lignes vont permettre de recontextualiser le texte au moyen d'une structuration sémantique de la donnée. L'intérêt est de conserver la structure diplomatique initiale en vue de l'encodage éditorial du texte.

Comme nous pouvons le constater à travers de cette image reprenant la codification des régions, la stratégie adoptée s'est appuyée sur une délimitation sémantique, au détriment de la segmentation centrée sur une perception visuelle (voir figure 1.3). Si cette méthode permet de conserver au maximum le sens de la donnée initiale, elle se fait au détriment d'une cohérence visuelle. On peut observer que de nombreuses lignes sont régulièrement rapprochées entre elles, voir superposées. Cette structuration graphique peut donner lieu à un certain nombre de confusions lors du processus de segmentation machine.

32. Noémie, Lucas et Doria Le Fur, *OCR / HTR et graphie arabe*, 3, GIS Moyen-Orient et Mondes musulmans, 2022, URL : <http://majlis-remomm.fr/72481> (visité le 19/08/2022), p. 25.

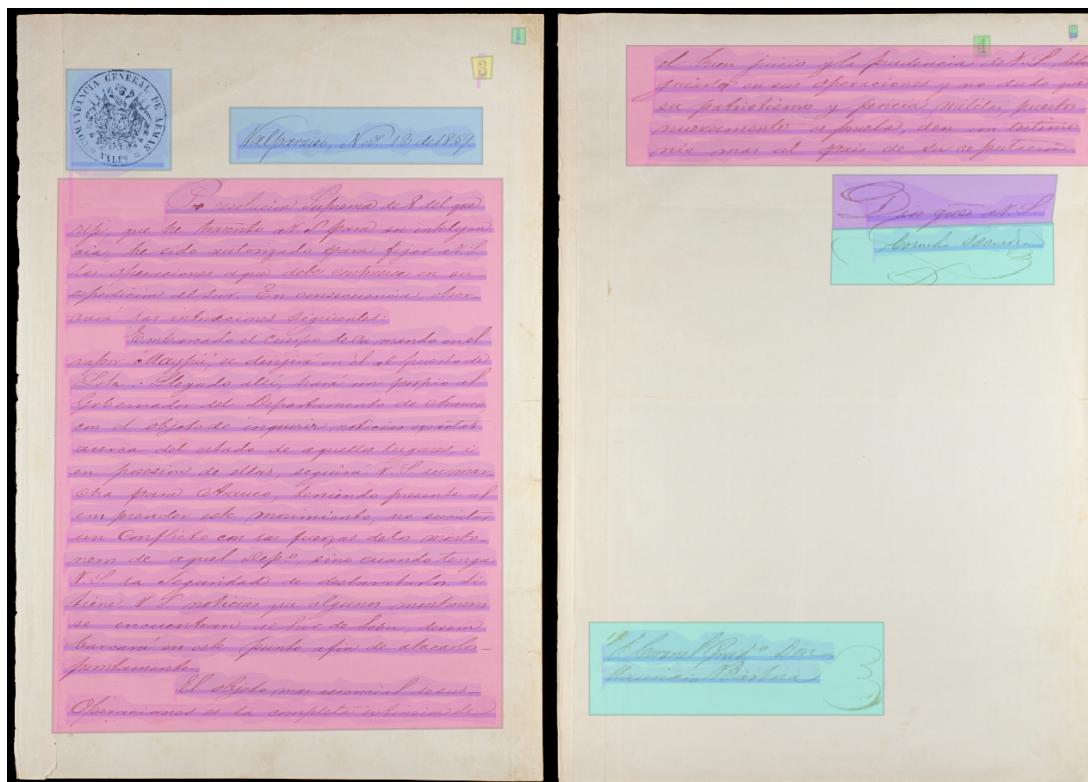


FIGURE 1.3 – Démonstration d'une segmentation d'image sur l'application eScriptorium

D'autres projets d'édition numérique ont eu pour enjeux la structuration de données autour de nature similaire à celui du projet du centre ACAB. Les travaux de Floriane Chiffoleau et Anne Baillot sur le projet de Dispositif de soutien à l'Archivistique et aux Humanités Numériques (DAHN) ont ainsi pu expérimenter l'élaboration d'une ontologie des corpus littéraires et mettre à disposition des *guidelines* décrivant les principes généraux³³. Sur ce constat et en s'appuyant sur une première appréciation du corpus documentaire, une ontologie a pu être dessinée dans le but d'annoter les différentes régions et les différentes lignes détectées . Celle-ci se décline sur 14 types de régions et 3 types de lignes présentées ci-dessous :

— Régions :

1. *CustomZone :Address* – Zone indiquant le destinataire
2. *CustomZone :Dateline* – Zone indiquant le contexte d'écriture (date et lieu)
3. *CustomZone :Object* – Zone indiquant l'objet du document
4. *MainZone :SaluteConclude* – Zone indiquant la salutation conclusive
5. *MainZone :SaluteIntro* – Zone indiquant la salutation introductory
6. *MainZone :Text* – Zone indiquant le corps du texte

33. F. Chiffoleau, DAHN Project...

7. *MarginTextZone :commentary* – Zone indiquant les commentaires en marge du texte
8. *MarginTextZone :note* – Zone indiquant les notes additionnelles
9. *NumberingZone :id* – Zone indiquant le numéro d'identification (*a posteriori*)
10. *NumberingZone :other* – Zone indiquant un numéro additionnel (*a posteriori*)
11. *NumberingZone :page* – Zone indiquant le numéro de page (*a posteriori*)
12. *QuireMarksZone :signature* – Zone indiquant la signature de l'auteur
13. *StampZone :graphic* – Zone indiquant une estampille graphique
14. *StampZone :manuscript* – Zone indiquant une estampille manuscrite

— **Lignes :**

1. *DefaultLine* – Ligne par défaut
2. *InterlinearLine :commentary* – Ligne indiquant un commentaire entre deux lignes
3. *InterlinearLine :correction* – Ligne indiquant une correction entre deux lignes

Cette composition doit permettre la description de trois natures de documents : les lettres, les notes et les circulaires. Si ces trois objets ont une portée sémantique différente, il en reste que cette sélection possède des caractéristiques structurelles très similaires pouvant rapidement être associées et ainsi alléger l'ontologie. Comme nous pouvons le constater, l'ontologie a été amplement enrichie au modèle initial du projet DAHN. Les particularités de notre corpus hétérogène et l'alignement sur la méthodologie de l'initiative A Controlled Vocabulary to Describe the Layout of Pages (SegmOnto) ont amené à reconstruire et remodeler cette base ontologique.

Avant tout créé pour parachever l'étude des manuscrits médiévaux et des imprimés anciens, le projet SegmOnto, né en 2021, est une réponse de différents chercheurs aux besoins d'un modèle efficace et de faire coexister la taxonomie des numérisations et l'approche éditoriale du texte³⁴. Cette initiative universitaire offre un vocabulaire contrôlé permettant une description graphique d'un texte sous sa forme la plus essentielle. L'ontologie est construite autour de 15 zones principales et 6 types de lignes ; auxquelles peut être additionné un complément d'information par l'utilisation des signes « : » ou « # » en suffixe³⁵. En outre, l'utilisation de ce lexique renforce l'interopérabilité des données entre

34. Simon Gabay, Jean-Baptiste Camps, A. Pinche et Claire Jahan, « SegmOnto : Common Vocabulary and Practices for Analysing the Layout of Manuscripts (and More) », dans *1st International Workshop on Computational Paleography (IWCP@ICDAR 2021)*, Lausanne, Suisse, 2021, URL : <https://hal.archives-ouvertes.fr/hal-03336528> (visité le 22/08/2022).

35. Id., *SegmOnto, A Controlled Vocabulary to Describe the Layout of Pages*, version 0.9, Ecole Nationale des Chartes, Université de Genève, 2021, URL : <https://github.com/SegmOnto> (visité le 22/04/2022).

les projets et ainsi permettre non seulement de créer des outils numériques pérennes, mais aussi de donner la possibilité de leur réutilisation.

1.3.2 XML-ALTO : un encodage structuré adapté à l’océrisation

Afin de comprendre la portée et l’utilisation possible des données possibles, il est nécessaire de comprendre le format d’expression de celle-ci. Au sein de l’univers de la reconnaissance d’écriture, il y a deux grands standards eXtensible Markup Language (XML) qui accompagnent principalement et structurent les résultats Optical Character Recognition (OCR) : le format XML ALTO et le format Page Analysis and Ground truth Elements (PAGE). Dans notre cas, nous avons choisi de nous appuyer sur le premier standard qui est le plus couramment utilisé, en particulier au sein des institutions patrimoniales³⁶. Il offre deux avantages : la conservation des données coordonnées géométriques et la superposition de l’image.

Pour revenir à la base, XML est un métalangage de structuration de données à balise publié en 1999 par le consortium World Wide Web Consortium (W3C). L’atout principal de ce langage est sa grande permissivité et sa grande extensibilité avec son système à chevron. ALTO est un standard dérivé de XML défini par un schéma aujourd’hui maintenu par la *Library of Congress*³⁷. Il a été développé à partir de 2004 afin de répondre au besoin naissant de l’OCR afin de pouvoir décrire la mise en page des documents et assurer la conservation à long terme des données.

Un fichier XML ALTO est composé de trois sections principales : <**Description**> permettant de décrire les métadonnées, <**Styles**> renseigne les données de styles (police, etc.) et <**Layout**> est la section décrivant les différents éléments de mise en page. Dans notre cas, les deux éléments les plus intéressants sont les éléments <**TextBlock**> et <**TextLine**> comme nous pouvons constater à travers cet extrait du code sources d’un résultat HTR (voir le code 1). L’élément <**TextBlock**> indique les différentes données liées à la région dont l’attribut **@TAGREFS** affiche la référence de la région, les autres attributs sont dédiés aux informations géométriques condensées. Le sous-élément <**Polygon**> permet de décrire le dessin du masque de la zone sur l’axe X/Y. De la même manière, <**TextLine**> est réservé aux données autour de la ligne. Au sein du sous-élément <**String**>, l’attribut **@CONTENT** rend compte de la transcription effectuée.

Récemment un certains nombres d’outils ont été mis en place afin d’appuyer la soli-

36. Bertrand Caron et Etienne Cavalié, *Formats de données pour la préservation à long terme : la politique de la BnF*, 1, Paris, France, Bibliothèque nationale de France, 2021, URL : https://www.bnf.fr/sites/default/files/2021-04/politiqueFormatsDePreservationBNF_20210408.pdf, p. 74.

37. *Ibid.*

```

^^I      <TextBlock HPOS="842"
          VPOS="3078"
          WIDTH="1161"
          HEIGHT="271"
          ID="eSc_textblock_493a77e7"
          TAGREFS="BT3114">
<Shape><Polygon POINTS="883 3128 842 3342 2003 3349 1999
→   3078"/></Shape>

<TextLine ID="eSc_line_1f8ebff3"
          TAGREFS="LT1061"
          BASELINE="1264 3203 1951 3209"
          HPOS="1261"
          VPOS="3079"
          WIDTH="690"
          HEIGHT="156">
<Shape><Polygon POINTS="1264 3203 1264 3140 1315 3140 1318
→   3140 1388 3079 1388 3079 1391 3079 1391 3079 1391 3079
→   1395 3079 1395 3079 1395 3079 1398 3079 1398 3079 1401
→   3079 1468 3095 1573 3079 1573 3079 1576 3079 1576 3079
→   1576 3079 1579 3079 1579 3079 1579 3079 1703 3140 1789
→   3111 1792 3111 1792 3111 1792 3111 1795 3111 1795 3111
→   1798 3111 1798 3111 1798 3111 1801 3111 1801 3111 1801
→   3114 1833 3143 1833 3146 1951 3146 1951 3209 1948 3235
→   1261 3235"/></Shape>
^^I      <String CONTENT=" J del C. Díaz"
          HPOS="1261"
          VPOS="3079"
          WIDTH="690"
          HEIGHT="156"></String>
</TextLine>

</TextBlock>

```

Listing 1 – Structuration d'un fichier ALTO

dité des données. Dans notre cas, la jeune application HTRVX développée et promue par HTR-United a permis de vérifier le schéma des données produites, la bonne concordance entre les zones et les lignes et de valider l'utilisation du vocabulaire SegmOnto³⁸. De ce fait, il fut très facilement possible de l'intégrer au sein du *workflow* GitHub afin de vérifier l'apport de nouvelles de données et la bonne synchronisation entre elles.

1.3.3 Définir un protocole de transcription et d'annotation

Au cours de la préparation de données, l'ensemble des différentes transcriptions ont été effectuées sur l'application eScriptorium afin d'en extraire des fichiers ALTO exploitable lors de nos futurs entraînements. Cette phase s'est appuyée sur la première approche de Cecilia del Carmen Ramallo Díaz et son analyse des sources autour de l'Araucanie³⁹. Néanmoins, les transcriptions des documents ont été faites à travers une modernisation de la langue. Cela reste un frein majeur qui a nécessité de s'employer à de nouvelles transcriptions littérales.

Au cours de la première phase, un ensemble de stratagèmes a été mis en place afin de récupérer un maximum de particularités depuis le texte original. Il s'agit des mots soulignés, barrés ou directement corrigés, des caractères illisibles ou partiellement lisibles et enfin des lettres suscrites. L'idée initiale était de pouvoir récupérer les différentes données et leurs évolutions grâce à l'aide d'un ensemble d'Expression régulière (REGEX). Une REGEX, comme son nom l'indique, est un motif de caractères permettant de capturer ou échapper un à plusieurs éléments souhaités. Elle s'appuie sur une syntaxe particulière permettant de décrire la fonctionnalité souhaitée. Par exemple, pour récupérer le contenu d'un mot barré tel que Saavedra signalé sous la forme *Saavedra*, cela se traduit par le motif suivant : *([A-Za-zA-ÖØ-öø-ÿ -]+)*.

Toutefois au vu des difficultés apparentes de l'interprétation machine lors des premiers essais effectués, il a été choisi de considérablement restreindre la codification des transcriptions. Une simplification qui s'est appuyée sur la méthodologie employée par le projet LECTure Automatique de REPertoire (LECTAUREP)⁴⁰. Si la première des raisons s'explique par l'intention d'optimiser les résultats, l'autre facteur était de pouvoir s'aligner sur les données du projet afin de les agréger à notre corpus et donc multiplier les données utilisables. Ainsi, seulement trois règles ont été appliquées :

- lettre suscrite : ê
- mot illisible : xxx

38. Thibault Clérice, A. Chagué et Pauline Jacsont, *HTR-United/HTRVX : HTR Validation with XSD*, version 0.0.10, HTR-United, mars 2022, URL : <https://github.com/HTR-United/HTRVX> (visité le 18/04/2022).

39. Cecilia Carmen Ramallo Díaz (del), *Transcripción de Los Documentos de La Serie “Pacificación de La Araucanía” (Colección Manuscritos)*, Santiago, Chili, Archivo Central Andrés Bello, 2014, p. 30.

40. Marc Durand, Aurélia Rostaing et A. Chagué, *Notaires de Paris - Répertoires, Ground Truth for Various Parisian Registries of Notary Deeds (French 19th and 20th Centuries)*, HTR-United, 2021, URL : <https://github.com/HTR-United/lectaurep-repertoires> (visité le 23/08/2022).

1.4. LES ENJEUX DU DROIT DU PATRIMOINE ET DE L'OPEN DATA AU CHILI⁴¹

— nouveau paragraphe : ¶.

Cette dernière règle a été ajoutée afin de pouvoir délimiter plus facilement les différents paragraphes lors du processus d'édition. Les différentes expériences HTR ont démontré le besoin de simplifier les différentes règles de transcriptions afin d'être le plus optimal possible. Toutefois, l'utilisation du pied-de-mouche renversé n'a pas été conservée lors de la récupération des données sous un format texte (txt)⁴¹.

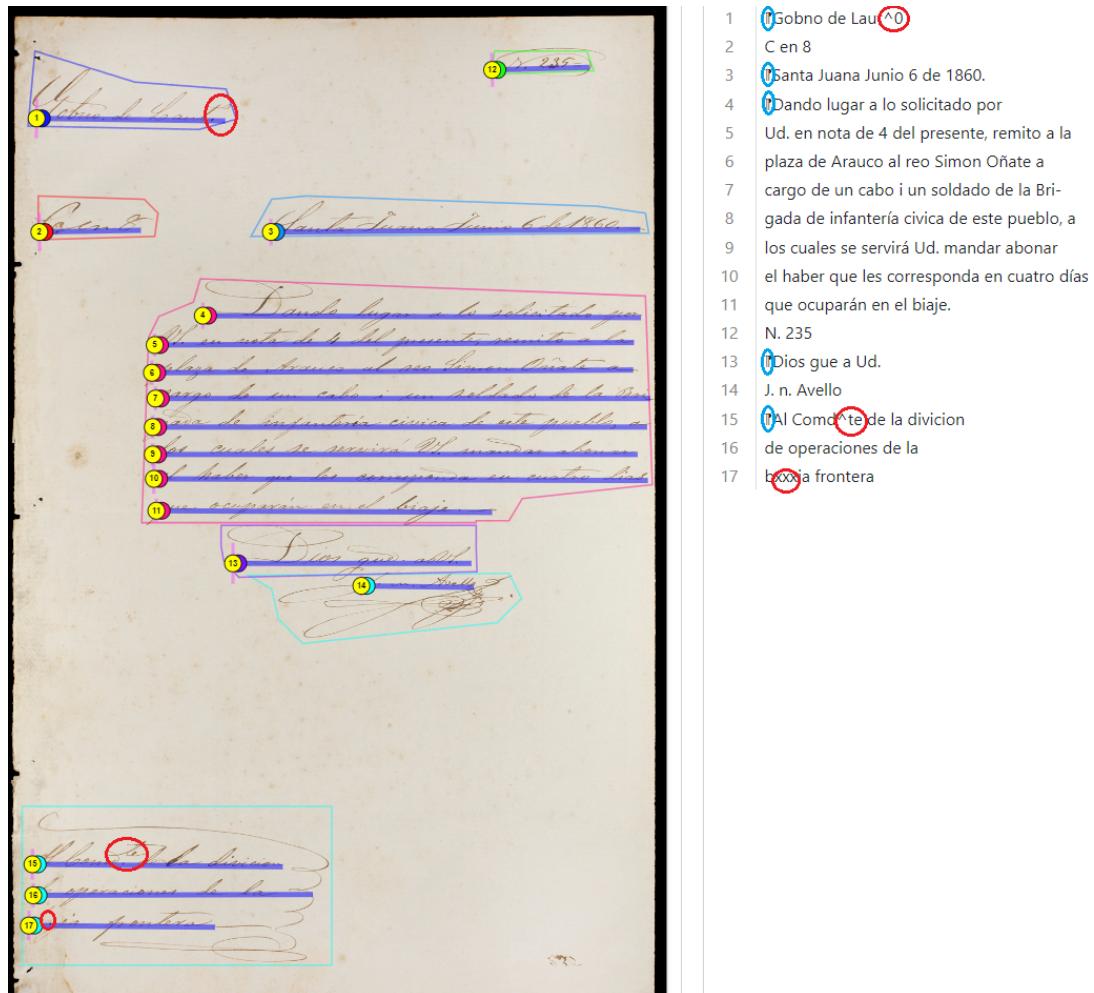


FIGURE 1.4 – Visualisation des règles de transcription

1.4 Les enjeux du droit du patrimoine et de l'*open data* au Chili

De nos jours, l'*Open data* est devenu un enjeu majeur des institutions patrimoniales en permettant de faciliter l'accessibilité, le partage et la valorisation des données produites.

41. Maxime Humeau, *Preprocessing HTR*, avec la coll. d'Alessandro Chiaretti, Archivo Central Andres Bello, avr. 2022, URL : <https://github.com/Proyecto-Ocupacion-Araucania-UChile/model-HTR/tree/main/Preprocess>.

Un mouvement scientifique est né autour de la volonté d'ouvrir la donnée afin de : « rendre la science plus accessible, démocratique, transparente et bénéfique pour tous⁴² ». Une note de l'UNESCO (Organisation des Nations unies pour l'éducation, la science et la culture) révèle cet engouement général en Amérique latine afin de faire face aux problèmes d'accès aux ressources scientifiques, mais aussi à la valorisation de leur propres travaux⁴³. Toutefois, l'*Open data* reste confronté à de nombreux besoins scientifiques, économiques et techniques et juridiques⁴⁴. À travers ces observations, il s'agit de revenir sur les défis que traverse le développement d'un projet numérique autour des archives chiliennes.

1.4.1 Retour sur la législation des archives au Chili

En 2019, Pierre Fabry, ancien élève archiviste-paléographe, amorce quelques pistes de réflexion autour des enjeux archivistiques, sur l'accès et la transparence de l'information par rapport au mouvement de l'*Estallido social* qui traverse de plein fouet l'ensemble du territoire chilien⁴⁵. En pleine réflexion, l'institution *Archivo Nacional* émet trois problèmes majeurs auxquels le Chili doit se confronter : la mise en place d'un service d'archive électronique centralisé, le développement des réseaux archivistiques sur l'ensemble du territoire (qui ne comprend que trois institutions publiques majeures) et surtout la mise en place d'une véritable loi sur les archives.

Le développement des institutions gestionnaires des archives publiques a été considérablement mis à mal suite au coup d'État du 11 septembre 1973 et la mise en place d'un système dictatorial sous l'égide du général Pinochet (1915-2006). Une grande part des documents étatiques a été détruit durant cette période afin d'effacer les traces des exactions du régime. La transition démocratique n'a pas été l'occasion pour les différents gouvernements successifs d'avoir la volonté ni l'autorité suffisantes pour ouvrir les archives et de légiférer autour⁴⁶.

Aujourd'hui, l'appareil juridique autour des archives est en réalité une législation composite regroupant un ensemble de décret et circulaires ministérielles encadrant les institutions publiques d'archivages, notamment le centre *Archivo Nacional*. Sur le plan législatif, le système des archives reste régi par la loi de 1929 dont les effets sont aujourd'hui

42. *Open Science in Latin America and the Caribbean : A Strong Tradition with a Long Journey Ahead*, UNESCO, 2020, URL : <https://en.unesco.org/news/open-science-latin-america-and-caribbean-strong-tradition-long-journey-ahead> (visité le 05/09/2022).

43. *Ibid.*

44. Bernard Jacquemin, Joachim Schöpfel et Renaud Fabre, « Libre accès et données de recherche. De l'utopie à l'idéal réaliste », *Études de communication. langages, information, médiations*–52 (52[2019]), p. 11-26, DOI : 10.4000/edc.8468.

45. Pierre Fabry, *Archives, Archivistes et Crise Au Chili, Quelques Réflexions*, Ecole des chartes, janv. 2020, URL : https://ecoledeschartes.tumblr.com/post/190494197307/archives-archivistes-et-crise-au-chili-quelques?is_related_post=1 (visité le 13/09/2022).

46. Bruno Groppo, « Chapitre v – Les archives des droits humains. Documenter la répression et la résistance au Chili et en Argentine », dans *Documenter les violences : Usages publics du passé dans la justice transitionnelle*, dir. Camille Goirand et Angélica Müller, Paris, 2020 (Travaux et mémoires), p. 131-149, URL : <http://books.openedition.org/iheal/8887> (visité le 13/09/2022).

obsoletes. Malgré la loi de 2008 sur la transparence de la vie publique, les institutions des archives condamnent les manquements persistant sur la définition des archives, de leur conservation et de leur accès. La directrice de l'époque de *Archivo Nacional*, Emma de Ramón déclare : « no existe un marco legal que los ampare, proteja, reglamente y ordene⁴⁷ ».

Principalement, les archives restent soumises à la volonté des institutions productrices ou de conservations qui vont définir leurs propres politiques patrimoniales. Seuls les fonds notamment classés « Monuments historiques » sont soumis à un certain nombre de restrictions avancées, prohibant notamment la destruction de documents⁴⁸. Aujourd’hui, l'accès aux archives est un enjeu majeur de cette transition démocratique et sociale du Chili, à l'image du premier projet de constitution de l'Assemblée constituante⁴⁹. Ce projet inscrit directement le droit d'accès aux archives au sein de la constitution, notamment avec les articles 24-5 et 162-2. Ils sont révélateurs d'une question en suspens et des besoins sociaux autour pour la transparence de la vie publique, mais aussi historique sur la question des droits indigènes et de la mémoire durant la dictature.

1.4.2 Libéraliser l'accès aux données numériques

La question des archives et de la production documentaire des écosystèmes électro-niques a fait le fruit d'une série d'encadrements juridiques entre les années 2002 et 2004. Ils explicitent les bases des enjeux de conservations et les compétences de l'institution *Archivo Nacional* sur ce domaine, bien que rudimentaire⁵⁰.

La publication des données et la mise à disposition des données numériques des archives restent à la libre appréciation des institutions archivistiques, restreintes par le droit de la propriété intellectuelle, la législation sur les données personnelles et la loi sur la transparence des données publiques. De plus, si le producteur reste propriétaire des données, il ne possède aucun droit sur la base de données et son utilisation⁵¹. Cette situation confuse ne facilite pas la propagation de l'*Open data*, qui reste souvent impopulaire auprès des institutions. Si les enjeux techniques, juridiques et politiques restent un problème

47. « il n'existe pas de cadre juridique pour les défendre, les protéger, les réglementer et les ordonner » in Billet de l'Universidad de Chile, *Archivos en Chile : la necesidad de una ley que proteja la memoria de nuestro país*, 20 juin 2016, <https://www.uchile.cl/noticias/122827/archivos-en-chile-y-la-necesidad-de-una-ley-que-proteja-la-memoria->, consulté le 5 septembre 2022.

48. Ley N° 18.845 *Establece sistemas de microcopia o micrograbacion de documentos*, promulgée le 19 octobre 1989 et publiée le 3 novembre 1989

49. Le projet a finalement été refusé au cours du référendum national sur l'approbation du projet de constitution pour la République du Chili le 3 septembre 2022.

50. Fernández Acevedo et Fernando J, « El Documento Electrónico En El Derecho Civil Chileno : Análisis de La Ley 19.799 », *Ius et Praxis*, 10-2 (2004), p. 137-167, DOI : 10.4067/S0718-00122004000200005.

51. N.C., *Manual de Datos Abiertos*, Santiago, Chili, Comisión Nacional de Investigación Científica y Tecnológica, 2014, URL : <http://datoscientificos.cl/files/manual-2014.pdf>, p. 23.

certain, elle est aussi d'ordre culturelle dans un pays où la culture de la propriété reste très forte⁵².

Mettre en place un projet sur les principes de la science ouverte et les principes FAIR (Trouvable, Accessible, Interopérable, Réutilisable) ne tient pas seulement au ressort des dispositions techniques. C'est aussi un travail pédagogique sur les enjeux et les bienfaits des données ouvertes pour la science et l'institution. Au contraire d'une négation du travail accompli, l'ouverture numérique est un instrument efficace pour la valorisation de celui-ci, tout en permettant de multiplier les efforts conjoints. Les licences internationales *Creative Commons* sur la protection de l'utilisation des données produites sont un excellent outil de confiance pour la mise à disposition en libre accès. Dans notre cas, nous avons privilégié la licence *Attribution-NonCommercial-ShareAlike 4.0 International* permettant le partage et l'utilisation des données sous condition de situation et à usage non commercial⁵³.

Pour la mise à disposition des fichiers ALTO produits, nous nous sommes appuyés sur l'initiative *HTR-UNITED* qui souhaite faciliter l'accès et la réutilisation des données HTR⁵⁴. Les différents outils mis en place tel que *HTRVX* permettent à la bonne interopérabilité entre les formats et les ontologies et ainsi vérifier la compatibilité avec ces propres données terrains. L'organisation souhaite ainsi favoriser le partage massif des données et des modèles autour de l'HTR grâce à la standardisation des dépôts (*via HTRUCS*), leurs recensements et faire correspondre à une charte de qualité⁵⁵. Dans ce même registre, nous avons publié également les données sur la plateforme *Zenodo* qui est un répertoire de travaux de recherche, de logiciel et de données.

Il est clair que la mise en place d'infrastructures et d'organisations communes facilite amplement la mise à profit des données, et ce tant du point de vue technique que morale. Ces efforts communs de référencement, d'accessibilité et d'interopérabilité sont d'autant plus importants, car ils sont une opportunité de pérenniser et démocratiser les projets numériques à travers une réduction des coûts qu'ils impliquent et une circulation des savoirs⁵⁶.

52. Id., *Estado del arte nacional e internacional en materia de gestión de datos de investigación e información científica y tecnológica y recomendaciones de buenas prácticas*, Santiago, Chili, Comisión Nacional de Investigación Científica y Tecnológica, 2010, p. 108, URL : <http://datoscientificos.cl/files/ufro-2010.pdf>, p. 43.

53. <https://creativecommons.org/licenses/by-sa/4.0/>, consulté le 01/0/2022.

54. *HTR-United*, url : <https://htr-united.github.io/>, consulté le 14/08/2022.

55. A. Chagué, « Conditions de La Mutualisation : Les Principes FAIR et HTR-United », dans *Humanistica 2022*, Montréal, Canada, 2022, URL : <https://hal.inria.fr/hal-03685731> (visité le 13/09/2022).

56. *Ibid.*

Chapitre 2

L'apprentissage machine et la reconnaissance de texte

Les technologies de reconnaissance d'écriture automatique s'ancrent dans une histoire longue. Elle remonte avant même la naissance de l'Intelligence Artificielle (IA) et l'article fondateur d'Alan Turing « Computing Machinery and Intelligence »¹. C'est en 1929 que l'ingénieur allemand Gustav Tauschek développe la première technologie pouvant être affiliée au domaine de la reconnaissance manuscrite optique. Toutefois, c'est véritablement à partir des années 1960-1970 que l'OCR connaît ses premiers résultats scientifiques satisfaisants, encourageant une première commercialisation en 1978².

L'intérêt suscité pour cette technologie se fait rapidement remarquer au sein des sphères scientifiques et patrimoniales tant la reconnaissance automatique de caractères ouvre de nouvelles possibilités. Malgré des avancées prometteuses, l'OCR est encore loin d'offrir des résultats suffisamment exploitables aussi bien du point de vue méthodologique qu'éditorial comme le relate l'historien Gian Piero Zarri³. Ce fantasme de la reconnaissance des écritures par la machine ne se réalise véritablement que depuis une dizaine d'années. Durant cette décennie, on observe une multiplication de projets patrimoniaux et scientifiques autour de la reconnaissance textuelle. Hugo Scheithauer rattache cette démocratisation de la reconnaissance de texte avec le développement de la plateforme Transkribus, émergeant au début des années 2010 dans le cadre d'un financement européen⁴.

1. Alan Turing, « Computing Machinery and Intelligence », *Mind*, LIX–236 (1^{er} oct. 1950), p. 433-460, DOI : 10.1093/mind/LIX.236.433.

2. Khalaf Alkhalf, « OCR-Based Electronic Documentation Management System », *International Journal of Innovation, Management and Technology*, 5–6 (2014), DOI : 10.7763/IJIMT.2014.V5.560.

3. Gian Piero Zarri, « Quelques aspects techniques de l'exploitation informatique des documents textuels : saisie des données et problèmes de sortie », *Publications de l'École Française de Rome*, 31–1 (1977), p. 399-413, URL : https://www.persee.fr/doc/efr_0000-0000_1977_act_31_1_2286 (visité le 24/08/2022).

4. Hugo Scheithauer, *La Reconnaissance d'entités Nommées Appliquées à Des Données Issues de La Transcription Automatique de Documents Manuscrits Patrimoniaux. Expérimentations et Préconisations à Partir Du Projet LECTAUREP*, mémoire de master "Technologies numériques appliquées à l'histoire",

À travers ces observations primaires, l'ébullition de la Reconnaissance d'Écriture Manuscrite (REM) remet profondément en question les anciennes pratiques et les anciens avis au sein du paysage scientifique et culturel. À partir du projet de l'Araucanie, nous allons observer comment la reconnaissance de texte a pu bénéficier au traitement de ces sources, mais aussi les limites atteintes de cette technologie.

Ce chapitre souhaite développer les différents enjeux et les défis qu'offre l'apprentissage machine autour de l'édition numérique, notamment par le prisme de la reconnaissance des écritures. Si initialement la généralisation de l'OCR s'exerce dans un objectif d'exploitation industrielle, l'intérêt des milieux archivistiques et patrimoniaux s'observe facilement avec la multiplication de l'impression numérique⁵. Les progrès exponentiels qui ont été réalisés depuis plus d'une dizaine d'années ont permis de donner un second souffle à la reconnaissance automatique de l'écriture et plus particulièrement l'écriture manuscrite. Ce renouveau favorise l'édification de projets d'envergures au sein des humanités afin de contribuer à l'extraction et l'exploitation de l'information numérique.

2.1 État scientifique et technique autour de la reconnaissance de texte

Aujourd'hui, la reconnaissance de texte est une catégorie générale se déclinant en un éventail de technologies et de techniques. On peut y distinguer deux grandes taxinomies : l'OCR et HTR (*alias* la REM). La reconnaissance automatique de texte est une catégorie du *machine learning* consistant à permettre à l'ordinateur l'interprétation de données textuelles à partir d'une image (imprimées ou manuscrites) en une suite de caractères numériques.

Ce domaine de pointe se rattache ainsi à l'écosystème de l'apprentissage machine et plus particulièrement au traitement automatique de l'image (*Digital image processing* en anglais), la reconnaissance de forme (*Pattern recognition* en anglais) et au Traitement Automatique des Langues (TAL) (*Natural language processing* en anglais).

2.1.1 Écriture, matrice et *deep learning*

Comme nous l'avons entrevu précédemment, on distingue deux grands types au sein de la reconnaissance de texte : l'OCR centré autour des textes imprimés et l'HTR pour les documents manuscrits. Au cours des années 1960, les premières avancées résident dans le changement ontologique de l'image en s'appuyant sur la puissance de calcul des

dir. Alix Chagué et Thibault Clérice, Paris, École nationale des chartes, 2021, URL : https://github.com/HugoSchtr/memoire_TNAH_M2_HugoScheithauer.

5. Jean-François Moufflet, « 5 ans d'expérimentation de la technologie HTR aux Archives nationales », dans *Futurs Fantastiques*, Paris, France, 2021, p. 39, URL : https://www.bnf.fr/sites/default/files/2022-01/futurs_fantastiques_moufflet.pdf.

ordinateurs. Grâce aux procédés du traitement de l'image, celle-ci est alors binarisée au travers de ses couleurs primaires ou des nuances de gris selon la méthode. C'est-à-dire que l'image est alors convertie en une suite de valeurs numériques (entiers ou décimaux) de multiples dimensions, dans ce qu'on appelle une matrice (voir figure 2.1)⁶. Comme le rappelle le groupe de chercheur de l'Université de Washington, les techniques de seuillage (*Thresholding*) sont alors essentielles dans le processus de traitement pictural puisqu'elle permettent de segmenter l'image à partir d'une valeur booléenne, obtenant ainsi un filtrage des pixels la composant⁷.

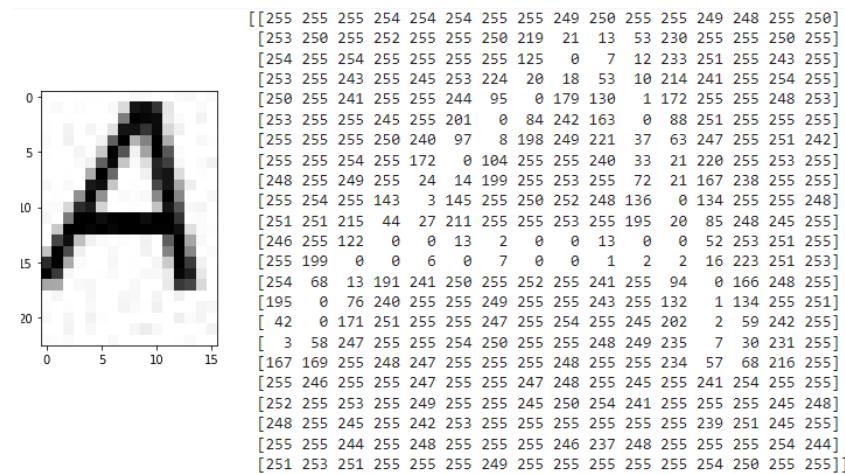


FIGURE 2.1 – Transposition d'un caractère en matrice (niveau de gris)

Cette transposition de l'image en une matrice décimale facilite la détermination d'un caractère *via* l'application de modèles statistiques et mathématiques. L'une des premières méthodes appliquées dans le domaine de l'apprentissage machine est la méthode *k-NN*(Méthode des *k* plus proches voisins)⁸. Pour la définir, cette technique d'apprentissage supervisé réside sur l'alignement de données étiquetées, en établissant la donnée équivalente la plus proche (la distance *k*) de notre caractère *x*. À partir des années 1980, les laboratoires travaillant sur la reconnaissance d'écriture réutilisent les avancées dans le domaine de la linguistique et de ses modèles arithmétiques (*Hidden Markov Model*) s'appuyant sur des systèmes d'occurrences et de nouveaux modèles de segmentation⁹.

6. Un exemple permettant de comprendre le seuillage et la binarisation est disponible. Voir annexe C

7. Maya R. Gupta, Nathaniel P. Jacobson et Eric K. Garcia, « OCR Binarization and Image Pre-Processing for Searching Historical Documents », *Pattern Recognition*, 40–2 (févr. 2007), p. 389-397, DOI : 10.1016/j.patcog.2006.04.043.

8. L. Terriel, *Représenter et Évaluer Les Données Issues Du Traitement Automatique d'un Corpus de Documents Historiques. L'exemple de La Reconnaissance Des Écritures Manuscrites Dans Les Répertoires de Notaires Du Projet LectAuRep.* mémoire de master « Technologies numériques appliquées à l'histoire », dir. Alix Chagué et Thibault Clérice, Paris, École nationale des chartes, 2020, URL : https://github.com/Lucaterre/L-TERRIEL_memoireDeStage_M2TNAH_ENC, p. 39.

9. *Ibid.*, p. 40.

Reconnaissance de texte et apprentissage profond

De nos jours, les techniques de reconnaissance de texte se sont considérablement améliorées. Ce perfectionnement est dû à l'explosion du *deep learning* (apprentissage profond) en 2012, suite à la victoire incontestable d'un de ces modèles lors d'un concours de classification d'images. Le procédé n'est pas nouveau puisqu'on peut remonter son origine à 1943, où le neurologue et le logicien Warren S. McCulloch et Walter Pitts décrivent une première application « neuronale » de la machine de Turing¹⁰. Toutefois, la parenté de l'apprentissage profond est encore contestée puisque le chercheur Charles C. Tappert rattache davantage sa première conceptualisation à Frank Rosenblatt et son invention du perceptron¹¹.

Le *deep learning* reste en marge au sein de l'intelligence artificielle pendant de nombreuses années en raison de la puissance de calcul nécessaire à son fonctionnement intrinsèque. Malgré tout, certaines équipes de chercheurs comme celle dirigée par Yann Le Cun continuent de développer la recherche fondamentale, permettant en 2012 d'exploser les records aux yeux du monde.

L'apprentissage profond se veut donc à la frontière de l'informatique, des mathématiques et de la neurobiologie en reprenant le modèle du neurone et du système de stimulation synaptique. Cette artificialisation numérique s'appelle le perceptron *alias* le neurone simple. Pour reprendre la métaphore, le neurone reçoit une quantité d'information (les poids) dont la somme pondérée doit atteindre un certain seuil afin d'être stimulée. Cette stimulation applique une fonction d'activation ou de transfert envoyant une information de sortie (voir figure 2.2). À partir de ce système, des réseaux de neurones sont alors mis en place afin de traiter des données complexes, et ce avec une spécialisation progressive. De nos jours, l'algorithme du perceptron simple est dépassé au profit de systèmes améliorés tel que le perceptron multicouche, consistant en une série de neurones simples et cachés, ou le *Support Vector Machine*¹².

À l'image de l'ensemble du domaine de l'intelligence artificielle, l'apprentissage profond a connu une véritable ébullition dans le secteur de la reconnaissance automatique de

10. Warren S. McCulloch et Walter Pitts, « A Logical Calculus of the Ideas Immanent in Nervous Activity », *The bulletin of mathematical biophysics*, 5–4 (1^{er} déc. 1943), p. 115-133, DOI : 10.1007/BF02478259.

11. Il met en place un système probabiliste de stockage et de classifications de l'information s'appuyant sur le modèle d'un réseau neuronal Charles C. Tappert, « Who Is the Father of Deep Learning ? », dans *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019, p. 343-348, DOI : 10.1109/CSCI49370.2019.00067.

12. Pour prolonger la curiosité, voici trois articles permettant d'expliquer certains principes généraux et des cas applicatifs : William S Noble, « What Is a Support Vector Machine ? », *Nature Biotechnology*, 24–12 (déc. 2006), p. 1565-1567, DOI : 10.1038/nbt1206-1565 ; M. W Gardner et S. R Dorling, « Artificial Neural Networks (the Multilayer Perceptron)—a Review of Applications in the Atmospheric Sciences », *Atmospheric Environment*, 32–14 (1^{er} août 1998), p. 2627-2636, DOI : 10.1016/S1352-2310(97)00447-0 ; Hassan Ramchoun, Mohammed Amine, Janati Idrissi, Youssef Ghanou et Mohamed Ettaouil, « Multilayer Perceptron : Architecture Optimization and Training », *International Journal of Interactive Multimedia and Artificial Intelligence*, 4–1 (2016), p. 26, DOI : 10.9781/ijimai.2016.415.

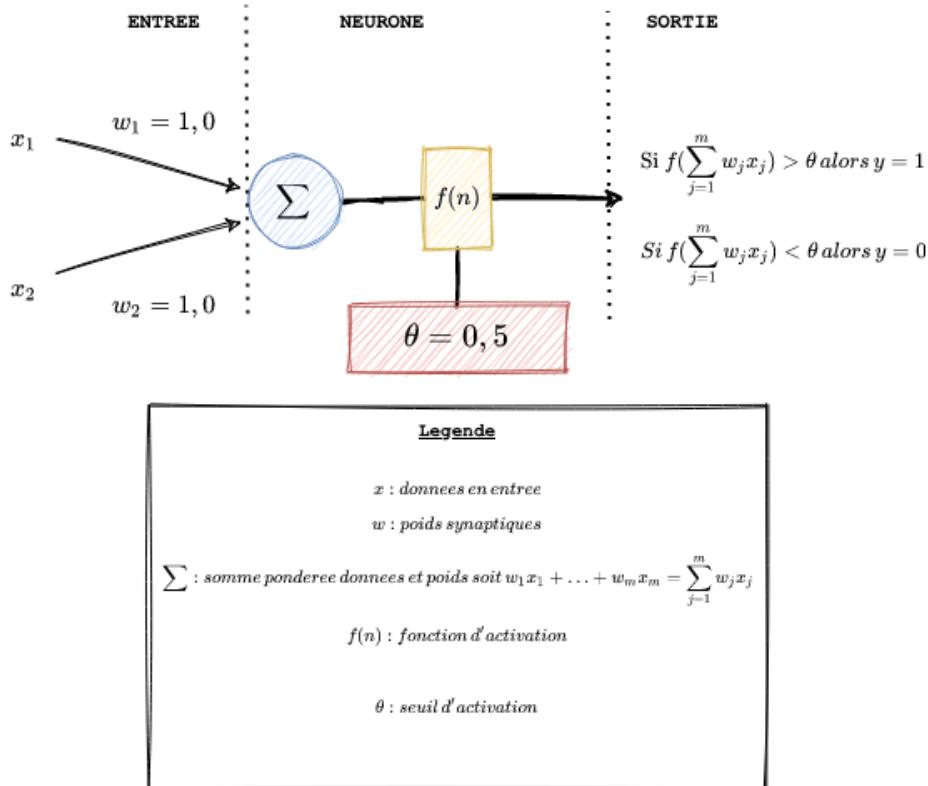


FIGURE 2.2 – Illustration simplifiée d'un neurone formel, ©Lucas Terriel, 2020

texte, et ce particulièrement après 2016¹³. Les résultats furent rapidement très satisfaisants grâce à la capacité des réseaux de neurones à traiter des données complexes et en sa capacité de mémorisation. C'est justement sur ce dernier point que certains laboratoires se sont appuyés pour améliorer les résultats concernant les écritures manuscrites en incitant l'utilisation d'une architecture Recurrent Neural Networks (RNN). Cette architecture traite l'information de manière cyclique et lui permet donc de la contextualiser. Elle a rapidement affiché des taux supérieurs à 90%, néanmoins au détriment d'une très imposante puissance de calcul¹⁴. Cette architecture reste encore aujourd'hui la plus commune, même si régulièrement présente sous forme hybride. Toutefois, plusieurs groupes de chercheurs ont signalé les problèmes sous-jacents de cette architecture à la fois sur le plan technique (la distorsion et l'explosion des très grandes images) et les besoins exponentiels en terme de puissance de calcul. Les systèmes hybrides couplés avec les architectures Convolutional Neural Networks (CNN) semblent remettre en cause la suprématie des réseaux RNN en permettant une réduction considérable du coût de calcul grâce à un système de segmentation des tâches, accentué par un traitement de l'image plus important (notamment le

13. Jamshed Memon, Maira Sami, Rizwan Ahmed Khan et Mueen Uddin, « Handwritten Optical Character Recognition (OCR) : A Comprehensive Systematic Literature Review (SLR) », *IEEE Access*, 8 (2020), p. 142642-142668, DOI : 10.1109/ACCESS.2020.3012542.

14. Alex Graves et Jürgen Schmidhuber, « Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks », dans *Advances in Neural Information Processing Systems*, 2008, t. 21, URL : <https://proceedings.neurips.cc/paper/2008/hash/66368270ffd51418ec58bd793f2d9b1b-Abstract.html> (visité le 25/08/2022).

processus d'augmentation)¹⁵.

2.1.2 La révolution technique de l'HTR

Comme nous l'avons vu précédemment, le *deep learning* a consolidé amplement les nombreuses avancées dans la reconnaissance de texte et plus particulièrement le secteur de l'HTR. Dans la majorité des projets, les techniques d'apprentissage se fondent sur l'apprentissage supervisé. C'est-à-dire qu'on fournit au processus d'apprentissage des données préalablement étiquetées afin de lui fournir un support d'apprentissage à partir duquel la machine va pouvoir déployer une méthode d'analyse, et ainsi classifier les caractères.

Néanmoins, la REM n'est pas simplement due à l'essor de l'apprentissage profond, même si les avancées techniques lui sont que corrélées. En effet, l'HTR correspond à un procédé de segmentation bien spécifique en comparaison des techniques portant sur l'OCR. Le point névralgique de cette technique de reconnaissance de texte réside dans la segmentation de l'image dans son intégralité, et non sur un centrage spécifique. À l'origine, le moteur établit un masque définissant la zone à reconnaître à partir d'une ligne de référence, souvent en bas¹⁶. Ce point d'appui permet d'exercer ce que l'on appelle la reconnaissance en-ligne ce qui signifie une conception des caractères cursifs dans l'espace et le temps par la machine. Le modèle, s'appuyant sur son expérience, va donc déchiffrer l'écriture selon ses mouvements et les comparer aux données terrain. D'autres méthodes tendent à remplacer le procédé par ligne par un système à polygone (*bounding boxes*) depuis quelques années à l'image des groupes de chercheurs de l'École Pratique des Hautes Études - Université PSL (EPHE) ou de l'Université autonome de Barcelone¹⁷. Comme l'expose la figure 2.3 s'appuyant sur le moteur kraken développé par l'EPHE, la reconnaissance des caractères s'appuie sur la détection de ligne afin de délimiter une zone de contour, un polygone permettant une meilleure détection et évaluation de l'objet¹⁸.

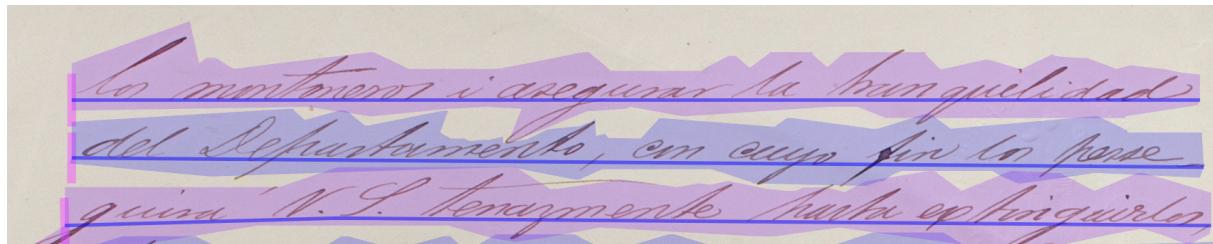
La reconnaissance de texte n'est donc qu'une étape au sein d'un moteur HTR. Si l'on regarde le schéma suivant (figure 2.4), on remarque trois étapes primordiales à l'obtention d'une prédiction qualitative. La détermination de cette chaîne de traitement a caractérisé

15. Joan Puigcerver, « Are Multidimensional Recurrent Layers Really Necessary for Handwritten Text Recognition ? », dans *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Kyoto, Japon, 2017, t. 01, p. 67-72, DOI : 10.1109/ICDAR.2017.20 ; Arthur Flor de Sousa Neto, Byron Leite Dantas Bezerra, Alejandro Hector Toselli et Estanislau Baptista Lima, « HTR-Flor : A Deep Learning System for Offline Handwritten Text Recognition », dans *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Recife/Porto de Galinhas, Brazil, 2020, p. 54-61, DOI : 10.1109/SIBGRAPI51738.2020.00016.

16. Noémie, Lucas et D. L. Fur, *OCR / HTR et graphie arabe...*, p. 25-26.

17. Manuel Carbonell, Alicia Fornés, M. Villegas et Josep Lladós, *A Neural Model for Text Localization, Transcription and Named Entity Recognition in Full Pages*, 4 mai 2020, arXiv : 1912.10016 [cs], URL : <http://arxiv.org/abs/1912.10016> (visité le 26/08/2022) ; Benjamin Kiessling, Daniel Stökl Ben Ezra et Matthew Thomas Miller, *BADAM : A Public Dataset for Baseline Detection in Arabic-script Manuscripts*, juill. 2019, URL : <https://hal.archives-ouvertes.fr/hal-02167164> (visité le 19/08/2022).

18. Noémie, Lucas et D. L. Fur, *OCR / HTR et graphie arabe...*, p. 26.

FIGURE 2.3 – Exemple de masque à partir d'une *baseline*

de nombreux progrès au sein de la REM et plus particulièrement au sein des écritures complexes usant de nombreux glyphes tel que l'arabe¹⁹. En outre, une phase de traitement des données de sorties est régulièrement ajoutée au sein des projets afin de maximiser les résultats.

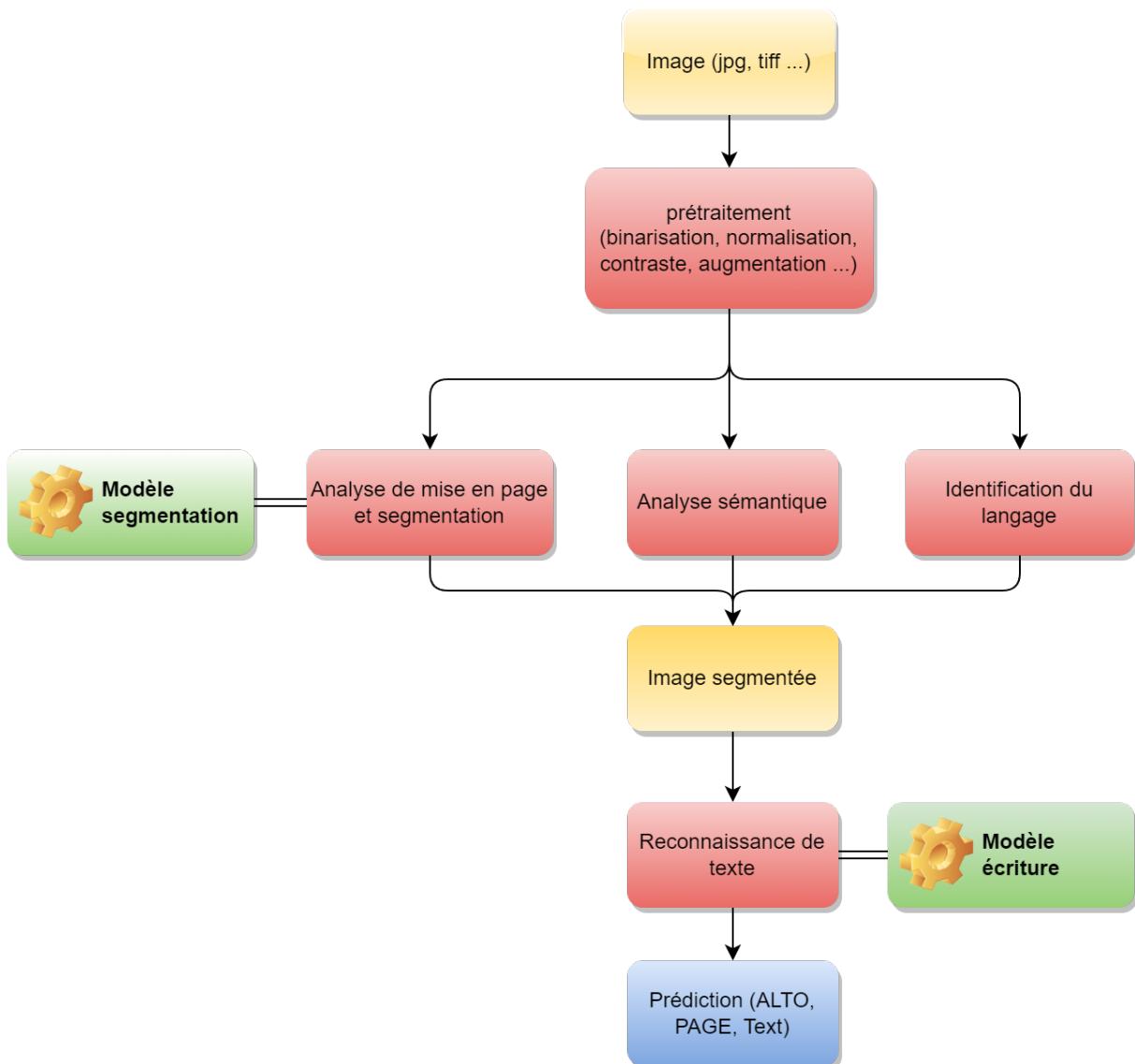


FIGURE 2.4 – Chaîne de traitement d'un procédé HTR

19. B. Kiessling, D. Stökl Ben Ezra et M. T. Miller, *BADAM...*

2.1.3 Kraken et eScriptorium : moteur et application pour l’HTR

Avec l’explosion du *deep learning* et de l’HTR, de nombreuses plateformes dédiées ont fait leur apparition au sein de la sphère de la REM. Une en particulier s’est rapidement imposée de part la qualité de ses résultats : *Transkribus*, avec le financement du projet européen READ-IT(Reading Europe Advanced Investigation Tools). Elle propose des modèles extrêmement complets, permettant d’obtenir des résultats très qualitatifs. Originalement gratuite, son utilisation est devenue payante au fil du temps, tournant de nombreux projets d’édition numérique vers d’autres outils *open source* tel qu’eScriptorium.

L’initiative du projet est née en 2019 par un groupe de chercheurs issus de l’institut de recherche Scripta de l’écosystème Paris Science et Lettres, et plus particulièrement de l’EPHE. Le but est de fournir une interface web de REM à destination de l’histoire scientifique et la plus ouverte possible, en réaction à l’opacité du projet Transkribus et son nouveau modèle économique²⁰. Le second objectif est de renforcer la REM sur un panel de langages beaucoup plus importants grâce au soutien des compétences du laboratoire. Ce besoin naît du constat que les principaux modèles sont essentiellement adaptés aux langues latines. Scripta doit permettre d’étendre le fonctionnement de la REM à des langues et des documents plus complexes comme le signale Peter A. Stokes :

« It has therefore been a crucial element of the project that the software must avoid, as far as possible, all assumptions about the nature of the writing and language that is in the system. The writing may be left to right, right to left, top to bottom or even bottom to top ; the support may be paper, parchment, but also stone, palm leaf, clay, wood, or many others ; it may be written with a pen, painted with a brush, inscribed with a chisel ; the writing system may be alphabetic, logographic, hieroglyphic ; and so on.²¹ »

La plateforme eScriptorium s’appuie sur le moteur OCR Kraken développé par Benjamin Kiessling. Kraken est un moteur clé en main interactif sous la forme d’un Interface en Ligne de Commande (CLI), optimisé pour les documents historiques et les textes en caractères non latins²². Basé sur le moteur Ocröpy, Kraken offre une très grande modularité lors de la conception de modèle de segmentation ou de reconnaissance de texte,

20. B. Kiessling, Robin Tissot, Peter Stokes et D. Stökl Ben Ezra, « eScriptorium : An Open Source Platform for Historical Document Analysis », dans *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, 2019, t. 2, p. 19-19, DOI : 10.1109/ICDARW.2019.90032.

21. « Un élément crucial du projet a donc été que le logiciel doit éviter, autant que possible, toute hypothèse sur la nature de l’écriture et de la langue qui se trouve dans le système. L’écriture peut être de gauche à droite, de droite à gauche, de haut en bas ou même de bas en haut ; le support peut être du papier, du parchemin, mais aussi de la pierre, de la feuille de palmier, de l’argile, du bois, ou bien d’autres encore ; elle peut être écrite à la plume, peinte au pinceau, inscrite au ciseau ; le système d’écriture peut être alphabétique, logographique, hiéroglyphique ; et ainsi de suite. » in Peter A. Stokes, B. Kiessling, D. Stökl Ben Ezra, R. Tissot et El Hassene Gargem, « The eScriptorium VRE for Manuscript Cultures », *Classics@Journal* (), URL : <https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/> (visité le 02/08/2022)

22. B. Kiessling, « Kraken - a Universal Text Recognizer for the Humanities », dans Utrecht , Pays-Bas, 8, DOI : 10.34894/Z9G2EX.

mais aussi de multiples schémas de sortie ce qui en fait un moteur rapidement intégral au sein d'une chaîne de traitement. Le moteur est construit sur deux architectures neuronales différentes. À l'origine, le système de détection des *baselines* utilise une architecture mixte CNN et LSTM (*long short-term memory*) qui est une catégorie des modèles RNN²³. Plusieurs études ont démontré la capacité de cette architecture neuronale mixte à limiter le nombre de vérités terrain nécessaires à la production d'un modèle²⁴. Ces expérimentations architecturales et ces observations empiriques ont permises une réelle amélioration de la REM.

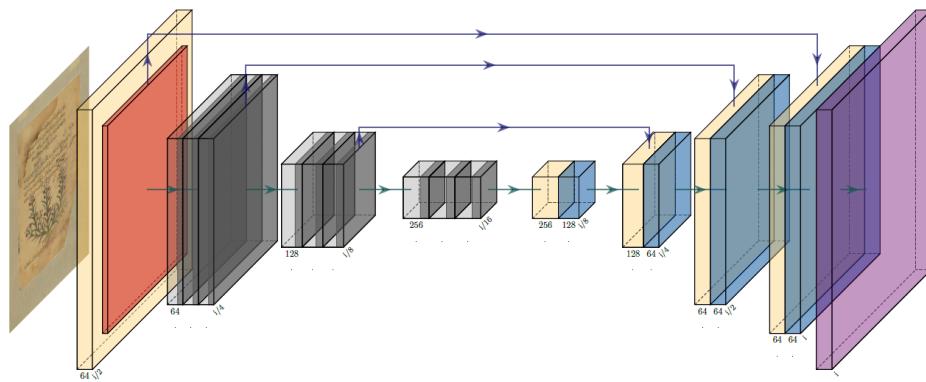


FIGURE 2.5 – Architecture CNN à échantillonnage et BLSTM de détection des *baselines* au sein du moteur Kraken - @Benjamin Kiessling, 2019

2.2 Méthodologie appliquée à la production d'un modèle HTR

La plateforme eScriptorium et le moteur Kraken ont été au centre du processus de transposition numérique des données autour des archives de l'Occupation de l'Araucania. La volonté initiale d'orienter son moteur vers le traitement des documents historiques ouvre ainsi de multiples possibilités pour la production d'une chaîne de traitement dédiée à l'édition numérique historique. En outre, le choix de procéder à une application ouverte, gratuite et modulable renforce l'accessibilité de cette technologie à des projets plus

23. A. H. Toselli, Si Wu et David A. Smith, « Digital Editions as Distant Supervision for Layout Analysis of Printed Books », dans *International Conference on Document Analysis and Recognition*. Springer, Lausanne, Suisse, 2021, t. 12822, p. 462-476, DOI : 10.1007/978-3-030-86331-9_30, arXiv : 2112.12703 [cs].

24. José Carlos Aradillas, Juan José Murillo-Fuentes et Pablo M. Olmos, « Boosting Handwriting Text Recognition in Small Databases with Transfer Learning », dans *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2018, p. 429-434, DOI : 10.1109/ICFHR-2018.2018.00081, arXiv : 1804.01527 [cs, stat] ; Adeline Granet, Emmanuel Morin, Harold Mouchère, Solen Quiniou et Christian Viard-Gaudin, « Transfer Learning for Handwriting Recognition on Historical Documents », dans *7th International Conference on Pattern Recognition Applications and Methods (IC-PRAM)*, Madère, Portugal, 2018, URL : <https://hal.archives-ouvertes.fr/hal-01681126> (visité le 27/08/2022).

modestes et expérimentaux. Dans ce cadre, nous allons observer comment a été conçu à modèle REM adapté à ces archives à partir du moteur Kraken.

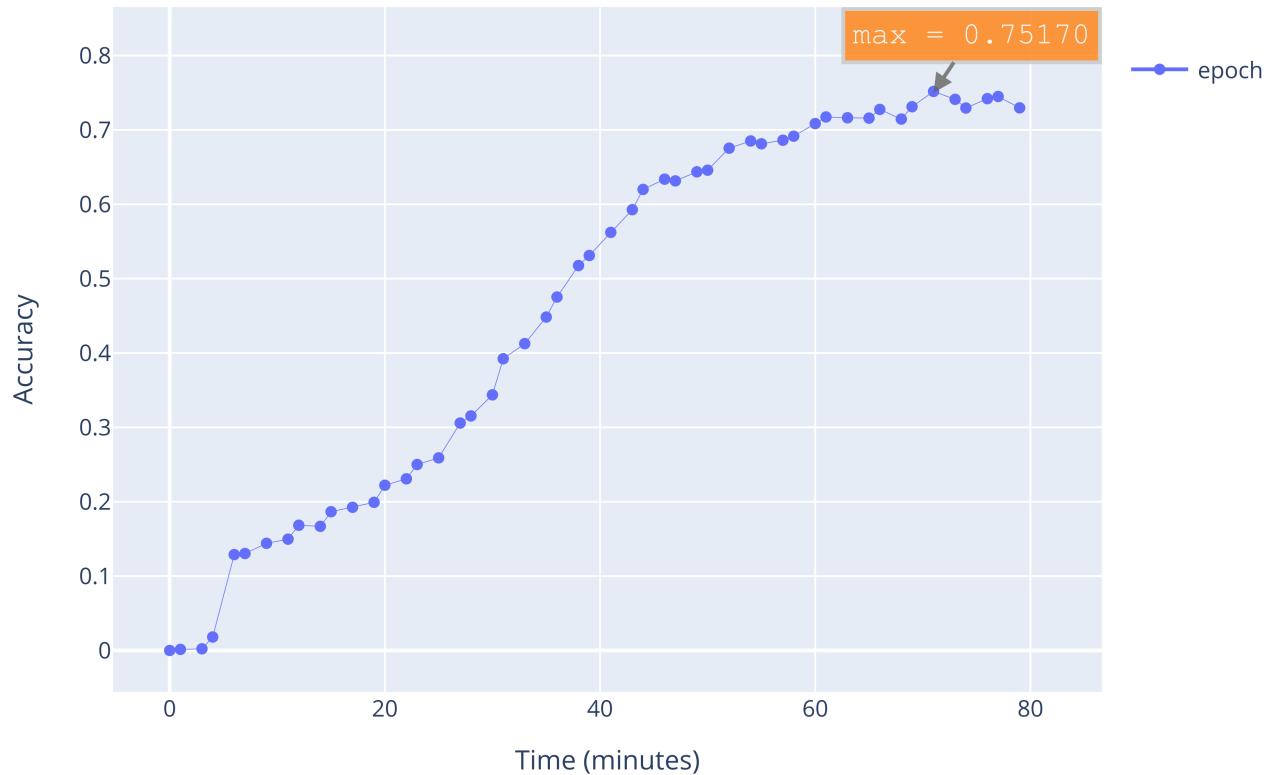
2.2.1 Produire un modèle HTR avec Kraken

Initialement le projet souhaitait s'appuyer directement sur les ressources *hardware* de la plateforme eScriptorium, car comme nous l'avons vu le moteur Kraken nécessite une très forte puissance de calcul en raison de son architecture neuronale RNN. Toutefois, cette puissance exigée a aussi un coût pour la structure de l'INRIA et le temps d'attentes peut être assez long en fonction de la demande. Il a donc rapidement été décidé de passer par des moyens alternatifs en s'appuyant sur un Graphics Processing Unit (GPU) personnel et d'utiliser directement le CLI Kraken.

Pour la méthodologie employée, il convient de définir quelques termes complexes. Le processus d'apprentissage supervisé s'appuie sur trois types de données qui ont été préalablement compilées afin d'améliorer les performances : les données d'entraînements, les données de validation qui vont être le support d'apprentissage lors de l'entraînement et enfin les données de tests qui vont permettre d'évaluer la qualité de l'apprentissage. En général et ici dans notre cas, les données sont réparties entre 80% pour les données d'entraînements, 10% pour les données de validation et 10% pour les données de tests. Chaque processus d'entraînement est nommé *epoch* et se termine par la procédure d'évaluation. Lors de chaque entraînement, les données sont subdivisées en différents lots d'apprentissage simultané (*batch*) permettant d'améliorer la qualité d'analyse du modèle²⁵. La fin générale du processus d'entraînement s'applique selon plusieurs critères. Une limite s'applique quand s'effectue une normalisation des scores et ainsi éviter le risque d'*overfitting*. Il s'agit d'un risque de surspécialisation du modèle qui tend à apprendre les données "par cœur". Il devient ainsi incapable d'agir correctement face à de nouvelles données.

Le moteur développé par Benjamin Kiessling permet de créer son modèle à partir de deux méthodes de *machine learning* : le modèle *skratch* dont les données d'apprentissages sont natives ce qui signifie un apprentissage *ad hoc*, et la technique du *transfert learning* qui correspond à la réutilisation d'un modèle pré-entraîné et d'en exploiter les connaissances basiques en reprenant son architecture et en y ajoutant des données *ad hoc*. L'entraînement *via fine-tuning* est donc une sous catégorie du *transfert learning*, assez propre aux architectures CNN, puisque cette méthode permet de réadapter les poids neuronaux au cours de l'apprentissage ce qui conduit à affiner le modèle. Il faut penser qu'au sein du réseau neuronal, les neurones situés en premières séries ont très souvent des tâches génériques alors que les dernières séries de neurones ont la plus grande part une fonction très spécifique. Tout l'enjeu de ce procédé est finalement d'affiner les poids de

²⁵. B. Kiessling, *The Kraken OCR System*, version 4.1.2, avr. 2022, URL : <https://kraken.re> (visité le 27/08/2022), Pour plus de détails, la documentation de Kraken est particulièrement explicite dans le fonctionnement de l'apprentissage *in*.


 FIGURE 2.6 – Entraînement d'un modèle avec la méthode *skratch*

ces derniers.

Les avantages du *transfert learning* puisqu'il est désormais possible de produire des modèles performants à moindre coût, en limitant le besoin de produire des données terrains²⁶. Coupler les données en puisant sur des neurones existants permet généralement d'obtenir de très bon, voir d'excellents résultat. Vincent Jolivet estime aujourd'hui que le fine-tuning est devenu la norme au sein du paysage de la REM, en permettant de créer des modèles à moindre coût²⁷. La multiplication des plateformes de partage de données et de modèles est ainsi indispensable au développement de cette stratégie, et *de facto* à la démocratisation de l'HTR.

2.2.2 Modélisation, résultats et interprétations

Afin d'observer concrètement les avantages et les inconvénients de ces deux méthodes, plusieurs modèles ont été réalisés afin de déterminer le réseau neuronal le plus

26. J. C. Aradillas, J. J. Murillo-Fuentes et P. M. Olmos, « Boosting Handwriting Text Recognition in Small Databases with Transfer Learning »...

27. Sergio Torres et Vincent Jolivet, « HTR Fine-tuning for Medieval Manuscripts Models : Strategies and Evaluation », dans 2022, URL : <https://dahtr.sciencesconf.org/> (visité le 21/08/2022).

Name	Quantity (GT)	Val_acc	Test_acc	CER	WER
ArSKR	180	0.75170	0.77860	0.26558	0.38640
ArLCTP	144	0.93328	0.83570	0.04410	0.18993
ArLCTP-NFKD	144	0.91871	0.84650	0.03668	0.15853
ArLCTP+pl	244	0.91697	0.83230	0.05045	0.21338
ArMcFR	180	0.90354	0.86730	0.05598	0.21423
ArMcFR-NFKD	180	0.89872	0.85630	0.06646	0.24963

TABLE 2.1 – Résultats des modèles HTR³¹

performant pour déchiffrer des sources manuscrites de multiples mains²⁸).

Datasets et modèles

En plus des données produites, deux jeux de données extérieurs ont été utilisés : les archives notariales du projet LECTAUREP et le modèle éponyme, ainsi que les archives des testaments de poilus²⁹. Ces deux *datasets* peuvent être rapprochés de nos données terrains, car ils sont composés de documents avec une écriture cursive contemporaine (XIX^e siècle).

La difficulté consiste à disposer de suffisamment de données terrains propres afin de « casser le modèle de langue », car on le rappelle, les moteurs HTR s'appuient sur le traitement du langage pour affiner ces prédictions. Enfin, nous nous sommes fondés sur le récent méga-modèle *Manu MacFrench* développé par Thibault Clérice et Alix Chagué³⁰.

Résultats et analyses

Comme nous pouvons le constater au sein du tableau 2.1, les différentes méthodes employées ont affichés des résultats assez contrastés. La valeur Précision proposée par kraken lors du processus d'entraînement a rapidement été insuffisante pour comparer et optimiser le caractère. Lors de la compilation préalable des fichiers, la désignation des fichiers tests est donc hasardeuse et les résultats obtenus sont alors intrinsèquement relatifs. Dans un premier temps, nous nous sommes appuyés sur l'application KaMi-Lib

28. Les lignes de commandes utilisées avec le moteur kraken sont observables en annexe (voir figure 5 et 6)

29. M. Durand, A. Rostaing et A. Chagué, *Notaires de Paris - Répertoires, Ground Truth for Various Parisian Registries of Notary Deeds (French 19th and 20th Centuries)...*; T. Clérice et A. Chagué, *CREMMA-AN-TestamentDePoilus*, 2022, URL : <https://github.com/HTR-United/CREMMA-AN-TestamentDePoilus>.

30. A. Chagué et T. Clérice, *HTR-United - Manu McFrench V1 (Manuscripts of Modern and Contemporary French)*, version 1.0.0, Zenodo, 17 juin 2022, URL : <https://zenodo.org/record/6657809> (visité le 27/08/2022).

31. ArSKR : modèle *skratch dataset ad hoc*; ArLCTP : *fine-tuning* avec modèle LECTAUREP ; ArLCTP+pl : *fine-tuning* avec modèle LECTAUREP et ajout du jeu de données des testaments de poilus ; ArMcFR : *fine-tuning* avec modèle *Manu MacFrench*.

développée par l'INRIA et plus particulièrement Lucas Terriel³². Elle permet d'étendre le nombre de mesures possibles afin d'étudier plus en profondeur les prédictions OCR en fonction des vérités terrains, mais aussi d'obtenir certains mesures relevant du TAL. Au sein de chaque main principale, un échantillon a été prélevé afin de constituer les données de terrains de l'évaluation par KaMi-Lib³³. Les principales mesures qui ont été utilisées pour comprendre l'efficacité des modèles REM sont les métriques CER et WER, permettant de donner le taux d'erreurs par caractères et par mots. En parallèle, un jeu de données tests a été constitué afin de déterminer le taux Précision à partir d'une main inconnue et de qualité moyenne.

À première vue, le modèle ArLCTP, *fine-tuned* à partir du modèle déployé par le projet LECTAUREP, semble donner les meilleurs résultats. En revanche, en le comparant aux données issues de l'évaluation test, le taux Précision est bien plus faible, laissant penser à un possible phénomène d'*overfitting* en raison de sa mauvaise adaptation. Le modèle ArMcFR semble ainsi offrir les résultats les plus polyvalents et indiquant une meilleure performance malgré des taux CER et WER légèrement plus élevés. Il reste tout de même perfectible en vue de ces résultats pouvant être qualifiés de « moyen-bon » en raison de ses taux d'erreurs³⁴. Lors du colloque « Documents anciens et reconnaissance automatique des écritures manuscrites » qui se déroulait en juin 2022, Vincent Jolivet et Sergio Torres ont estimé qu'un taux Précision de 90% doit être considéré comme le seuil sur l'échelle du rapport coût de production et efficacité du modèle³⁵. De même, Maciej Eder estime que le bruit des données au sein des corpus textuels n'affecte pas singulièrement les recherches autour, la stylométrie dans ce cas précis, si le taux de corruption des données est inférieur à 20%³⁶.

Toutefois, il reste à évaluer la capacité du modèle à être affiné selon les situations. Sans avoir pu être réalisée en l'état, l'agrégation de quelques nouvelles données terrains sur une main bien particulière pourrait permettre d'atteindre des taux extrêmement satisfaisants comme le souligna Ariane Pinche au cours de ce même colloque³⁷.

32. L. Terriel et A. Chagué, *KaMI-lib*, version 0.1.3, 8 août 2022, DOI : 10.5281/zenodo.1234.

33. M. Humeau, *HTR Evaluation*, avec la coll. d'Alessandro Chiaretti, Archivo Central Andres Bello, avr. 2022, URL : https://github.com/Proyecto-Ocupacion-Araucania-UChile/model-HTR/tree/main/test_kami (visité le 28/08/2022).

34. Ciprian Tomoia, Paul Feng, Mathieu Salzmann et Patrick Jayet, *Field Typing for Improved Recognition on Heterogeneous Handwritten Forms*, 22 sept. 2019, arXiv : 1909.10120 [cs], URL : <http://arxiv.org/abs/1909.10120> (visité le 28/08/2022), Dans cette article, il est estimé que le CER doit être inférieur à 2 afin d'être qualifié de très bon voir d'excellent.

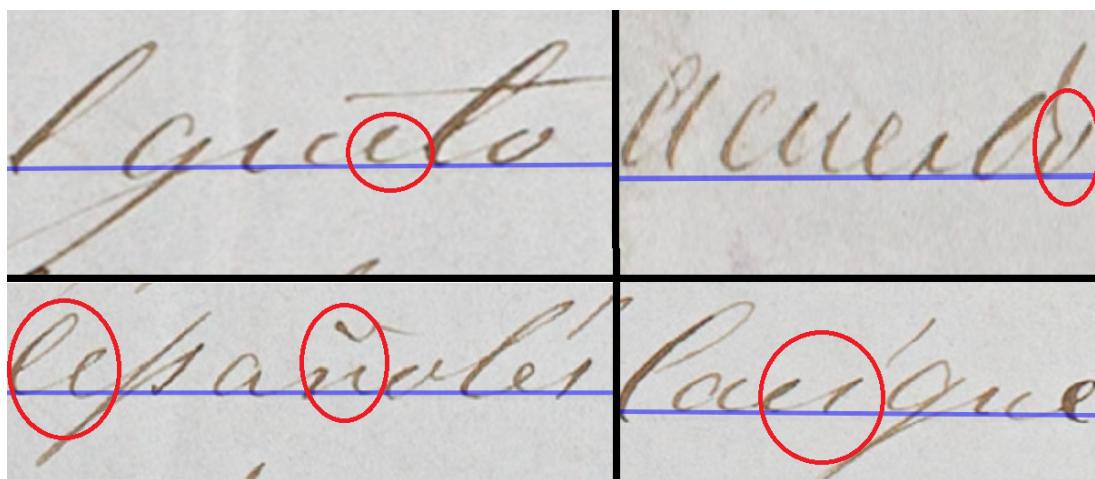
35. S. Torres et V. Jolivet, « HTR Fine-tuning for Medieval Manuscripts Models : Strategies and Evaluation »...

36. Maciej Eder, « Mind Your Corpus : Systematic Errors in Authorship Attribution », *Literary and Linguistic Computing*, 28–4 (1^{er} déc. 2013), p. 603-614, DOI : 10.1093/linc/fqt039.

37. A. Pinche, S. Gabay et Kelly Christensen, « SegmOnto – A Controlled Vocabulary to Describe Historical Textual Sources », dans *Documents Anciens et Reconnaissance Automatique Des Écritures manuscrites/Documents Anciens et Reconnaissance Automatique Des Écritures Manuscrites*, Paris, France, 2022.

Nombre	Correct	Généré
40	s	c
24	SPACE	NONE
22	i	e
21	o	a
21	r	s
20	ACCENT	NONE
18	s	r

TABLE 2.2 – Les 7 erreurs les plus courantes du modèle ArMcFR

FIGURE 2.7 – Exemple d’erreurs courantes pour le modèle ArMcFR³⁸

Difficultés et encodage

En reprenant le tableau 2.1, on peut constater la présence de modèles NFKD (*Normalization Form Canonical Decomposition*). Il s’agit d’un système d’uniformisation au sein des caractères Unicode. La méthode NFKD permet dans ce cas de décomposer les caractères diacritiques, en un ensemble de caractères Unicode permettant une appréhension fondamentale, en conservant sa relation canonique et ordonnée. Les divers essais montrent une appropriation relative des modèles HTR à cette uniformisation. On remarque au sein du tableau 2.2 que la gestion des accents représente le sixième type d’erreurs les plus fréquentes.

En examinant plus attentivement les erreurs récurrentes du modèle, on peut rapidement identifier plusieurs lacunes et confusions, en particulier la reconnaissance des caractères Unicode ‘r’ et ‘s’. Cette répétitivité amène à un WER moyen indiquant un mot erroné tous les cinq mots. Ces confusions peuvent s’expliquer par la fréquence élevée de ces caractères au sein de la langue espagnole et de la proximité graphologique entre les caractères.

38. Image haut gauche : guito/ gusto ; image haut droite : acuerd/acuerdo ; image bas gauche : lepanoles/españoles ; image bas droite : Caesque/Cacique.

En observant plus en détail, la distance des mots, selon l'algorithme de distance de Levenshtein, se situe autour d'une moyenne tronquée autour de 43, mais avec un écart-type plus conséquent concernant les mains Villalon et Saavedra. Cette distance semble *a priori* suffisamment faible indiquant une relative concordance entre les mots prédits et les mots corrects, limitant le risque d'altérer le sens initial³⁹. À partir de cette distance, Thi-Tuyet-Hai Nguyen et al. classe deux types d'erreurs : les erreurs simples et les erreurs complexes ayant un nombre d'erreurs supérieur à 1⁴⁰. Ces erreurs vont plus ou moins être influencées selon la longueur des mots et la gestion des espacements des modèles OCR et ainsi augmenter le risque de cette distance.

Il est donc toujours difficile d'évaluer correctement un modèle, et plus encore, de sélectionner le modèle le plus performant dans le cadre d'une application contrainte et spécifique. Dans notre cas, nous sommes appuyés sur des métriques basées sur le lexique dont le modèle ArMcFR semble ressortir comme le plus complet. Cependant, à l'heure de la massification des solutions TAL certaines mesures pourraient reprendre les évaluations par système de masque sur le modèle des architectures encodeurs-décodeurs comme le constatent Phillip Benjamin Strobel et al.⁴¹.

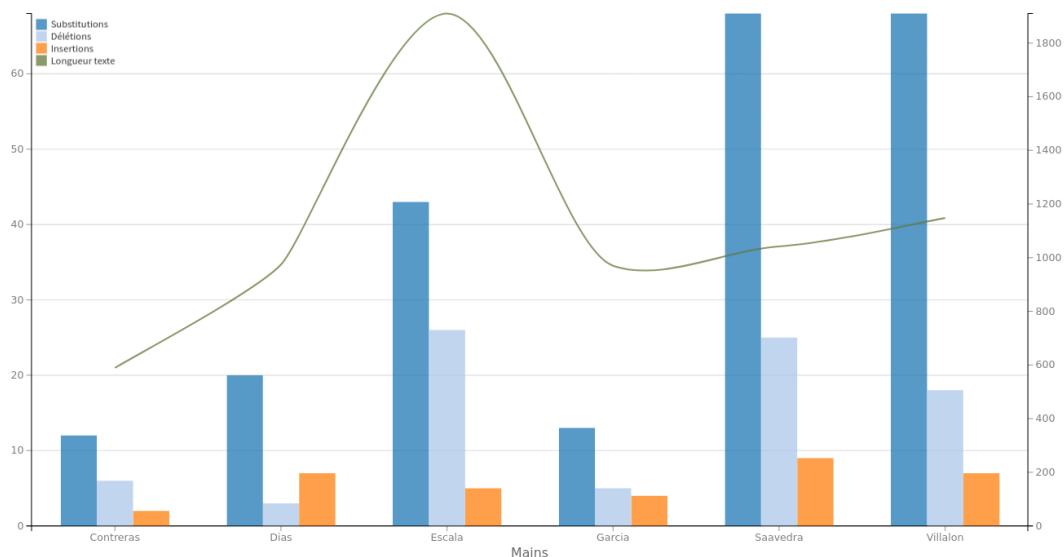


FIGURE 2.8 – Évaluation Délétion Insertion Substitution (DIS) du modèle ArMcFR avec KaMi-Lib

39. L. Terriel, « Atelier : Production d'un Modèle Affiné de Reconnaissance d'écriture Manuscrite Avec eScriptorium et Évaluation de Ses Performances. Évaluer Son Modèle HTR/OCR Avec KaMI (Kraken as Model Inspector) », dans *Les Futurs Fantastiques - 3e Conférence Internationale Sur l'Intelligence Artificielle Appliquée Aux Bibliothèques, Archives et Musées*, Paris, France, 2021, URL : <https://hal.archives-ouvertes.fr/hal-03495762> (visité le 29/08/2022).

40. Thi-Tuyet-Hai Nguyen, Adam Jatowt, Mickael Coustaty, Nhu-Van Nguyen et Antoine Doucet, « Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing », dans *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 2019, p. 29-38, DOI : [10.1109/JCDL.2019.00015](https://doi.org/10.1109/JCDL.2019.00015).

41. P. B. Ströbel, S. Clematide, M. Volk, Raphael Schwitter, Tobias Hodel et David Schoch, *Evaluation of HTR Models without Ground Truth Material*, 29 avr. 2022, arXiv : 2201.06170 [cs], URL : [http://arxiv.org/abs/2201.06170](https://arxiv.org/abs/2201.06170) (visité le 29/08/2022).

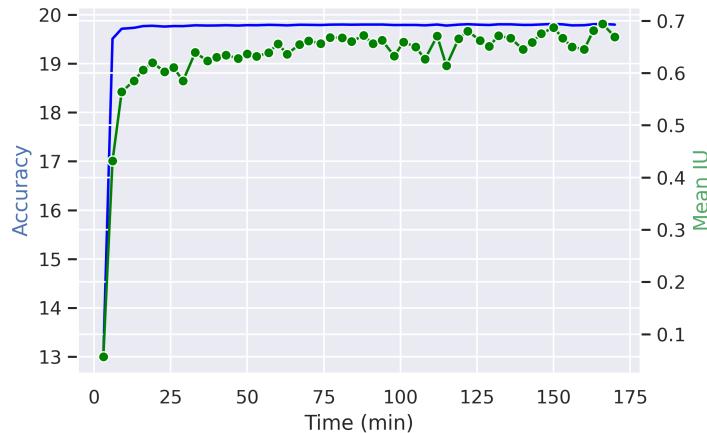


FIGURE 2.9 – Le processus d’entraînement du modèle ArSeg

Model	val_acc	mean_acc	mean_iu	freq_iu
ArSeg	19.81395	19.81395	0.69404	0.69404

TABLE 2.3 – Evaluation des performances du modèle ArSeg

2.2.3 Les difficultés d’un modèle de segmentation

À partir de notre lot de transcriptions, nous avons dans le même temps essayer de produire un modèle de segmentation HTR à partir du moteur Kraken. Il fait suite à la volonté d’automatiser la segmentation des zones et des lignes à partir de l’ontologie prédefinie ultérieurement.

Pour se faire, nous avons repris le modèle de segmentation par défaut développé par l’équipe de Scripta afin de procéder à un entraînement par *finetuning*. Le développement du modèle **b11a** s’est appuyé sur le jeu de donnée qui a remporté la compétition ICDAR de 2017. Il propose un modèle minimalist possédant de hauts taux de performance sur la détection de ligne et de zones (bounding boxes) et permettre à des modèles sémantiques de s’appuyer sur cette base⁴².

L’entraînement sur nos documents historiques a été exécuter sur la base des recommandations établis par Juliette Janès et al. pour l’entraînement d’un modèle de mise en page⁴³. La commande shell (retrouvable au sein de l’annexe D, voir 7) reprend ainsi le réseau neuronal défini préalablement.

Néanmoins, les résultats affichés durant l’entraînement démontrent de nombreuses fragilités. L’évaluation d’un modèle de segmentation s’est axé sur deux systèmes mesures : le *mean Intersection-Over-Union* (Mean IU) qui reprend l’ensemble de l’indice de Jacard

42. B. Kiessling, D. Stökl Ben Ezra et M. T. Miller, *BADAM...*

43. Juliette Janès, A. Pinche, C. Jahan et S. Gabay, « Towards Automatic TEI Encoding via Layout Analysis », dans *Fantastic Future 21, 3rd International Conference on Artificial Intelligence for Librairies, Archives and Museums*, Paris, France, 2021, URL : <https://hal.archives-ouvertes.fr/hal-03527287> (visité le 21/08/2022).

2.3. LA GESTION DES ERREURS : LE POST-TRAITEMENT COMME SECOND SOUFFLE À L'HTR

calculant la similarité entre deux éléments et le taux d'alignement ; le mean Accuracy calcule la moyenne de toutes les classes sur le principe de la métrique Précision⁴⁴.

Comme le démontre la figure 2.10, les prédictions affichent un manque de régularité sur l'appréciation des *baselines* et des zones. Les lignes en marge du document sont quant à elles ignorées ou très mal déterminées. De même, on observe que seules les zones *Main :text* et *QuireMarksZone :signature* sont identifiées. Ce dernier point peut démontrer le besoin de simplifier l'ontologie.

Face à ce constat, le modèle ne permet pas d'obtenir une segmentation efficiente et exploitable afin de l'intégrer au sein du flux éditorial. Une segmentation manuelle à partir du modèle par défaut proposé sur la plateforme eScriptorium reste privilégiée au cours de cette chaîne de traitement afin de réévaluer les données d'entraînements.

Les récents résultats présentés par Thibault Clérice peuvent laisser un espoir d'amélioration du système de segmentation⁴⁵. Il propose de déplacer, par souci d'efficacité, la reconnaissance non plus sur une polygonisation basée sur la classification des pixels, mais à une détection d'objet utilisant des rectangles isothétiques. Pour cela, il appuie l'incorporation du moteur YOL0v5 au sein du moteur Kraken.

2.3 La gestion des erreurs : le post-traitement comme second souffle à l'HTR

La reconnaissance automatique de texte ne s'arrête pas aux seuls prédictions du modèle HTR. Depuis les années 1990, de nombreux projets se sont appuyés sur le post-traitement des prédictions afin d'améliorer la qualité, en s'appuyant notamment sur des modèles statistiques aux frontières de la linguistique, en particulier les systèmes d'occurrences séquentielles (n-grammes)⁴⁶.

Nous allons observer comment le projet Araucania a tenté de prolonger cette chaîne de traitement HTR en procédant à une mise en place d'une post-correction automatique. Le but est de réduire significativement les erreurs issues des transcriptions grâce à l'aide d'un corpus témoin.

44. Pour plus de détails, les explications faites par Hugo Scheithauer sont très explicites. H. Scheithauer, *La Reconnaissance d'entités Nommées Appliquées à Des Données Issues de La Transcription Automatique de Documents Manuscrits Patrimoniaux. Expérimentations et Préconisations à Partir Du Projet LECTAUREP...*, p. 79-80

45. T. Clérice, *You Actually Look Twice At It (YALTAi) : Using an Object Detection Approach Instead of Region Segmentation within the Kraken Engine*, juill. 2022.

46. H. Takahashi, N. Itoh, T. Amano et A. Yamashita, « A Spelling Correction Method and Its Application to an OCR System », *Pattern Recognition*, 23–3 (1^{er} janv. 1990), p. 363-377, DOI : 10.1016/0031-3203(90)90023-E; Xiang Tong et David A. Evans, « A Statistical Approach to Automatic OCR Error Correction in Context », dans *Fourth Workshop on Very Large Corpora*, Herstmonceux Castle, Sussex, UK, 1996, URL : <https://aclanthology.org/W96-0108> (visité le 30/08/2022).

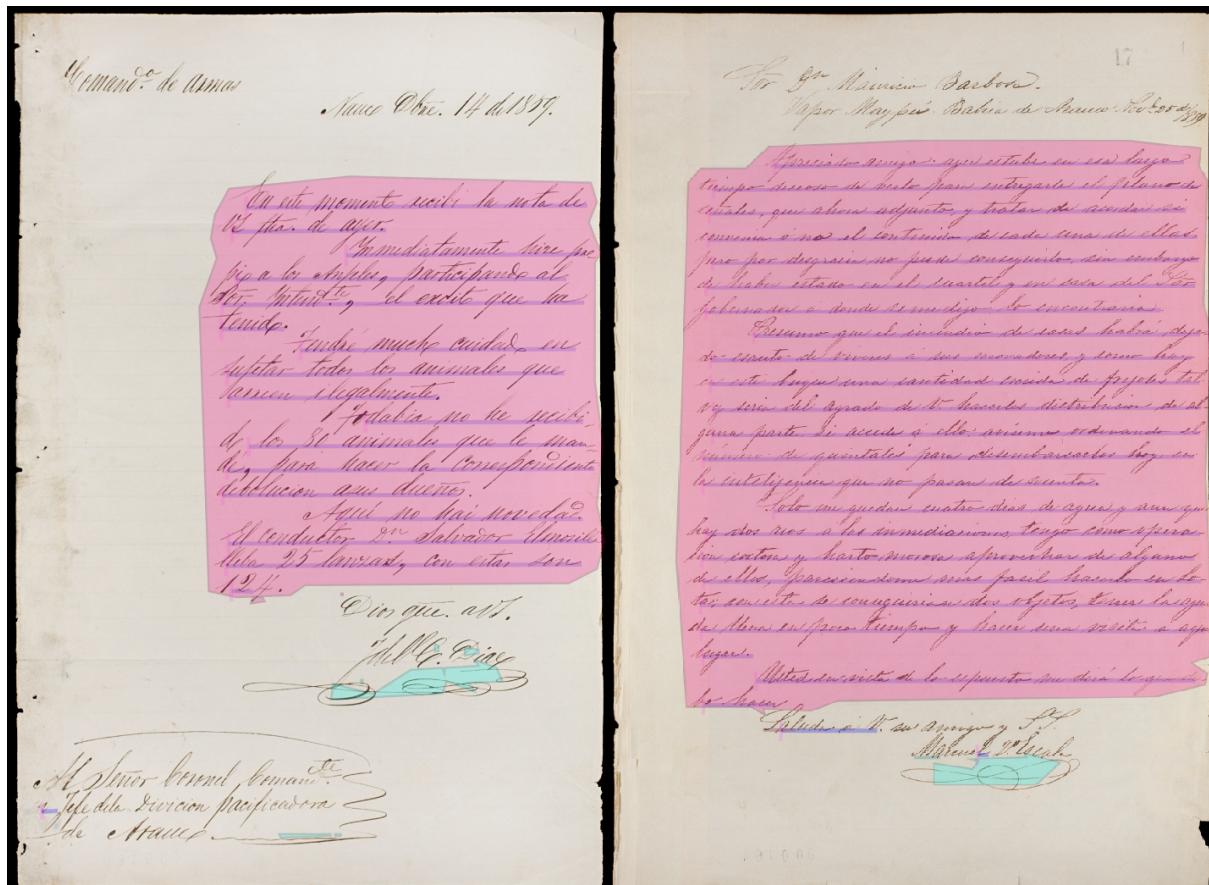


FIGURE 2.10 – Exemple de segmentation à partir du modèle ArSeg

2.3.1 Principe de la Distance de Levenshtein

Les erreurs des prédictions OCR et HTR se résument à des erreurs de substitutions, d'insertions ou de suppressions de caractères au sein d'un ou plusieurs mots. Ces altérations au sein d'un mot transcrit peuvent être décrit à travers la distance de Levenshtein. Cette mesure de la différence entre deux chaînes de caractères a été proposée sous la forme d'un algorithme linguistique par le mathématicien russe Vladimir Levenshtein en 1965. Le principe y est assez simple, plus la distance est élevée, plus le mot transcrit a donc subi d'altérations.

Dès la fin des années 1990, certains ingénieurs se sont essayé à intégrer cette distance dans leur chaîne de traitement OCR afin d'estimer les candidats correctifs les plus prometteurs à partir de calculs probabilistes selon des séquences de caractères et de mots (n-grammes) texte⁴⁷. Depuis, d'autres projets ont essayé d'intégrer cette mesure grâce à l'appui de dictionnaires d'occurrences, classification de mots, séquence de caractères notamment pour les problèmes de segmentations entre les mots⁴⁸. Les différentes méthodes ont permis de réduire le nombre d'erreurs des prédictions OCR jusqu'à 30%. En ce sens, nous remarquons l'initiative du projet DAHN dirigé par Floriane Chiffolleau qui introduit une correction *via* la distance de Levenshtein⁴⁹.

$$\text{lev}(a, b) = \begin{cases} \max(|a|, |b|) & \text{si } \min(|a|, |b|) = 0, \\ \text{lev}(a - 1, b - 1) & \text{si } a[0] = b[0], \\ 1 + \min \begin{cases} \text{lev}(a - 1, b) \\ \text{lev}(a, b - 1) \\ \text{lev}(a - 1, b - 1) \end{cases} & \text{sinon.} \end{cases}$$

FIGURE 2.11 – Algorithme de Levenshtein - @wikipedia

2.3.2 Mise en place d'une correction automatisée

Sur le modèle du projet DAHN, nous avons expérimenté la mise en place d'un CLI correctif au sein de la chaîne de traitement. Le script s'est fondé plus exactement sur l'incorporation de la librairie python **PySpellchecker**, et secondairement la librairie de TAL SpaCy⁵⁰.

PySpellchecker intègre un système de correction statistique à partir d'un dictionnaire d'occurrence, en appliquant la distance de Damerau-Levenshtein pour identifier les altérations possibles et une détermination des candidats grâce au théorème de Bayes, permettant de déterminer la probabilité d'un évènement par rapport à d'autres⁵¹. L'idée est

47. *Ibid.*

48. Ido Kissos et Nachum Dershovitz, « OCR Error Correction Using Character Correction and Feature-Based Word Classification », dans *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, Santorini, Greece, 2016, p. 198-203, DOI : 10.1109/DAS.2016.44 ; Rishin Haldar et Debjyoti Mukhopadhyay, « Levenshtein Distance Technique in Dictionary Lookup Methods : An Improved Approach » (), p. 5, DOI : 10.48550.

49. F. Chiffolleau, *DAHN Project...*

50. Tyler Barrus, *Pyspellchecker*, version 0.6.3, 29 août 2022, URL : <https://github.com/barrust/pyspellchecker> (visité le 30/08/2022).

51. Peter Norvig, *How to Write a Spelling Corrector*, févr. 2007, URL : <https://norvig.com/spell->

donc de générer l'ensemble des possibilités dans la limite de cette distance, et choisir le candidat le plus plausible. En ce sens, le dictionnaire d'occurrences permet de sélectionner le candidat à partir de ces calculs probabilistes, l'inférence bayésienne c'est-à-dire la démarche logique permettant de calculer ou actualiser la probabilité d'une hypothèse.⁵².

$$P(\alpha|\beta) = \frac{P(\beta|\alpha)P(\alpha)}{P(\beta)} \quad (2.1)$$

La complexité de la correction est décrite par un article publié en 2019 qui étudie statistiquement les erreurs récurrentes au sein des mécanismes OCR⁵³. Comme évoqué en amont, le groupe de chercheurs décrivent deux types d'erreurs fondamentales. La détection de ces erreurs multiples est corrélée à une distance plus grande, multipliant par conséquent le nombre de candidats possibles et donc accroissement du bruit. Au vu des résultats du modèle ArMcFR, nous avons estimé qu'une distance de 2 serait suffisante au vue d'une distance moyenne par mot assez faible (43) sur l'ensemble de la transcription, indiquant une répartition plus forte des erreurs (WER de 21%).

Comme l'indique le schéma du script suivant (voir figure 2.12), le processus s'est basé sur les recommandations de Daniel Lopresti dans la construction d'une chaîne de traitement post-OCR⁵⁴. Ligne par ligne, les phrases subissent un processus de tokénisation afin d'en révéler les informations essentielles, et de pouvoir traiter les mots individuellement. Le dictionnaire d'occurrences comme corpus de contrôle a ainsi été construit autour des données terrains produites précédemment afin d'aligner le niveau de langue de nos documents historiques avec nos prédictions HTR⁵⁵. Les dictionnaires par défaut sont davantage adaptés aux corrections des productions très contemporaines.

Une des premières étapes est de sélectionner les tokens dont la nature (POS, *part of speech* en anglais) est un nom propre afin d'être identifiée au sein d'un répertoire de noms propres géographiques⁵⁶. S'il n'est pas détecté il est donc retourné comme erreur

`correct.html` (visité le 31/08/2022).

52. Récemment cette technique est encore recommandée, avec une amélioration de près de 30% des cas comme le révèle cette étude. R. Haldar et D. Mukhopadhyay, « Levenshtein Distance Technique in Dictionary Lookup Methods : An Improved Approach »...

53. T.T.H. Nguyen, A. Jatowt, M. Coustaty, *et al.*, « Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing »...

54. Daniel Lopresti, « Optical Character Recognition Errors and Their Effects on Natural Language Processing », *International Journal on Document Analysis and Recognition (IJDAR)*, 12–3 (1^{er} sept. 2009), p. 141-151, DOI : 10.1007/s10032-009-0094-8.

55. M. Humeau, *Postprocess HTR*, avec la coll. d'Alessandro Chiaretti, Archivo Central Andres Bello, mai 2022, URL : https://github.com/Proyecto-Ocupacion-Araucania-UChile/postprocess_alto.

56. Le dictionnaire JSON s'appuie sur un dictionnaire du XIX^e siècle : Francisco Solano Asta-Buruaga, *Diccionario Geográfico de la República de Chile*, Segunda edición corregida y aumentada, Santiago, Chili, 1899, URL : https://es.wikisource.org/wiki/Diccionario_Geogr%C3%A1fico_de_la_Rep%C3%BAblica_de_Chile (visité le 31/08/2022). Le dictionnaire a été produit via *webscraping* dont le script est disponible ici : M. Humeau, *Enrichment Wikisource*, avec la coll. d'Alessandro Chiaretti, Archivo Central Andres Bello, mai 2022, URL : https://github.com/Proyecto-Ocupacion-Araucania-UChile/data_enrichment

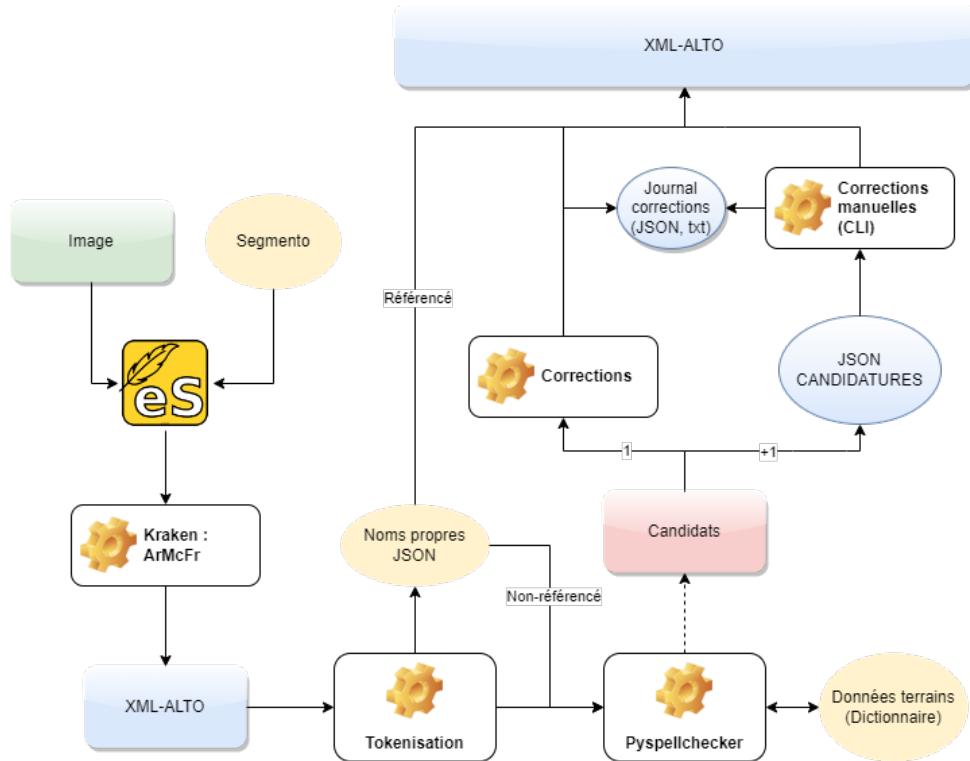


FIGURE 2.12 – Schéma de traitement des prédictions HTR

au sein du processus de correction. Ce procédé permet de pallier au limite du dictionnaire d'occurrences construit sur des données préalablement identifiées. Certains lieux lui sont donc encore inconnus, relevant ainsi d'une erreur.

Selon les erreurs détectées et retournées, les candidats sont alors soumis à un processus de sélection restrictif, et ce en fonction d'une distance limite de 2 entre le mot identifié et les propositions faites. Si les propositions sont au nombre de 1 alors, le fichier ALTO est automatiquement édité avec la correction sélectionnée. En revanche, si la liste des propositions est égale à zéro ou supérieure à 1 alors les données sont enregistrées au sein d'un fichier JavaScript Object Notation (JSON) afin d'être corrigées manuellement.

2.3.3 Résultats et améliorations

Afin d'évaluer la performance du traitement correctif des fichiers HTR, nous avons réalisé une analyse textuelle comparative grâce à la librairie KaMi-Lib à partir d'un échantillon de quatre fichiers de notre jeu de données test. Les prédictions basiques ont été effectuées avec le modèle ArMcFr, puis comparées aux mêmes données transformées par le script de corrections. Malgré quelques tentatives d'améliorations, les résultats se sont montrés assez décevants pour le moment comme le montre le tableau de comparaison 2.4.

Étonnamment, les premières mesures indiquent une augmentation des métriques CER et WER pouvant signaler une forte distorsion provoquée par la correction automatique. En revanche, les évaluations complémentaires indiquent une très légère progression

concernant la distance entre la prédiction HTR et la correction proposée et les altérations de caractères. L'augmentation des mesures CER et WER pourraient donc s'expliquer par une réduction des caractères totaux et une plus forte disparité entre les mots corrigés positifs et les mots corrigés négatifs notamment à cause d'une mauvaise gestion des accents.

	Prediction	Correction	Différence
CER (%)	13.58	13.77	+0.19
WER (%)	45.48	42.56	+2.92
D _{mots}	77	72	-5
Insertions	13.25	11.25	-2
Deletions	19.25	17	-2.25
Substitutions	91	87.75	-3.25

TABLE 2.4 – Analyse des effets du traitement automatique des erreurs HTR

Nous pouvons émettre plusieurs constats d'améliorations à la suite de notre analyse de cas. La première difficulté réside dans la gestion des noms propres, et ce malgré le filtrage à partir du POS, dont la détermination se révèle souvent erronée. Dans de nombreux cas, les noms géographiques ont été abrégés au sein des transcriptions, complexifiant grandement l'opération de correction. L'autre point est la gestion dans sa globalité des noms propres comme le renseigne Hugo Scheithauer, après quelques expérimentations pour le projet LECTAUREP⁵⁷.

La seconde difficulté, et sans doute la plus grande, réside dans la détection des erreurs au sein des mots réels (40%)⁵⁸. À l'inverse d'un mot non-réel, le mot réel signifie que les altérations DIS ont transformés le mot initial en un autre mot existant et référencé. Nous l'avons vu, l'orthographe est très variable et ponctuée de fautes au sein des archives. Ces différences linguistiques amènent donc une difficulté à identifier les erreurs HTR avec les variations orthographiques et les erreurs orthographiques originales. De plus, ces variations, volontaires ou non, augmentent artificiellement le nombre de mots existants, biaisant le dictionnaire d'occurrences et donc les probabilités bayésiennes. Une solution pourrait s'imaginer avec l'introduction d'un processus de lemmatisation entre le produit HTR et le corpus contrôle, et une meilleure utilisation de l'étiquetage de la parole (POS).

Comme signalé par Youness Chaabi et Fadoua Ataa Allah, le système Norving employé par la librairie PySpellchecker possède de nombreuses limites⁵⁹. Le système de

57. H. Scheithauer, *La Reconnaissance d'entités Nommées Appliquées à Des Données Issues de La Transcription Automatique de Documents Manuscrits Patrimoniaux. Expérimentations et Préconisations à Partir Du Projet LECTAUREP...*, p. 88-89.

58. T.T.H. Nguyen, A. Jatowt, M. Coustaty, et al., « Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing »...

59. Youness Chaabi et Fadoua Ataa Allah, « Amazigh Spell Checker Using Damerau-Levenshtein

classification par dictionnaire d'occurrences est éprouvé depuis très longtemps qui bien que conçu pour être effectué au travers de calcul simple, celui-ci est connu pour générer un trop gros nombre de résultats⁶⁰. De plus, l'effet de contextualisation du mot erroné au sein de la phrase est assez limité. En ce sens, une combinaison entre le système Norvik et l'utilisation de l'analyse séquentielle par N-grammes pourrait permettre l'obtention de meilleurs résultats comme le proposent les chercheurs de l'Université du Roi-Saoud⁶¹.

Lors du concours de l'IDCAR 2019, l'utilisation de cette architecture est remarquée puisque le modèle BERT démontre un fort potentiel dans la résolution de correction contextuelle⁶². Par la suite, plusieurs groupes de chercheurs ont mis à profit ces résultats primaires afin de construire de nouvelles expérimentations avec les architectures TAL encodeurs-décodeurs (BERT, RoBERTa, etc.)⁶³. Les projets mettent à contribution la recherche d'entités-nommées et les vecteurs de mots afin de prédire le mot correct caché par le système du masque. Les résultats obtenus notamment par l'Université de l'Essex indiquent une réduction substantielle des erreurs faites par la prédition OCR⁶⁴. Toutefois, la mise en place de ce système exige des ressources matérielles bien plus importantes.

Algorithm and N-gram », *Journal of King Saud University - Computer and Information Sciences*, 34 (8, Part B[2022]), p. 6116-6124, DOI : 10.1016/j.jksuci.2021.07.015.

60. Même si le nombre est alors limité par la définition de la distance de Levenshtein.

61. *Ibid.*

62. Christophe Rigaud, A. Doucet, Mickaël Coustaty et Jean-Philippe Moreux, « ICDAR 2019 Competition on Post-OCR Text Correction », dans *15th International Conference on Document Analysis and Recognition*, Sydney, Australia, 2019, p. 1588-1593, URL : <https://hal.archives-ouvertes.fr/hal-02304334> (visité le 31/08/2022).

63. Aditya Pal et Abhijit Mustafi, *Vartani Spellcheck – Automatic Context-Sensitive Spelling Correction of OCR-generated Hindi Text Using BERT and Levenshtein Distance*, 14 déc. 2020, DOI : 10.48550/arXiv.2012.07652, arXiv : 2012.07652 [cs] ; Srinidhi Karthikeyan, Alba G. Seco de Herrera, Faiyaz Doctor et Asim Mirza, « An OCR Post-Correction Approach Using Deep Learning for Processing Medical Reports », *IEEE Transactions on Circuits and Systems for Video Technology*, 32-5 (mai 2022), p. 2574-2581, DOI : 10.1109/TCSVT.2021.3087641.

64. *Ibid.*

Deuxième partie

Produire et enrichir une édition numérique

Chapitre 3

De l’HTR à XML-TEI : mise en place d’une chaîne de transformation

Le déploiement d’une chaîne de traitement automatisée depuis les données océrisées vers le format XML-TEI (*Text Encoding Initiative*) a fait l’objet de nombreux investissements de la part des laboratoires de recherches comme en témoigne les projets Artl@s, Katabase, DAHN ou encore le projet LECTAUREP¹. L’objectif est d’assembler et structurer différentes briques de transformations afin de passer de l’image numérisée à une édition numérique native par le prisme du schéma TEI.

L’intérêt pour l’application de ce format aux documents historiques résulte de sa grande polyvalence et adaptabilité, et ce peut importe l’origine ou la nature du document numérisé. Ce standard souhaite être le plus universel possible tout en rendant compte des singularités intrinsèques de chaque document et de son contexte d’origine. Lucas Terriel évoque ce schéma comme un « format pivot² » pour souligner les multiples exploitations possibles.

Toutefois, l’automatisation de données brutes vers ce format reste complexe, car elle fait face à un problème méthodologique et humaniste majeur. L’application d’un fichier

1. Les travaux de Juliette Janès, de Lucie Rondeau du Noyer ou encore corbieresCatalogueAuFichier2020 sont particulièrement révélateur de velléités scientifiques. Lucie Rondeau du Noyer, *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la revue des autographes de la librairie Charavay*, mémoire pour le diplôme de master « Technologies numériques appliquées à l’histoire », dir. Thibault Clérice et Simon Gabay, Paris, France, Ecole nationale des chartes, 2019, URL : https://github.com/lairaines/M2TNAH/blob/429079b2d9723dc85a3e8ba07f84c0b7e18717d1/RondeauduNoyer_M2TNAH.pdf (visité le 01/09/2022) ; Caroline Corbières, Béatrice Joyeux-Prunel (dir.) et Thibault Clérice (dir.), *Du Catalogue Au Fichier TEI : Création d’un Workflow Pour Encoder Automatiquement En Xml-Tei Des Catalogues d’exposition*, mémoire pour le diplôme de master « Technologies numériques appliquées à l’histoire », dir. Thibault Clérice et Béatrice Joyeux-Prunel, Paris, France, Ecole nationale des chartes, 2020 ; Juliette Janès, *Du catalogue papier au numérique : Une chaîne de traitement ouverte pour l’extraction d’informations issues de documents structurés*, mémoire de master « Technologies numériques appliquées à l’histoire », dir. Thibault Clérice et Simon Gabay, Paris, France, École nationale des chartes, 2021, URL : https://github.com/Juliettejns/Memoire_TNAH.

2. L. Terriel, *Représenter et Évaluer Les Données Issues Du Traitement Automatique d’un Corpus de Documents Historiques. L’exemple de La Reconnaissance Des Écritures Manuscrites Dans Les Répertoires de Notaires Du Projet LectAuRep....*, p. 59.

TEI n'a de sens qu'en la joignant à une politique éditoriale ou scientifique et conjointement au document d'origine. La mise en place d'un protocole d'automatisation n'est valide qu'au sein d'un projet précis.

À travers ce chapitre, nous cherchons à étudier les besoins, les réflexions et les défis auxquels s'est affronté ce projet d'édition des documents autour de l'Occupation de l'Araucanie. Ce projet s'est appuyé sur les précédentes expériences afin de concilier la retranscription des données aussi bien du point de vue quantitatif que qualitatif.

3.1 L'intérêt d'une édition numérique native

Avant de décrire plus en détail la mise en place de la chaîne de traitement éditorial, il convient de revenir sur les caractéristiques et les enjeux scientifiques de ce format pivot qu'est XML-TEI. Cette ébullition pour ce format dans le domaine des humanités numériques nécessite d'en considérer l'ensemble des aspects afin d'évaluer les ambitions scientifiques tout autant que les perspectives de valorisation autour de ce fonds.

3.1.1 L'édition numérique pour les humanités

De nombreuses préoccupations et interrogations sont apparues avec la prolifération des éditions électroniques à la fin des années 2000, notamment à la suite du mouvement *Digital scholarship* britannique. La question du numérique remet pertinemment en cause l'ontologie du texte et sa textualité. Les projets d'ampleur tels que les éditions des *Chroniques* de Jean Froissart³, le roman *Partonopeus de Blois*⁴ ou les légendes arthuriennes⁵ ont été le support de développement de multiples textes autour de leurs apports et de leurs contraintes.

Si l'hésitation entre le choix de l'orientation vers l'auteur ou le document demeure insoluble, de nombreux travaux linguistiques et philologiques considèrent que ces éditions numériques ne sont pas une altération, mais bien une solution pérenne pour les chercheurs. L'hypertexte devient un nouveau lieu de connaissances pour le scientifique et la constitution de son appareil critiques⁶. La définition faite par la chercheuse humaniste Joana Casenave est particulièrement révélatrice des enjeux autour de l'édition électronique pour les humanités. Selon elle « L'édition numérique est [...] l'occasion d'un positionnement éditorial nouveau. Elle permet une diversification des informations présentées et une prise

3. *The Online Froissart*, version 1.4, éd. P. Ainsworth et G. Croenen, Sheffield, <https://www.hrionline.ac.uk/onlinefroissart/>.

4. *Partonopeus de Blois*, Sheffield, <http://www.hrionline.ac.uk/partonopeus>.

5. *Queste del saint Graal*, éd. C. Marchello-Nizia et A. Lavrentiev, Lyon, Équipe BFM, publié en ligne par la Base de Français Médiéval, <http://portal.textometrie.org/bfm/?command=documentation&path=/GRAAL>.

6. Frédéric Duval, « Pour des éditions numériques critiques. L'exemple des textes français », *Médiévales. Langues, Textes, Histoire*–73 (73[2017]), p. 13-29, DOI : 10.4000/medievales.8165.

en compte de plus en plus forte des lecteurs au cours du travail philologique.⁷ ».

La mise en place de ces projets a abondamment été commentée au gré des résultats des précédentes expériences éditoriales et scientifiques. Peter Robinson rappelle que la mise en place de ce type d'édition doit se faire dans le cadre d'une politique spécifique et s'appuyant sur cinq besoins : projets scientifiques particuliers, possession d'une transcription originale complète ou l'ouverture de ces transcriptions en données ouvertes, la valorisation et le changement de perception auprès du lecteur et enfin le projet d'établir des analyses quantitatives autour des textes⁸.

Toutefois, Joana Casenave souhaite compléter cette définition autour de la notion de données ouvertes⁹. La mise en place de ces éditions sous le format numérique permet d'augmenter les échanges autour de celle-ci. À la fois ouvert et en libre accès, il donne une nouvelle dimension au travail éditorial puisque celle-ci peut être axée sur des méthodes collaboratives, mais aussi sur la notion d'édition évolutive. La dimension socio-politique et scientifique des archives autour de l'Occupation de l'Araucanie a incité à mettre en place cette approche et expérience éditoriale.

La construction d'un projet d'édition électronique native s'appuie aussi sur une volonté d'ouverture aux apports numériques au sein de la recherche fondamentale et la valorisation scientifique. Le développement des éditions numériques s'est fait en corrélation avec l'essor des humanités numériques en permettant l'exploration textuelle au travers de nouvelles méthodes scientifiques. La fouille de texte, la stylométrie, les analyses quantitatives ou la philologie numérique s'installent progressivement comme des méthodes heuristiques prépondérantes au sein des humanités¹⁰. Elles incarnent un rapport nouveau par rapport au texte et à la donnée avec un nouveau regard méthodologique introduisant la notion de reproductibilité¹¹. Ces données ouvertes offrent de même la possibilité de développer des outils numériques d'analyses, autant que la valorisation du corpus.

3.1.2 XML-TEI : retour sur un format pivot

Avec la croissance des projets d'éditions numériques, le format XML-TEI s'est imposé comme un standard pour le développement de ce type de projet. La TEI est fondée sur la structuration de données XML et une syntaxe précise permettant de définir et contraindre l'ensemble des éléments utilisés.

7. Joana Casenave, « Le positionnement éditorial dans l'édition critique numérique », *Digital Studies / Le champ numérique*, 9–1 (1[2019]), DOI : 10.16995/dscn.348.

8. Robinson, Peter, *What is a Critical Digital Edition ?*, Variants. The Journal of the European Society for Textual Scholarship, 1, 2002. via (Caroline Blanc Feracci et Marthe Vertongen, *Qu'est-ce qu'un projet d'édition critique numérique ?*, DLIS, 1^{er} mars 2022, URL : <https://dlis.hypotheses.org/5790> [visité le 20/08/2022])

9. J. Casenave, « Le positionnement éditorial dans l'édition critique numérique »...

10. J.B. Camps, « Où va la philologie numérique ? », *Fabula-LhT-20* (29 janv. 2018), URL : <https://www.fabula.org:443/lht/20/camps.html> (visité le 20/08/2022).

11. *Ibid.*

En réalité, la TEI désigne initialement une communauté scientifique qui a ensuite donné son nom à cet encodage. La communauté académique a été fondée en 1987 afin de répondre aux nouveaux défis numériques, en proposant une première grammaire sous la forme d'une Document type definition (DTD) pour le format SGML (*Standard Generalized Markup Language*), puis à son langage descendant XML. Les objectifs de cet encodage sont de faciliter le traitement et l'échange des données textuelles, en préférant des solutions générales et modulaires. Cette initiative s'inscrit dans une volonté de proposer un format de données conservant le sens informationnels et contextuels tout en affichant une indépendance claire avec les logiciels propriétaires¹².

L'association a depuis considérablement multiplié et enrichi son encodage depuis sa création, multipliant par cinq le nombre de balises disponibles. Ainsi, Jean-Baptiste Camps parle d'une « communauté de pratiques¹³ » pour évoquer le rôle de cette communauté et de cet encodage. La souplesse et l'accessibilité de ce standard en font des qualités indéniables puisqu'elles permettent aux différents projets de pouvoir l'adapter à leurs besoins et tout en conservant une certaine interopérabilité. Ces nombreuses qualités et sa constante amélioration ont été permises grâce à une communauté active qui en a fait un format pivot majeur pour les humanités numériques. Ce format est aussi essentiel pour le lecteur car il offre une lecture accessible mais aussi la possibilité de déployer des outils numériques annexes. Le centre ACAB a donc souhaité s'ouvrir et répondre à ces nouveaux enjeux numériques avec une expérimentation autour de ces archives en proposant des données ouvertes et exploitables, mais aussi valoriser les données contenues par ces documents.

3.1.3 Les outils numériques de valorisation

Avec l'ébullition de l'encodage TEI au sein des humanités numériques, une pluralité d'outils a accompagné la mise en place d'éditions nativement numériques. Afin de préparer ce projet d'édition et de valorisation des sources du conflit entre l'État chilien et les tributs Mapuches, trois applications ont été étudiées. Afin d'en déterminer les avantages et les inconvénients. Par manque de temps, le développement n'a pu être mis en application restant pour le moment à l'état de prospection.

La première étude fut centrée sur l'utilisation d'un Content Managing System (CMS) très connue au sein des institutions archivistiques : la plateforme Omeka. Elle permet de mettre en place rapidement et aisément une application dédiée à l'exploration et la visualisation des collections patrimoniales, ainsi qu'à leurs notices associées. Omeka possède

12. Lou Burnard, « Introduction » in Lou Burnard, *Qu'est-ce que la Text Encoding Initiative ?*, trad. par Marjorie Burghart, Marseille, 2015 (Encyclopédie numérique), URL : <http://books.openedition.org/oep/1237> (visité le 02/09/2022).

13. J.B. Camps, « La TEI, Une Communauté de Pratiques », dans *Édition Électronique et TEI : Enjeux, Pratiques et Perspectives - _dayClic()*, Le Mans, 2017.

une architecture minimalisté fondé sur le système LAMP (Linux Apache, MySQL, PHP) permettant d'être facilement exploitable ainsi que d'être sous licence libre et gratuite¹⁴. Néanmoins, le nombre de possibilités permis par ce CMS reste très limité de par sa structure volontairement rigide et simpliste. Ces caractéristiques ne permettent pas une exploitation exhaustive des données transcrites et n'offre qu'une très faible modularité¹⁵.

La seconde solution logicielle a été développée par la société internationale à but non lucratif *e-editiones* sous le nom de *TEI Publisher*¹⁶. Cette application *open-source* s'appuie sur *exist-db* qui est un système de gestion de base de données XML. *TEI Publisher* offre une interface extrêmement complète avec la présence de nombreux modules très intéressants dont des fonctions de recherche plein texte ou à filtres, un système de pagination, la reconnaissance d'entités nommées, géolocalisation ou encore la compatibilité avec l'API IIIF (*International Image Interoperability Framework*). À partir de ces expérimentations, Hugo Scheitauer estime que cette plateforme offre une excellente solution moyen-terme pour la valorisation des corpus TEI¹⁷. La transformation XML vers le langage web HTML s'effectue selon les règles définies au sein de l'One Document Does (ODD) accompagnant le corpus TEI. *TEI publisher* fournit ainsi une application extrêmement complète et à moindre coût technique et financier pour les institutions patrimoniales et scientifiques.

Enfin, les prémisses d'une application web ont été élaborées au sein de centre afin de dessiner les fonctionnalités nécessaires à la valorisation et l'exploration scientifique du corpus océrisé. Le développement d'une telle application permettrait *de facto* d'établir et d'étoffer des fonctionnalités propres à l'exploration des données éditées. Toutefois, elle nécessite une maîtrise avancée de *framework* web et de modules adaptés¹⁸. L'idée initiale est de centrer le projet auprès de la richesse des entités géographiques présentes au sein des différentes sources, en établissant une carte dynamique et des fonctionnalités de recherche à filtre permettant de retracer l'évolution narrative de la correspondance. Les travaux de Ludovic Moncla sont particulièrement éclairants sur l'automatisation des dynamiques spatiales à partir de la Reconnaissance des entités nommées (REN)¹⁹. Le projet à long

14. OMEKA, url : <https://omeka.org/about/project/>, consulté le 19/08/2022.

15. Ces problèmes ont été amplement commentés par Cécile Boulaire et Romeo Carabelli, « Du digital naïve au bricoleur numérique : les images et le logiciel Omeka », in Étienne Cavalié, *et al.* (éd.), *Expérimenter les humanités numériques. Des outils individuels aux projets collectifs*, Nouvelle édition [en ligne], Montréal, 2018 (Parcours numérique), URL : <http://books.openedition.org/pum/11091> (visité le 02/09/2022)

16. *TEI Publisher*, version 7.1.0, e-editiones, 3 juill. 2022, URL : <https://github.com/eeditiones/tei-publisher-app> (visité le 02/09/2022).

17. H. Scheithauer, *La Reconnaissance d'entités Nommées Appliquées à Des Données Issues de La Transcription Automatique de Documents Manuscrits Patrimoniaux. Expérimentations et Préconisations à Partir Du Projet LECTAUREP...*, p. 139-143.

18. Les frameworks *Flask* ou *Django* paraissent les plus aptes à ce type de projet puisque de nombreux modules complémentaires comme *Lxml* ou *folium* leurs sont compatibles.

19. Ludovic Moncla, Mauro Gaio, Ekaterina Egorova et Christophe Claramunt, « An Automatic Extraction Method of Static and Dynamic Spatial Contexts from Texts », dans *Atelier Science Des Données et Humanités Numériques (SDHN)*, Conférence Internationale Francophone Sur l'Extraction et La Gestion de Connaissance (EGC 2018), Paris, France, 2018, URL : <https://hal.archives-ouvertes.fr/hal-01833232>

terme souhaiterait développer cette contextualisation des dynamiques spatiales et de la géomatique en s'appuyant sur le georéférencement d'une carte historique du Chili.

3.2 XML-TEI et les archives océrisées du conflit Mapuche

Au regard de ces observations préliminaires, il doit être évoquer la méthodologie employée afin de transformer les fichiers ALTO vers le format XML-TEI. La mise en place d'un script python a nécessité un ensemble de réflexions sur la flexibilité de la transformation des transcriptions en fonction de la typologie du document et du schéma édité. Nous nous concentrerons ensuite sur la validation du schéma et la documentation associée.

3.2.1 Convertir les données océrisées au format TEI

Comme nous l'avons vu, de nombreux projets se sont appliqués à transformer les données ALTO vers un schéma TEI préalablement défini. Selon les différentes méthodologies employées, la chaîne de traitement a été développé sur la base du langage python, et plus particulièrement la librairie `Lxml`, ou XSLT spécialisée dans la transformation XML.

Nous avons finalement pris le parti de nous appuyer sur le langage python en raison de ses performances, d'une communauté très active permettant de faciliter la maintenance et l'amélioration de l'application. De ce fait, nous avons repris la structure du CLI `Alto2tei` développée par Kelly Christensen²⁰. Ce script a été développé dans le cadre du projet Gallic(opora) au cours de l'année 2022 dans le but de transformer les documents numérisés, hébergés au sein de la plateforme Gallica, vers le format XML-TEI à partir de leurs manifestes IIIF. L'application possède deux atouts majeurs qui sont la forte structuration des documents de sorties et la conservation des données d'océrisations originales.

Nous avons donc décidé de remanier l'application originale afin d'adopter la structure et la transformation au projet éditorial conçu par le centre ACAB, permettant un gain de temps considérable et fondé sur une une expérience antérieure solide²¹. Comme présenté sur le schéma (voir figure 3.1), la première étape s'applique à regrouper les fichiers ALTO selon leur identifiant et appliquer la transformation par groupe, en créant la racine `<TEI>` et de récupérer les labels des régions et des lignes au sein du fichier ALTO. Les

ouvertes.fr/hal-01695643 (visité le 02/09/2022).

20. K. Christensen, *Alto2tei*, Gallic(opora), 2022, URL : <https://github.com/kat-kel/alto2tei> (visité le 03/09/2022).

21. application développée pour le projet Araucania est disponible ici. M. Humeau, *Tei Transformation*, avec la coll. d'Alessandro Chiaretti, Archivo Central Andres Bello, juill. 2022, URL : https://github.com/Proyecto-Ocupacion-Araucania-UChile/TEI_transformation

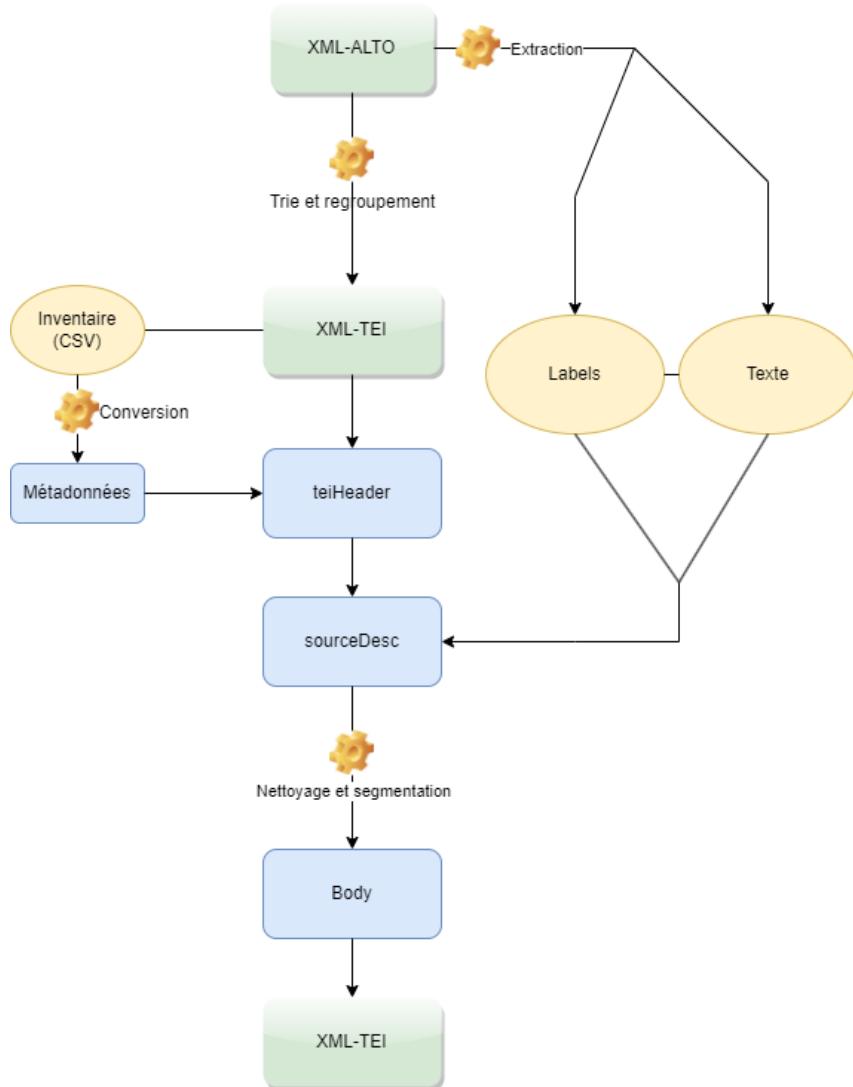


FIGURE 3.1 – Schéma de conversion des fichiers ALTO vers le format XML-TEI

différentes métadonnées retranscrites au sein de l'inventaire CSV sont alors enregistrés au sein du `<teiHeader>` grâce à la librairie Pandas. Ensuite, le texte contenu dans les **@CONTENT** est extrait des fichiers ALTO et réassocié aux labels précédemment recensés puis structuré selon les pages, les zones puis les lignes au sein du `<sourceDesc>`. Les données récoltées sont enfin nettoyées et restructurées au sein du `<text>` en fonction des attributs **@type** et **@subtype** des différentes lignes et des différentes zones afin d'appliquer le schéma éditorial.

3.2.2 Documenter et valider un schéma

Au sein du chapitre 1, nous avons pu constater la variété typologique de notre fonds documentaire et la forte représentativité des sources épistolaires. Pour le moment, le projet c'est avant tout axé sur l'édition de ces documents dont la quantité et la similarité diplomatique justifie une automatisation. Selon l'évolution du traitement du fonds, d'autres

```

<constraintSpec scheme="schematron" ident="refVal">
  <constraint>
    <s:rule context="tei:placeName[@ref and
      ↳ ancestor::tei:body]">
      <s:let name="ref" value="@ref"/>
      <s:assert test="//tei:place[@xml:id =
        ↳ substring-after($ref, '#')] and
        ↳ starts-with($ref, '#)"> The value of @ref
        ↳ has not been declared |<s:value-of
        ↳ select="$ref"/>| </s:assert>
    </s:rule>
  </constraint>
</constraintSpec>

```

Listing 2 – Application d'une règle *Schematron*

schémas pourront être développés.

Afin d'assurer la pérennité de cette édition et le sens sémantique de sa structuration, la conception d'une ODD est primordiale puisqu'elle permet de documenter l'ensemble des choix éditoriaux entrepris au cours de cette édition²². Cette documentation devient en quelque sorte le révélateur des enjeux scientifiques, des réflexions éditoriales et de la méthodologie critique du document. Dans notre cas, cette ODD à propos de la correspondance a été générée à partir du modèle de transformation *ODD by example* grâce à la création d'un scénario sur le logiciel **Oxygen** et le processeur Saxon 9. L'ODD a par la suite été enrichie manuellement afin de décrire l'encodage utilisé.

L'autre enjeu de la construction est la validation et la spécification de notre schéma TEI. Cette spécification peut se faire grâce aux éléments **<schemaSpec>**, **<moduleRef>**, **<elementSpec>** et **<classSpec>** permettant de modifier la grammaire ou restreindre les règles applicables au schéma XML, voir une DTD (Document type definition)²³. La déclaration de certaines règles *Schematron* a permis de préciser l'usage de certains attributs et de leurs contenus au travers de règles. Comme nous avons pu le voir ci-dessus (voir code 2, la règle *Schematron* décrit l'obligation pour tous éléments **<place-Name>** d'avoir son identifiant référencé au sein d'un élément **<place>** présent au sein du **<profileDesc>**). La conversion de l'ensemble de ces règles et de cette documentation sous le format Regular Language for XML Next Generation (RNG) permet d'automatiser le travail de validation et d'évaluation de la conformité des documents TEI produits. Cette vérification est exécutée à la fin de l'application du script de conversion avec la fonction de classe **validate** de la librairie **lxml**.

22. La consultation de l'ODD pour la correspondance est consultable au sein du dossier documentation, *Ibid.*

23. L. Burnard, *Qu'est-ce que la Text Encoding Initiative ?...*

3.3 Mise en place de l'application autour d'un schéma

Comme l'amorçait nos premières observations sur l'utilisation d'un script de transformation, la mise en place de cet encodage TEI nécessite de revenir aux choix employés lors de la production. L'emploi de la librairie `Lxml` a été indispensable dans l'extraction, la structuration et l'enrichissement des données.

Ainsi, les axes de nos observations se fondent sur les entités fondamentales de la TEI : le `<teiHeader>`, le `<sourceDesc>` et le `<text>`.

3.3.1 Le teiHeader et ses métadonnées

Le `<teiHeader>` correspond à l'élément en-tête d'un fichier XML-TEI. Il regroupe et structure l'ensemble des métadonnées recensées au cours de sa production, permettant d'indexer et de contextualiser le document source. La granularité du schéma des métadonnées doit être ainsi clairement définie en amont en se basant sur la transformation du fichier ALTO lui-même, le ou les documents numériques originelles (fac-similés, ect.) et l'inventaire décrit ultérieurement par la section archivistique du centre ACAB. Le point d'appui pour leurs recensements repose sur l'identifiant extrait du fac-similé numérique afin d'aligner les données avec l'inventaire sous le format CSV²⁴.

Dans un premier temps, les métadonnées décrivent l'ensemble des informations de productions du fichier TEI au sein de l'élément `<fileDesc>` à partir des métadonnées extraites de l'inventaire²⁵. Le `<titleStmt>` renseigne alors les informations essentielles telles que le titre et l'auteur du document. À l'inverse, les éléments `<publicationStmt>`, `<editionStmt>` et `<notesStmt>` sont pratiquement inamovibles en raison de leur fonction descriptive. Ces balises regroupent les informations de l'autorité éditoriale et de publication. Nous avons choisis de rattacher ces deux éléments à la description du centre ACAB étant, pour le moment, la seule institution véritablement engagée au sein de ce projet d'édition. En revanche, les responsables directs sont identifiables à partir d'un fichier JSON aisément modifiable par les intervenants, puis extrait au cours de la transformation.

Le décompte des documents rattachés au fichier TEI est décrit au sein de l'élément `<extent>` affichant la quantité de fac-similés dénombrés sous cet identifiant, avant d'être identifié au sein du `<sourceDesc>` décrivant le fonds documentaire de l'Araucania. Les métadonnées d'ordre technique sont davantage disposées au sein du `<encodingDesc>` permettant de décrire succinctement les pratiques autour de cet encodage. L'élément `<appinfo>` permet de détailler les différentes applications utilisées sur l'ensemble de la chaîne de traitement. Cette description permet de renforcer l'usabilité et la compré-

24. voir le fichier `./src/opt/inventory.py` ; M. Humeau, *Tei Transformation...*

25. La structuration de l'arbre du `<teiHeader>` est disponible en suivant le chemin suivant : `./src/teiheader.py` . *Ibid.*

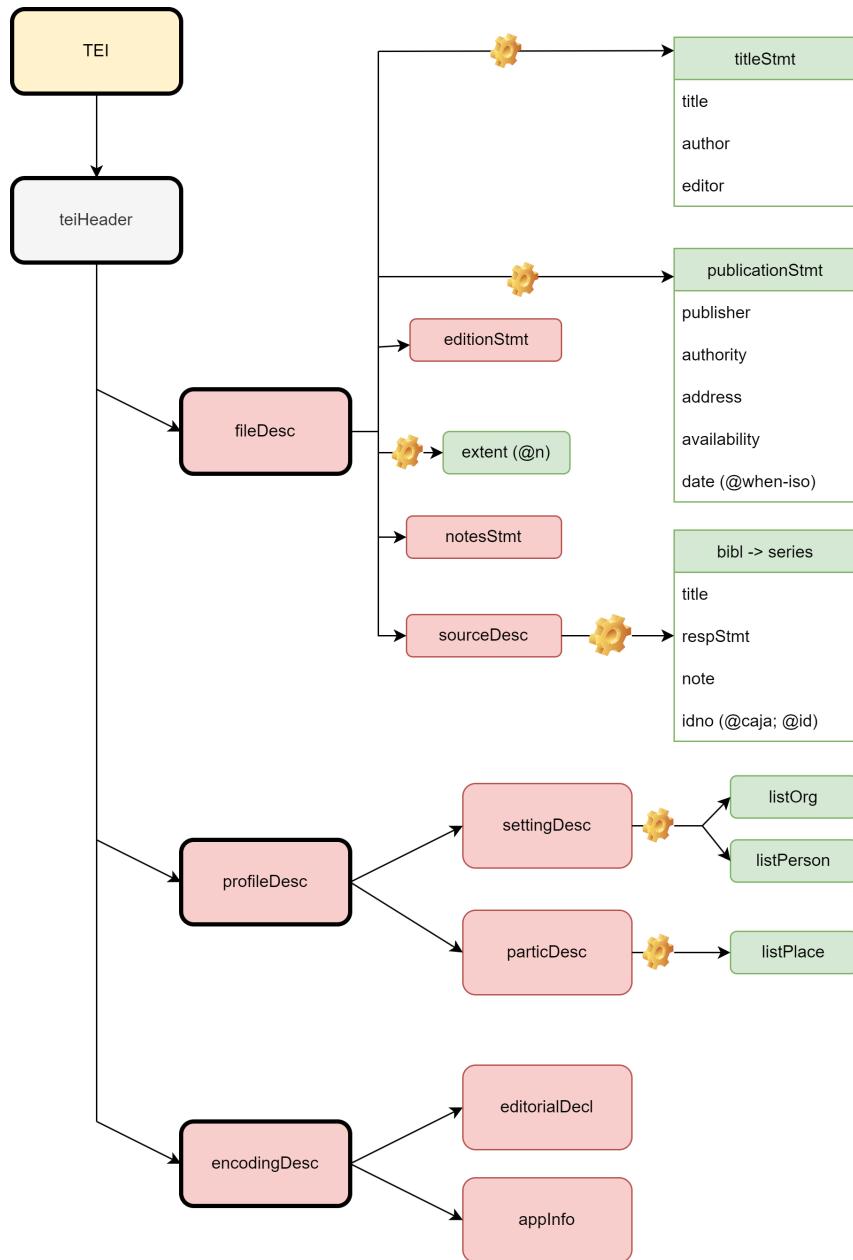


FIGURE 3.2 – Structuration du teiHeader

hension des données édités en retracant l'ensemble des processus subit au sein du fichier électronique.

3.3.2 <sourceDoc> : une trace originale

Dans sa définition officielle, le <sourceDoc> est un élément qui contient une transcription ou autre représentation d'un document source unique faisant potentiellement partie d'un dossier génétique ou d'une collection de sources. Plus concrètement, cet élément devient le lieu de description du fichier ALTO d'origine en reprenant les éléments de mise en page et du texte océrisé. La structuration est directement reprise depuis le

```

~~I<surfaceGrp>
  <surface xml:id="299_a" n="16" ulx="0" uly="0" lrx="2327"
    ↵ lry="2970">
    <graphic url=".//input/299_a.jpg"/>
    <zone xml:id="299_a_z1" type="CustomZone" subtype="Dateline"
      ↵ n="1" points="1515,217 [...] 1128,178 1294,178 1425,247"
      ↵ source=".//input/299_a.jpg" corresp="eSc_textblock_18faefaf">
      <zone xml:id="299_a_z1_11" type="DefaultLine" subtype="none"
        ↵ n="1" points="1037,388 1133,413 [...] 1034,289 1037,348"
        ↵ source=".//input/299_a.jpg" corresp="eSc_line_d13c289a">
          <path xml:id="299_a_z1_11_p" points="1037,348 [...]
            ↵ 1953,334"/>
          <line xml:id="299_a_z1_11_t">Valparaiso, Julio 101860</line>
        </zone>
      [...]
    </surface>
  </surfaceGrp>

```

Listing 3 – Exemple de structuration du sourceDoc

modèle établi par l’application **Alto2tei**²⁶.

En observant plus attentivement, la structuration est une réadaptation du modèle ALTO vers l’encodage TEI en reprenant les données essentielles. L’ensemble des fac-similés numérisés est ainsi décrit au sein de l’élément **<surface>** en rassemblant l’ensemble des zones et des lignes rattachées. Elles sont décrites au sein de la balise **<zone>** dont les types et les sous-types issus d’une segmentation préalable indiquent la typologie²⁷. L’ensemble des coordonnées ALTO ont été conservées au sein des **@points**, laissant ainsi la possibilité d’exploiter ses coordonnées directement au sein de l’image numérisée. Enfin le contenu textuel est identifiable parmi l’élément **<line>** et le masque HTR associé, présent dans l’élément précédent **<path>**.

La bonne extraction et l’attention particulière à la propreté et la validité des données sont fondamentales puisqu’elles constituent la base de la future structuration du **<text>**. Du reste, cette conservation exhaustive des données initiales accorde une grande facilité à entretenir l’interopérabilité et la pérennité des éditions numériques, et ainsi leurs réexploitations.

3.3.3 Structurer et éditer un texte

Comme nous l’avons vu au cours du chapitre 1, la définition d’une ontologie et la mise en corrélation avec le vocabulaire contrôle *SegmOnto* au cours du processus de

26. K. Christensen, *Alto2tei...*

27. CustomZone :Dateline est décousu en deux attributs : type='CustomZone' et subtype="Dateline")

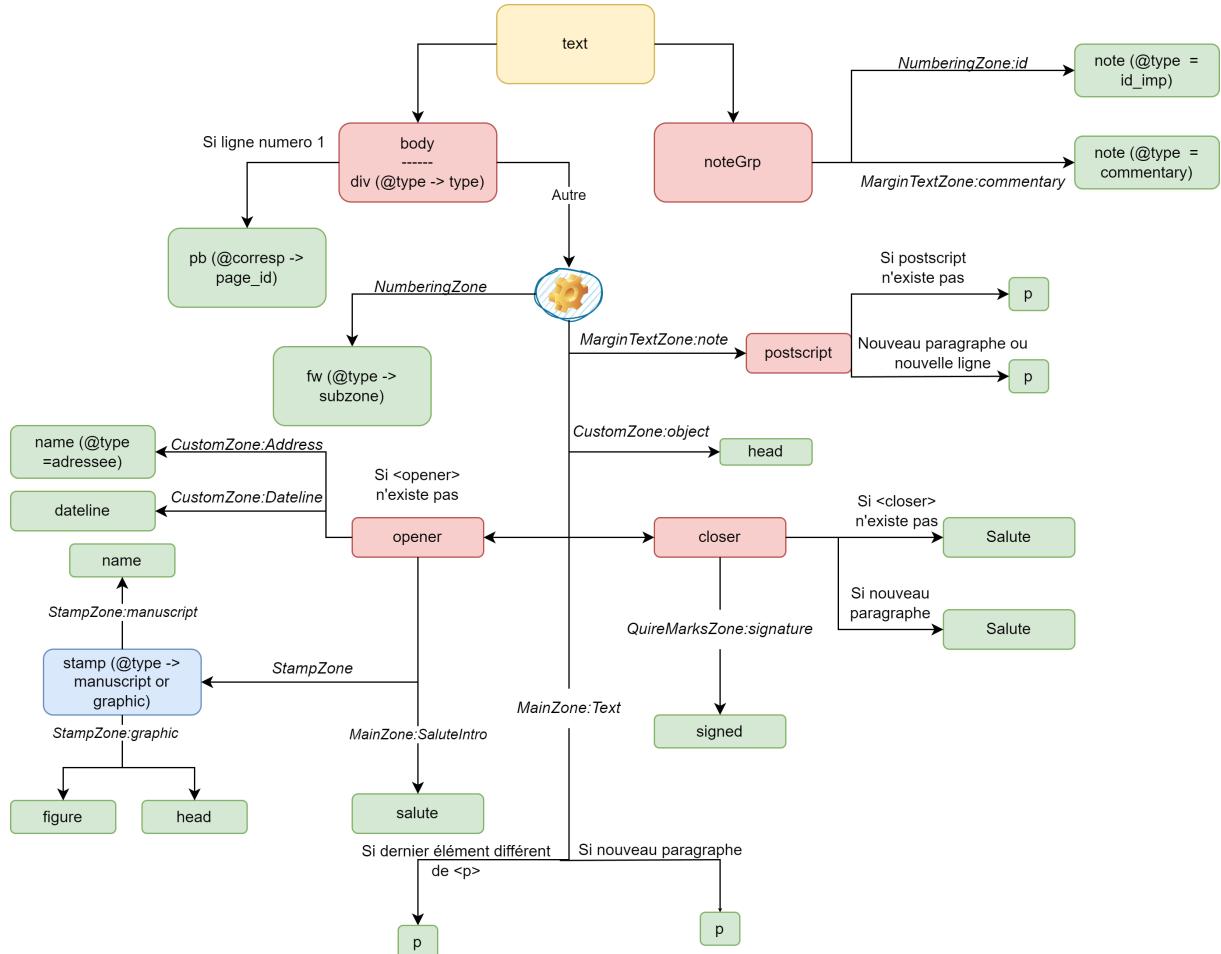


FIGURE 3.3 – Arbre décisionnel de structuration de <text>

transcription automatique a permis de conserver la valeur sémantique et sémiotique de l'image originale. Cette structuration en amont a ainsi constitué l'arc de voûte de la transformation et l'adaptation de l'ALTO à l'encodage TEI. À partir de là, l'extraction des différents labels présents au sein du fichier ALTO nous permet d'organiser la structuration du texte en fonction des différentes zones et des différentes lignes présentes.

Au sein de l'élément <text>, il nous faut établir deux subdivisions entre le <body> qui représente le contenu initial du document et le groupe <noteGrp> permettant d'y recenser la plupart des écrits ultérieurs. Seules les lignes contenues au sein d'une zone *NumberingZone* sont référencées au sein de la balise <fw>, à l'intérieur du <body>, indiquant des éléments de mises en pages²⁸. La seconde étape consiste à repérer le commencement d'une nouvelle page en se référant au numéro de ligne. Si elle est la première, elle indique la présence d'une nouvelle page soulignée par l'incrémentation de l'élément <pb> et son identifiant au sein <@corresp>.

Le schéma et la granularité du <body> sont ainsi organisés en fonction de l'ontologie HTR développée précédemment. Ce schéma reprend les recommandations du modèle

28. Hormis le sous-groupe *id* qui rejoint les notes postscript

établi pour l'édition de la correspondance de l'édition DAHN²⁹. Ce <**body**> est en réalité subdivisé en 3 sections majeures et une section optionnelle :

- <**opener**> regroupe l'ensemble du discours introductif et les éléments de contextualisation présentés par l'auteur. Il comprend les ontologies SegmOnto suivantes : les éléments *StampZone*, les sous-groupes *CustomZone* comprenant l'adresse, la date (spatiale et temporelle) et la salutation introductory présente au sein du label *MainZone :SaluteIntro*.
- A l'inverse, le <**closer**> représente l'élément général de conclusion du document. Il rassemble le salut final, parfois sectionné, et la suscription de l'auteur vers son homologue.
- Le corps principal du texte est quant à lui directement injecté au sein du <**body**> à travers l'élément <**p**>. Lors de l'énumération des différentes lignes du sourceDoc présentes au sein du *MainZone :text*, la rencontre du signe '¶' signale un nouveau paragraphe qui est alors décompté.
- La mise en place d'un élément <**postscript**> est parfois nécessaire, en raison de la présence d'une zone possédant le label *MarginTextZone*. Ces notes marginales au texte principal peuvent être divisées en différents paragraphes selon leurs mises en pages.

Cette structuration n'est pas totalement représentative de la mise en page du facsimilé original en raison de la propriété *tail*, mise en place avec la librairie **Lxml**. De cette façon, l'API *ElementTree* ne nécessite pas de noeuds de texte spéciaux en plus de la classe *Element*, qui ont parfois tendance à se mettre en travers du chemin. L'inclusion d'une balise autofermante <**lb**> permet de signaler la présence d'une nouvelle ligne sans déformer la structuration initiale. Cette technique permet de plus facilement ingérer ou modifier le fichier *a posteriori*³⁰.

Une telle structuration souhaite ainsi répondre aux besoins d'intelligibilité et d'interopérabilité des données éditées comme le rappel l'ingénieur Syd Bauman³¹. Elle doit faire ressortir aisément les identités du texte à la fois pour l'œil humain et la machine. L'automatisation de ces éditions renforce ainsi cette accessibilité en facilitant le futur traitement et la future exploitation du fichier TEI édité.

29. F. Chiffolleau, *DAHN Project...*

30. Voir le site web de la librairie **Lxml** : *Lxml*, url : <https://lxml.de/tutorial.html>, consulté le 21/08/2022.

31. Syd Bauman, « Interchange vs. Interoperability » In *Balisage : The Markup Conference 2011 Proceedings*, Montréal, 2011 via Trevor Muñoz et Raffaele Viglianti, « Texts and Documents : New Challenges for TEI Interchange and Lessons from the Shelley-Godwin Archive », *Journal of the Text Encoding Initiative*-Issue 8 (8[2014]), DOI : 10.4000/jtei.1270

Chapitre 4

Indexer et enrichir une édition numérique

Alors que les humanités numériques, évoquées au travers de la philologie, incarnent un progrès méthodologique considérable pour le travail en amont, Jean-Baptiste Camps rappelle que cette mise en place d'une édition numérique reste un effort fastidieux, chronophage et coûteux pour les institutions¹. Si la sérialisation des données ne peut-être une fin en soi, l'enrichissement des textes, primordial à la recherche, peut tout de même être considérablement allégé grâce à l'ingénierie.

Aujourd'hui, les outils associés au traitement automatique des langues sont devenus suffisamment matures et fiables pour fournir une analyse à la fois quantitative et qualitative. Ces avancées ont permis de créer un véritable engouement du TAL au sein des humanités numériques et de l'édition numérique². De nombreux écosystèmes ont été développés dans le but de permettre un enrichissement quantitatif de bases de données XML-TEI à partir de la fouille de texte et l'extraction de l'information³. Ces différents processus permettent régulièrement la correction des données produites par les modèles HTR⁴, l'indexation des entités ou encore de faciliter la recherche au sein du texte⁵.

Durant ce projet d'édition des archives de l'Occupation de l'Araucania, l'implantation d'un système d'enrichissement fondé sur les concepts du traitement automatique du langage et la reconnaissance d'entités nommées a été un des axes majeurs du processus éditorial. La mise en place d'un tel procédé nécessite de revenir sur les enjeux épistémolo-

1. J.B. Camps, « Où va la philologie numérique ? »...

2. Barbara McGillivray et Thierry Poibeau, « Digital Humanities and Natural Language Processing : “Je t'aime... Moi Non Plus” », dans 2020, t. 14–2, p. 11.

3. Marie Bisson et Anne Goloubkoff, « Les notices d'autorité en XML-TEI : un outil pour l'accroissement collaboratif de connaissances et l'indexation d'éditions de sources », *Tabularia. Sources écrites des mondes normands médiévaux* (, 3 mars 2020), DOI : 10.4000/tabularia.4176.

4. Nous l'avons partiellement vu avec la mise en place d'une application correctrice au travers de l'algorithme de Levenshtein

5. L. Terriel, *Représenter et Évaluer Les Données Issues Du Traitement Automatique d'un Corpus de Documents Historiques. L'exemple de La Reconnaissance Des Écritures Manuscrites Dans Les Répertoires de Notaires Du Projet LectAuRep....*

giques et techniques d'un processus lourd et complexe. Dans ce cadre, nous allons observer au cours de ce chapitre les différentes réflexions et observations qui ont accompagné et nourrie cette insertion technique d'un pré-enrichissement automatique.

4.1 L'ébullition du traitement du langage naturel

Pour comprendre les progrès qui ont entouré l'incorporation du TAL dans le domaine des humanités numériques, il est nécessaire d'en décliner les principes généraux et les innovations récentes permises par la généralisation du *deep learning* et des nouvelles architectures affiliées.

En ce sens cette partie se veut comme un aparté pédagogique et introductif autant sur le plan théorique que technique.

4.1.1 Principes généraux du TAL

Le traitement automatique des langues est en réalité une discipline regroupant un ensemble de pratiques autour de l'analyse et l'interprétation des langues naturelles à travers l'outil informatique. Ce domaine pluridisciplinaire est ainsi à la croisée de la linguistique, de l'informatique, des mathématiques et de l'intelligence artificielle. Aujourd'hui, le TAL regroupe un ensemble de technologies de la vie courante et scientifique tel que les traducteurs automatiques, l'analyse marketing et des sentiments, le *data mining* (fouille de texte, en français), la correction orthographique, la reconnaissance vocale, etc... Quatre catégories peuvent être définies : le traitement syntaxique et sémantique, l'extraction d'information et le traitement du signal⁶.

Le TAL provient d'un très long processus de recherche remontant aux débuts des années 1960 avec un véritable engouement probabiliste à partir des années 1983 et 1984⁷. Dans un article paru en 2014, Ludovic Tanguy y distingue deux méthodes : la première dite « traditionnelle » ou linguistique consistant à formaliser les règles à appliquer et les ressources linguistiques. La seconde réside dans la mise en place de modèles mathématiques et statistiques applicables à l'intelligence artificielle⁸.

À l'heure actuelle, ce domaine de l'apprentissage machine s'appuie sur un ensemble de pratiques syntaxo-sémantiques bien précises, bâtit autour des modèles linguistiques, dont les principales fonctionnalités sont les suivantes :

6. Cette catégorisation reprend la déclinaison faite au sein de la page Wikipedia dédiée au TAL. *Wikipédia / Traitement automatique des langues*, url : https://fr.wikipedia.org/wiki/Traitement_automatique_des_langues, consulté le 20/08/2022.

7. Cette éclosion d'intérêt au sein des milieux scientifiques est permis avec la publication des premiers arbres de décisions et de d'un système d'étiquetage probabiliste. Ludovic Tanguy et Cécile Fabre, « Évolutions de la linguistique outillée : méfaits et bienfaits du TAL », *L'information grammaticale*-142 (2014), p. 15, URL : <https://halshs.archives-ouvertes.fr/halshs-01057493> (visité le 05/09/2022).

8. *Ibid.*

- La tokenisation qui consiste en la segmentation d'une phrase en mot ou d'un mot en caractères.
- Le *stemming* et la lemmatisation. La première technique désigne le découpage de la fin du mot pour en conserver la racine. La lemmatisation consiste quant à elle à supprimer les terminaisons « en isolant la forme canonique du mot⁹ ».
- L'étiquetage morpho-syntaxique (POS) est un processus reliant le mot à sa fonction grammaticale.
- La suppression des mots vides (*stop words* en anglais), c'est-à-dire les mots courants et vides de sens propres.
- Le Plongement de mots ou lexical (*word embedding* en anglais) est une méthode de représentation d'un mot sous forme de vecteur. Elle donne une valeur numérique au sens et au contexte du mot en établissant un vecteur contextuel.

4.1.2 Les réseaux neuronaux et le TAL : la révolution des modèles encodeurs-decodeurs

Aujourd’hui, le TAL s’est totalement inséré dans nos vies quotidiennes avec la renaissance du *machine learning*, après la crise qu’il a connu au cours des années 2000. Plus particulièrement, cette démocratisation a été permise par l’amélioration notable des performances avec l’utilisation du *deep learning*. Ces dernières avancées ont su puiser dans l’approche du *Word Embedding* afin de constituer des algorithmes statiques au texte à l’image du système *Word2Vec*, notamment le modèle Skip-Gram¹⁰.

En 1957, le linguiste John Rupert Firth émet l’hypothèse suivante : « You shall know a word by the company it keeps!¹¹ ». Cette hypothèse distributionnelle qui reconnaît un mot à partir de son contexte va être le socle du système *Word2Vec*. Alors, l’algorithme de transformation vers des valeurs algébriques tend à étendre la définition originale de la distribution en y incluant une valeur sémantique et contextuelle au travers d’espaces multidimensionnels. La distance vectorielle indique ainsi la proximité sémantique entre deux mots. Ces transformations des mots sous des modèles algébriques (matriciels et vectoriels) permettent ainsi de définir les poids des réseaux neuronaux appliqués.

Afin de traiter les données en séries, deux architectures neuronales sont principalement utilisées ; les premiers furent les réseaux neuronaux acycliques (RNN ; LSTM et GRU), aujourd’hui dépassés en raison des problèmes de « rétropropagation »¹². C'est-à-

9. Dominique Labbé, « Normalisation et lemmatisation d'une question ouverte » (), p. 21.

10. D’autres valorisent des approches moins lourdes de plongement de mots tel que le *Singular Value Decomposition* Chris Moody, *Stop Using Word2vec / Stitch Fix Technology – Multithreaded*, multithreaded, 18 oct. 2017, URL : <https://multithreaded.stitchfix.com/blog/2017/10/18/stop-using-word2vec/> (visité le 06/09/2022).

11. « Vous reconnaîtrez un mot à la compagnie qu'il garde ! » in Henry Widdowson, « J.R. Firth, 1957, Papers in Linguistics 1934–51 », *International Journal of Applied Linguistics*, 17–3 (2007), p. 402–413, DOI : 10.1111/j.1473-4192.2007.00164.x

12. T. Clérice, *Détection d'isotopies Par Apprentissage Profond : L'exemple de La Sexualité En Latin*

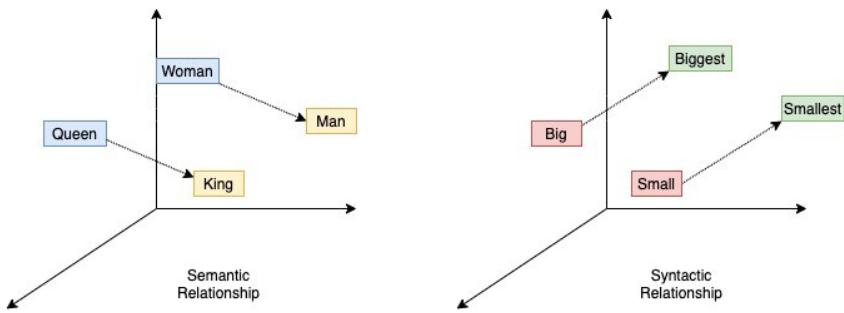


FIGURE 4.1 – Représentation vectorielle des relations sémantiques et syntaxiques - ©towardsdatascience, 2021

dire que l'algorithme concentre davantage son attention sur la fin d'un élément. Ce phénomène est réglé par les réseaux CNN en prenant l'élément dans sa globalité¹³. Ce réseau s'exerce en deux étapes : la partie convolutive qui se tâche de l'extraction des caractéristiques propres à travers différents algorithmes de transformation (*pooling*, échantillonage, etc...), puis la classification des données.

Depuis la fin des années 2010, le TAL s'est dirigé vers les architectures encodeurs-décodeurs. Cette innovation se fonde sur les réseaux acycliques et l'application du mécanisme de l'attention, du système *Masked LM*, l'application d'un masque aléatoire permettant une approche simultanée du traitement, et l'approche du *Dynamic Embedding*¹⁴. En 2018, Google présente le modèle BERT (*Bidirectional Encoder Representations from Transformers*) qui étalonne en de multiples étapes cycliques le traitement textuel *via* une architecture encodeur-décodeur¹⁵. Ces modèles de langage appellent plus couramment appelés modèles *Transformers*. Ces modèles de langages sont des outils de pré-entraînement de modèles plus perfectionnés, à travers la méthode du *fine-tuning*.

De nombreux modèles de langage ont fait leurs apparitions à la suite de cette révolution au sein du TAL et la démocratisation des modèles *Transformers*¹⁶. En 2020, l'Universidad de Chile développe et publie le modèle BETO qui reprend les principes de cette architecture pour l'adapter à la langue espagnole, en utilisant la méthode du

Classique et Tardif, thèse de doctorat en lettres et civilisations antiques, dir. Christian Nicolas, Lyon, Université Lyon 3, 2022, URL : <https://github.com/PonteIneptique/these-redaction/releases/tag/1.0.1>, p. 161.

13. *Ibid.*, p. 164.

14. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser et Illia Polosukhin, *Attention Is All You Need*, 5 déc. 2017, DOI : 10.48550/arXiv.1706.03762, arXiv : 1706.03762 [cs].

15. Jacob Devlin, Ming-Wei Chang, Kenton Lee et Kristina Toutanova, *BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding*, 24 mai 2019, DOI : 10.48550/arXiv.1810.04805, arXiv : 1810.04805 [cs].

16. Il est à noter que cette architecture démontre aussi des capacités prometteuses pour le cas de la reconnaissance d'écriture. Dmitrijs Kass et Ekta Vats, *AttentionHTR : Handwritten Text Recognition Based on Attention Encoder-Decoder Networks*, 1^{er} avr. 2022, arXiv : 2201.09390 [cs], URL : <http://arxiv.org/abs/2201.09390> (visité le 06/09/2022)

TASK	BETO-cased	BETO-uncased	Best Multilingual BERT
POS	98.97	98.44	97.10
NER-C	55.43	82.67	87.38
MLDoc	95.60	96.12	95.70
PAWS-X	89.05	89.55	90.70
XNLI	82.01	80.15	78.50

TABLE 4.1 – Évaluation des performances du modèle BETO (selon standard GLUE)

fine-tuning avec le modèle BERT¹⁷. Le *dataset* a été constitué à partir des données de Wikipedia et du projet OPUS. Comme l'indique le tableau 4.1, les résultats du modèle sont très satisfaisants, plus particulièrement pour le modèle prenant en compte la case dans son analyse. Ce modèle constituera notre point d'appui pour le développement du projet d'enrichissement de nos éditions numériques.

4.1.3 Le TAL, la reconnaissance des entités nommées et les humanités

L'extraction d'information est une catégorie générale de la linguistique computationnelle dont le principe repose sur l'analyse, une sélection et une classification pertinentes d'une donnée textuelle¹⁸. Elle regroupe de nombreuses tâches permettant de déstructurer et reconstruire l'information telle que l'identification des relations et des évènements, l'extraction de propriétés, la résolution des coréférences ou dans ce qui nous concerne, la reconnaissance des entités nommées.

Le concept linguistique des entités nommées définit le principe d'une classification des mots selon son fonctionnement référentiel. Dans sa thèse sur les processus d'extractions des entités nommées, Maud Ehrmann les décrit comme « Étant donné 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.¹⁹ ». Cette définition décrit des propriétés malléables dépendantes d'une ontologie définie préalablement, dont la référence sémantique est alors primordiale et la référence syntaxique secondaire (exemple : les noms propres). Autrement dit, la nature d'une entité nommée ne réside pas dans son essence, mais bien dans son existence applicative²⁰.

Comme le présente la figure 4.2, la reconnaissance d'entités nommées est une tech-

17. José Canete, Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang et Jorge Pérez, *Spanish Pre-Trained BERT Model and Evaluation Data*, Santiago (Chili), 2020, URL : <https://github.com/dccuchile/beto>.

18. T. Poibeau et Adeline Nazarenko, « L'extraction d'information, Une Nouvelle Conception de La Compréhension de Texte ? », *Traitement Automatique des Langues*, 2-40 (1999), p. 87-115.

19. Maud Ehrmann, *Les Entités Nommées, de La Linguistique Au TAL : Statut Théorique et Méthodes de Désambiguisation*, thèses d'informatique et langage, Paris Diderot University, 2008, URL : <https://hal.archives-ouvertes.fr/tel-01639190> (visité le 06/09/2022), p. 167-168.

20. *Ibid.*

Robespierre PERS a été député à la Convention Nationale ORG et membre du Comité de Salut Public ORG. Il est né à Arras LOC en 1758.

FIGURE 4.2 – Exemple de reconnaissance d’entités nommées à partir du modèle camem-BERT

nique consistant à repérer ces chaînes de caractères au sein d’un texte et d’en déterminer son référentiel. Ici, l’ontologie employée se décline en quatre grandes natures : LOC pour les lieux géographiques, ORG pour les organisations, PERS pour les personnes et MISC pour les entités diverses et regroupées au sein de cette catégorie générale (non référencée dans notre cas). Cette approche de classification est parmi les plus communes puisqu’elle reprend les règles définies par CoNLL 2003, une conférence annuelle organisée par la SIGNLL (*ACL’s Special Interest Group on Natural Language Learning*)²¹. Toutefois, cette normalisation ne peut faire l’objet d’un véritable consensus puisque l’étiquetage reste dépendant des sujets traités en prenant en compte les limites, la portée et la granularité d’une entité²².

Les techniques de reconnaissances des entités nommées

Le processus de reconnaissance des entités nommées peut s’exercer à travers deux approches techniques, non nécessairement dichotomiques. La première est l’étiquetage des entités par l’intervention humaine en définissant des règles appliquées à partir d’une ontologie préalable.

Cette méthode majoritairement répandue dans les années 1990 est décrite comme la mise en forme de « patrons d’extraction » en exploitant les indices morpho-syntaxique et un ensemble de ressources externes²³. Pour reprendre l’exemple de Maud Ehrmann, si Maximilien est un nom connu au sein des ressources utilisées et que le nom suivant Robespierre est inconnu, mais qu’il possède une majuscule et suit la première entité ; alors l’approche statistique peut facilement déduire que Robespierre appartient à l’entité PERS²⁴. Toutefois, la mise en place de motifs probabilistes reste parfois insuffisant en raison d’un paramétrage intrinsèquement trop large comme le signale Damien Nouvel²⁵.

Cette approche a été confortée avec l’utilisation de l’intelligence artificielle qui connaît un nouveau souffle après les années 2000. Elle nécessite l’emploi d’une base de données afin de pouvoir y entraîner différents modèles. Parmi les approches courantes, on

21. Jens Lemmens, *CoNLL 2022 / CoNLL*, CoNLL, 2022, URL : <https://conll.org/> (visité le 28/08/2022).

22. Simon Hengchen, Seth van Hooland, Ruben Verborgh et Max De Wilde, « L’extraction d’entités nommées : une opportunité pour le secteur culturel ? », *I2D - Information, données & documents*, 52–2 (2015), p. 70-79, DOI : 10.3917/i2d.152.0070.

23. M. Ehrmann, *Les Entités Nommées, de La Linguistique Au TAL...*, p. 32.

24. *Ibid.*, p. 33.

25. Damien Nouvel, *Reconnaissance Des Entités Nommées Par Exploration de Règles d’annotation : Interpréter Les Marqueurs d’annotation Comme Instructions de Structuration Locale*. These de doctorat, Tours, 2012, URL : <http://www.theses.fr/2012TOUR4011> (visité le 24/08/2022), p. 146.

peut y recenser :

- L'apprentissage non-supervisé en utilisant des techniques de *clustering*, c'est-à-dire de regroupement de données selon une interprétation machine²⁶.
- L'apprentissage supervisé, utilisant donc des données préalablement étiquetées dont les principaux algorithmes utilisés sont les arbres de décisions, le *Machine Support Vector* (SVM) ou le système Markov caché²⁷. Ces modèles sont assez similaires aux modèles classiques de HTR.
- Les algorithmes d'apprentissages profonds se sont imposés comme l'architecture dominante depuis ces dernières années dans les tâches de reconnaissances d'entités nommées²⁸. L'ébullition des architectures *Transformers* a amplifié ce phénomène en raison de ces résultats (voir tableau 4.1). De plus, l'apprentissage profond permet de traiter facilement des données complexes et l'obtention d'un modèle multi-tâche, facilement intégrable au sein d'une chaîne de traitement.

La REN et les humanités

L'intégration des technologies REN a susciter de nombreux émois au sein des secteurs patrimoniaux et des humanités. La place de plus en plus grande qu'occupent les humanités numériques au sein de la recherche a multiplié l'intérêt pour cette tâche. La préparation de données statiques ou le renouveau historiographique permis par la méthode de *Distant reading*²⁹ ont généralisée la fouille de texte, dont l'extraction des entités nommées permet une annotation efficace des documents.

Dans le même temps, le catalogage et l'indexation manuel sont mis à rude épreuve depuis plusieurs années en raison des restrictions budgétaires au sein des secteurs culturels et la multiplication des documents archivés³⁰. Face à ce constat, les institutions patrimoniales s'essayent à réorganiser le processus d'indexation avec une utilisation systémique des technologies REN. Le projet NER4Archives développés par les Archives Nationales en partenariat avec l'INRIA démontre cette volonté d'automatiser la classification au sein des instruments de recherche archivistique³¹.

La volonté d'incorporer cette tâche de classification de l'information n'est pas nouvelle et s'appuie sur de nombreuses expériences antérieures d'application sur les documents

26. Jing Li, Aixin Sun, Jianglei Han et Chenliang Li, *A Survey on Deep Learning for Named Entity Recognition*, 18 mars 2020, arXiv : 1812.09449 [cs], URL : <https://arxiv.org/abs/1812.09449> (visité le 07/09/2022).

27. *Ibid.*

28. *Ibid.*

29. Méthode d'analyse historique imaginé par Franco Moretti. Elle consiste en l'application des méthodes calculs et d'observations de motifs à l'échelle de grandes bases de données. Marie Puren, *La Lecture Distante : Introduction et Exemples d'application*, France, nov. 2020, URL : <https://hal.archives-ouvertes.fr/hal-03152747> (visité le 07/09/2022)

30. S. Hengchen, S. van Hooland, R. Verborgh, *et al.*, « L'extraction d'entités nommées... ».

31. F. Clavaud, L. Romary, P. Charbonnier, *et al.*, « NER4Archives (Named Entity Recognition for Archives)... ».

historiques, notamment au sein d'une chaîne de traitement d'océrisation. Les modèles *Transformers* ont démontré une certaine capacité dans le traitement et l'extraction des informations au sein de corpus numérisés bruyants sans pour autant en dégrader les performances³². En 2020, Manuel Carbonell et Al. vont encore plus loin et démontrent la capacité d'amélioration de la reconnaissance de caractères manuscrits au sein d'un modèle unifié avec la reconnaissance d'entités nommées et la localisation de texte³³. Dans notre cas, plusieurs projets ont intégré l'automatisation de la REN au sein des chaînes de traitements d'éditions numériques à l'image des projets AGODA (Analyse sémantique et Graphes relationnels pour l'Ouverture et l'étude des Débats à l'Assemblée nationale) ou encore le LECTAUREP³⁴.

Ces différents constats prometteurs, nous amenés à considérer et évaluer l'intégration de la REN au cours de l'édition du corpus autour de l'Occupation de l'Araucanie.

4.2 Entrainer la reconnaissance d'entités nommées

Face au constat de la difficile adaptation des modèles REN généralistes sur les données océrisées d'un espagnol latino-américain du XIX^e siècle³⁵, il a été rapidement décidé d'expérimenter la mise en place d'un modèle propre au projet et d'enrichir l'annotation. Néanmoins, ce choix ne peut être effectué sans prendre en considération le coût humain, mais aussi énergétique et écologique que représente la génération d'un modèle *Transformers*. La prise de conscience de cet aspect est une nécessité face à la massification des modèles et méga-modèles TAL et de *machine learning* dans leur globalité³⁶.

À travers cette aspiration, nous allons observer comment peut on développer un modèle REN et les enjeux qui ont entouré la production des données et l'apprentissage.

32. Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere et A. Doucet, « Alleviating Digitization Errors in Named Entity Recognition for Historical Documents », dans *Proceedings of the 24th Conference on Computational Natural Language Learning*, Online, 2020, p. 431-441, DOI : 10.18653/v1/2020.conll-1.35.

33. M. Carbonell, A. Fornés, M. Villegas, et al., *A Neural Model for Text Localization, Transcription and Named Entity Recognition in Full Pages...*

34. Nicolas Bourgeois, Aurélien Pellet et M. Puret, « Using Topic Generation Model to Explore the French Parliamentary Debates during the Early Third Republic (1881-1899) », dans *DiPaDA 2022 Digital Parliamentary Data in Action 2022. Proceedings of the Digital Parliamentary Data in Action (DiPaDA 2022) Workshop Co-Located with 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, dir. Matti La Mela, Fredrik Norén et Eero Hyvönen, 2022 (CEUR Workshop Proceedings), t. 3133, p. 35-51, URL : <https://hal.archives-ouvertes.fr/hal-03526254> (visité le 07/09/2022) ; H. Scheithauer, *La Reconnaissance d'entités Nommées Appliquées à Des Données Issues de La Transcription Automatique de Documents Manuscrits Patrimoniaux. Expérimentations et Préconisations à Partir Du Projet LECTAUREP...*

35. Différents modèles (BETO, Flair) ont été expérimentés sur des vérités terrains via l'interface HuggingFace.

36. Emma Strubell, Ananya Ganesh et Andrew McCallum, *Energy and Policy Considerations for Deep Learning in NLP*, 5 juin 2019, DOI : 10.48550/arXiv.1906.02243, arXiv : 1906.02243 [cs].

4.2.1 Annoter un jeu de données pour la reconnaissance d'entités nommées

Le développement d'un modèle de reconnaissance REN est le fruit d'un long processus d'annotation des entités à partir d'un corpus de données. La mise en place d'un système de reconnaissance d'entités nommées à partir de documents historiques rencontre trois défis majeurs selon une étude dirigée par Maud Ehrmann et al.³⁷. Le premier est ce qu'on appelle la « dynamique de langues » c'est-à-dire la prise en compte des variations orthographiques historiques (et du niveau de langue dans notre cas), les conventions de noms ou encore les entités contextuelles dont le sens évolue à travers le temps. Les autres points de difficultés sont la gestion des espaces, notamment pour les documents historiques médiévaux, et la gestion du bruit suite aux procédés de reconnaissances HTR. Dans de nombreux cas, il est possible de s'appuyer sur des corpus annotés existant au sein d'un « lac de ressources » afin d'adapter des modèles ayant déjà fait leur preuve. Néanmoins, la langue espagnole n'est présente que marginalement et ne peut pas répondre à nos besoins spécifiques³⁸. Les données de vérité terrain ont été directement reprises de notre jeu utilisé pour l'entraînement d'un modèle HTR.

Afin de saisir notre jeu de données final, il nous faut préalablement revenir sur le protocole d'annotations et le pré-traitement des données ALTO. Tout d'abord, les transcriptions ont été nettoyées des méthodes utilisées pour la transformation du format ALTO vers le format TEI avec l'utilisation de REGEX, avant d'être converti sous le format texte qui définit comme notre standard de donnée. Le pré-traitement est volontairement minimaliste, car l'objectif de ce modèle est d'être capable d'analyser correctement à partir de données bruitées issues de la REM. Il doit être capable de poursuivre son analyse au-delà des dynamiques de langues ou des erreurs d'océrisations, malgré un impact minimal. Confirmant les premières constatations observées lors de la conférence *24th Conference on Computational Natural Language Learning*³⁹, Les expériences de Hugo Scheithauer au sein du projet LECTAUREP démontrent l'impact très relatif du CER sur les performances des modèles TAL en dessous de 20%⁴⁰. Le point primordial est donc de former à la compréhension des abréviations et la segmentation de certaines entités, notamment à la suite de saut de ligne.

37. M. Ehrmann, A. Hamdi, Elvys Linhares Pontes, Matteo Romanello et A. Doucet, *Named Entity Recognition and Classification on Historical Documents : A Survey*, 23 sept. 2021, arXiv : 2109.11406 [cs], URL : <http://arxiv.org/abs/2109.11406> (visité le 08/09/2022).

38. Seul un corpus de transcriptions médiévales espagnols et un jeu concernant des données bibliographiques multilingues ont été identifiés. *Ibid.*

39. E. Boros, A. Hamdi, E. Linhares Pontes, *et al.*, « Alleviating Digitization Errors in Named Entity Recognition for Historical Documents »...

40. H. Scheithauer, *La Reconnaissance d'entités Nommées Appliquées à Des Données Issues de La Transcription Automatique de Documents Manuscrits Patrimoniaux. Expérimentations et Préconisations à Partir Du Projet LECTAUREP...*, p. 102-104.

Par la suite, nous avons établi un protocole d'annotation à la fois concernant l'ontologie à appliquer et le processus. Pour ce dernier, nous avons donc établi qu'une première personne effectue une annotation avant d'être vérifiée par une seconde personne afin d'obtenir un consensus sur les étiquetages ambigus. Certaines initiatives scientifiques ont publié certains schémas possibles en fonction de la nature du référentiel. En reprenant certaines recommandations du Manuel d'annotation linguistique publié par les chercheurs Simon Gabay et al.⁴¹, nous avons établi 5 groupes d'entités selon les conditions suivantes :

- LOC – Lieux géographiques ou administratifs (564 entités annotées)
- ORG – Organisations administratives et institutionnelles (398 entités annotées)
- PERS – Personne ou groupes de personnes (1022 entités annotées)
- DATE – Dates absolues ou relatives, voir événement (262 entités annotées)
- MISC – Entités d'intérêts et inclassables au sein des référentiels précédents : navires, objets de ravitaillements, stratégies, etc... (430 entités annotées)

L'entité MISC est volontairement englobante et peu précise. L'intention était d'observer la capacité d'identification d'éléments plus précis notamment le ravitaillement (PROD) ou encore les bateaux à vapeur dont les premières hypothèses appuient l'utilisation d'un sous-groupe à ORG ou l'ajout d'une entité propre (BAT par exemple). Souvent un compromis doit-être fait entre précision et simplification afin d'améliorer la performance du modèle et la capacité d'extraction de données⁴².

Si il est possible d'exercer une pré-annotation automatique à partir de modèle générique de NER ou d'appliquer des systèmes de règles, nous avons préféré annoter manuellement afin de s'approprier plus facilement le corpus documentaire, mais aussi les stratégies d'annotations possibles pour un corpus relativement modeste (180 documents). Pour ce faire, nous avons privilégiée la plateforme d'annotation *open-source Doccano* pour son accessibilité et son orientation vers la librairie de TAL SpaCy⁴³. Cet outil multifonctions d'annotation de données a été installé en local à partir du système Docker. À l'image de la figure 4.3, les différentes séquences de caractères sont étiquetées de manière exclusive, c'est-à-dire à dire qu'une entité ne peut pas appartenir à une entité plus large afin d'éviter les futures confusions possibles. Cette définition a déjà été faite au sein de la conférence CoNLL 2003 où les chercheurs Erik F. Tjong Kim Sang et Fien de Meuder considéraient les entités nommées comme « non-récursive » et « *non-overlapping* »⁴⁴.

41. S. Gabay, J.B. Camps et T. Clérice, *Manuel d'annotation Linguistique Pour Le Français Moderne (XVIe - XVIIIe Siècles)*, avr. 2022, URL : <https://hal.archives-ouvertes.fr/hal-02571190> (visité le 07/09/2022).

42. Cyril Grouin, « Simplification de Schémas d'annotation : Un Aller sans Retour ? », dans *Actes de TALN*, Rennes, France, 2018, URL : <https://hal.archives-ouvertes.fr/hal-01831221> (visité le 08/09/2022).

43. Cette plateforme a été programmé avec le langage python et le framework Django. Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi et Xu Liang, *Doccano*, version 1.8.0, doccano, 8 sept. 2022, URL : <https://github.com/doccano/doccano> (visité le 08/09/2022).

44. Erik F. Tjong Kim Sang et Fien De Meulder, *Introduction to the ConLL-2003 Shared Task : Language-Independent Named Entity Recognition*, 12 juin 2003, arXiv : cs/0306050, URL : <http://arxiv.org/abs/cs/0306050>

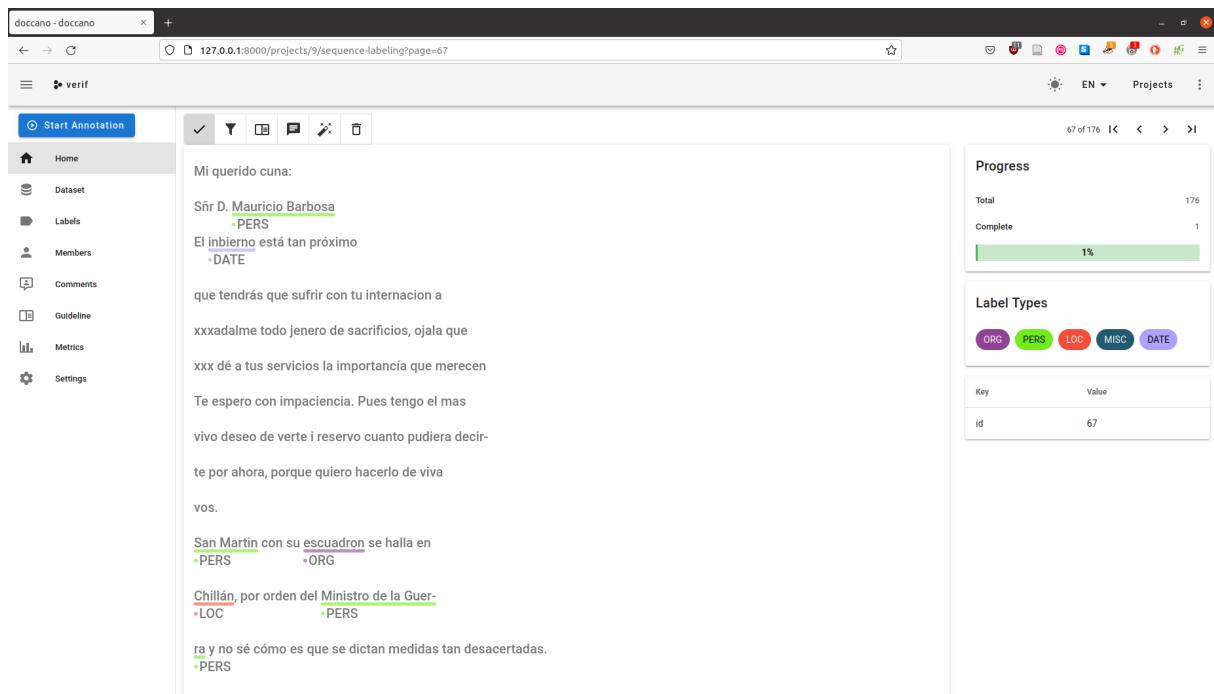


FIGURE 4.3 – Démonstration de la plateforme Doccano

À l’heure actuelle, Doccano limite l’exportation au format JSONL (pour JSON Lines), c’est-à-dire un fichier JSON dont la structuration est décomposée en plusieurs lignes. Si ce n’est pas le plus courant contrairement au format CoNLL ou autres, il permet d’obtenir une annotation simple, lisible, sans marquage et ainsi sans dissection des entités.

Les schèmes IOB⁴⁵, BIOES⁴⁶, BILOU⁴⁷ pour les plus connus ont pour fonction d’offrir ce que l’on appelle une analyse syntaxique de surface. Le choix n’est pas anodin, car il peut avoir une incidence à la fois sur l’interopérabilité des données, mais aussi sur les performances du modèles⁴⁸. Nous avons choisi de laisser les annotations sans ce pré-marquage afin de conserver une facilité d’interagir avec les données, mais aussi d’éviter un impact négatif sur l’entraînement de notre modèle.

arxiv.org/abs/cs/0306050 (visité le 09/09/2022).

45. I : Inside (token à l’intérieur), O : Output (token à l’extérieur), B : Beginning (token au début de l’entité). Le schème standard du format CoNLL

46. B : Beginning (début de l’entité), I : Inside (token à l’intérieur), O : Output (token à l’extérieur), E : Ending (token de fin de l’entité), S : Single element (entité constituée d’un seul token)

47. B : Beginning (token au début de l’entité), I : Inside (token à l’intérieur), L : Last (token de fin), O : Output (token à l’extérieur), U : Unit (entité constituée d’un seul token). Il est format de préférence utilisé par SpaCy.

48. Nasser Alshammari et Saad Alanazi, « The Impact of Using Different Annotation Schemes on Named Entity Recognition », *Egyptian Informatics Journal*, 22–3 (1^{er} sept. 2021), p. 295-302, DOI : 10.1016/j.eij.2020.10.004.

4.2.2 Produire un modèle appliquée à la reconnaissance des entités nommées

La généralisation du *deep learning* a permis à l'indexation REN des documents historiques de faire bondir les scores (F1-score) de 10 à 20%⁴⁹. Avec les innovations entourant les architectures encodeurs-décodeurs, cette augmentation des résultats est encore plus probante sur la reconnaissance des entités nommées au sein des documents historiques au delà des problèmes qu'ils impliquent, et ce malgré un corpus pouvant être rudimentaire (50 documents annotés)⁵⁰.

Entraîner un modèle NER en validation croisée

Au regard de ces observations, nous nous sommes essayé au développement d'un modèle REN à partir du modèle pré-entraîné BETO que nous avons vu précédemment⁵¹. Ce pré-entraînement permet la modification du modèle initiale afin de lui donner une tâche précise grâce au *finetuning* et ainsi modifier les poids au sein de l'architecture neuronale selon nos besoins. Pour ce faire, de nombreuses librairies python permettent de mettre en œuvre ce type d'entraînement à partir de modèles pré-entraînés *Transformers*. SpaCy et HuggingFace sont sans doute à l'heure actuelle les plus populaires, car elles offrent des solutions TAL clés en main, ergonomiques et accessibles. Si elles sont par nature restrictives, elles restent suffisantes à l'ambition du projet qui ne nécessite pas une refonte neuronale totale contrairement à des projets amplement plus complexes. Nous avons privilégié l'utilisation de la librairie SpaCy en raison de la galaxie de solutions *softwares* qui l'entourent, le maintien d'une documentation exhaustive et extrêmement pédagogique avec la possibilité d'utiliser les modèles *Transformers* grâce à la mise en place d'une *pipeline* pour l'utilisation des modèles originalement disponibles au sein de la librairie HuggingFace⁵².

Au préalable, il est à noter que le processus d'apprentissage de modèles *Transformers* nécessite une grande puissance de calcul. Dans un premier temps, nous avons souhaité utilisé la plateforme Google Colab⁵³ mettant à disposition l'utilisation de GPU par l'intermédiaire de leurs serveurs. Par la suite, nous avons utilisé un GPU personnel pouvant

49. M. Ehrmann, A. Hamdi, E. L. Pontes, *et al.*, *Named Entity Recognition and Classification on Historical Documents...*

50. N. Abadie, E. Carlinet, J. Chazalon et B. Duméniel, « A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories », dans *Document Analysis Systems. DAS 2022*. Dir. S. Uchida, E. Barney et V. Eglin, La Rochelle, France, 2022 (Document Analysis Systems. DAS 2022. 13237), DOI : 10.1007/978-3-031-06555-2_30.

51. J. Canete, Chaperon, R. Fuentes, *et al.*, *Spanish Pre-Trained BERT Model and Evaluation Data...*

52. *Spacy-Transformers : Use Pretrained Transformers like BERT, XLNet and GPT-2 in spaCy*, version 1.1.8, Explosion, 7 sept. 2022, URL : <https://github.com/explosion/spacy-transformers> (visité le 08/09/2022).

53. *Google Colab*, <https://colab.research.google.com/>, consulté le 04/09/2022.

avoir eu une incidence sur les résultats du modèles⁵⁴.

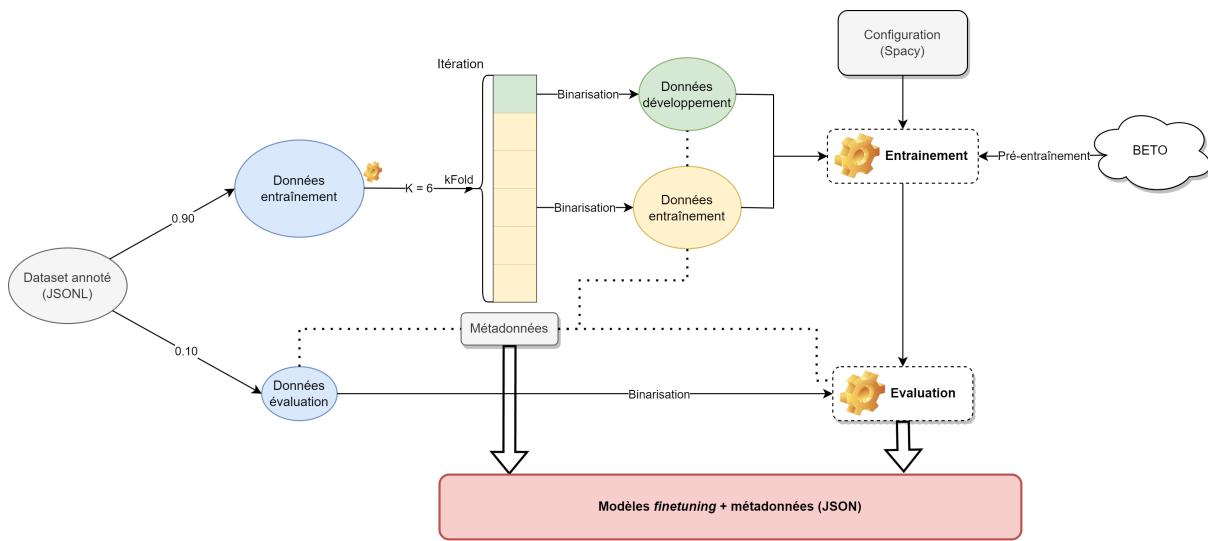


FIGURE 4.4 – Schéma du processus d'entraînement d'un modèle NER en validation croisée⁵⁵

Afin de mettre en oeuvre ce processus d'entraînement, nous sommes aidés de la librairie **scikit-learn** mettant à disposition un grand nombre d'outils appliqués au *machine learning*⁵⁶. Plus précisément, le processus s'appuie sur la séparation des données en plusieurs dépôts utilisable au cours de l'entraînement (un jeu d'entraînement et un jeu d'évaluation) afin d'offrir des jeux avec le moindre biais humain.

Dans un second temps, nous pouvons observer au sein de la figure 4.4 la présence d'une itération autour de *kFold*. Cet élément correspond à la mise en place d'un procédé de validation croisée. Cette méthode K-fold, permet de tester et améliorer les performances d'un modèle prédictif en exploitant l'ensemble de notre jeu de données d'entraînement, en constituant des *samples* (κ) de données partagées pour l'entraînement et sa validation. Cet technique d'apprentissage permet de maximiser le potentiel du jeu original en le prenant dans son intégralité, particulièrement quand celui-ci est restreint⁵⁷. Les données d'évaluation restent les mêmes tout au long du processus afin de fournir des données tests stables et ainsi fournir une évaluation fiable.

Chaque itération détermine un set de données d'entraînement et de validation (qui ne pourra donc plus l'être par la suite) qui sont alors binarisées pour être exploitables par le CLI de SpaCy. Le programme va effectué une première modélisation avec le modèle pré-

54. Nvidia RTX 3080, CUDA 11.3, cuDNN

55. M. Humeau, *Entraînement et Annotations NER*, avec la coll. d'Alessandro Chiaretti, Archivo Central Andres Bello, juin 2022, URL : https://github.com/Proyecto-Ocupacion-Araucania-UChile/NER_Araucania

56. *Scikit-Learn/Scikit-Learn*, version 1.1.2, scikit-learn, 9 sept. 2022, URL : <https://github.com/scikit-learn/scikit-learn> (visité le 09/09/2022).

57. Adam Persson, « The Effect of Excluding Out of Domain Training Data from Supervised Named-Entity Recognition », dans *Proceedings of the 21st Nordic Conference on Computational Linguistics*, Gothenburg, Sweden, 2017, p. 289-292, URL : <https://aclanthology.org/W17-0240> (visité le 09/09/2022).

entraîné BETO avant d'incorporer les données d'entraînement au sein de l'apprentissage. Chaque étape se conclut par un processus d'évaluation qui détermine ses performances à chaque *epoch* à partir des données d'évaluation. Une fois le meilleur modèle estimé, celui-ci est évalué à partir des données tests dont les résultats sont associés à un fichier JSON.

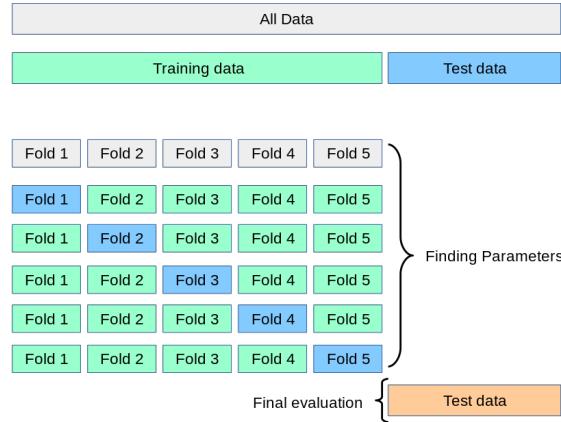


FIGURE 4.5 – Explication du système de validation croisée à partir de la méthode K-fold – ©scikit-learn

Interprétation des résultats et améliorations possibles

À partir de ces résultats, nous pouvons comparer les différents modèles afin de déterminer le plus performant d'entre eux⁵⁸. Cette phase d'évaluation consiste à estimer la capacité du modèle à sélectionner la bonne chaîne de caractères et d'attribuer à cette chaîne le référentiel adéquat. Pour cela, trois mesures basées sur la matrice de confusion sont considérées les plus pertinentes par la communauté scientifique : Précision, Rappel, F1-score⁵⁹. Elles permettent d'avoir une perception globale du modèle en évaluant le taux de bonnes réponses sur les supposées entités repérées. Le score F1-score décrit une valeur globale en conciliant les deux mesures⁶⁰.

Selon les mesures présentées, les différents modèles ont été analysés selon leurs performances comme nous pouvons l'observer au sein de la figure 4.6. Au regard des données qui ont été évaluées, deux modèles sont apparus comme les plus solides : le modèle κ_2 et le modèle κ_6 . À première vue, le modèle κ_2 apparaît comme le plus satisfaisant, avec un taux de surpositivité plus faible comme le montre la valeur de Rappel. À la suite d'une comparaison plus approfondie entre les deux modèles sélectionnés, nous avons pu observer une lacune générale sur les entités MISC qui peut s'expliquer par la forte diversité de celle-ci. Toutefois, certaines entités semblent plus accessibles que d'autres à l'image des

58. L'ensemble des données sont consultables au sein de l'annexe F.

59. J. Li, A. Sun, J. Han, *et al.*, *A Survey on Deep Learning for Named Entity Recognition...*

60. D'autres métriques plus complexes existent aussi pour l'évaluation des modèles REN tel que la mesure SER, mais nous avons conserver les mesures proposées par spaCy. M. Ehrmann, A. Hamdi, E. L. Pontes, *et al.*, *Named Entity Recognition and Classification on Historical Documents...*

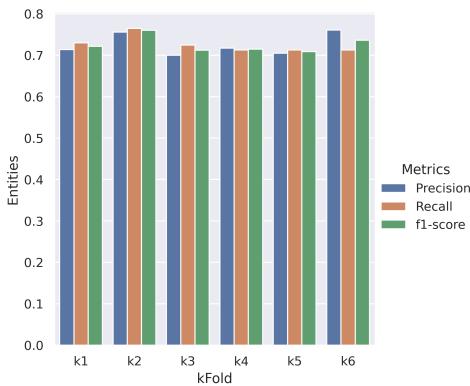


FIGURE 4.6 – Résultats globaux par κ selon les trois mesures principales

navires qui sont facilement identifiables⁶¹. En revanche, on observe une légère supériorité du modèle κ_2 concernant les entités de type LOC, PERS et plus légèrement ORG. Sur l'ensemble, le modèle κ_2 semble produire plus de bruit dans la reconnaissance des données, ce qui s'observe avec le fort contraste entre le taux de précision et le taux de rappel.

Face à ce constat, nous avons privilégié l'emploi du modèle κ_2 qui semblent présenter des performances globales plus pertinentes dans le cadre du projet, avec une attention particulière sur les entités LOC et PERS estimés les plus indispensables à l'édition en vue de l'enrichissement des données. Sans être excellente, les performances du modèle REN peuvent être classé comme positives si on le compare à d'autres expérimentations, plus particulièrement en raison de la faible quantité de données ajoutées⁶². De plus le réseau neuronal semble avoir rapidement assimilé l'entité DATE qui est souvent présentée de manière structurée. En revanche, on peut remarquer une forte fébrilité sur la reconnaissance de certains notamment « Usted » pouvant être associé à la fois au référentiel LOC, MISC ou ORG. De même, il semble parfois amalgamer certaines entités avec leurs prépositions.

Afin de perfectionner ce modèle, deux solutions possibles ont été décrites afin de faire face aux problèmes constatés. La première est de spécifier plus en détail les entités MISC qui sont trop généralistes. L'autre axe d'amélioration est ce que l'on appelle la désambiguïsation et la relation des entités nommées (*entity linking* en anglais) qui n'ont pas pu être mises en place pour l'instant faute de temps⁶³. Le modèle doit être capable de reconnaître la référence d'un mot au-delà de sa polysémie. Par exemple, « Arauco » peut signifier à la fois une ville et une administration régionale. En général, cette désambiguïsation s'appuie sur l'entraînement d'un modèle à partir d'une base de connaissance extérieure en identifiant les éléments polysémiques⁶⁴. Cette stratégie est appelée *entity linking* et elle permet

61. Voir l'annexe G.

62. H. Scheithauer, *La Reconnaissance d'entités Nommées Appliquées à Des Données Issues de La Transcription Automatique de Documents Manuscrits Patrimoniaux. Expérimentations et Préconisations à Partir Du Projet LECTAUREP...*, p. 105-109.

63. M. Ehrmann, A. Hamdi, E. L. Pontes, et al., *Named Entity Recognition and Classification on Historical Documents...*

64. Razvan Bunescu et Marius Pașca, « Using Encyclopedic Knowledge for Named Entity Disambi-

l’identification du mot, par une approche vectorielle ou syntaxique, en l’alignant au près de grandes bases de connaissances accessibles *via* le web sémantique⁶⁵.

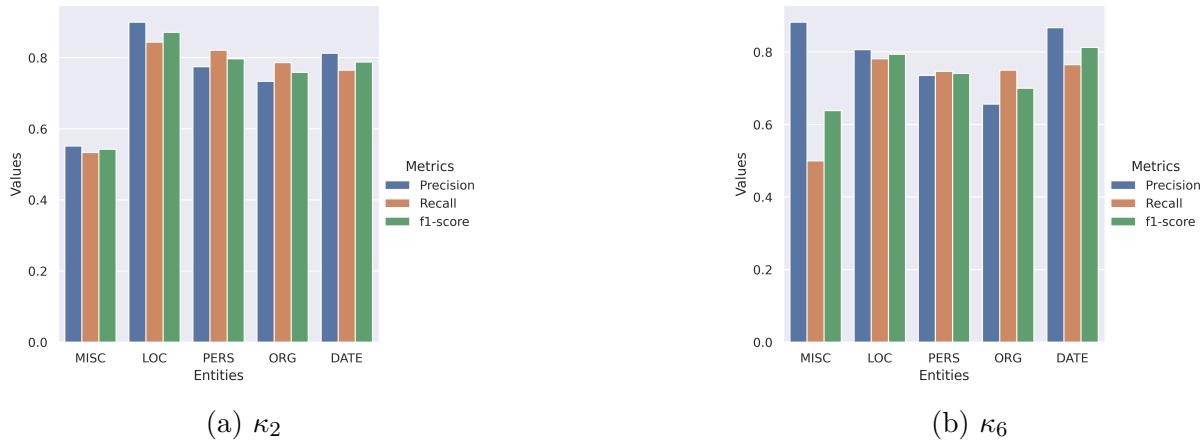


FIGURE 4.7 – Résultats détaillés par entité

	Precision	Rappel	F1-score
Général	0.7556	0.7643	0.7600
MISC	0.5517	0.5333	0.5423
LOC	0.9000	0.8437	0.8709
DATE	0.8125	0.7647	0.7878
PERS	0.7746	0.8208	0.7971
ORG	0.7333	0.7857	0.7586

TABLE 4.2 – Évaluation des performances du modèle κ_2

4.3 Indexer, enrichir et exploiter les données

La reconnaissance des entités nommées offre une double opportunité. Elle permet dans un premier temps d’indexer l’ensemble des entités et ainsi faciliter l’accès à cette donnée pour l’utilisateur ou l’utilisatrice. L’autre aspect est la capacité de renforcer les mécanismes d’enrichissement au cours du processus éditorial. Cet enrichissement est en général un exercice conséquent et pour autant nécessaire dans la constitution d’un appareil critique⁶⁶.

guation », dans *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, 2006, p. 9-16, URL : <https://aclanthology.org/E06-1002> (visité le 09/09/2022).

65. José G. Moreno, Romaric Besançon, Romain Beaumont, Eva d’Hondt, Anne-Laure Ligozat, Sophie Rosset, Xavier Tannier et Brigitte Grau, « Apprendre Des Représentations Jointes de Mots et d’entités Pour La Désambiguisation d’entités », dans *24ème Conférence Sur Le Traitement Automatique Des Langues Naturelles - TALN 2017*, Orléans, France, 2017, URL : <https://hal.archives-ouvertes.fr/hal-01626197> (visité le 09/09/2022).

66. J.B. Camps, « Où va la philologie numérique ? »...

Dans cette perspective nous allons observer la mise en place d'un processus d'indexation et d'enrichissement automatique grâce à l'aide de l'*open data* et les problématiques que ce procédé impose.

4.3.1 Indexer les entités nommées au sein d'une édition numérique TEI

La première difficulté qui a été rencontré dans l'application d'un schéma d'indexation a été simplement d'appliquer la reconnaissance au sein d'un fichier structuré et de conserver cette structuration, dans notre cas XML et de l'encodage TEI. Au retour des campagnes d'annotations ESTER, Solenn Le Pevedic et Denis Maurel approuvent la capacité de l'encodage à manipuler les entités nommées en permettant une normalisation des éléments descriptifs⁶⁷. Au sein de cet article, les auteurs vont émettre de premières recommandations sur le balisage de ces entités en fonction de leur référentiel, et les attributs adéquats, en reprenant le guide Quaero.

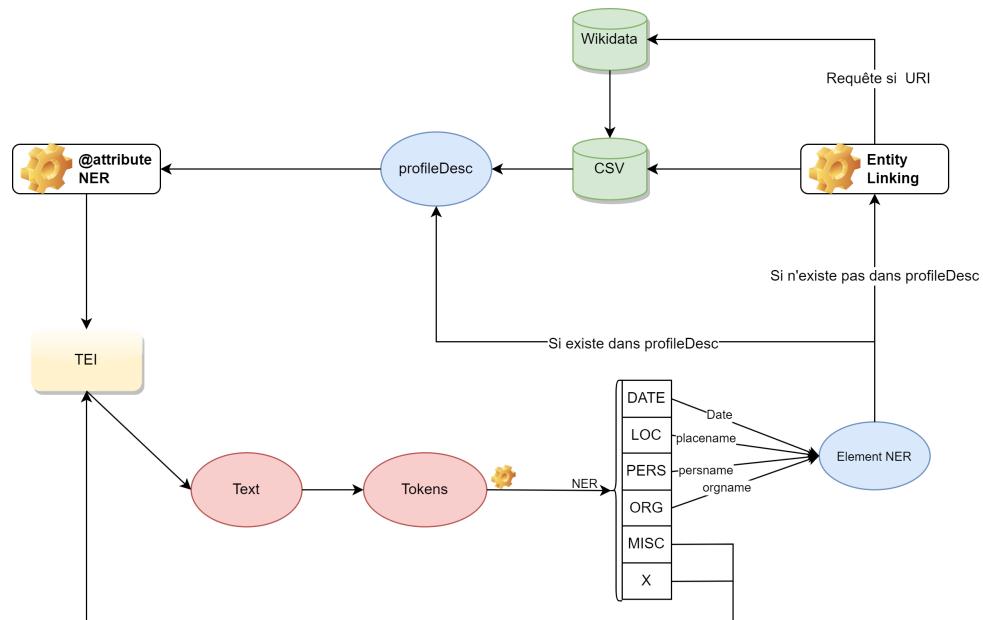


FIGURE 4.8 – Système d'indexation des entités nommées au sein d'un fichier XML-TEI

Pour manipuler nos fichiers XML-TEI déjà existants, nous nous sommes appuyés sur la librairie **StandoffConverter** développée récemment par David Lassner⁶⁸. L'utilisation de son API permet le balisage des entités détectées par le modèle développé⁶⁹. Pour

67. Solenn Le Pevedic et Denis Maurel, « Retour sur les annotations des entités nommées dans les campagnes d'évaluation françaises et comparaison avec la TEI », *Corela. Cognition, représentation, langage*, 2–14–2 (14[2016]), DOI : 10.4000/corela.4644.

68. David Lassner, *Standoffconverter*, standoff-nlp, 2021, URL : <https://github.com/standoff-nlp/standoffconverter> (visité le 10/09/2022).

69. Le script est visible au chemin suivant : TEI_transformation/src/enrichment/nlp.py in M. Humeau, *Tei Transformation...*

ce faire, la classe StandoffConverter va analyser l'ensemble du <**text**>, le tokeniser et repérer son référentiel au sein des catégories définies par le modèle REN. Cette librairie a permis de faciliter ce travail d'extraction et les problèmes sous-jacents en limitant l'impact de la structuration XML en permettant de convertir certains éléments sous la syntaxe REGEX, par exemple : <**lb**> pour \n.

Pour chaque entité, nous avons remis les recommandations de balisage émises par S. Le Pevedic et D. Maurel en les simplifiant afin de permettre l'automatisation de l'indexation⁷⁰. Ainsi, l'entité DATE est balisée sous l'élément <**date**>, LOC est balisée sous l'élément <**placename**>, PERS est balisée sous l'élément <**persname**> et ORG est balisé sous l'élément <**orgname**>. À l'inverse, les tokens MISC ou sans labélisation ne sont pas référencés dans le fichier XML-TEI final.

Les entités balisées sont ensuite analysées par la librairie **spaCy Fishing** afin de déterminer les possibles ambiguïtés et pouvoir les aligner au sein du web sémantique. Cette librairie a été développée par Lucas Terriel, inspiré par les travaux de Patrice Lopez sur la désambiguisation des entités nommées et les fonctions d'*entity linking*⁷¹. Elle est une fonctionnalité rattachable aux fonctions sommaires de SpaCy, faisant la liaison entre la détection d'entités nommées et l'identification au sein d'une base de connaissances, ici *Wikidata*. Si le score de l'évaluation de cette requête est supérieur à 0.8, l'Uniform Resource Identifier (URI) de l'entité (ex : Q4233309 pour Cornelio Saavedra Rodríguez) est alors conservée et enregistrée au sein de la base de données CSV, sinon un identifiant est généré si l'entité n'est pas déjà référencée au sein de la base de données.

Cette chaîne d'extraction d'informations a dans le même temps été appliquée au sein du projet NER4Archives⁷². Néanmoins, dans notre cas les résultats donnés par cette pratique de *fishing* ont démontrés de nombreuses lacunes. Les premiers essais ont montré une réelle difficulté à déterminer et aligner la bonne entité avec un haut niveau de certitude. La gestion des abréviations n'ayant pu être pour l'instant incorporée au sein du processus de transformation, l'alignement avec la base de données *wikidata* ne peut se faire correctement.

4.3.2 Le web de données : SPARQL et Wikidata

L'identifiant des entités nommées permet l'accès d'un point de référence à la fois au sein de la base de données CSV, mais aussi à l'intérieur de notre encodage TEI. Au sein des éléments, cet identifiant est indiqué à l'intérieur d'un attribut **@ref** le reliant

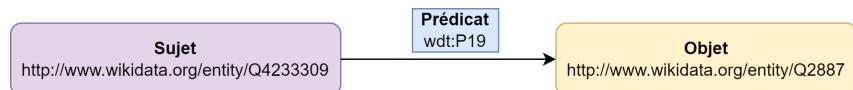
70. S. Le Pevedic et D. Maurel, « Retour sur les annotations des entités nommées dans les campagnes d'évaluation françaises et comparaison avec la TEI »...

71. L. Terriel, *spaCy Fishing*, version 0.1.7, Inria, 24 août 2022, URL : <https://github.com/Lucaterre/spacyfishing> (visité le 10/09/2022).

72. F. Clavaud, L. Romary, P. Charbonnier, *et al.*, « NER4Archives (Named Entity Recognition for Archives)... ».

directement à son élément parent présent au sein des index définis par la TEI. Pour le cas des entités LOC, celles-ci sont enregistrées au sein du `<settingDesc>`, alors que les entités ORG et PERS sont quant à elles recensées au sein du `<particDesc>`. Ce référencement permet de décrire plus en détail les différents éléments indexés.

Comme le rappelle Simon Hengchen et al., l'effervescence du web de données a eu un réel impact auprès des secteurs culturels en donnant des points d'accès à des bases de données permettant d'enrichir considérablement leurs propres collections⁷³. Une méta-étude recense justement que la recherche et l'exploration de données sont les principaux facteurs dans le choix de mettre en place ces bases de données liées⁷⁴. La REN sert de point d'accroche avec les données présentes au sein de web de données (*linking data*). Cette notion est apparue dans les années 1990 sous l'égide de Tim Berner-Lee qui la définit comme un espace d'échange de documents permettant d'accéder à leurs contenus et effectuer des raisonnements⁷⁵. Les contenus sont communément appelés des ressources qui doivent être organisées de manière structurée afin d'être intelligible, souvent à partir de la syntaxe RDF. La relation sémantique entre les ressources est possible sous la forme de « triplet » : Sujet (ressource principale) – Prédicat (relation) – Objet (ressource secondaire). La représentation RDF va s'appuyer sur un ensemble de classes (Sujets) et de propriétés intrinsèques permettant d'être reliée entre elles.



```

<?xml version="1.0"?>
<rdf:RDF xmlns:wdt="http://www.wikidata.org/prop/direct/"
  xmlns:schema="http://schema.org/">
  <rdf:Description rdf:about="http://www.wikidata.org/entity/Q4233309">
    <wdt:P373>Cornelio Saavedra Rodriguez</wdt:P373>
    <wdt:P19 rdf:resource="http://www.wikidata.org/entity/Q2887"/>
  </rdf:Description>
  <rdf:Description rdf:about="http://www.wikidata.org/entity/Q2887">
    <schema:name xml:lang="en">Santiago</schema:name>
    <schema:description xml:lang="en">capital city of
      <Chile></schema:description>
    </rdf:Description>
</rdf:RDF>
  
```

FIGURE 4.9 – Structuration d'un triplet au sein d'un fichier XML-RDF

73. S. Hengchen, S. van Hooland, R. Verborgh, *et al.*, « L'extraction d'entités nommées... ».

74. Edie Davis et Bahareh Heravi, « Linked Data and Cultural Heritage : A Systematic Review of Participation, Collaboration, and Motivation », *Journal on Computing and Cultural Heritage*, 14–2 (10 mai 2021), 21 :1-21 :18, DOI : 10.1145/3429458.

75. Tim Berners-Lee, James Hendler et Ora Lassila, « The Semantic Web », *Scientific American*, 284–5 (2001), p. 34-43, JSTOR : 26059207.

En partant de notre recherche prélimaines à partir de `spaCy Fishing`⁷⁶, nous avons donc pu récupérer l’URI de cette ressource au sein de la Base de données orientée graphe *Wikidata*⁷⁷. Au sein du web de données, cette base de données est sans doute la plus populaire et la plus active puisqu’elle provient directement de l’activité Wikipedia⁷⁸. À l’aide de l’URI, il a été possible d’émettre une requête, différente selon la nature de l’entité, au sein de *Wikidata* grâce au langage SPARQL. “SPARQL Protocol and RDF Query Language” est un langage standard pour interroger les données de graphes représentés par des triplets RDF. Comme nous pouvons le voir au sein de l’application développée⁷⁹, il a été possible de récupérer les informations suivantes :

- Label
- Fate de naissance et de mort
- Identifiant numérique (VIAF ou Geoname)
- Coordonnées géographiques
- Adresse (région, pays)
- Nationalité
- Description

Une fois les informations récupérées (dans leur intégralité ou non), celles-ci sont corrigées et traitées afin de pouvoir être alignées et enregistrées au sein de la base de données CSV. Enfin les données sont alors modélisées parmi l’élément TEI correspondant au sein du `<particDesc>` pour les personnes identifiées et `<settingDesc>` pour les entités géographiques (voir les exemples au sein de l’annexe H).

4.3.3 Limites et solutions aux web de données

Les premiers essais d’enrichissement automatique avec l’utilisation de requêtes SPARQL ont pu démontrer le bon fonctionnement global du processus. Néanmoins, il a été relevé une certaine légèreté des données acquises, notamment pour les catégories descriptives. *Wikidata* reste une base de données généraliste qui n’a pas pour vocation ni à prétendre à l’exhaustivité, ni à la valeur d’une autorité scientifique.

Il est à rappeler que *Wikidata* n’est pas l’unique base de données s’appuyant sur le web sémantique et disponible au grand public. D’autres bases de données graphes s’appuient sur le même modèle à l’image de *DBpedia* développé par l’Université de Leipzig

76. L. Terriel, *spaCy Fishing...*

77. Une base de données orientée graphe (et plus exactement orienté objet) reprend la théorie des graphes en mathématique, en permettant de représenter et stocker et de mettre en relation les données à travers des graphes.

78. En 2012, on recense plus de 15 millions d’entités, dont plus de 34 millions de déclarations, et plus de 80 millions d’étiquettes et de descriptions dans plus de 350 langues. Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez et Denny Vrandečić, « Introducing Wikidata to the Linked Data Web », dans *The Semantic Web – ISWC 2014*, dir. Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz et Carole Goble, Cham, 2014 (Lecture Notes in Computer Science), p. 50-65, DOI : 10.1007/978-3-319-11964-9_4

79. Voir le fichier `TEI_transformation/src/enrichment/sparql.py` . M. Humeau, *Tei Transformation...*

et l’Université libre de Berlin à partir de 2007⁸⁰; *esDBpedia* mis en place en 2011 et maintenu par l’Ontology Engineering Group (OEG), ETSI Informáticos et l’Universidad Politécnica de Madrid⁸¹. Ces bases de données graphes conçues sur le modèle des Linked Open Data s’appuie sur l’extraction des données structurées depuis l’encyclopédie participative Wikipedia.

Une autre solution a été imaginée selon la conversion d’une édition web du *Diccionario Geográfico de la República de Chile* en fichier JSON⁸². Ce dictionnaire géographique publié à la fin du XIX^e siècle recense plus de 5197 lieux géographiques du Chili et à ses frontières. Cette transformation JSON a pour but de retrouver les données de l’époque et ainsi de pouvoir s’aligner les variations orthographiques historiques ou les nominations historiques.

Enfin, la Biblioteca del Congreso Nacional de Chile (Bibliothèque du Congrès national du Chili, en français) qui, à l’image de nombreuses institutions culturelles, a mis en place sa propre base de données reprenant les principes du Linked Open Data. Ce projet *datos.bnc.cl* dont le développement a débuté en 2011, est l’une des premières bases de données à initié ce type de projet en Amérique latine avec pour objectif initial de mettre à disposition du public le travail législatif⁸³. Dans le même temps, un point d’accès API (un *endpoint*) a été mis en place afin d’offrir la possibilité de requêter la base de données graphe avec le langage SPARQL⁸⁴. Deux types de données pourraient permettre d’étayer le projet d’enrichissement : les bases de données biographiques recensant un grand nombre de personnalités de la vie politique et plus secondairement, militaires ; et les données géographiques (à l’heure actuelle, l’ontologie est en cours de mise à jour). Dans un autre registre, il est à considérer que l’État chilien a lui aussi créé une infrastructure numérique nationale recensant et donnant accès à un ensemble de *dumps* (dépôt de données) dans un objectif de transparence au sein de la gouvernance des données. Ils sont issus des données officielles enregistrés par le gouvernement et l’administration nationale du Chili⁸⁵.

Si dans ce dernier cas il fut difficile de trouver des données pertinentes dans le cadre de notre projet, l’essor de la pratique du Linked Open Data au sein de la vie scientifique, mais aussi civique chilienne est symptomatique des nombreux enjeux qui ont entouré la mise en place du projet autour des archives de l’Araucania. Les chercheurs Felipe Gonzalez-Zapata et Richard Heeks ont explicité le rôle du *Open government data* pour

80. *DBpedia*, url : <https://www.dbpedia.org/>, consulté le 09/09/2022.

81. *esDBpedia*, url :<https://es.dbpedia.org/>, consulté le 09/09/2022.

82. F. Solano Asta-Buruaga, *Diccionario Geográfico de la República de Chile...*

83. Francisco Cifuentes-Silva, Christian Sifaqui et Jose Emilio Labra-Gayo, « Towards an Architecture and Adoption Process for Linked Data Technologies in Open Government Contexts : A Case Study for the Library of Congress of Chile », dans *Proceedings of the 7th International Conference on Semantic Systems*, New York, NY, USA, 2011 (I-Semantics '11), p. 79-86, DOI : 10.1145/2063518.2063529.

84. Le site est consultable à l’adresse suivante : <http://datos.bcn.cl/es/>, consulté le 10/09/2022.

85. Le site de dépôt est consultable à cette adresse url : <https://datos.gob.cl/>

continuer de rompre avec un passé trouble en redonnant de la valeur à l'information⁸⁶.

86. Felipe Gonzalez-Zapata et Richard Heeks, « The Multiple Meanings of Open Government Data : Understanding Different Stakeholders and Their Perspectives », *Government Information Quarterly*, 32–4 (1^{er} oct. 2015), p. 441-452, DOI : 10.1016/j.giq.2015.09.001.

Conclusion

Ce mémoire et ce stage ont été l'occasion de mettre en évidence ou de confirmer certaines appréciations sur le développement d'une chaîne de traitement automatique pour la production d'édition numérique native. De la production d'un modèle HTR à l'analyse, l'annotation et l'enrichissement des éditions restent un processus complexe qui doit s'inscrire dans une dynamique globale. Le projet sur les archives autour de l'« Occupation de l'Araucania » est particulièrement révélateur de ces attentes.

Si la scientificité des données repose l'apport quantitatif, mais aussi critique que le chercheur y apporte. Il en reste que l'automatisation du processus de traitement représente des avantages considérables pour les institutions patrimoniales et scientifiques, sans pour autant prétendre à une exhaustivité. Le travail de l'ingénierie se tâche de concevoir et d'ajuster du mieux possible les différents programmes afin de procéder au traitement le plus complet possible, mais qui se confrontent bien souvent à de nombreuses problématiques techniques et intellectuelles.

En premier lieu, la mise en place d'un tel projet nécessite de mettre en œuvre un travail conséquent de préparation de données sur lequel va se centrer l'ensemble du programme. Il faut donc pouvoir saisir l'essence des documents qui constituent notre réservoir et notre objet d'études. L'ensemble de cette *pipeline* se nourrit de ses données afin de pouvoir alimenter, poursuivre et améliorer le rendu de celles-ci. L'apprentissage automatique est ainsi devenu un axe majeur de notre réflexion et de notre pratique. Son introduction doit être considérée non pas comme finalité, et ce malgré la plus-value esthétique que ces techniques représentent, mais comme un moyen avec ces faiblesses, ces biais et ses réussites.

Ces méthodes d'apprentissages ont permises de numériser un grand nombre de nouvelles données sous l'encodage TEI et ce à partir de données océrisées. Les technologies HTR ont été des outils efficaces pour la numérisation native des documents historiques et ainsi servir leur pérennité. Pour cela, la mise en place d'un prototype d'une transformation vers le format pivot a été nécessaire.

Toutefois, il en est ressorti de nombreuses difficultés quant à leurs traitements postérieurs, ce qui induit le besoin de nouvelles réflexions sur les données présentant des défis variations orthographiques historiques. La faible homogénéité des données produites

a amenée à utiliser et mettre en place un modèle de reconnaissance des entités nommées afin de procéder à l'indexation des entités majeures et laisser l'opportunité d'explorer plus en profondeur les documents ou de l'enrichir. De même, l'automatisation de la segmentation reste une étape difficile à automatiser. Cette chaîne de traitement ne peut se passer de l'intervention directe de l'humain afin d'en contrôler la qualité et la cohérence.

En observant le filigrane de ce projet, nous avons pu de même observer l'enjeu que représente l'ouverture des données, de la cohérence de leurs partages et la mise en place de plateformes communes. Ces initiatives permettent une véritable mutualisation des efforts permettant de soutenir des projets de moindre envergure qui peuvent s'appuyer sur des programmes plus vastes. En outre, ces données ouvertes interrogent sur notre rapport à la construction et l'appropriation de la science « en adoptant une posture humble et ouverte devant nos sources, sans appropriation indue et sans dissimuler la possibilité de l'erreur — mieux, en fournissant à la communauté non seulement les résultats de nos recherches, mais aussi la manière dont ils ont été produits et peuvent être reproduits⁸⁷ ». Ces données doivent pour s'émanciper de l'autorité productrice afin de dévoiler de nouvelles perceptions au travers de nouveaux regards éthiques, méthodologiques ou socio-politiques.

Le projet de ce stage a conclu au développement d'un mille-feuille dont chaque étape ne prend sens que collectivement si qui nécessite un travail en amont. Le numérique ne peut s'éloigner ou s'émanciper des méthodes humanistes, devant au contraire s'accommoder afin de faire face aux enjeux du *big data* ou des nouvelles méthodes numériques. Pour Dominique Boullier, les sciences sociales ne peuvent pas être le fruit d'une numérisation de leurs méthodes, de leurs données et de leur objet d'étude, mais en apporter une valeur ajoutée⁸⁸. Cette volonté d'enrichir s'appuie sur cette volonté d'utiliser ces nouvelles méthodes afin d'inciter ces nouvelles perceptions épistémologiques, mais aussi des nouveaux enjeux autour de la valorisation numérique. Víctor Gayol et Jairo Antonio Melo Flórez appellent, justement, les institutions à développer ces projets, notamment autour de l'histoire digital afin de sortir des expériences individuelles allant au détriment d'une véritable cohérence⁸⁹.

Ces humanités numériques restent en plein essor et entourés d'espoirs sur la compréhension d'une histoire complexe où la mémoire est encore douloureuse. Du Chili au Mexique, ce champs des humanités s'organise et se construit progressivement afin de faciliter la circulation des savoir-faire et des données à l'image du réseau RED⁹⁰. La multiplication des efforts conjoints et de mise en parallèle avec des projets vont permettre révéler

87. J.B. Camps, « Où va la philologie numérique ? »...

88. Dominique Boullier, *Sociologie du numérique*, Paris, 2016.

89. Víctor Gayol et Jairo Antonio Melo Flórez, « Presente y perspectivas de las humanidades digitales en América Latina », *Mélanges de la Casa de Velázquez. Nouvelle série*, 2–47-2 (47[2017]), p. 281-284, DOI : 10.4000/mcv.7907.

90. Red de Humanidades Digitales, url : <http://humanidadesdigitales.net/>, consulté le 30/08/2022

et encourager les archives autour de l'« Occupation de l'Araucanie » afin d'encourager l'émergence d'une nouvelle historiographie sur ces évènements⁹¹.

91. Pedro Canales Tapia et Jorge Pinto Rodríguez, « Historiografía Mapuche : balances y perspectivas de discusión en el Chile reciente », *Izquierdas*-24 (24[2015]), URL : <https://journals.openedition.org/izquierdas/358?lang=fr> (visité le 15/09/2022).

Acronymes

ACAB Archivo Central Andres Bello.

CLI Interface en Ligne de Commande.

CMS Content Managing System.

CNN Convolutional Neural Networks.

DAHN Dispositif de soutien à l'Archivistique et aux Humanités Numériques.

DIS Délétion Insertion Substitution.

DTD Document type definition.

EPHE École Pratique des Hautes Études - Université PSL.

GPU Graphics Processing Unit.

HTR Handwritten Text Recognition.

IA Intelligence Artificielle.

INRIA Institut national de recherche en sciences et technologies du numérique.

JPG Joint Photographic Group.

JSON JavaScript Object Notation.

LECTAUREP LECTure Automatique de REPertoire.

OCR Optical Character Recognition.

ODD One Document Does.

PAGE Page Analysis and Ground truth Elements.

REGEX Expression régulière.

REM Reconnaissance d'Écriture Manuscrite.

REN Reconnaissance des entités nommées.

RNG Regular Language for XML Next Generation.

RNN Recurrent Neural Networks.

SegmOnto A Controlled Vocabulary to Describe the Layout of Pages.

TAL Traitement Automatique des Langues.

TIFF Tagged Image File Format.

URI Uniform Resource Identifier.

W3C World Wide Web Consortium.

XML eXtensible Markup Language.

ÉNC École nationale des chartes.

Glossaire

ALTO *Analysed Layout and Text Object* – Standard XML permettant de rendre compte de la mise en page physique et de la structure logique d'un texte transcrit par reconnaissance optique de caractères.

API *Application Programming Interface* (Interface de programmation d'application, en français) – C'est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d'échanger des données et des fonctionnalités. (©CNIL).

CER *Character Error Rate* (en français Taux d'erreur de caractères) métrique évaluant le taux de d'erreurs (DIS) entre la prédiction d'un modèle et la donnée terrain en fonction du nombres de caractères. Le taux est déterminé par la formule suivante :

$$CER = \frac{S + D + I}{N_{total de caracteres}}$$

CSV *Comma-separated value* – format texte ouvert représentant des données tabulaires sous forme de valeurs déterminées par un séparateur, en général une virgule, point-virgule ou une tabulation.

eScriptorium Application web *open source* dédiée à la transcription automatique des documents, utilisant le moteur HTR Kraken.

F1-score (ou F-score) – Au sein de la matrice de confusion, il correspond à la moyenne harmonique de la précision et du rappel dont le score maximum possible est de 1. Le taux est déterminé par la formule suivante :

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Git Logiciel libre et gratuit de gestion de versions décentralisé.

GitHub Plateforme web propriétaire s'appuyant sur le logiciel de gestion de versions Git, tout en permettant l'hébergement de code source pour des logiciels et autres applications..

Kraken Système d'HTR clé en main, optimisé pour les documents historiques et les textes en caractères non latins..

Linked Open Data Les données liées sont une méthode de publication de données structurées, de sorte qu'elles puissent être interconnectées et deviennent plus utiles au moyen de requêtes sémantiques. Il s'appuie sur des technologies web standard telles que HTTP, RDF et URI, mais plutôt que de les utiliser pour desservir des pages web pour les lecteurs humains, elle les étend au partage d'informations de manière à pouvoir être lues automatiquement par des ordinateurs.

Précision (*Accuracy* en anglais), métrique évaluant la performance d'un modèle de *machine learning* selon la matrice de confusion en calculant le pourcentage de prédictions valides. Le taux est déterminé par la formule suivante :

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Rappel (*Recall* en anglais) – Au sein de la matrice de confusion, il permet d'évaluer le nombre de vrai positif sur l'ensemble des éléments évalués positifs par le modèle. Le taux est déterminé par la formule suivante :

$$R = \frac{TP}{TP + FN}$$

RDF Resource Description Framework – Syntaxe pour représenter des données sur le Web de manière générale, et proposer un schéma de description des ressources à partir des modèles graphes..

SpaCy Librairie Python de traitement automatique des langues développée par Matt Honnibal et Ines Montani. SpaCy est un logiciel libre publié sous licence MIT. (©Wikipedia).

TEI *Text Encoding Initiative* – Standard XML qui met l'accent sur le contenu et le sens informationnel des documents. A travers ce format, on parle souvent.

WER *Word Error Rate* (Taux d'erreur de mots en français métrique évaluant le taux le mots possédant une erreur de caractères (DIS) entre la prédiction d'un modèle et la donnée terrain. Le taux est déterminé par la formule suivante :

$$WER = \frac{S_w + D_w + I_w}{N_w}$$

Annexes

Annexe A

Le fonds Araucania

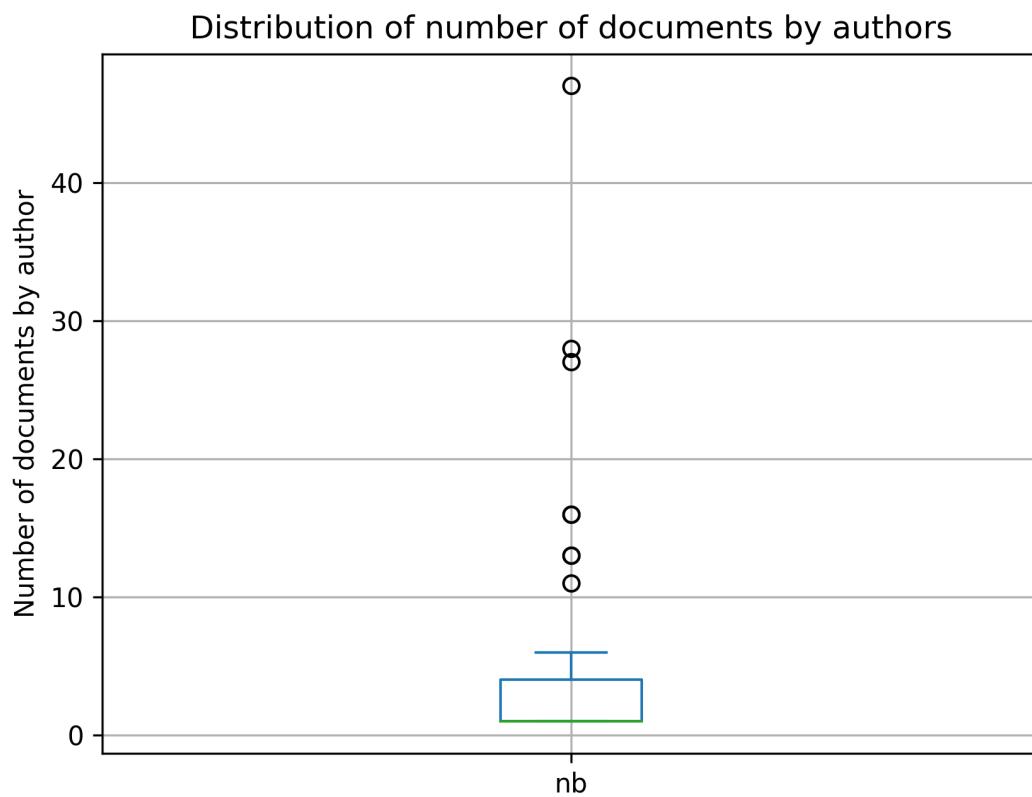


FIGURE A.1 – Distribution du nombres de documents par auteurs

Representation of the type of documents

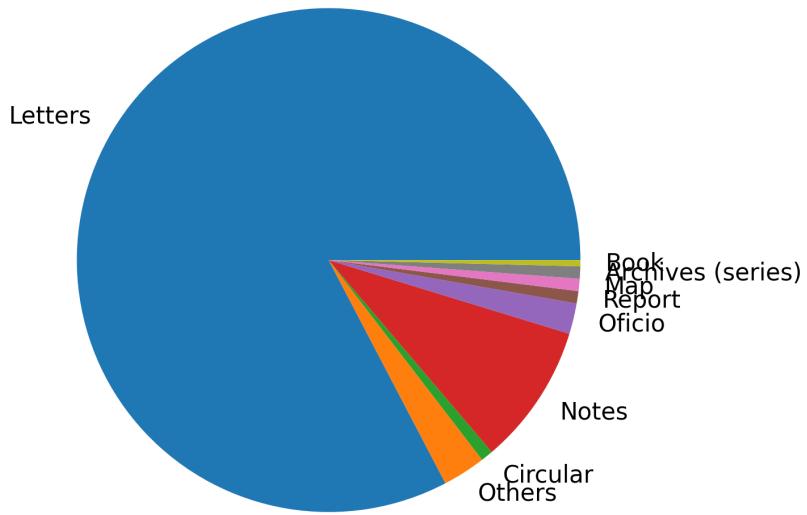


FIGURE A.2 – Représentation des documents selon leur nature

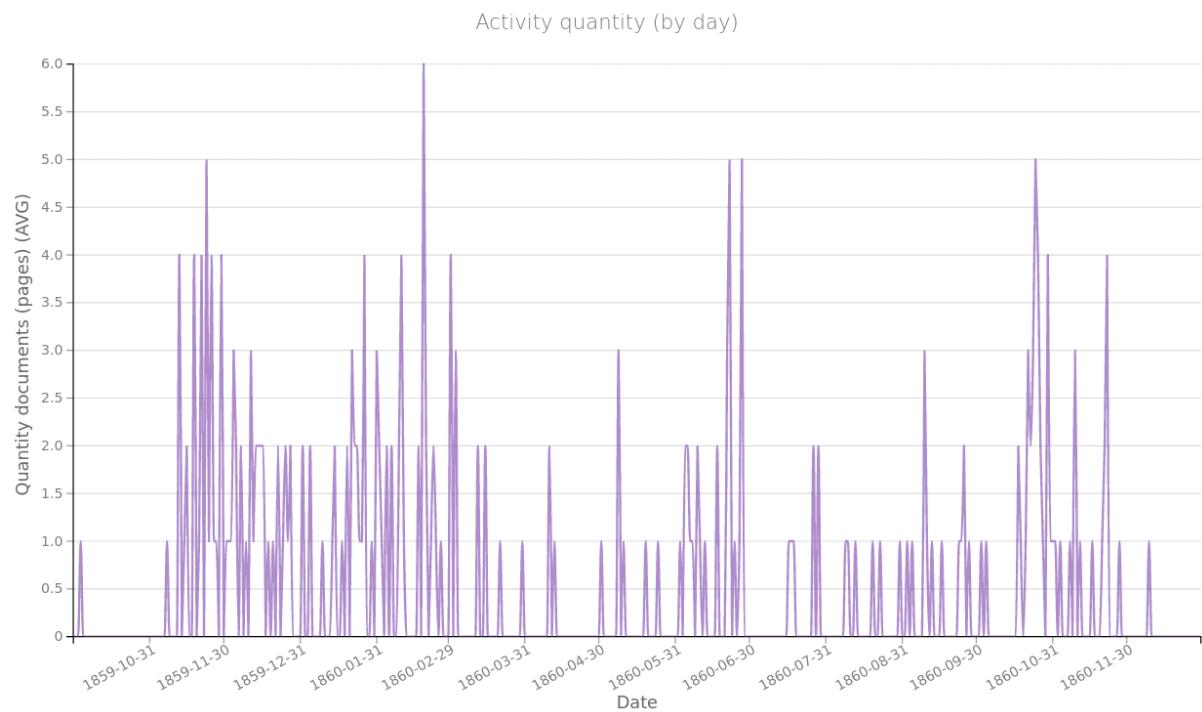


FIGURE A.3 – Activités documentaires sur les années 1859-1860 (par jours)

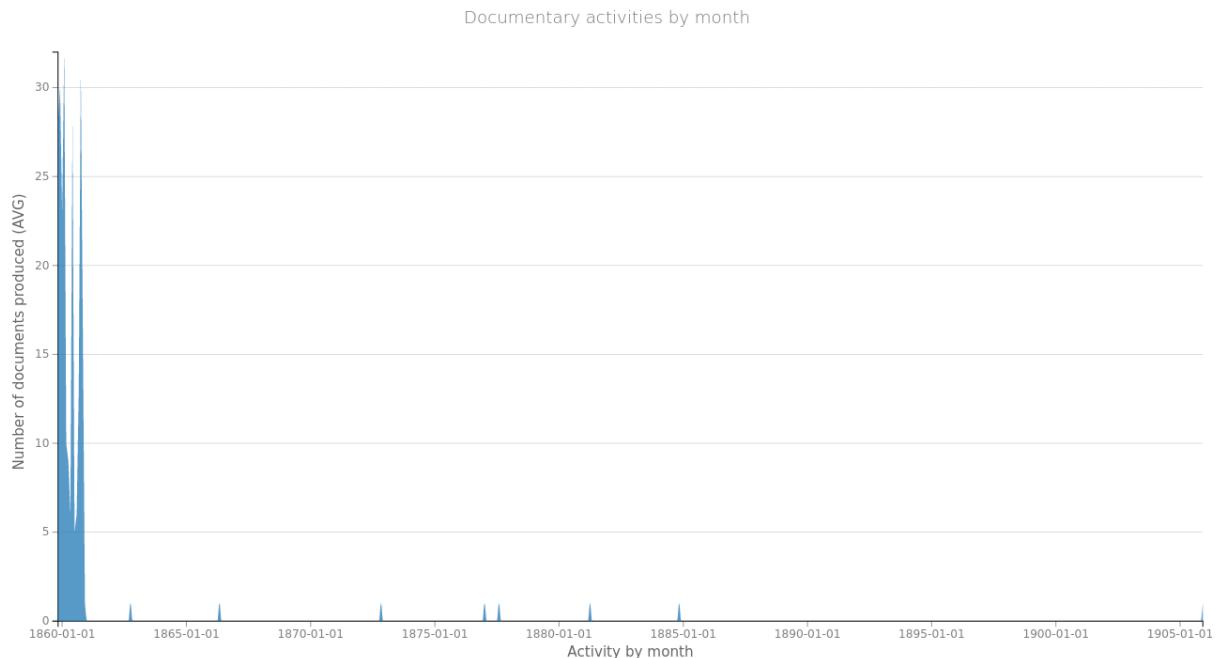


FIGURE A.4 – Activités documentaires (par mois)

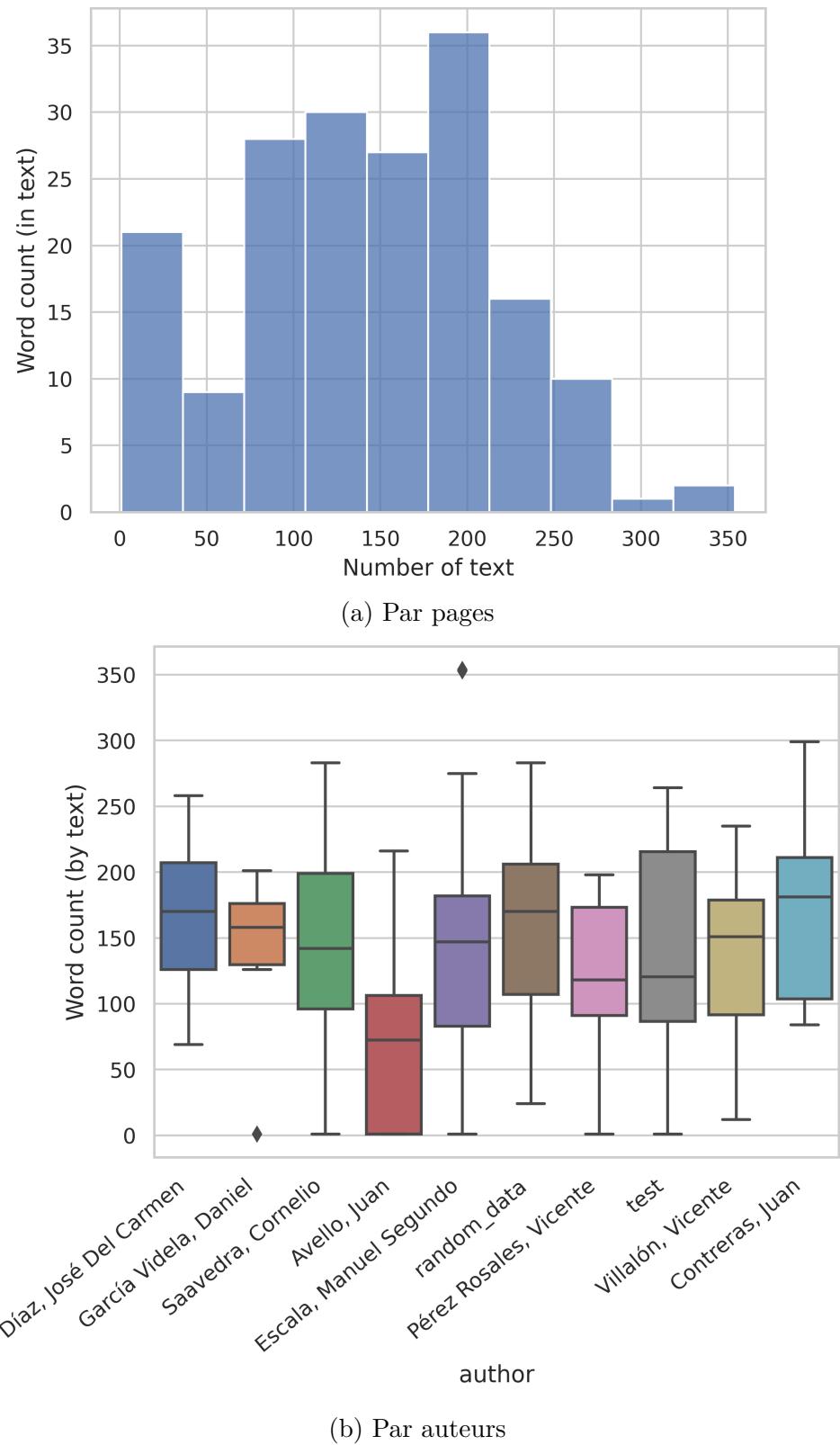


FIGURE A.5 – Distribution des mots

Annexe B

Automatisation d'une conversion de format d'image

```
^^I      #!/bin/bash

FOLDER=~/Bureau/data/jpg
ERROR="$FOLDER/error.txt"

if ! [[ -d "$FOLDER" ]] || ! [[ -e $ERROR ]]
then
  ^Imkdir -p $FOLDER
  ^Itouch $ERROR
else
  rm -f "$FOLDER/*.jpg"
fi

for img in "$PWD"/*.tif; do
  filename=$(basename "${img%.tif}")
  convert "$img" "$FOLDER/$filename.jpg"
done 2>> $ERROR
```

Listing 4 – Script de conversion d’images TIFF vers le format JPG

Annexe C

Exemple de binarisation et de seuillage d'une image

```
In [1]: import numpy as np  
import cv2  
from matplotlib import pyplot as plt
```

ANNEXE C.

```
In [7]: img = 'data/LetterA.jpg'
```

Binarisation

```
In [8]: #Binarisation par niveau de gris  
gray_img = cv2.imread(img, cv2.IMREAD_GRAYSCALE)
```

```
In [9]: # Matrice Image  
gray_img
```

```
Out[9]: array([[255, 255, 255, 254, 254, 254, 255, 255, 249, 250, 250, 255, 255, 249,  
    248, 255, 250],  
   [253, 250, 255, 252, 255, 255, 250, 219, 21, 13, 53, 230, 255,  
    255, 250, 255],  
   [254, 255, 254, 255, 255, 255, 255, 125, 0, 7, 12, 233, 251,  
    255, 243, 255],  
   [253, 255, 243, 255, 245, 253, 224, 20, 18, 53, 10, 214, 241,  
    255, 254, 255],  
   [250, 255, 241, 255, 255, 244, 95, 0, 179, 130, 1, 172, 255,  
    255, 248, 253],  
   [253, 255, 255, 245, 255, 201, 0, 84, 242, 163, 0, 88, 251,  
    255, 255, 255],  
   [255, 255, 255, 250, 240, 97, 8, 198, 249, 221, 37, 63, 247,  
    255, 251, 242],  
   [255, 255, 254, 255, 172, 0, 104, 255, 255, 240, 33, 21, 220,  
    255, 253, 255],  
   [248, 255, 249, 255, 24, 14, 199, 255, 253, 255, 72, 21, 167,  
    238, 255, 255],  
   [255, 254, 255, 143, 3, 145, 255, 250, 252, 248, 136, 0, 134,  
    255, 255, 248],  
   [251, 251, 215, 44, 27, 211, 255, 255, 253, 255, 195, 20, 85,  
    248, 245, 255],  
   [246, 255, 122, 0, 0, 13, 2, 0, 0, 13, 0, 0, 52,  
    253, 251, 255],  
   [255, 199, 0, 0, 6, 0, 7, 0, 0, 1, 2, 2, 16,  
    223, 251, 253],  
   [254, 68, 13, 191, 241, 250, 255, 252, 255, 241, 255, 94, 0,  
    166, 248, 255],  
   [195, 0, 76, 240, 255, 255, 249, 255, 255, 243, 255, 132, 1,  
    134, 255, 251],  
   [42, 0, 171, 251, 255, 255, 247, 255, 254, 255, 245, 202, 2,  
    59, 242, 255],  
   [3, 58, 247, 255, 255, 254, 250, 255, 255, 248, 249, 235, 7,  
    30, 231, 255],  
   [167, 169, 255, 248, 247, 255, 255, 248, 255, 255, 234, 57,  
    68, 216, 255],  
   [255, 246, 255, 255, 247, 255, 255, 247, 248, 255, 245, 255, 241,  
    254, 255, 255],  
   [252, 255, 253, 255, 249, 255, 255, 245, 250, 254, 241, 255, 255,  
    255, 245, 248],  
   [248, 255, 245, 255, 242, 253, 255, 255, 255, 255, 255, 255, 239,  
    251, 245, 255],  
   [255, 255, 244, 255, 248, 255, 255, 246, 237, 248, 255, 255,  
    255, 254, 244],  
   [251, 253, 251, 255, 255, 255, 249, 255, 255, 255, 255, 255, 254,  
    250, 255, 255]], dtype=uint8)
```

```
In [10]: # Nombres d'éléments par dimension  
gray_img.shape
```

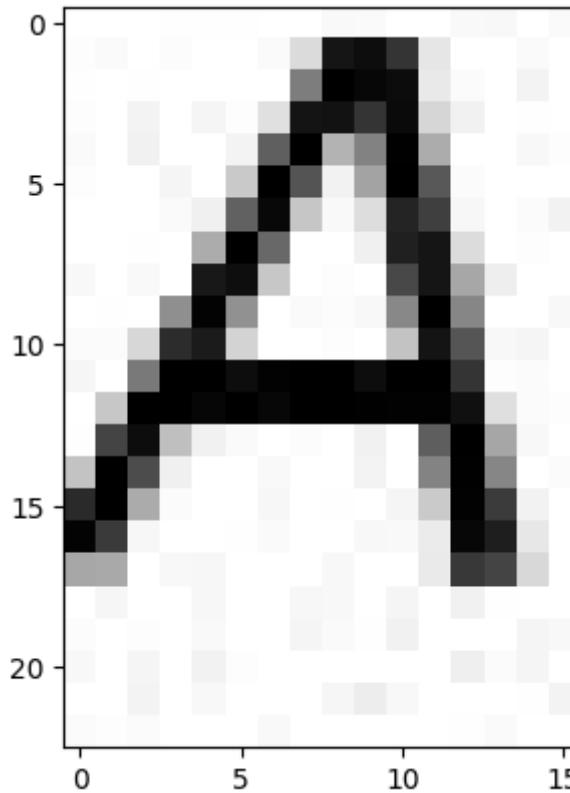
```
Out[10]: (23, 16)
```

105

```
In [11]: # Les quatre premiers éléments des 3 premières dimensions  
print(gray_img[:3,:4])
```

```
[[255 255 255 254]  
 [253 250 255 252]  
 [254 255 254 255]]
```

```
In [12]: plt.imshow(gray_img, cmap="gray")  
plt.show()  
print(np.array_str(gray_img, precision=2, suppress_small=True))
```



```
[[255 255 255 254 254 254 255 255 249 250 255 255 249 248 255 250]  
 [253 250 255 252 255 255 250 219 21 13 53 230 255 255 250 255]  
 [254 255 254 255 255 255 255 125 0 7 12 233 251 255 243 255]  
 [253 255 243 255 245 253 224 20 18 53 10 214 241 255 254 255]  
 [250 255 241 255 255 244 95 0 179 130 1 172 255 255 248 253]  
 [253 255 255 245 255 201 0 84 242 163 0 88 251 255 255 255]  
 [255 255 255 250 240 97 8 198 249 221 37 63 247 255 251 242]  
 [255 255 254 255 172 0 104 255 255 240 33 21 220 255 253 255]  
 [248 255 249 255 24 14 199 255 253 255 72 21 167 238 255 255]  
 [255 254 255 143 3 145 255 250 252 248 136 0 134 255 255 248]  
 [251 251 215 44 27 211 255 255 253 255 195 20 85 248 245 255]  
 [246 255 122 0 0 13 2 0 0 13 0 0 52 253 251 255]  
 [255 199 0 0 6 0 7 0 0 1 2 2 16 223 251 253]  
 [254 68 13 191 241 250 255 252 255 241 255 94 0 166 248 255]  
 [195 0 76 240 255 255 249 255 255 243 255 132 1 134 255 251]  
 [42 0 171 251 255 255 247 255 254 255 245 202 2 59 242 255]  
 [3 58 247 255 255 254 250 255 255 248 249 235 7 30 231 255]  
 [167 169 255 248 247 255 255 255 248 255 255 234 57 68 216 255]  
 [255 246 255 255 247 255 255 247 248 255 245 255 241 254 255 255]  
 [252 255 253 255 249 255 255 245 250 254 241 255 255 245 248]  
 [248 255 245 255 242 253 255 255 255 255 255 239 251 245 255]  
 [255 255 244 255 248 255 255 255 246 237 248 255 255 254 244]  
 [251 253 251 255 255 255 249 255 255 255 255 254 250 255 255]]
```

Le seuillage

106en original

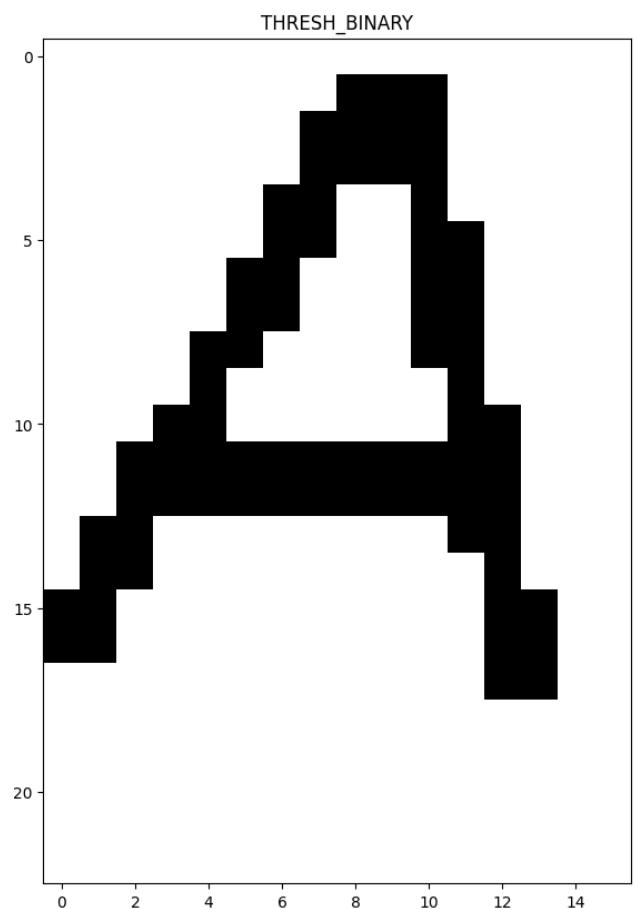
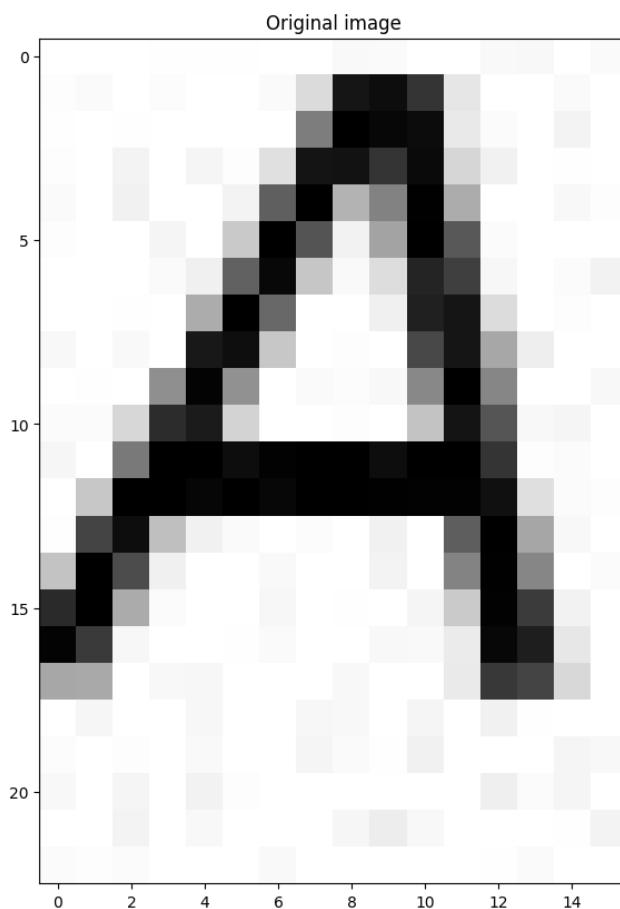
ANNEXE C.

```
In [13]: # Binarisation  
img = cv2.imread('data/LetterA.jpg', 0)
```

```
In [14]: ## Function pour afficher plusieurs images  
def plot_img(images, titles):  
    fig, axs = plt.subplots(nrows = 1, ncols = len(images), figsize = (15, 15))  
    for i, p in enumerate(images):  
        axs[i].imshow(p, 'gray')  
        axs[i].set_title(titles[i])  
        #axs[i].axis('off')  
    plt.show()
```

```
In [15]: #On supprime les valeurs en dessous du seuil 127 (niveau de gris)  
#cv2.threshold(img, thresh_value, maxVal, style)  
ret, img_binary = cv2.threshold(img, 127, 255, cv2.THRESH_BINARY)
```

```
In [16]: # Afficher image  
images = [img, img_binary]  
titles = ['Original image', 'THRESH_BINARY']  
plot_img(images, titles)
```



Annexe D

Entraînement d'un modèle HTR avec Kraken

```
^I      ketos train --augment --workers 8 -d cuda:0 -f binary
↪ --min-epochs 20 -w 0 -s '[1,120,0,1Cr3,13,32 Do0.1,2 Mp2,2 Cr3,13,32
↪ Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 Mp2,2 Cr3,9,64 Do0.1,2 S1(1x0)1,3
↪ Lbx200 Do0.1,2 Lbx200 Do.1,2 Lbx200 Do]' --optimizer Adam -B 20 -r
↪ 0.0001 dataset.arrow > train_skratch.txt
```

Listing 5 – Commande shell d'entraînement selon la méthode *skratch*

```
^I      ketos train -f binary --augment -B 20 -d cuda:0 -o
↪ araucania_finetuning_McFrench --resize both --load
↪ HTR-United-Manu_McFrench.mlmodel dataset.arrow --lrate 0.0001
↪ --workers 8 > train_finetuning.txt
```

Listing 6 – Commande shell d'entraînement selon la méthode *finetuning*

```
^I      ketos segtrain --augment -d cuda:0 -s '[1,1800,0,3
↪ Cr7,7,64,2,2 Gn32 Cr3,3,128,2,2 Gn32 Cr3,3,128 Gn32 Cr3,3,256 Gn32
↪ Cr3,3,256 Gn32 Lbx32 Lby32 Cr1,1,32 Gn32 Lby32 Lbx32]' -f alto -bl
↪ data/**/*.xml -r 0.0001 --resize both --optimizer Adam -i
↪ models/blla.mlmodel --merge-baselines DefaultLine:default
↪ --merge-regions MainZone:text > segtrain.txt
```

Listing 7 – Commande shell d'entraînement selon la méthode *finetuning* d'un modèle de segmentation

Annexe E

Exemple de fichier XML-TEI : AH0299

```
<?xml version='1.0' encoding='UTF-8'?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Carta del 10 de junio de 1860 de Cornelio Saavedra a
          ↳ Mauricio Barbosa</title>
        <author>Saavedra, Cornelio</author>
        <editor>
          <orgName>Archivo Central Andres Bello</orgName>
        </editor>
      </titleStmt>
      <editionStmt>
        <respStmt>
          <resp xml:id="DirSc">Scientific Director of the edition</resp>
          <persName>
            <forename>Alessandro</forename>
            <surname>Chiaretti</surname>
            <roleName>Professor and archivist, responsable Área de
              ↳ Información Bibliográfica y Archivística</roleName>
            <affiliation>Archivo Central Andres Bello | Universidad de
              ↳ Chile</affiliation>
          </persName>
        </respStmt>
        <respStmt>
          <resp xml:id="Enc">In charge of digital encoding</resp>
```

```

<persName>
    <forename>Maxime</forename>
    <surname>Humeau</surname>
    <roleName>Student, trainee</roleName>
    <affiliation>Ecole nationale des chartes | PSL</affiliation>
</persName>
</respStmt>
</editionStmt>
<extent>
    <measure unit="images" n="2"/>
</extent>
<publicationStmt>
    <publisher xml:id="ACAB">Archivo Central Andres Bello</publisher>
    <authority>Área de Información Bibliográfica y
        ↳ Archivística</authority>
    <address>
        <country key="CL"/>
        <region>Región Metropolitana</region>
        <settlement type="city">Santiago Centro</settlement>
        <postCode>8320000</postCode>
        <street>Arturo Prat</street>
        <street>#23</street>
    </address>
    <availability status="restricted">
        <licence
            ↳ target="https://creativecommons.org/licenses/by-sa/3.0/deed.fr">Attribu
            ↳ 4.0 International (CC BY-NC-SA 4.0)</licence>
        <p>Share - copy and redistribute the material in any medium or
            ↳ format</p>
        <p>Adapt - remix, transform, and build upon the material</p>
        <p>Attribution - You must give appropriate credit, provide a
            ↳ link to the license, and indicate if changes were made. You
            ↳ may do so in any reasonable manner, but not in any way that
            ↳ suggests the licensor endorses you or your use.</p>
        <p>NonCommercial - You may not use the material for commercial
            ↳ purposes.</p>
        <p>ShareAlike - If you remix, transform, or build upon the
            ↳ material, you must distribute your contributions under the
            ↳ same license as the original.</p>
    </availability>
</publicationStmt>

```

```

<p>The license is restricted to the use of XML-TEI files. The
↳ exploitation, distribution or publication of the attached
↳ images is subject to the approval of the institution. The
↳ full rights of the archive are reserved. The request can be
↳ made to the following address:
↳ &lt;email&gt;archivo.central@uchile.cl&lt;/email&gt;</p>
</availability>
<date when-iso="2022-08-15"/>
</publicationStmt>
<notesStmt>
  <note>Digital editing done as part of an international
    ↳ internship.</note>
  <note>A first transcription of part of the collection was made in
    ↳ &lt;date when-iso="2014-06"&gt;2014&lt;/date&gt; by
    ↳ &lt;persName&gt;Cecilia del Carmen Ramallo
    ↳ Díaz&lt;/persName&gt;.</note>
  <note>HTR scanning done at Universidad de Chile with kraken
    ↳ engine and the application eScriptorium. The HTR models from
    ↳ the transcript are available at this address &lt;ref
    ↳ target="https://github.com/Proyecto-Ocupacion-Araucania-UChile/model-HTR"></ref>
</notesStmt>
<sourceDesc>
  <bibl>
    <series xml:lang="es">
      <title series="s" type="principal">Colección
        ↳ Manuscritos</title>
      <title type="subtitle">Pacificacion de la Araucania</title>
      <respStmt>
        <resp>Classification and conservation by</resp>
        <persName ref="#DirSc">Alessandro Chiaretti</persName>
        <persName xml:id="M_Parra">Marcos Parra</persName>
        <orgName ref="#ACAB">Área de Información Bibliográfica y
          ↳ Archivística</orgName>
      </respStmt>
      <idno type="caja" n="3"/>
      <idno type="id" n="AH0299"/>
      <note>
        <unit type="documents" quantity="255"/>
      </note>
    </series>
  </bibl>
</sourceDesc>

```

```

    </series>
    </bibl>
  </sourceDesc>
</fileDesc>
<profileDesc>
  <particDesc>
    <listOrg>
      <head>List of organizations</head>
      <org xml:id="ORG_16832"><orgname>Gob^otu</orgname></org>
      <org xml:id="ORG_84978"><orgname>tropa</orgname></org>
      <org xml:id="ORG_89363"><orgname>cuerpo</orgname></org>
      <org xml:id="ORG_56624"><orgname>Co-misaría</orgname></org>
      <org xml:id="ORG_84978"><orgname>tropa</orgname></org>
      <org xml:id="ORG_74844"><orgname>Gobierno de
        ↳ Chile</orgname></org>
      <org xml:id="ORG_96712"><orgname>división</orgname></org>
    </listOrg>
    <listPerson>
      <head>List of persons</head>
      <person xml:id="PERS_41113" xml:base="N.C." xml:lang="es"
        ↳ sex="0"><persname>Mauricio Barbosa</persname></person>
      <person xml:id="PERS_86612" xml:base="N.C." xml:lang="es"
        ↳ sex="0"><persname>General</persname></person>
      <person xml:id="PERS_70914" xml:base="N.C." xml:lang="es"
        ↳ sex="0"><persname>doctor</persname></person>
      <person xml:id="Q1327"
        ↳ xml:base="https://viaf.org/viaf/26785781" xml:lang="en"
        ↳ sex="1.0"><persname>José Joaquín Pérez</persname><birth
        ↳ when-iso="1801-05-06">6 de mayo de 1801</birth><death
        ↳ when-iso="1889-07-01">1 de julio de 1889</death><note
        ↳ type="description">Chilean politician and President
        ↳ (1801-1889)</note></person>
      <person xml:id="PERS_46502" xml:base="N.C." xml:lang="es"
        ↳ sex="0"><persname>Ocha-gavía</persname></person>
    
```

```

<person xml:id="Q5945826"
    ↳ xml:base="https://viaf.org/viaf/50032889" xml:lang="en"
    ↳ sex="1.0"><persname>José Tomás Urmenate</persname><birth
    ↳ when-iso="1808-10-08">8 de octubre de 1808</birth><death
    ↳ when-iso="1878-10-20">20 de octubre de 1878</death><note
    ↳ type="description">Chilean politician</note></person>
<person xml:id="PERS_47745" xml:base="N.C." xml:lang="es"
    ↳ sex="0"><persname>Boonen</persname></person><person
    ↳ xml:id="Q4233309"
    ↳ xml:base="https://viaf.org/viaf/5,34145858091523E+020"
    ↳ xml:lang="en" sex="1.0"><persname>Cornelio Saavedra
    ↳ Rodríguez</persname><birth when-iso="1821-01-01">1 de
    ↳ enero de 1821</birth><death when-iso="1891-04-07">7 de
    ↳ abril de 1891</death><note type="description">Chilean
    ↳ general</note></person>
</listPerson>
</particDesc>
<settingDesc>
    <listPlace>
        <head>List of places</head>
        <place xml:id="Q33986"
            ↳ xml:base="https://www.geonames.org/3868626" xml:lang="en"
            ↳ type="city_in_Chile"><placename>Valparaíso</placename><region>Valparaíso</region>
            ↳ -33.046111111</geo><note type="description">city in
            ↳ Chile</note></place>
        <place xml:id="LOC_88528" xml:base="N.C." xml:lang="es"
            ↳ type="None"><placename>Tucapel</placename></place><place
            ↳ xml:id="LOC_18849" xml:base="N.C." xml:lang="es"
            ↳ type="None"><placename>Los Angeles</placename></place>
        <place xml:id="Q3632" xml:base="https://www.geonames.org/3899462"
            ↳ xml:lang="en"
            ↳ type="city_in_Chile"><placename>Arauco</placename><region>Arauco</region>
            ↳ -37.2463</geo><note type="description">Chilean
            ↳ commune</note></place>
        </listPlace>
    </settingDesc>
</profileDesc>
<encodingDesc>
    <editorialDecl>

```

```

<p>Encoding with XML-TEI P5</p>
<correction>
    <p>There are no corrections for spelling or grammatical errors.
        ↵ The transcription is as original as possible. A
        ↵ post-process HTR correction was performed via spellchecker
        ↵ and Levenshtein's Distance algorithm.</p>
</correction>
<punctuation>
    <p>The punctuation has been transcribed as found.</p>
</punctuation>
<segmentation target="https://github.com/segmonto">
    <p>The segmentation is done via the kraken segmentation model
        ↵ and restructured from the XML-ALTO files and Segmonto
        ↵ ontology.</p>
</segmentation>
<normalization>
    <p>Words that are crossed out, illegible or only interpretable
        ↵ have not been transcribed.</p>
</normalization>
</editorialDecl>
<appInfo>
    <application version="4.1.1" ident="kraken">
        <label>Kraken HTR</label>
        <ptr target="https://github.com/mittagessen/kraken"/>
    </application>
    <application version="1.0.0" ident="escriptorium">
        <label>eScriptorium</label>
        <ptr target="https://gitlab.com/scripta/escriptorium"/>
    </application>
    <application version="0.6.3" ident="pyspellchecker">
        <label>PYspellchecker</label>
        <ptr target="https://github.com/barrust/pyspellchecker"/>
    </application>
    <application version="3.4" ident="spacy">
        <label>spaCy</label>
        <ptr target="https://github.com/explosion/spaCy"/>
    </application>
</appInfo>
</encodingDesc>

```

```

</teiHeader>
<sourceDoc>
[...]
</sourceDoc>
<text>
<body>
  <div type="Letters">
    <pb corresp="#299_a"/>
    <opener>
      <dateline corresp="#299_a_z1"><placename resp="spacy"
        ↵ ref="#Q33986"><lb
        ↵ corresp="#299_a_z1_11"/>Valparaiso</placename>, <date
        ↵ resp="spacy">Julio 101860</date></dateline>
      <dateline corresp="#299_a_z1"><placename resp="spacy"
        ↵ ref="#LOC_88528"><lb
        ↵ corresp="#299_a_z1_12"/>Tucapel</placename></dateline>
      <name corresp="#299_a_z2" type="addressee"><lb
        ↵ corresp="#299_a_z2_11"/>Sr D. <persname resp="spacy"
        ↵ ref="#PERS_41113">Mauricio Barboza</persname></name>
      <salute corresp="#299_a_z4"><lb corresp="#299_a_z4_11"/>Mi
        ↵ estimado amigo:</salute>
    </opener>
    <p corresp="#299_a_z3" n="1"><lb corresp="#299_a_z3_11"/>Tus
      ↵ cartas del <date resp="spacy">17 de Junio</date> son las
      ↵ únicas que<lb corresp="#299_a_z3_12"/>he recibido desde que
      ↵ fuiste a <placename resp="spacy" ref="#LOC_18849">Los
      ↵ Angeles</placename> y con tanto atraso han llega-<lb
      ↵ corresp="#299_a_z3_13"/>do a mi poder que hacen solo <date
      ↵ resp="spacy">dos días</date> las he recibido y en el mo-<lb
      ↵ corresp="#299_a_z3_14"/>mento he tomado medidas para alistar
      ↵ un buque y avisar al <orgname resp="spacy"
      ↵ ref="#ORG_16832">Gob^o<lb
      ↵ corresp="#299_a_z3_15"/>tu</orgname> situacion. <date
      ↵ resp="spacy">Hoi</date> se ordena ya la salida del "Maule" y
      ↵ embarque de<lb corresp="#299_a_z3_16"/>viveres y el <date
      ↵ resp="spacy">viernes 13</date> estará de viaje este vapor
      ↵ para proporcionarte<lb corresp="#299_a_z3_17"/>los auxilios
      ↵ necesarios.</p>

```

<p corresp="#299_a_z3" n="2"><lb corresp="#299_a_z3_18"/>Los
 ↵ techos están listos para mandartelos cuando tu creas<lb
 ↵ corresp="#299_a_z3_19"/>convenientes emprender el trabajo y
 ↵ tengas los elementos necesa-<lb corresp="#299_a_z3_110"/>rios
 ↵ para conducir el fierro. De todos modos te los había<lb
 ↵ corresp="#299_a_z3_111"/>mandado por el Maule, pero este
 ↵ vapor es tan pequeño que<lb corresp="#299_a_z3_112"/>apenas
 ↵ puede llevarte los víveres.</p>

<p corresp="#299_a_z3" n="3"><lb corresp="#299_a_z3_113"/>En
 ↵ cuanto a la remesa de artículos, te mando lo que<lb
 ↵ corresp="#299_a_z3_114"/>creo puedes necesitar tanto para la
 ↵ <orgname resp="spacy" ref="#ORG_84978">tropa</orgname> como
 ↵ para los ofi-<lb corresp="#299_a_z3_115"/>ciales y lo que no
 ↵ necesites, bien puedes realizarlo en esa con<lb
 ↵ corresp="#299_a_z3_116"/>ventaja y evitar el cargo que irá
 ↵ contra tu <orgname resp="spacy"
 ↵ ref="#ORG_89363">cuerpo</orgname>. La <orgname resp="spacy"
 ↵ ref="#ORG_56624">Co-<lb
 ↵ corresp="#299_a_z3_117"/>misaría</orgname> ha sido la
 ↵ encargada para esta compra.</p>

<p corresp="#299_a_z3" n="4"><lb corresp="#299_a_z3_118"/>Debes
 ↵ pues estar prevenido de la llegada del "Maule" y si<lb
 ↵ corresp="#299_a_z3_119"/>encuentras mas prudente retirarte
 ↵ sobre <placename resp="spacy"
 ↵ ref="#Q3632">Arauco</placename>, debes hacerlo<lb
 ↵ corresp="#299_a_z3_120"/>a pesar que sería nuevo sacrificio
 ↵ llevar tu <orgname resp="spacy"
 ↵ ref="#ORG_84978">tropa</orgname> a esa locali-<lb
 ↵ corresp="#299_a_z3_121"/>dad pasado el invierno, el que
 ↵ estando ya mui adelantado<lb corresp="#299_a_z3_122"/>hace
 ↵ desaparezca luego tu penosa situacion; sinembargo tu<lb
 ↵ corresp="#299_a_z3_123"/>veras lo mas conveniente.</p>

<p corresp="#299_a_z3" n="5"><lb corresp="#299_a_z3_124"/>El
 ↵ <persname resp="spacy" ref="#PERS_86612">General</persname>
 ↵ me dice que te dirijas oficialmente al <orgname resp="spacy"
 ↵ ref="#ORG_74844">Gob^o</orgname></p>

<pb corresp="#299_b"/>

<p corresp="#299_b_z1" n="6"><lb corresp="#299_b_z1_11"/>pidiendo
 ↵ las medicinas y médico que necesitas y ya tengo<lb
 ↵ corresp="#299_b_z1_12"/>encargo de buscarte un <persname
 ↵ resp="spacy" ref="#PERS_70914">doctor</persname> para que
 ↵ esté con tu <orgname resp="spacy"
 ↵ ref="#ORG_96712">división</orgname>. </p>

<p corresp="#299_b_z1" n="7"><lb corresp="#299_b_z1_13"/>En
 ↵ cuanto al abono de real diario no será posible y te<lb
 ↵ corresp="#299_b_z1_14"/>lo aviso para que procures la
 ↵ economía en el rancho de tropa.</p>

<p corresp="#299_b_z1" n="8"><lb corresp="#299_b_z1_15"/>Por acá
 ↵ no ocurre novedad ninguna que pueda co-<lb
 ↵ corresp="#299_b_z1_16"/>municarte, todo sigue tranquilo y con
 ↵ mas que segurida-<lb corresp="#299_b_z1_17"/>des que
 ↵ continuaremos del mismo modo.</p>

<p corresp="#299_b_z1" n="9"><lb corresp="#299_b_z1_18"/>En
 ↵ materia de política hai mucho silencio y en cuanto<lb
 ↵ corresp="#299_b_z1_19"/>a candidatos se habla de Don
 ↵ <persname resp="spacy" ref="#Q1327">José Joaquín
 ↵ Pérez</persname>, <persname resp="spacy"
 ↵ ref="#PERS_46502">Ocha-<lb
 ↵ corresp="#299_b_z1_110"/>gavía</persname> y Don <persname
 ↵ resp="spacy" ref="#Q5945826">José Tomás Urmeneta</persname>,
 ↵ mas probablemente<lb corresp="#299_b_z1_111"/>se fijará la
 ↵ atención sobre los dos primeros: todavia esto<lb
 ↵ corresp="#299_b_z1_112"/>es un problema que se decidirá en
 ↵ pocos meses mas.</p>

<p corresp="#299_b_z1" n="10"><persname resp="spacy"
 ↵ ref="#PERS_47745"><lb
 ↵ corresp="#299_b_z1_113"/>Boonen</persname> recibió el recibo
 ↵ que me mandaste y me dice<lb corresp="#299_b_z1_114"/>que
 ↵ tiene en su poder \$ 300 poco mas ó menos de los que<lb
 ↵ corresp="#299_b_z1_115"/>te dará cuenta o pondrá a tu
 ↵ disposicion.</p>

<fw corresp="#299_b_z2" type="n_page"><lb
 ↵ corresp="#299_b_z2_11"/>2</fw>

<closer>

```
<signed><persname resp="spacy" ref="#Q4233309"><lb
    ↵  corresp="#299_b_z4_11"/>Cornelio
    ↵  Saavedra</persname></signed>
<salute corresp="#299_b_z5" n="11"><lb
    ↵  corresp="#299_b_z5_11"/>Como siempre me repito tu amigo y
    ↵  S.S.</salute>
</closer>
</div>
</body>
<noteGrp>
    <note corresp="#299_b_z3_11" type="id_imp">000299</note>
</noteGrp>
</text>
</TEI>
```

Annexe F

Résultats des évaluation NER par *cross-validation*

```
[  
  {  
    "token_acc":1.0,  
    "token_p":1.0,  
    "token_r":1.0,  
    "token_f":1.0,  
    "ents_p":0.7134831461,  
    "ents_r":0.7298850575,  
    "ents_f":0.7215909091,  
    "ents_per_type":{  
      "MISC":{  
        "p":0.7647058824,  
        "r":0.4333333333,  
        "f":0.5531914894  
      },  
      "LOC":{  
        "p":0.77777777778,  
        "r":0.875,  
        "f":0.8235294118  
      },  
      "DATE":{  
        "p":0.52,  
        "r":0.7647058824,  
        "f":0.619047619  
      },  
      "PERS":{
```

```

    "p":0.75,
    "r":0.7611940299,
    "f":0.7555555556
},
"ORG":{
    "p":0.6875,
    "r":0.7857142857,
    "f":0.7333333333
}
},
"speed":2742.5599353071
},
{
    "token_acc":1.0,
    "token_p":1.0,
    "token_r":1.0,
    "token_f":1.0,
    "ents_p":0.7556818182,
    "ents_r":0.7643678161,
    "ents_f":0.76,
    "ents_per_type":{
        "MISC":{
            "p":0.5517241379,
            "r":0.5333333333,
            "f":0.5423728814
        },
        "LOC":{
            "p":0.9,
            "r":0.84375,
            "f":0.8709677419
        },
        "DATE":{
            "p":0.8125,
            "r":0.7647058824,
            "f":0.7878787879
        },
        "PERS":{
            "p":0.7746478873,
            "r":0.8208955224,

```

```
    "f":0.7971014493
},
"ORG":{
    "p":0.7333333333,
    "r":0.7857142857,
    "f":0.7586206897
},
},
"speed":2522.5018865445
},
{
    "token_acc":1.0,
    "token_p":1.0,
    "token_r":1.0,
    "token_f":1.0,
    "ents_p":0.7,
    "ents_r":0.724137931,
    "ents_f":0.7118644068,
    "ents_per_type":{
        "MISC":{
            "p":0.6,
            "r":0.6,
            "f":0.6
        },
        "LOC":{
            "p":0.7878787879,
            "r":0.8125,
            "f":0.8
        },
        "DATE":{
            "p":0.6875,
            "r":0.6470588235,
            "f":0.6666666667
        },
        "PERS":{
            "p":0.6944444444,
            "r":0.7462686567,
            "f":0.7194244604
        },
    }
}
```

```

    "ORG": {
        "p": 0.724137931,
        "r": 0.75,
        "f": 0.7368421053
    },
    "speed": 2520.9481323492
},
{
    "token_acc": 1.0,
    "token_p": 1.0,
    "token_r": 1.0,
    "token_f": 1.0,
    "ents_p": 0.7167630058,
    "ents_r": 0.7126436782,
    "ents_f": 0.7146974063,
    "ents_per_type": {
        "MISC": {
            "p": 0.7083333333,
            "r": 0.5666666667,
            "f": 0.6296296296
        },
        "LOC": {
            "p": 0.8333333333,
            "r": 0.78125,
            "f": 0.8064516129
        },
        "DATE": {
            "p": 0.5416666667,
            "r": 0.7647058824,
            "f": 0.6341463415
        },
        "PERS": {
            "p": 0.7868852459,
            "r": 0.7164179104,
            "f": 0.75
        },
        "ORG": {
            "p": 0.6176470588,

```

```
    "r":0.75,
    "f":0.6774193548
  },
},
"speed":2756.2619204097
},
{
  "token_acc":1.0,
  "token_p":1.0,
  "token_r":1.0,
  "token_f":1.0,
  "ents_p":0.7045454545,
  "ents_r":0.7126436782,
  "ents_f":0.7085714286,
  "ents_per_type":{
    "MISC":{
      "p":0.6363636364,
      "r":0.4666666667,
      "f":0.5384615385
    },
    "LOC":{
      "p":0.7352941176,
      "r":0.78125,
      "f":0.7575757576
    },
    "DATE":{
      "p":0.7647058824,
      "r":0.7647058824,
      "f":0.7647058824
    },
    "PERS":{
      "p":0.7571428571,
      "r":0.7910447761,
      "f":0.7737226277
    },
    "ORG":{
      "p":0.5757575758,
      "r":0.6785714286,
      "f":0.6229508197
    }
  }
}
```

```

        },
    },
    "speed":2619.3018869685
},
{
    "token_acc":1.0,
    "token_p":1.0,
    "token_r":1.0,
    "token_f":1.0,
    "ents_p":0.7607361963,
    "ents_r":0.7126436782,
    "ents_f":0.7359050445,
    "ents_per_type":{
        "MISC":{
            "p":0.8823529412,
            "r":0.5,
            "f":0.6382978723
        },
        "LOC":{
            "p":0.8064516129,
            "r":0.78125,
            "f":0.7936507937
        },
        "DATE":{
            "p":0.8666666667,
            "r":0.7647058824,
            "f":0.8125
        },
        "PERS":{
            "p":0.7352941176,
            "r":0.7462686567,
            "f":0.7407407407
        },
        "ORG":{
            "p":0.65625,
            "r":0.75,
            "f":0.7
        }
    },
}

```

```
    "speed":2471.587894989  
}  
]
```


Annexe G

Visualisation des modèles NER

Comandancia de Armas LOC
Arauco LOC Diciembre 14 de 1859 DATE
En este momento recibi la nota de Usted fecha de ayer DATE .
Inmediatamente LOC hice propio a Los Ángeles LOC , participando al Señor Intendente PERS , el éxito que ha tenido.
Tendré mucho cuidado en sujetar todos los animales MISC que arrean ilegalmente.
Todavía no he recibido los 80 animales MISC que le mandé, para hacer la correspondiente devolución a sus dueños PERS .
Aquí no hay novedad. El conductor don Salvador [Hermosilla] lleva 25 lanzas, con éstas son 124.
Dios guarde a Usted LOC .
José del Carmen Diaz PERS
Al Señor Comandante en Jefe PERS de la División Pacificadora de ORG Arauco LOC

Gobierno Interino de ORG
Arauco LOC Enero 2 de 1860 DATE
Por el Gobernador ORG del Departamento ORG de Lautaro LOC se me comunica lo que sigue:
Santa Juana LOC . Enero 2 de 1860 DATE
Por la Intendencia de mi PROVINCIA LOC en nota oficial fecha 29 del mes DATE próximo pasado N° 472, se me ordena poner a disposición de Usted al reo Juan Hermosilla PERS .
titulado Sargento Mayor PERS de la montonera ORG de Patricio Silva PERS ; para que allí sea juzgado, y en su consecuencia se lo remita bajo segura custodia,
y Usted ORG se servirá acusar recibo. Dios guarde a Usted LOC Pascual Ruiz PERS
Yo lo transcribo a Usted PERS para su conocimiento.
Mientras Usted PERS se sirva determinar de dicho reo, he dispuesto mandarlo a bordo del Vapor "Maipú" MISC para la mayor seguridad.
Dios guarde a Usted LOC .
José del Carmen Diaz PERS

FIGURE G.1 – Visualisation du modèle NER κ_2

Comandancia de Armas LOC
Arauco LOC Diciembre 14 de 1859 DATE
En este momento recibi la nota de Usted MISC fecha de ayer.
Inmediatamente hice propio a Los Ángeles LOC , participando al Señor Intendente PERS , el éxito que ha tenido.
Tendré mucho cuidado en sujetar todos los animales MISC que arrean ilegalmente.
Todavía no he recibido los 80 animales MISC que le mandé, para hacer la correspondiente devolución a sus dueños.
Aquí no hay novedad. El conductor don Salvador [Hermosilla] lleva 25 lanzas PERS , con éstas son 124.
Dios guarde a Usted PERS .
José del Carmen Diaz PERS
Al Señor Comandante en Jefe PERS de la División Pacificadora ORG de Arauco LOC

Gobierno Interino de ORG
Arauco LOC Enero 2 de 1860 DATE
Por el Gobernador PERS del Departamento ORG de Lautaro LOC se me comunica lo que sigue:
Santa Juana LOC . Enero 2 de 1860 DATE
Por la Intendencia ORG de mi provincia en nota oficial fecha 29 del mes próximo pasado N° 472 DATE , se me ordena poner a disposición de Usted al reo Juan Hermosilla PERS .
titulado Sargento Mayor PERS de la montonera ORG de Patricio Silva PERS ; para que allí sea juzgado, y en su consecuencia se lo remita bajo segura custodia,
y Usted se servirá acusar recibo. Dios guarde a Usted PERS Pascual Ruiz PERS
Yo lo transcribo a Usted para su conocimiento.
Mientras Usted MISC se sirva determinar de dicho reo, he dispuesto mandarlo a bordo del Vapor "Maipú" MISC para la mayor seguridad.
Dios guarde a Usted PERS .
José del Carmen Diaz PERS

FIGURE G.2 – Visualisation du modèle NER κ_6

Annexe H

Modélisation d'entité XML-TEI à partir d'une requête SPARQL

```
<person xml:id="Q4233309"
↪  xml:base="https://viaf.org/viaf/534145858091523021888" xml:lang="en"
↪  sex="1.0">
  <persname>Cornelio Saavedra Rodríguez</persname>
  <birth when-iso="1821-01-01">1 de enero de 1821</birth>
  <death when-iso="1891-04-07">7 de abril de 1891</death>
  <note type="description">Chilean general</note>
</person>
```

Listing 8 – Structuration du <person> au sein du <particDesc>

```
<place xml:id="Q33986" xml:base="https://www.geonames.org/3868626"
↪  xml:lang="en" type="city_in_Chile">
  <placename>Valparaíso</placename>
  <region>Valparaíso</region>
  <country>Chile</country>
  <geo>-71.619722222 -33.046111111</geo>
  <note type="description">city in Chile</note>
</place>
```

Listing 9 – Structuration du <place> au sein du <settingDesc>

Table des figures

1.1	Carte de l'usurpation progressive du territoire Mapuche (1810-1885) ¹	5
1.2	Carte schématique des principaux peuples autochtones existants ou ayant existés au sein du territoire chilien actuel - ©Wikipedia	6
1.3	Démonstration d'une segmentation d'image sur l'application eScriptorium .	14
1.4	Visualisation des règles de transcription	19
2.1	Transposition d'un caractère en matrice (niveau de gris)	25
2.2	Illustration simplifiée d'un neurone formel, ©Lucas Terriel, 2020	27
2.3	Exemple de masque à partir d'une <i>baseline</i>	29
2.4	Chaîne de traitement d'un procédé HTR	29
2.5	Architecture CNN à échantillonnage et BLSTM de détection des <i>baselines</i> au sein du moteur Kraken - @Benjamin Kiessling, 2019	31
2.6	Entraînement d'un modèle avec la méthode <i>skratch</i>	33
2.7	Exemple d'erreurs courantes pour le modèle ArMcFR ²	36
2.8	Évaluation DIS du modèle ArMcFR avec KaMi-Lib	37
2.9	Le processus d'entraînement du modèle ArSeg	38
2.10	Exemple de segmentation à partir du modèle ArSeg	40
2.11	Algorithme de Levenshtein - @wikipedia	41
2.12	Schéma de traitement des prédictions HTR	43
3.1	Schéma de conversion des fichiers ALTO vers le format XML-TEI	55
3.2	Structuration du teiHeader	58
3.3	Arbre décisionnel de structuration de <text>	60
4.1	Représentation vectorielle des relations sémantiques et syntaxiques - ©towardsdatascience, 2021	66
4.2	Exemple de reconnaissance d'entités nommées à partir du modèle camembERT	68
4.3	Démonstration de la plateforme Doccano	73
4.4	Schéma du processus d'entraînement d'un modèle NER en validation croisée ³ .	75
4.5	Explication du système de validation croisée à partir de la méthode K-fold – ©scikit-learn	76

4.6	Résultats globaux par κ selon les trois mesures principales	77
4.7	Résultats détaillés par entité	78
4.8	Système d'indexation des entités nommées au sein d'un fichier XML-TEI .	79
4.9	Structuration d'un triplet au sein d'un fichier XML-RDF	81
A.1	Distribution du nombres de documents par auteurs	98
A.2	Représentation des documents selon leur nature	98
A.3	Activités documentaires sur les années 1859-1860 (par jours)	99
A.4	Activités documentaires (par mois)	99
A.5	Distribution des mots	100
G.1	Visualisation du modèle NER κ_2	128
G.2	Visualisation du modèle NER κ_6	128

Liste des tableaux

1.1	Détails du jeu de données sélectionné	12
2.1	Résultats des modèles HTR ⁴	34
2.2	Les 7 erreurs les plus courantes du modèle ArMcFR	36
2.3	Evaluation des performances du modèle ArSeg	38
2.4	Analyse des effets du traitement automatique des erreurs HTR	44
4.1	Évaluation des performances du modèle BETO (selon standard GLUE) . .	67
4.2	Évaluation des performances du modèle κ_2	78

Liste des codes sources

1	Structuration d'un fichier ALTO	17
2	Application d'une règle <i>Schematron</i>	56
3	Exemple de structuration du sourceDoc	59
4	Script de conversion d'images TIFF vers le format JPG	102
5	Commande shell d'entraînement selon la méthode <i>skratch</i>	107
6	Commande shell d'entraînement selon la méthode <i>finetuning</i>	107
7	Commande shell d'entraînement selon la méthode <i>finetuning</i> d'un modèle de segmentation	107
8	Structuration du <person> au sein du <particDesc>	129
9	Structuration du <place> au sein du <settingDesc>	129

Table des matières

Résumé	i
Remerciements	iii
Bibliographie	v
Études sociopolitiques autour du Chili	v
Histoire et historiographie des relations Chileno-Mapuche	vi
Humanités numériques et science ouverte	vii
Éditions numériques	ix
Généralités autour de l'apprentissage Machine	xii
Reconnaissance d'écriture manuscrite	xiii
Traitement automatique du langage	xviii
Reconnaissance d'entités nommées	xxi
Le web sémantique	xxiii
Codes et scripts produits durant le stage	xxiv
Introduction	xxvii
I Production d'une <i>pipeline</i> de transcription automatisée	1
1 Les archives et ses données : un enjeu technique, méthodologique et juridique	3
1.1 <i>Araucanía</i> : conflit, histoire et archives	3
1.1.1 Une brève histoire d'un conflit	4
1.1.2 La question Mapuche : symbole des problèmes socio-politiques du Chili contemporain	7
1.1.3 La sous-collection de l' <i>Araucanía</i>	8
1.2 La construction d'un jeu de donnée : vers une utilisation tout terrain	9
1.2.1 Préparation des données numérisées et leur inventorisation	9
1.2.2 Constitution d'un corpus interopérable	11

1.3	Disséquer l'image et reconstruire l'information : une étape préliminaire à la transformation éditoriale	13
1.3.1	Ontologie d'une procédure de segmentation d'un document	13
1.3.2	XML-ALTO : un encodage structuré adapté à l'océrisation	16
1.3.3	Définir un protocole de transcription et d'annotation	18
1.4	Les enjeux du droit du patrimoine et de l' <i>open data</i> au Chili	19
1.4.1	Retour sur la législation des archives au Chili	20
1.4.2	Libéraliser l'accès aux données numériques	21
2	L'apprentissage machine et la reconnaissance de texte	23
2.1	État scientifique et technique autour de la reconnaissance de texte	24
2.1.1	Écriture, matrice et <i>deep learning</i>	24
2.1.2	La révolution technique de l'HTR	28
2.1.3	Kraken et eScriptorium : moteur et application pour l'HTR	30
2.2	Méthodologie appliquée à la production d'un modèle HTR	31
2.2.1	Produire un modèle HTR avec Kraken	32
2.2.2	Modélisation, résultats et interprétations	33
2.2.3	Les difficultés d'un modèle de segmentation	38
2.3	La gestion des erreurs : le post-traitement comme second souffle à l'HTR .	39
2.3.1	Principe de la Distance de Levenshtein	41
2.3.2	Mise en place d'une correction automatisée	41
2.3.3	Résultats et améliorations	43
II	Produire et enrichir une édition numérique	47
3	De l'HTR à XML-TEI : mise en place d'une chaîne de transformation	49
3.1	L'intérêt d'une édition numérique native	50
3.1.1	L'édition numérique pour les humanités	50
3.1.2	XML-TEI : retour sur un format pivot	51
3.1.3	Les outils numériques de valorisation	52
3.2	XML-TEI et les archives océrisées du conflit Mapuche	54
3.2.1	Convertir les données océrisées au format TEI	54
3.2.2	Documenter et valider un schéma	55
3.3	Mise en place de l'application autour d'un schéma	57
3.3.1	Le teiHeader et ses métadonnées	57
3.3.2	<sourceDoc> : une trace originale	58
3.3.3	Structurer et éditer un texte	59

TABLE DES MATIÈRES	139
4 Indexer et enrichir une édition numérique	63
4.1 L'ébullition du traitement du langage naturel	64
4.1.1 Principes généraux du TAL	64
4.1.2 Les réseaux neuronaux et le TAL : la révolution des modèles encodeurs-decodeurs	65
4.1.3 Le TAL, la reconnaissance des entités nommées et les humanités . .	67
4.2 Entrainer la reconnaissance d'entités nommées	70
4.2.1 Annoter un jeu de données pour la reconnaissance d'entités nommées	71
4.2.2 Produire un modèle appliquée à la reconnaissance des entités nommées	74
4.3 Indexer, enrichir et exploiter les données	78
4.3.1 Indexer les entités nommées au sein d'une édition numérique TEI .	79
4.3.2 Le web de données : SPARQL et Wikidata	80
4.3.3 Limites et solutions aux web de données	82
Conclusion	85
Acronymes	89
Glossaire	91
Annexes	93
A Le fonds Araucania	97
B Automatisation d'une conversion de format d'image	101
C Exemple de binarisation et de seuillage d'une image	103
D Entraînement d'un modèle HTR avec Kraken	107
E Exemple de fichier XML-TEI : AH0299	109
F Résultats des évaluation NER par <i>cross-validation</i>	119
G Visualisation des modèles NER	127
H Modélisation d'entité XML-TEI à partir d'une requête SPARQL	129