Chaînes d'acquisition, de traitement et de publication du texte

Alix Chagué, Floriane Chiffoleau, Matthias Gille Levenson, Hugo Scheithauer, Ariane Pinche

Chaînes d'acquisition, de traitement et de publication du texte : des images à la mise en ligne

Alix Chagué[1,2,3], Floriane Chiffoleau[3,4,5], Matthias Gille Levenson[6,7], Hugo Scheithauer[8], et Ariane Pinche[7,9]

[1]École Pratique des Hautes Études, [2]Université de Montréal,
[3]ALMAnaCH - Automatic Language Modelling and ANAlysis & Computational Humanities,
[4]Le Mans Université,
[5]3L.AM - Langues, Littératures, Linguistique des Universités d'Angers et du Mans,
[6]École nationale des chartes, [7]CIHAM (UMR 5648)
[8]Inria de Paris, [9]Centre Nationale de Recherche Scientifique

Ce livrable a été rédigé dans le cadre du consortium Huma-Num Ariane, axe 1 – GT3 : « Éditions numériques de qualité (coord. Anaïs Chambat et Nathalie Rousseau)

TABLE DES MATIÈRES

Au fil des ans, l'édition numérique s'est solidement établie comme un domaine majeur des Humanités numériques. Depuis l'avènement du standard *XML*[1] *TEI*[2] en 1987, une communauté scientifique s'est formée, partageant régulièrement ses pratiques à travers des publications telles que le journal de la TEI[3], RIDE[4], ainsi que lors de la conférence annuelle de la TEI[5]. Toutefois, la *Text Encoding Initiative*, en raison de sa souplesse inhérente, permet une grande diversité d'approches scientifiques. Ainsi, chaque édition numérique, bien qu'encodée en TEI, présente le texte de manière différente : éditions critiques, diplomatiques, hyperéditions basées sur des réseaux d'hyperliens et autres. L'importance de la modélisation des données dans les éditions est donc capitale, car elle reflète une interprétation scientifique et un travail philologique original, conférant toute sa valeur à l'objet produit. Bien que garantissant une certaine qualité scientifique, cette liberté herméneutique ne favorise pas le développement d'outils génériques pour accompagner les chercheur·se·s dans la constitution des corpus numériques jusqu'à leur publication en ligne. Les solutions sont multiples, la plupart des projets optant pour des solutions personnalisées qui s'adaptent parfaitement à leurs objectifs à l'aide de chaînes de traitement du texte sur mesure[6]. Ainsi, comme le souligne E. Pierazzo[7], la plupart des solutions actuelles sont des solutions « haute couture », spécifiques à chaque projet. Elles exigent un haut niveau d'expertise technique pour maîtriser le flux de texte numérique, rendant la création de nouvelles éditions fastidieuse, chronophage et coûteuse, sans parler des défis de maintenance à long terme. Aujourd'hui, cette diversité complique, voire entrave, le développement d'outils génériques pour accompagner la modélisation et la publication en ligne des éditions, bien que ces dernières années aient vu l'émergence d'outils de publication d'édition en XML TEI. Ainsi, *Leaf Writer* fournit une suite d'outils indépendants, mais interopérables pour créer, encoder et publier des sources historiques[8]. Quant à TEI Publisher[9], il fonctionne à partir du téléchargement des fichiers et d'un fichier ODD spécifique[10] pour la mise en place d'une application Web[11]. Toutefois, à ce jour, aucune solution de « prêt (et facile) à porter » n'est véritablement disponible pour des projets qui ne disposent pas des ressources nécessaires pour mettre en place une solution sur mesure pour leur corpus. Se pose également

[1] *eXtensible Markup Language*
[2] *Text Encoding Initiativ*e
[3] <https://journals.openedition.org/jtei/>
[4] <https://ride.i-d-e.de/issues/>
[5] <https://members.tei-c.org/Events/meetings/>
[6] Par chaîne, nous entendons le processus de constitution d'un texte numérique en plusieurs étapes distinctes : transcription et établissement du texte, annotation du texte, visualisation et exploitation du texte.
[7] Elena Pierazzo, « What future for digital scholarly editions? From Haute Couture to Prêt-à-Porter *», International Journal of Digital Humanities*, vol. 1 / 2, juillet 2019, p. 210.
[8] <https://www.leaf-vre.org/docs/features/leaf-commons>
[9] Floriane Chiffoleau. *TEI Publisher, a platform for sustainable digital editions*. 2023. hal-04247980.
[10] ODD, *One Document Does it All*, fichier qui contient le schéma TEI d'une édition, à la fois sous forme d'un document rédigé et d'un schéma technique qui permet de régie le balisage des fichiers auquel le schéma sera associé. Ce procédé dans teiPublisher permet de lier les éléments à des comportements/types d'affichage en fonction des formats de sortie voulus.
[11] Toutefois, pour personnaliser l'outil, il faut avoir quelques compétences en langage WEB et une bonne connaissance des schémas XML (ODD), ainsi que de Xquery.

la question de savoir si une telle solution est vraiment envisageable sans entraîner une approche éditoriale injonctive du type « *we-know-what's-best-for-you* », pour reprendre les termes de Lou Burnard, rendant toute liberté herméneutique impossible[12].

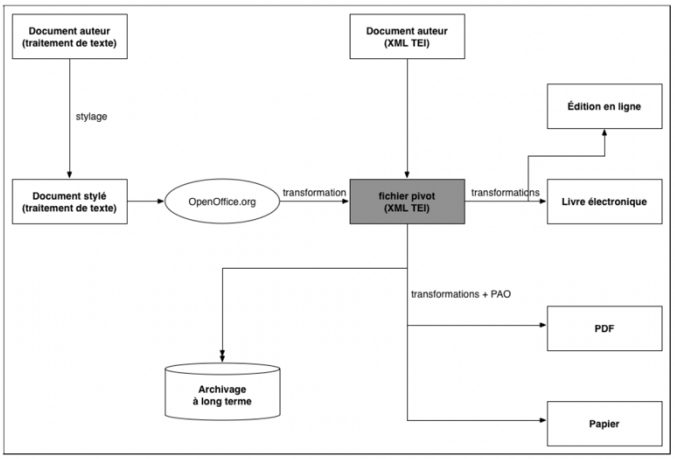


Figure 1: Schéma extrait de BUARD, Pierre-Yves, Modélisation des sources anciennes et édition numérique, thèse de doctorat, Université de Caen, 2015, [En ligne : https://hal.science/tel-01279385], p. 117.

À ce jour, des chaînes éditoriales numériques telles que celles de l'AEDRES (Association des éditeurs de la Recherche et de l'Enseignement supérieur)[13] ou du CLEO (Centre pour l'Édition Électronique Ouverte)[14] utilisent en entrée des feuilles de style Word ou LibreOffice, suivant le modèle d'organisation éditoriale inspiré par le concept de *single source publishing* (publication à partir d'une source unique) du *Chicago Manual of Style* (voir figure 1). Ce processus permet de produire des fichiers XML TEI à partir des fichiers stylisés, qui sont ensuite transformés pour générer des éditions en ligne, voire des publications papier. Cependant, ce schéma relativement simple n'est pas conçu pour répondre aux besoins spécifiques des éditions scientifiques, tels que l'annotation avancée du texte ou la gestion de niveaux de notes, par exemple. La modélisation adoptée ici ne découle pas d'une démarche herméneutique complexe, mais plutôt de la réduction du XML à un balisage éditorial de mise en page. Ce modèle est applicable à une édition numérique scientifique, mais au prix d'une simplication poussée de la complexité du document source pour obtenir un balisage éditorial générique[15].

Parmi les chaînes existantes adaptées aux projets de recherche, citons la chaîne Métopes développée à la Maison de la Recherche en Sciences humaines (MRSH) de Caen[16]. Cette chaîne intègre plusieurs outils, notamment un environnement de travail *XMLmind XML Editor* (XXE) qui, associé avec un scénario d'encodage, offre une expérience proche du traitement de texte en masquant les balises XML. Différents environnements sont proposés en fonction de la typologie des sources, permettant ainsi des solutions adaptées à chaque type de projet[17] sur le modèle de ce que Lou Burnard a appelé les principes d'encodage : « *we-know-what-we're-*

[12] Lou Burnard, « What is TEI Conformance, and Why Should You Care? », *Journal of the Text Encoding Initiative,* May 2020.
[13] Jean-Michel Henny, « Politique numérique (questions 33-42) », *L'édition scientifique institutionnelle en France : État des lieux, matière à réflexions, recommandation*, Association des Éditeurs De la Recherche et de l'Enseignement Supérieur, 2015, p. 85-92.
[14] Voir la documentation du logiciel Lodel : <https://github.com/OpenEdition/lodel/wiki>.
[15] Voir Pierre-Yves Buard, *Modélisation des sources anciennes et édition numérique*, Thèse de doctorat, Université de Caen, 2015, p. 88-101.
[16] <https://www.metopes.fr>.
[17] <https://mrsh.unicaen.fr/pluridisciplinaire/poles-pluridisciplinaires/pole-document-numerique/>.

doing » à partir de standards établis par une communauté de recherche sur un sujet précis[18]. Un plugin*, PluCo*[19], est également disponible pour faciliter le travail collaboratif sur des fichiers XML TEI. Enfin, en bout de chaîne, la solution MAX[20] offre la possibilité de visualiser le corpus à l'aide d'une base de données baseX avec un client graphique, ce qui facilite les requêtes et l'exploration des données.

Bien que ces chaînes permettent d'encoder et de visualiser le texte, elles ne permettent pas de l'enrichir avec des annotations linguistiques telles que les lemmes, les parties du discours (PoS, pour *Part of Speech*), ou encore des entités nommées, qui sont essentielles pour une exploitation plus efficace du texte. La chaîne éditoriale de la *Base de Français Médiéval* (BFM)[21] répond aux besoins de balisage, d'enrichissement et d'affichage du corpus en s'appuyant sur les ressources du logiciel TXM[22]. Tout comme la chaîne de l'AEDRES et du CLEO, elle repose sur le stylage de documents Word pour produire des fichiers XML TEI. Ces fichiers sont ensuite annotés à l'aide du lemmatiseur RNNtagger[23] pour l'ancien français, puis transformés pour être consultables dans le portail du logiciel TXM. Ce portail permet non seulement de lire le texte, mais aussi de mener des fouilles textuelles et des analyses, qu'il s'agisse de comptages de mots, de séquences de mots, d'inférences statistiques entre contextes et occurrences, ou encore d'analyses qualitatives à l'aide de concordanciers. Cependant, cette chaîne est conçue spécifiquement pour les éditions de textes médiévaux et ne comprend pas d'étapes d'acquisition automatique du texte. De plus, le texte doit nécessairement suivre les recommandations de stylage de la BFM pour être affiché de manière optimale, ou nécessitera un développement *ad hoc* pour intégrer les variations.

En outre, il existe également des chaînes non pas d'éditorialisation, mais d'enrichissement du texte afin de permettre une fouille plus efficace et plus précise des textes. Pour le traitement des textes anciens, citons le *Classical Language Toolkit.* Cette bibliothèque Python permet d'appliquer des outils de traitement automatique du langage à des langues anciennes (au total 19 langues, dont le latin et le grec), comprenant des étapes de tokenisation (séparation des mots), de lemmatisation et d'annotation linguistique, en incluant dans ses mises à jour les plus récentes des plugins de reconnaissance automatique d'écriture[24]. Cependant, cette boîte à outils ne propose pas d'éditeur de texte XML ni de solution d'affichage Web du corpus. Par ailleurs, pour la comparaison textuelle et l'analyse des variantes, des outils tels que CollateX et

---

[18] Lou Burnard, « What is TEI Conformance, and Why Should You Care? », *Journal of the Text Encoding Initiative, Text Encoding Initiative Consortium*, mai 2020, p. 3: "we-know-what-we're-doing (WKWWD) kind of encoding standard made up and maintained by the leading lights of a particular research community)".

[19] <https://mrsh.unicaen.fr/pluco/>.

[20] « MaX MRSH Maison de la Recherche en Sciences Humaines », https://mrsh.unicaen.fr/max/ ; consulté le 11 mars 2024.

[21] Alexei Lavrentiev et Céline Guillot-Barbance, « La BFM 2022 : un corpus pour les recherches diachroniques en français médiéval et au-delà », *Corpus, Bases, corpus et langage - UMR 6039*, janvier 2024.

[22] Serge Heiden, Jean-Philippe Magué et Bénédicte Pincemin, « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement », vol. 2 / 3*, Edizioni Universitarie di Lettere Economia Diritto*, 2010, p. 1021.

[23] « RNNTagger », <https://www.cis.uni-muenchen.de/~schmid/tools/RNNTagger/>, consulté le 15 mars 2024.

[24] Kyle P. Johnson, Patrick J. Burns, John Stewart, [et al.], « The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages », *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, éds. Heng Ji, Jong C. Park et Rui Xia, Online, Association for Computational Linguistics, 2021, p. 20-29.

StemmaWeb[25] ont été developpés. Toutefois, ces outils ne sont pas mis à jour régulièrement, et certains exigent des connaissances techniques avancées pour leur intégration dans des chaînes éditoriales, comme l'outil FALCON[26] ou le pipeline IRSB-integrator développé par V. Seretan, qui permet de passer des transcriptions XML TEI à un graphe de représentation des variations textuelles[27].

Ces dernières années ont vu l'émergence d'une première phase d'acquisition automatique du texte à partir des numérisations des corpus, ainsi que le développement de l'annotation automatique des corpus, qu'il s'agisse d'annotations linguistiques ou d'autres types, grâce à l'utilisation de lemmatiseurs fondés sur l'apprentissage profond tel que *Pie*[28], ou encore de grands modèles de langue comme BERT[29]. Face à cette avancée technologique, il est désormais possible de produire plus de texte, plus rapidement, permettant ainsi la création de corpus à grande échelle de manière (semi)automatisée. Ce changement est perceptible dans les productions scientifiques et influence même la TEI qui cherche à produire des encodages minimaux adaptés à ce type de corpus, certains étant destinés à des usages relevant de la lecture distante (*distant reading*[30]). C'est pourquoi nous envisageons ici des chaînes de production textuelle qui englobent le processus depuis l'acquisition du texte jusqu'à sa publication dans la continuité des protocoles expérimentaux proposés par les projets AGODA[31] et Gallic(orpor)a[32].

Il existe aujourd'hui des scripts plus aboutis, à l'instar de ceux élaborés dans le cadre du projet DAHN (Dispositif de soutien à l'Archivistique et aux Humanités Numériques)[33]. Ce projet mobilise des équipes interdisciplinaires à travers une collaboration entre l'Inria, l'EHESS et l'Université du Mans, visant à développer une chaîne d'acquisition textuelle depuis la reconnaissance automatique jusqu'à la production de fichiers XML TEI de fonds d'archives,

[25] Voir les travaux de T. Andrews, Tara L. Andrews, « The Third Way: Philology and Critical Edition in the Digital Age », *Variants,* vol. 10, Rodopi, 2013, 16 p., p. 61-76. Anahit Safaryan, Tara L. Andrews et Tatevik Atayan, « Continuous Integration Systems for Critical Edition: The Chronicle of Matthew of Edessa », Zenodo, 2019.

[26] Jean-Baptiste Camps, Lucence Ing et Elena Spadini, « Flacon : A processing workflow for automated collation », 2021.

[27] Violetta Seretan, « Sharing the Experience: Workflows for the Digital Humanities », *Digital Critical Edition of Apocryphal Literature: Sharing the Pipeline*, DARIAH-Campus, juin 2020.

[28] Enrique Manjavacas, Thibault Clérice et Mike Kestemont, « Emanjavacas/pie v0.2.3 », Zenodo, 2019.

[29] <https://aclanthology.org/N19-1423/>

[30] Lou Burnard, Christof Schöch et Carolin Odebrecht, « In search of comity: TEI for distant reading », *Journal of the Text Encoding Initiative,* mars 2021.

[31] Marie Puren, Aurélien Pellet, Nicolas Bourgeois, [et al.], « Between History and Natural Language Processing: Study, Enrichment and Online Publication of French Parliamentary Debates of the Early Third Republic (1881–1899) », *European Language Resources Association (ELRA), (éd.). Proceedings of The Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, éd. European Language Resources Association (ELRA), 2022, p. 16-24.

[32] Simon Gabay, Ariane Pinche, Kelly Christensen, [et al.], « Gallic(orpor)a : extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue », 2022. Ce dernier projet a également permis la mise en ligne de scripts d'acquisition textuelle permettant la génération de fichiers XML TEI dans le cadre du projet FONDuE pour des projets de petite ampleur : https://github.com/FoNDUE-HTR/Documentation/blob/master/notebook_pipeline.ipynb

[33] https://github.com/FloChiff/DAHNProject, https://github.com/DiScholEd/pipeline-digital-scholarly-editions

adaptables et réutilisables[34]. L'intégralité de la chaîne repose sur des outils *open source* et des standards largement diffusés dans la communauté des humanités numériques. Elle s'appuie sur les étapes de base des chaînes d'acquisition du texte : numérisation des images, transcription, encodage, extraction d'informations et publication. Elle nécessite peu de prérequis si ce n'est une modélisation des données en TEI. Ce projet, comme les exemples qui suivront, vise à établir un protocole de conception d'édition numérique et scientifique prêt à l'emploi, constitué de briques indépendantes et aisément maintenables, allant de l'archive matérielle à une publication en ligne. Ce protocole a par ailleurs été appliqué et enrichi avec succès dans le cadre du projet européen EHRI (*European Holocaust Research Infrastructure*)[35] des archives de l'Holocauste[36].

Né dans le contexte du consortium Ariane-HN[37], et face à l'émergence de l'intégration de l'intelligence artificielle dans la production de textes en sciences humaines, ce livrable vise à proposer une approche médiane qui ne soit pas limitée par des principes éditoriaux stricts, en proposant un protocole souple, adaptable et indépendant d'outils particuliers. En effet, enfermer cette réflexion dans une chaîne logicielle présente des risques, notamment en raison de leur obsolescence, de la diversité des besoins et du niveau de complexité des tâches liées aux particularités des corpus. C'est pourquoi nous préférons nous en tenir à une chaîne théorique adaptable en fonction des solutions techniques disponibles. En outre, nous tenons à souligner l'importance d'intégrer ces projets dans des infrastructures existantes et de respecter les standards des différents domaines afin de répondre aux critères scientifiques de son champ d'étude, mais aussi d'améliorer la visibilité et la pérennité du projet. Ainsi, en fonction des ressources à disposition et des objectifs des projets, nous proposons deux voies : une voie simple qui demandera peu de compétences en ingénierie et une voie plus complexe qui ajoutera un certain nombre de tâches d'automatisation dans l'acquisition du texte et son enrichissement, nécessitant une plus grande maîtrise des outils techniques ainsi qu'une compréhension plus approfondie de leurs enjeux scientifiques.

[34] Floriane Chiffoleau et Anne Baillot, « Le projet DAHN : une pipeline pour l'édition numérique de documents d'archives », 2022.
[35] Voir https://www.ehri-project.eu/. Consulté le 24/07/2024.
[36] Sarah Bénière, Floriane Chiffoleau et Hugo Scheithauer, « Streamlining the Creation of Holocaust-related Digital Editions with Automatic Tools », 2024.
[37] Le Consortium-HN ARIANE (Analyses, Recherches, Intelligence Artificielle et Nouvelles Editions numériques) a été labellisé par Huma-Num en 2023 pour une période de 4 ans. Il réunit des spécialistes du texte et de l'informatique en vue de créer un espace de dialogue scientifique interdisciplinaire afin de progresser dans la connaissance et le raffinement des méthodes informatiques appliquées aux objets et données des sciences humaines. Idmhand Fatiha, Ioana Galleron, Sabine Loudcher. « Consortium-HN ARIANE ». *Synthèse du projet scientifique*. 2023. ⟨halshs-04060828⟩

I-Chaîne de production simple

Nous présentons ici un modèle conceptuel de ce qu'une chaîne d'acquisition, de structuration, d'enrichissement et de publication de texte peut être, en liant ces étapes théoriques et indépendantes les unes des autres à des outils existants afin d'orienter chercheur·se·s et ingénieur·e·s dans la recherche de solutions adéquates pour leurs projets de recherche. Cette approche ne fournira pas de solutions prêtes à l'emploi. Elle exigera de combiner des compétences philologiques et des compétences techniques et/ou d'ingénierie plus ou moins avancées. Les premières seront nécessaires pour la composition et l'acquisition du corpus, ainsi que la modélisation des données qui relèvent d'objectifs herméneutiques. Les deuxièmes permettront la maîtrise de l'encodage, des outils d'automatisation de l'annotation, ou encore la mise en ligne, ainsi que le chaînage des différentes étapes. Il nous a paru essentiel de fournir une vue d'ensemble pour permettre aux lecteurs de réaliser l'étendue des connaissances nécessaires : modélisation d'un corpus en XML TEI, utilisation de modèles d'IA, mise en place de serveurs web pour la mise en ligne. Ces processus sont le fruit d'expertises variées et d'un travail collectif exigeant en termes de connaissances et de temps.

A. Acquisition du corpus

La constitution et l'acquisition d'un corpus sont des étapes cruciales pour la mise en place d'une chaîne d'acquisition textuelle. Elles doivent répondre à une problématique de recherche bien définie, et le choix des textes ainsi que leur traitement dépendent de plusieurs facteurs tels que la taille, le type de sources, l'hétérogénéité des documents, ainsi que des futures exploitations des données textuelles.

1. Constitution du corpus

Un projet n'est pas mené de la même manière s'il se concentre sur un extrait ou un texte relativement court, un corpus du même auteur ou homogène ou s'il recouvre un corpus très vaste et diversifié tant au niveau des sources, que des genres, périodes historiques et aires linguistiques. Un corpus homogène, constitué de textes similaires, est plus facile à traiter de manière cohérente, de même qu'un petit corpus peut souvent être traité manuellement, permettant alors une analyse détaillée, qualitative, ainsi qu'une méthodologie d'acquisition du corpus plus lâche.

En revanche, un grand corpus nécessite généralement une automatisation des tâches et des stratégies de normalisation/standardisation pour être géré de manière efficace et systématique. Dans ce cas, des outils d'acquisition automatique du texte ou de traitement automatique des langues (TAL) peuvent être utilisés, demandant de trier en amont les sources pour appliquer sur des corpus homogènes et cohérents les outils et modèles les plus adaptés (types de documents, d'écriture, langue du corpus). Afin de proposer une approche adaptée à chacun des documents, des critères de tri des documents peuvent être mis en place pour faciliter par la suite

un traitement automatique. Les critères choisis doivent être en lien avec l'organisation du projet et ses objectifs scientifiques. Une réflexion approfondie sur leur établissement pourra conduire à une analyse ontologique, permettant de définir le nommage des différentes métadonnées associées au corpus. Ces métadonnées pourront ensuite être intégrées dans le <teiHeader> des fichiers de publication des données (voir I.B.1). Les critères de tri peuvent inclure :

- La période de production de la source ;
- Le type de support : imprimé ou manuscrit ;
- Support sur lequel le texte est écrit ;
- La mise en page des sources ;
- Le genre des textes ;
- La langue des documents[38] ;

On peut également prendre en considération la chaîne de travail ou d'organisation du projet et inclure les questions suivantes :

- Les sources ont-elles déjà été numérisées ?
- Quelles personnes ou équipes supervisent cette partie du corpus ?
- Les sources sont-elles déjà transcrites ?
    - Les transcriptions disponibles peuvent être utilisées comme telles ?
    - Le texte doit-il être retravaillé pour harmoniser les normes de transcription au niveau du projet[39] ?

Si une étape d'annotation linguistique automatique du texte est envisagée, le critère de la langue des documents devient primordial pour choisir le modèle adéquat (voir II.C.2). Si on projette d'utiliser des outils d'acquisition automatique du texte, il est absolument nécessaire que le corpus soit numérisé au préalable. S'il n'est pas numérisé, il faut contacter l'institution détentrice des sources pour lancer une campagne de numérisation qui peut s'avérer longue et coûteuse. En plus de garantir une numérisation de haute qualité — avec un cadrage précis, une résolution de 300 dpi ou plus, une netteté adéquate, un éclairage approprié et une absence de distorsion de l'objet sur l'image — il est également crucial d'obtenir les droits de diffusion des images auprès de l'institution détentrice si jamais on souhaite les mettre en ligne ou encore publier les données d'entraînement de l'étape de reconnaissance automatique d'écriture (voir II.A.2). Toutefois, le copyright ne concerne que l'image numérisée, et non pas le document source. Ainsi, si, dans le cadre d'un projet de recherche, seule la publication du contenu textuel est prévue, aucune restriction ne peut être appliquée[40]. Enfin, les mêmes outils/modèles d'acquisition ne seront pas utilisés en fonction qu'on travaille sur des imprimés ou des manuscrits pour lesquels le type d'écriture ou de main peut également être déterminant, car on n'utilisera pas le même modèle pour reconnaître une main du XIX^{e} et un copiste médiéval (voir II.A.1). Enfin, la mise en page des documents peut avoir son importance pour l'étape d'analyse automatique de la mise en page (voir II.B.1) où il sera plus facile de traiter ensemble des mises en pages relativement similaires (correspondance, données tabulaires, texte en deux colonnes, etc.). En outre, une bonne reconnaissance de ces zones peut faciliter une pré-éditorialisation des documents en TEI en s'appuyant sur les données de la phase de segmentation de l'ATR (voir II.B.2).

[38] Critère important si l'on veut mettre en place un protocole de traitement automatique de la langue
[39] Voir I.A.1.
[40] Anna Busch, David Lassner et Aneta Plzáková, « ATR étape 2 : Où et comment obtenir des numérisations », iAnne Baillot, Mareike König (eds.), *Automatic Text Recognition. Harmonising ATR Workflows*, xx.05.2024, https://harmoniseatr.hypotheses.org/3780.

2. Transcription du corpus

La transcription manuelle des textes constitue une étape cruciale dans la préservation et l'exploitation des documents historiques. Cette section vise à détailler les différentes approches de la transcription manuelle, à expliquer les environnements de saisie utilisés, et à mettre en avant les règles fondamentales nécessaires pour une transcription rigoureuse et efficace. Il faut en premier lieu distinguer deux approches de la transcription manuelle :

1. La première approche que l'on appellera « traditionnelle » consiste à transcrire le texte de la source sans viser à proposer une transcription imitative qui tient compte de la représentation du texte sur son support. Ainsi certaines graphies pourront être normalisées, les abréviations développées, etc. Ces transcriptions servent souvent de base pour des éditions numériques où l'objectif est d'obtenir un texte numérique pour le publier en ligne à partir de fichiers TEI ou d'un simple fichier PDF, accompagné ou non des images du manuscrit, d'un apparat de notes explicatives et/ou d'un apparat critique. Cette transcription manuelle peut ensuite être utilisée pour des traitements simples comme de la mise en page dans un éditeur de texte ou alimenter une base de données en collectant les données au fil de la transcription.
2. Une deuxième approche de la transcription, moins limitée, consiste à conserver les informations de mise en forme présentes dans le document original : on garde alors trace des sauts de ligne, des sauts de page, voire, dans le meilleur des cas, on conserve l'alignement avec l'image originale. Cette approche est plus rigoureuse et respecte davantage la structure du document original, permettant une meilleure fidélité de la transcription et des perspectives d'exploitation du texte plus riche : analyse de la mise en page pour identifier un atelier, analyse statistique des abréviations pour identifier un changement de mains ou mieux saisir le public visé par le texte, faire des analyses comparées de leurs utilisations en fonction de la langue du texte.

Transcrire manuellement n'empêche pas de saisir la distinction qui s'opère entre texte brut et texte enrichi. Le texte brut désigne un texte dépourvu de toute mise en forme tandis que le texte enrichi contient des éléments de balisage (visibles ou non) qui structurent et stylisent le texte. Les éditeurs de texte se divisent alors en deux catégories : ceux qui interprètent le balisage et affichent directement le rendu final de la mise en forme contenue dans un texte enrichi, et ceux qui affichent le texte sous forme brute et laissent donc apparaître les balises qui portent les indications de mise en forme si elles existent, sans les interpréter. Dans le premier cas, on parle d'éditeurs *WYSIWYG*[41], dans le second, d'éditeurs *WYSIWYM*[42] ou simplement d'éditeurs de texte brut. Les éditeurs *WYSIWYM* peuvent demander une plus grande expertise puisqu'il faut pouvoir comprendre ce que les balises de mise en forme signifient. En revanche, ils donnent un plus grand contrôle sur le texte et sa mise en forme. Un logiciel de traitement de texte comme Microsoft Word, ou son équivalent applicatif Google Docs, entre dans la catégorie

[41] *What You See Is What You Get,* <https://www.google.com/url?q=https://theconversation.com/les-chercheurs-en-shs-savent-ils-ecrire-93024&sa=D&source=docs&ust=1722436942786621&usg=AOvVaw0-SLs5Y85s_8VloA_i3i6D>, prononcé *wi-zi-wig*.
[42] *What You See Is What You Mean*, prononcé *wi-zi-wim*.

des éditeurs *WYSIWYG*, tandis que des logiciels ou applications web comme *Oxygen*, *Stylo* ou encore *Sublime Text* sont des éditeurs *WYSIWYM*.

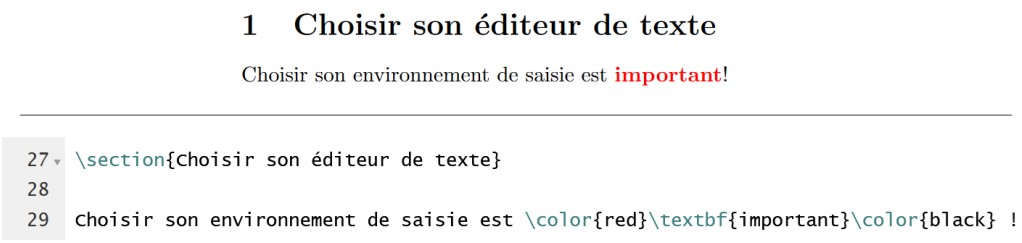


Figure 2 : Le même texte, en haut présenté en mode WYSIWYG, en bas en mode WYSIWYM, avec des annotations (en vert) qui reprennent la syntaxe du langage LaTeX (pour en savoir plus sur LaTeX: https://fr.overleaf.com/learn/latex/Tutorials).

Dans sa forme brute, le texte peut être encodé de plusieurs manières. L'encodage du texte désigne la manière dont les ensembles de bits (grâce auxquels la machine enregistre les informations comme le texte) sont décodés pour être affichés sous la forme d'un ensemble de caractères lisibles sur nos écrans. Il existe plusieurs standards d'encodage (ASCII et ISO 8859-1 en font partie), mais le plus commun aujourd'hui est UTF-8 qui est un format d'encodage édité par Unicode permettant de couvrir un très large ensemble de caractères représentant différents alphabets et signes diacritiques.

Certains signes sont cependant trop spécifiques pour être inclus dans le jeu de caractères défini par UTF-8 et des extensions peuvent être installées afin d'ajouter plus de caractères. C'est le cas de la MUFI[43] (*Medieval Unicode Font Initiative*), qui définit un ensemble de caractères spéciaux propres aux documents médiévaux. Afin de visualiser aux mieux ces caractères spéciaux, il est recommandé d'installer une police de caractères adaptée, comme *Junicode*[44] ou *PalemonasMufi,* qui permet d'associer à chaque caractère une représentation à l'écran[45]. S'assurer qu'un texte est saisi en utilisant la norme d'encodage UTF-8, ou l'étendre grâce à la MUFI, permet de garantir la compatibilité du fichier textuel avec les autres logiciels tout en préservant les caractères spéciaux nécessaires à la transcription d'un document.

En plus des éditeurs de texte traditionnels, les applications de saisie de transcription comme *eScriptorium*[46] et *Transkribus*[47]*,* développées pour automatiser la transcription (voir II.A), permettent de conserver l'alignement entre la transcription et l'image originale. C'est-à-dire que chaque ligne de texte est associée à des coordonnées sur l'image (voir II.B.1). Il est donc parfaitement envisageable d'utiliser ces applications uniquement pour l'environnement de saisie qu'ils proposent.

[43] <https://mufi.info>
[44] <https://junicode.sourceforge.io/>
[45] Liste des polices recommandées : <https://mufi.info/q.php?p=mufi/fonts>.
[46] B. Kiessling, R. Tissot, P. Stokes, [et al.], « eScriptorium: An Open Source Platform for Historical Document Analysis », *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2, 2019, p. 19-19.
[47] Philip Kahle, Sebastian Colutto, Günter Hackl, [et al.], « Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents », *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 04, 2017, p. 19-24.

Une bonne transcription devrait être homogène pour faciliter ensuite son exploitation, comme le repérage d'éléments structurels ou de certains mots. Par exemple, on pourra comptabiliser la fréquence d'apparition d'un signe dans une source si on l'a toujours transcrit de la même manière. Si les guillemets sont tantôt transcrits par < « >, par < “ > ou encore par < ‘ >, pour les comptabiliser, il faudra faire autant de requêtes qu'il y a de variations, au risque d'en oublier. Autre exemple, la transcription des césures en fin de ligne (figure 3) : si le ou les caractères employés pour les transcrire sont homogènes et documentés, cela facilite leur résolution ultérieure, tâche d'une certaine complexité technique (voir II.C.1)[48]. C'est pourquoi il est important d'établir des règles de transcription pour guider la saisie du texte et garantir son homogénéité, d'autant plus quand plusieurs personnes sont impliquées dans la transcription et/ou quand la transcription se fait sur un temps long.

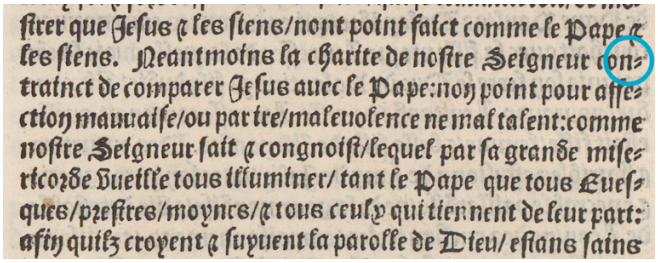


Figure 3: Exemple de texte comportant des traits de césure (entouré sur l'image), que l'on pourrait décider de transcrire par « ⸗ », « - » ou encore « ¬ », en fonction de la manière dont la tâche de transcription est envisagée.

Selon les principes énoncés notamment par Dominique Stutzmann[49], ces règles de transcription doivent répondre à plusieurs questions fondamentales :

1. Que transcrit-on ?
    - Il est important de déterminer si l'on transcrit uniquement le texte principal, les annotations, les marques de conservation du document, ou les marques pré-imprimées. Par exemple, pour une page donnée, il faut choisir clairement ce qui sera transcrit et ce qui sera ignoré, en fonction du corpus traité.
2. Comment transcrit-on ?
    - Un même phénomène manuscrit peut être transcrit de différentes manières puisque la transposition vers un format numérique est forcément la rationalisation d'un tracé libre sous la forme d'un caractère standard. Il est donc nécessaire de définir en amont la manière dont on traduit un phénomène manuscrit en une représentation numérique. Cette rationalisation permet d'assurer la cohérence et la qualité de la transcription. Par exemple si le texte transcrit contient un <d abréviatif> tel qu'on le trouve dans « duꝺ » (dudit), il faudra en amont décider si cela sera transcrit par « dudit », « dud. » ou encore « duꝺ ».
3. Pourquoi ? Et pour quoi faire ?

[48] Notez que, même lorsque l'homogénéité de la transcription d'une source est garantie par ces règles de transcription, si l'objectif est ensuite de réaliser de analyses lexicales, il faut éviter de s'appuyer sur une recherche plein texte et préférer utiliser le résultat d'un pré-traitement de lemmatisation (voir II.C.1).

[49] Dominique Stutzmann, « Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? », Franz Fischer, Christiane Fritze, Georg Vogeler, (éds.). *Kodikologie und Paläographie im digitalen Zeitalter 2 = Codicology and Palaeography in the Digital Age 2*, BoD, 2011, (« Schriften des Instituts für Dokumentologie und Editorik »), p. 247-277.

- La définition de règles claires est essentielle pour garantir la généricité et l'interopérabilité des transcriptions. Si les règles de transcription de plusieurs projets d'édition se rejoignent, alors il est possible de les mettre en commun pour créer de nouvelles ressources. De la même manière, si les règles sont clairement établies et décrites, il devient possible de créer des scénarios de transformation qui permettront de rendre les transcriptions compatibles entre elles. Une documentation détaillée permet également de partager les données de manière efficace et de faciliter la collaboration entre chercheur·se·s.

Pour approfondir ces questions, des ressources bibliographiques telles que « Les potentialités du texte numérique » de Sinclair & Rockwell peuvent être consultées[50]. Ces ouvrages offrent des perspectives détaillées sur les méthodes et les outils de transcription, ainsi que sur les normes d'encodage et les meilleures pratiques dans le domaine. Un bon exemple de règles d'annotation, pour les documents en langue romane, peut être trouvé dans l'initiative CATMuS (*Consistent Approaches to Transcribing ManuScripts*) [51].

La transcription manuelle des textes anciens est un travail exigeant qui nécessite une approche méthodique et rigoureuse. En distinguant les différents types de transcription, en choisissant les outils adaptés, et en définissant des règles claires, il est possible de produire des transcriptions précises et fidèles, essentielles pour la recherche académique et la préservation du patrimoine historique.

Quelques références pour aller plus loin :
– Ariane Pinche. *Guide de transcription pour les manuscrits du X^e au XV^e siècle*, 2022. ⟨hal-03697382⟩
– Peter Robinson et Elizabeth Solopova, « Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue », juillet 1993, 10.5281/zenodo.4050360.
– Stéfan Sinclair et Geoffrey Rockwell, « Chapitre 12. Les potentialités du texte numérique », in Marcello Vitali-Rosati, Michael E. Sinatra, (éds.). *Pratiques de l'édition numérique*, Presses de l'Université de Montréal, 2014, (« Parcours numérique »), p. 191-204.
– Dominique Stutzmann, « Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? », Franz Fischer, Christiane Fritze, Georg Vogeler, (éds.). *Kodikologie und Paläographie im digitalen Zeitalter 2 = Codicology and Palaeography in the Digital Age 2*, BoD, 2011, p. 247-277.

B. Modélisation du corpus en XML TEI

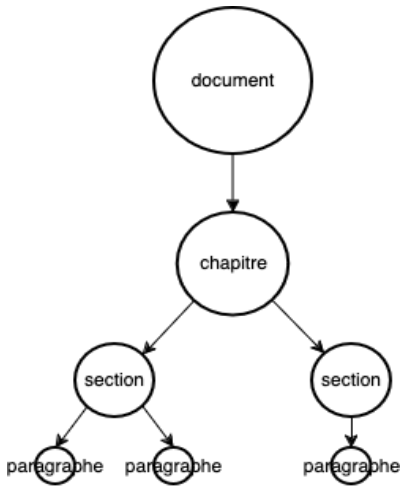
L'une des étapes essentielles dans la création d'une édition numérique est l'encodage des fichiers avec un standard qui permettra leur interopérabilité, tout en offrant la possibilité de les enrichir autant que nécessaire. Dans le domaine des humanités numériques, le standard le plus couramment utilisé pour la représentation du texte sous forme numérique est la *Text Encoding Initiative* (TEI). Ses directives constituent une liste détaillée de près de 600 éléments, spécifiant les règles d'encodage selon la structure du texte ou son genre, permettant ainsi l'enrichissement ciblé de son contenu. Elle permet non seulement de reproduire fidèlement la structure et les

[50] Stéfan Sinclair et Geoffrey Rockwell, « Chapitre 12. Les potentialités du texte numérique », Marcello Vitali-Rosati, Michael E. Sinatra, (éds.). *Pratiques de l'édition numérique*, Presses de l'Université de Montréal, 2014, (« Parcours numérique »), p. 191-204.
[51] <https://catmus-guidelines.github.io/>.

spécificités du document original, mais aussi d'ajouter des informations supplémentaires. Cela peut inclure des détails sur certains éléments mentionnés, une mise en contexte, ou encore d'autres types d'enrichissements : linguistiques, ecdotiques, philologiques, etc.

La TEI utilise la syntaxe XML, héritière de SGML[52]. Ces langages de balisage descriptifs s'appuient sur une organisation hiérarchique des éléments selon un système d'imbrication, appelé « *Ordered Hierarchies of Content Objects* » (OHCO). Ce système considère que les textes peuvent être représentés par des éléments imbriqués les uns dans les autres : chapitres, sections, paragraphes, listes, etc.[53]

```
<document>
      <chapitre>
        <section>
          <paragraphe>Texte du
    paragraphe 1.</paragraphe>
          <paragraphe>Texte du
    paragraphe 2.</paragraphe>
        </section>
        <section>
          <paragraphe>Texte du
    paragraphe 3.</paragraphe>
        </section>
      </chapitre>
</document>
```



Exemple de structuration en XML en regard de sa représentation en arbre hiérarchique.

Cette structure présente de nombreux avantages, car elle s'appuie sur la logique du contenu textuel et permet de construire un document facile à parcourir pour récupérer des informations ou créer des index. Les balises permettent ainsi de traiter les documents textuels comme des bases de données[54]. Cependant, dans la pratique, il n'est pas rare de se trouver face à un texte présentant plusieurs hiérarchies de différents types. Par exemple, une organisation matérielle (pages, notes marginales, titres courants) et une organisation logique (chapitres, sections, paragraphes) peuvent entraîner des chevauchements, comme un changement de page au milieu d'un paragraphe. Il faudra alors choisir une hiérarchie principale, en accord avec les objectifs scientifiques, pour l'encodage et adapter les autres en fonction de ce choix, notamment en

[52] Standard Generalized Markup Language (« langage de balisage généralisé normalisé » - SGML) est un langage de description à balises de norme ISO (ISO 8879:1986), <https://fr.wikipedia.org/wiki/Standard_Generalized_Markup_Language>. Le langage SGML est un langage standard pour créer des définitions de langage de balisage descriptif comme XML ou HTML. C'est un métalangage, c'est-à -dire un langage pour définir des langages. Créé dans les années 1980, son objectif était de créer un standard pour les langages à balise afin de publier des textes, voir Allen H. Renear, « Text Encoding », in Susan Schreibman, Raymond Georges Siemens, John M. Unsworth, (éds.). *A companion to digital humanities*, Malden, MA, Blackwell Publishing, 2004, p. 225-227.

[53] Steven J. DeRose, David G. Durand, Elli Mylonas, [et al.], « What is text, really? », *Journal of Computing in Higher Education,* vol. 1 / 2, décembre 1990, p. 3-26.

[54] Allen H. Renear, « Text Encoding », Susan Schreibman, Raymond Georges Siemens, John M. Unsworth, (éds.). *A companion to digital humanities*, Malden, MA, Blackwell Publishing, 2004, p. 218-270.

utilisant des balises auto-fermantes pour signaler les sauts de lignes et de pages[55]. Si subdiviser l'organisation de la matière textuelle en fonction d'une hiérarchie principale et des sous-hiérarchies permet de concilier différents aspects du texte et plusieurs types de visualisation, des cas plus complexes nécessitent des ajustements en respectant le principe de stricte imbrication des balises. Dans un encodage qui suit la structuration logique d'un texte[56], si, par exemple, un discours direct commence au milieu d'un paragraphe et continue sur un autre, il faudra soit utiliser des *milestones* (balises auto-fermantes), soit le diviser en fonction des paragraphes.

Il existe d'autres stratégies de modélisation du texte qui ne nécessitent pas ce système d'imbrication hiérarchique, traitant le texte comme un flux ou un graphe, « mais alors il n'y a pas de listes, mais seulement des entrées, pas de chapitres, mais seulement des titres de chapitres »[57], sur le modèle de certains éditeurs de texte grand public qui sont généralement plus faciles à prendre en main. Ainsi, la structure en XML induite par le choix de la TEI impose des contraintes dans la modélisation des données d'un projet en raison de la structure OHCO. Il est donc crucial de réfléchir aux différentes hiérarchies textuelles de l'objet à étudier pour optimiser l'encodage. Nous mettons, ici, l'accent sur l'importance de la construction d'un schéma d'encodage bien défini, en soulignant la nécessité d'une documentation rigoureuse pour de faciliter de travail et d'assurer l'homogénéité du corpus. Dans la suite de cette section, nous présenterons quelques éléments de la TEI dans le cadre d'un encodage de base, puis nous explorerons des utilisations plus avancées de ce standard.

1. Structurer son corpus en TEI

La TEI est là pour servir le texte grâce à ses multiples éléments et attributs qui permettent une annotation très fine et précise des documents. Généralement, en raison de l'étendue de ce standard, chaque élément à recenser est pourvu d'une balise. Un document TEI se divise en plusieurs parties, dans lequel chacune contient des informations spécifiques. Ils se composent de deux parties principales : (1) le <teiHeader>, pour encoder les métadonnées associées au document et (2) le <text> qui sera divisé lui-même en trois parties distinctes : le <front>, le <body> et le <back>, dont seule la balise <body> est obligatoire. Le *front* et le *back* contiennent respectivement les annexes d'avant et d'après texte (préface, note liminaire, épilogue, etc.), tandis que le *body* contient le corps du texte. Enfin, il existe aussi d'autres balises directement enfants de la balise racine TEI, mais facultatives, dont l'utilisation dépend du type d'enrichissement voulu. Il est possible d'ajouter à son encodage des détails issus de l'analyse de la mise en page du document avec les coordonnées des zones de texte et des lignes sur

[55] Pour une vue plus détaillée sur les avantages et les désavantages d'une approche OHCO, voir Allen Renear, Elli Mylonas et David G. Durand, « Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies », *Research in Humanities Computing*, 1996.
[56] Une hiérarchie qui s'appuie sur la structuration logique du texte est souvent à préférer pour faciliter par la suite la comparaison de différentes versions ou encore assurer la citabilité du texte en utilisant le standard DTS, Bridget Almas, Hugh Cayless, Thibault Clérice, [et al.], « Distributed Text Services (DTS) : A Community-Built API to Publish and Consume Text Collections as Linked Data », *Journal of the Text Encoding Initiative*, janvier 2023.
[57] Traduit de l'anglais, Steven J. DeRose, David G. Durand, Elli Mylonas, [et al.], « What is text, really? », *Journal of Computing in Higher Education*, vol. 1 / 2, décembre 1990, p. 11.

l'image, si l'on a procédé en amont à une acquisition automatique du texte (voir II.A.), avec le <sourceDoc>. Il est également possible d'ajouter des informations contextuelles ou d'autres types d'annotation (entités nommées, annotations linguistiques) avec le <standOff>.

```
<TEI xmlns=« http://www.tei-c.org/ns/1.0 »>
 <teiHeader>
    […]
 </teiHeader>
 <sourceDoc> [...] </sourceDoc>
 <text>
  <front>[…]</front>
  <body>
   <div>
    <p>Texte de l'exemple de document TEI.</p>
   </div>
  </body>
  <back>[...]</back>
 </text>
 <standOff></standOff>
</TEI>
```

Exemple de code avec les éléments enfants de l'élément racine <TEI>

Encoder les métadonnées de l'œuvre et du projet avec le <teiHeader>

Dans le cadre d'un encodage TEI, un même fichier permet de structurer le texte et d'ajouter des métadonnées : source(s), méthode d'encodage, etc. grâce au <teiHeader>[58]. Il est composé de quatre grandes parties, apportant chacune des informations différentes. Le <fileDesc>, élément obligatoire, est la partie du <teiHeader> qui est souvent la plus longue, elle peut contenir de très nombreuses informations, parfois très détaillées, sur la composition du fichier. Cela peut avoir trait au titre de l'œuvre, à son édition, à sa publication, la collection auquel il appartient ou même à une description physique de la source elle-même (figure 4). L'<encodingDesc> permet de décrire les principes et pratiques éditoriales appliqués lors de l'encodage, ainsi que les informations relatives au projet ou l'initiative dans lequel s'inscrit cette édition. Le <profileDesc> est là pour fournir une description détaillée des aspects non bibliographiques du texte, tels que sa date de création, les langues dans lequel il est écrit ou des détails de contexte. Enfin, comme pour tout fichier numérique, il est important de conserver la trace de ses avancées et de ses étapes de travail. Cela est possible grâce au <revisionDesc>. Il fournit un résumé de l'historique de révisions d'un fichier de toutes les modifications effectuées sur le fichier TEI depuis sa création (figure 5). Renseigner le <teiHeader> enrichit les fichiers numériques et facilite leur catalogage, leur moissonnage et leur exploitation. Le choix des métadonnées à inclure dépend des objectifs scientifiques du projet. Par exemple, pour une recension de corpus manuscrits visant l'étude d'un fonds documentaire, les données codicologiques stockées dans le <msDesc> et leur homogénéité au niveau de la collection

[58] Voir https://tei-c.org/release/doc/tei-p5-doc/fr/html/HD.html.

seront cruciales. En revanche, pour un corpus destiné à des études quantitatives, des métadonnées telles que le nombre de tokens ou les outils utilisés pour générer le corpus seront nécessaires.

```
<titleStmt>
    <title type="main">Orgueil et préjugé</title>
    <title type="sub">Par l'auteur de Raison et Sensibilité</title>
    <title type="sub">Traduit de l'anglois</title>
    <author>Jane Austen</author>
    <respStmt>
        <resp>Transcription by</resp>
        <persName>Floriane Chiffoleau</persName>
    </respStmt>
    <respStmt>
        <resp>Encoded by</resp>
        <persName>Floriane Chiffoleau</persName>
    </respStmt>
</titleStmt>
```

Figure 4: Exemple d'un <titleStmt>

```
<revisionDesc>
    <change when-iso="2021-02-17" who="#floriane.chiffoleau">Added a particDesc and
    settingDesc in the profileDesc</change>
    <change when-iso="2021-02-16" who="#floriane.chiffoleau">Encoding of the file</change>
</revisionDesc>
```

Figure 5 : Exemple d'un <revisionDesc>

Structurer le contenu du corps de texte

L'édition numérique passe principalement par l'encodage et la structuration du texte à l'aide de la balise <body>. Au sein de cette balise, la modélisation des données doit répondre aux exploitations futures du corpus. On peut par exemple consigner les aspects matériels du texte sur son support, tels que les mises en valeur (soulignement, gras, italique, etc.), les corrections (ratures, exponctuation, grattages, etc.) ou des notes ajoutées par l'auteur ou une autre main. Enfin, il est également possible d'enrichir le texte structuré en chapitres, sections, paragraphes et autres, avec des annotations linguistiques pour permettre, par exemple, une analyse textométrique.

De manière générale, au sein de la balise <body>, il est préférable d'encoder le texte à l'aide de divisions et subdivisions, qui seront faites à l'aide de la balise et qui peuvent, à l'aide d'attributs, se voir ajouter un identifiant, un type et un numéro. Une fois dans la division et avant de mettre le texte même, on peut donner un titre, s'il y en a un, à l'aide d'un <head> qui contient tout type d'en-tête (titre de section, intitulé de liste, description de manuscrit). D'autres balises permettent également d'ajouter des informations sur la disposition du texte sur son support originel tel que <lb>, pour les sauts de ligne, et <pb> pour les changements de page. Selon le type de texte et les objectifs scientifiques à atteindre, il sera préférable de choisir comme hiérarchie principale dans son encodage, un type d'organisation plutôt qu'un autre pour éviter le plus possible les phénomènes de chevauchement des balises. Il est possible d'encoder de simples paragraphes avec la balise <p>. Si on travaille sur un poème, l'utilisation de <lg>

et <l> pour les vers est recommandée. Autrement, on peut aussi encoder des listes (<list> et <item>), des tables (<table>, <row>, <cell>), etc.

```
<div type="book" n="3">
    <div type="chapter" n="9">
        <p>Elisabeth avoit arrangé dans sa tête<lb/> que Mr. Darcy lui amèneroit sa
        sœur le<lb/> lendemain de son arrivée à Pemberley
            ,<lb/> et en conséquence elle étoit bien déci-<lb break="no"/>dée à ne
            pas s'éloigner de l'auberge de<lb/> toute la matinée;
            mais elle avoit mal<lb/> calculé, car il l'amena à Lambton le<lb/> jour
            même. Elisabeth et sa tante qui<lb/> s'étoient
            promenées près de là avec quel-<lb break="no"/>ques-uns de leurs nouveaux
            amis, ren-<lb break="no"/>troient à l'auberge pour
            faire leur toi-<lb break="no"/>lette et aller dîner, lorsque le bruit d'un
            <lb/> équipage les attira vers la fenêtre, et
            elles<lb/> virent un monsieur et une dame dans un<lb/> Carriole qui
            s'arrêta à leur porte. Elisabeth,<lb/> reconnoissant la
            livrée, ne causa pas à son<lb/> oncle et à sa tante une légère surprise,<
            pb n="113"/><note type="foliation" place="top(right)"
                >113</note> en leur annonçant la visite qu'elle atten-<lb break="no"/>
                doit. Jusqu'alors ils n'avoient eu aucun<lb/>
            soupçon de ce qui se passoit ; mais com-<lb break="no"/>ment expliquer
            l'embarras d'Elisabeth, et<lb/> toutes les attentions
            de Darcy? Il falloit<lb/> qu'il eût du penchant pour leur nièce.</p>
    </div>
</div>
```

Figure 6: Exemple d'un texte régulier encodé

Outre la balise <body> qui contient le corps du texte, on peut trouver dans la balise <text> des annexes, qui se divisent en deux types : les annexes d'avant-texte ou <front> et les annexes d'après texte ou <back>. Une édition numérique pouvant être faite à partir d'une grande diversité de documents, la TEI a créé des options dans le <front> et <back> qui englobe beaucoup d'éléments qui peuvent se retrouver dans une annexe. Cela peut contenir des éléments tels qu'une préface, un résumé ou une table des matières pour le <front>, et un glossaire, un index ou des citations bibliographiques pour le <back>.

2. Enrichissement de l'encodage

Une fois que la structuration du texte est faite, il est possible d'enrichir son encodage avec d'autres éléments pour préciser tel ou tel détail.

Les entités nommées

Au sein de son texte, les noms de personnes, de lieux, d'organisations peuvent être signalés. Dans ce cas-là, il est nécessaire d'encoder ces entités avec des balises spécifiques[59]. Il existe deux techniques pour encoder les entités nommées, une version assez générique avec les balises <rs> et <name>, qui va encoder respectivement des noms génériques et des noms propres, et une version plus précise avec des balises spécifiques, telles que <persName>, <placeName> et <orgName>. Il est également possible d'ajouter des attributs qui permettront de lier ces entités à des notices d'autorité. Par exemple, il est possible de mettre un lien URL, voir un URN, grâce à l'attribut *ref* pour renvoyer vers un fichier d'autorité tel que VIAF[60], wikidata[61], GND[62] ou Geonames[63].

[59] Voir les chapitres <https://tei-c.org/release/doc/tei-p5-doc/fr/html/CO.html#COPU> et <https://tei-c.org/release/doc/tei-p5-doc/fr/html/ND.html>.
[60] <https://viaf.org/>
[61] <https://www.wikidata.org/wiki/Wikidata:Main_Page>
[62] <https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html>
[63] <http://www.geonames.org/>

```
<div type="named_entities">
    <div type="generic">
        <p><name type="person" ref="#p0001">Elisabeth</name> avoit arrangé dans sa
        tête que <name type="person" ref="#p0002">Mr.
                Darcy</name> lui amèneroit <rs type="person" ref="#p0003">sa sœur</rs>
                 le lendemain de son arrivée à <name type="place"
                ref="#l0001">Pemberley</name> , et en conséquence elle étoit bien
                décidée à ne pas s'éloigner de <rs type="place"
                ref="#l0003">l'auberge</rs> de toute la matinée; mais elle avoit mal
                calculé, car il l'amena à <name type="place"
                ref="#l0002">Lambton</name> le jour même. <name type="person" ref="
                #p0001">Elisabeth</name> et <rs type="person"
                ref="#p0004">sa tante</rs> qui s'étoient promenées près de là.</p>
    </div>
    <div type="specific">
        <p><persName ref="#p0001">Elisabeth</persName> avoit arrangé dans sa tête que
        <persName ref="#p0002">Mr. Darcy</persName> lui
            amèneroit <rs type="person" ref="#p0003">sa sœur</rs> le lendemain de son
            arrivée à <placeName ref="#l0001"
                >Pemberley</placeName> , et en conséquence elle étoit bien décidée à
                ne pas s'éloigner de <rs type="place" ref="#l0003"
                >l'auberge</rs> de toute la matinée; mais elle avoit mal calculé, car
                il l'amena à <placeName ref="#l0002"
                >Lambton</placeName> le jour même. <persName ref="#p0001">Elisabeth</
                persName> et <rs type="person" ref="#p0004">sa
                tante</rs> qui s'étoient promenées près de là.</p>
    </div>
</div>
```

Figure 7: Exemples d'utilisation de balises générales ou spécifiques d'entités nommées

Il est aussi possible de créer un fichier d'index séparé, avec des balises dédiées (<listPerson>, <person>, <listPlace>, <place>, <listOrg>, <org>), dans lequel chacune des entités mentionnées dans l'édition numérique seront inscrites et identifiées à l'aide d'un attribut *xml:id*, qui sera ensuite rappelé dans le corps même du texte, à l'aide d'un attribut *ref* et d'un identifiant précédé du signe #.

Ajouter un paratexte

```
<standOff>
   <listRelation type="family">
      <relation name="father" active="#père_Delmare" passive="#Charles_Delmare"/>
   </listRelation>
   <listRelation type="work">
      <relation name="employer" active="#Charles_Delmare" passive="#Geneviève"/>
   </listRelation>
   <listRelation type="friendship">
      <relation name="friends" mutual="#Charles_Delmare, #famille_Dumirail"/>
   </listRelation>
   <list type="certainty">
      <item>
         <certainty locus="value" target="#CE_head1" cert="high">
            <desc>Nous supposons ici que le chapitre 2 était numéroté, à l'instar du
            chapitre 1,
               à l'aide d'un chiffre romain. Donc "II".</desc>
         </certainty>
      </item>
   </list>
</standOff>
```

Figure 8 : Exemple d'utilisation du <standOff> pour préciser les relations entre les entités nommées

En plus des éléments mentionnés ci-dessus, il est possible d'enrichir davantage son édition pour contextualiser ou désambiguïser certains éléments. Cependant, entre les balises et les attributs pour la structure ou le contenu du texte, le corps du texte est déjà bien fourni, et rajouter encore plus de données pourrait rendre l'encodage hasardeux à contrôler et à lire. Pour éviter ces cas de figure, la balise <standOff>[64] permet d'encoder dans un élément à part les ajouts informationnels qui relèvent du paratexte : contextualisation, annotation grammaticales,

[64] https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-standOff.html

entités nommées (dates, coordonnées, notices), ou encore les relations entre les entités nommées (figure 8).

Aligner le texte et l'image : le <sourceDoc>

Dans certains cas, l'objectif est de conserver les informations sur la représentation du texte sur son support. Cela peut impliquer de vouloir conserver les éléments liés à la mise en page avec les coordonnées des lignes et des différentes zones de texte, ou encore à la transcription diplomatique du texte comme les abréviations. Il est possible de récupérer les informations quant à la mise en page et aux graphies originales (en fonction des normes de transcription du modèle utilisé) grâce aux informations générées pendant la phase de reconnaissance automatique du texte (voir II.A et II.B).

```
<surface xml:id="eSc_textblock_e94aafae" type="structure_{type:line_group;}"
    points="1770,370 1995,420 2093,471 2126,644 2013,690 1872,867 1849,1003
    <zone xml:id="eSc_line_c4880d79" type="mask"
        points="486,479 480,358 549,364 549,364 552,364 552,364 552,364 555,
        <path type="baseline" points="486,479 1993,465"/>
        <line>Sous le pont Mirabeau coule la Seine.</line>
    </zone>
    <zone xml:id="eSc_line_8af91efd" type="mask"
        points="515,635 509,537 636,534 639,534 639,534 642,534 642,534 656,
        <path type="baseline" points="515,635 2080,624"/>
        <line>Et nos amours, faut-il qu'il m'en souvienne ?</line>
    </zone>
    <zone xml:id="eSc_line_4ab1ac7a" type="mask"
        points="529,792 523,696 627,693 630,693 630,693 633,693 633,693 633,
        <path type="baseline" points="529,792 1924,783"/>
        <line>La joie venait toujours après la peine.</line>
    </zone>
    <zone xml:id="eSc_line_0d2948fe" type="mask"
        points="604,962 598,890 706,867 706,867 706,867 708,867 708,867 711,
        <path type="baseline" points="604,962 1255,948 1762,959"/>
        <line>Vienne la nuit, sonne l'heure,</line>
    </zone>
    <zone xml:id="eSc_line_00a0d9bd" type="mask"
        points="653,1037 651,979 743,962 743,962 746,962 746,962 746,962 749,
        <path type="baseline" points="653,1037 1768,1037"/>
        <line>Les jours s'en vont, je demeure.</line>
    </zone>
</surface>
```

Figure 9: Exemple d'un <sourceDoc>

Ces diverses coordonnées et données, récupérables depuis les exports XML PAGE ou ALTO (figure 9), peuvent aisément être transposées en TEI, grâce à l'élément <sourceDoc>[65] et ses divers enfants, tels que <graphic> pour mentionner l'image, <zone> pour mentionner le polygone qui comprend une ligne de texte, avec ses attributs variés mentionnant les coordonnées, et <line> pour retranscrire la ligne de texte.

3. Documenter son schéma d'encodage

Comme on a pu le voir, créer l'encodage d'une édition numérique peut être une tâche compliquée et chronophage. Selon la taille du corpus à traiter, il est possible que le projet s'étale sur plusieurs mois ou années, et il n'est pas toujours facile de se rappeler les choix qui ont pu être faits dans certains cas, surtout face à la grande variété d'éléments fournis par la TEI, et à la possibilité d'avoir plusieurs balises pour un même élément de texte. Il est donc important de pouvoir spécifier ses choix afin d'obtenir un encodage homogène. La TEI a créé un type particulier de fichier qui permet de faire explicitement cela : une ODD pour *One Document Does it all*. Un même fichier XML permet de définir le schéma d'encodage (enchaînement des

[65] https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-sourceDoc.html

éléments, valeurs d'attribut autorisées, etc.), mais aussi de le documenter. Il est possible de choisir d'exclure certaines balises, ce qui signifie que le fichier ne sera pas valide si ces balises sont utilisées dans les documents auquel le schéma personnalisé aura été associé. Définir sa propre ODD permet de restreindre les possibilités d'encodage afin d'assurer la cohérence des fichiers d'un même projet. L'intégralité des choix de modélisation des données, ainsi que le processus d'élaboration de cette dernière peuvent être documentés. L'intérêt de cette documentation est double. Elle permet non seulement de préciser les modalités d'encodage du corpus dans le cadre d'un projet collectif ou sur le long terme, mais également de faciliter la réutilisation et l'adaptation des données en fournissant une description complète de la méthode appliquée[66].

```
<head>Page breaks and facsimilies: <gi>pb</gi></head>
<p>The only element which refers to the picture is the page break
element <gi>pb</gi>. Each <gi>pb</gi> has a <att>n</att>
    attribute which indicates the page number. For technical
    reasons it is required to always start with page number 1.
    The differing page number of an archivist etc. must be marked
    up as a note. The <att>facs</att> attribute is used to
    link to the picture on the server. The first child element of
    <tag>div type="transcription"</tag> must be
    <gi>pb</gi>!</p>
<p>For example: <egXML xmlns="http://www.tei-c.org/ns/Examples">
        <body>
            <div type="transcription">
                <pb n="1" facs="00000356.jpg"/>
                <opener/> ... </div>
        </body>
    </egXML></p>
```

Figure 10 : Exemple de documentation de la balise <pb>

Quelques références pour aller plus loin :
– Marjorie Burghart er Elena Pierazzo, *Digital Scholarly Editions: Manuscripts, Texts and TEI Encoding*, DARIAH Teach, https://teach.dariah.eu/course/view.php?id=32, 2017
– Formation TEI, URFIST Rennes 2022 :
https://github.com/FloChiff/Introduction_TEI_2022
– Cours TEI, M2 TNAH : https://github.com/Segolene-Albouy/XML-TEI_M2TNAH
– Burnard, Lou, *Qu'est-ce que la Text Encoding Initiative ?* Traduit par Marjorie Burghart, OpenEdition Press, 2015, https://doi.org/10.4000/books.oep.1237.
– Allen Renear, Elli Mylonas et David G. Durand, « Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies », *Research in Humanities Computing*, 1996.
– Allen H. Renear, « Text Encoding », in S. Schreibman, R. Georges Siemens, J. M. Unsworth (éds.), *A companion to digital humanities*, Malden, Blackwell Publishing, 2004, p. 218-270.

C. Mettre en ligne son édition numérique

Une fois l'édition numérique encodée avec le standard TEI, l'étape suivante est la visualisation et la mise en ligne du corpus. Mettre en ligne son texte ne signifie pas simplement afficher son édition dans un navigateur web local, mais signifie rendre le corpus accessible à tous sur Internet. Cela demande donc la mise en place d'un serveur et l'utilisation de langages

[66] Voir en exemple l'ODD du projet DAHN (https://github.com/DiScholEd/pipeline-digital-scholarly-editions/tree/master/encoding/guidelines), ou du projet EHRI (https://github.com/EHRI/ehri-online-editions/tree/main/ODD).

web plus avancés tel que Python, JavaScript ou PHP. Un serveur web collecte et présente le contenu d'un site en temps réel à partir d'une base de données ou d'un serveur d'application, demandant ainsi un savoir-faire spécifique en ingénierie de projet. Il faut également trouver un hébergeur, institutionnel ou privé, ce qui peut entraîner des coûts à long terme[67]. En fonction du budget alloué au projet ou des ressources disponibles, il peut être judicieux d'éviter de développer une solution sur mesure et de préférer réutiliser des solutions existantes dont nous présenterons certaines dans la suite de cet exposé. Pour ce faire, il n'existe pas de solution unique, mais un large éventail de possibilités offrant divers degrés de personnalisation et d'interactions, nécessitant des ressources et une expertise en ingénierie plus ou moins poussées.

1. Pourquoi publier son édition numérique ?

Le terme de *publication* d'un corpus revêt ici deux sens. (1) Il peut s'agir de la publication des données brutes prêtes à être utilisées en dehors du projet qui leur a donné naissance. Cette mise en ligne est très importante pour assurer l'accès et la réutilisation aux données dans le respect des principes de la science ouverte. Afin d'optimiser ce partage des données, il faudra s'assurer dès le début du projet de les constituer dans un format ouvert, et standard si possible. Ainsi, il est important de déposer ses données brutes sur un quelconque outil de versionnage, qui permettra de conserver et de partager son travail. Ce dépôt peut être fait sur des outils tels que Nakala[68], Zenodo[69], GitHub[70], et GitLab[71]. De cette manière, images, fichiers de transcription, des fichiers XML, ou tout autre type de fichiers peuvent être centralisés et pérennisés. En outre, la publication des données brutes permettra d'assurer leur bonne citabilité grâce à l'attribution d'un DOI (*Digital Object Identifier*). (2) Il peut s'agir de la publication proprement dite, non plus des données, mais de la version consultable du corpus, qui permet de visualiser une ou plusieurs versions du texte et de consulter, voire exploiter (création d'index, analyse de réseau, *clustering*, interprétations statistiques) l'enrichissement du texte fourni par l'encodage. C'est à cette partie de la publication que ce chapitre est dédié.

La mise à disposition peut recouvrir plusieurs objectifs, dont le plus courant est la diffusion d'un corpus au sein d'une communauté scientifique. La mise en ligne permet aussi de mettre en avant l'enrichissement apporté aux documents, et de proposer des parcours de lecture différents. En effet, une édition numérique peut être aussi bien destinée aux spécialistes, qu'à un public plus large sans expertise particulière. Sa consultation en ligne permet un accès plus facile au texte, notamment si elle offre plusieurs parcours de lecture en fonction des intérêts des lecteurs et des options de personnalisation pour la visualisation. En outre, les corpus numériques offrent une exploration des textes aisée et rapide, grâce à des recherches plein texte ou des systèmes de requête plus précis en fonction des enrichissements proposés dans les fichiers XML, économisant de longues heures de recherche en archives à feuilleter des

[67] L'infrastructure Huma-Num peut accompagner la diffusion et la conservation sur le long terme des données numériques, voir <https://www.huma-num.fr/>.
[68] <https://nakala.fr/>
[69] <https://zenodo.org/>
[70] <https://github.com/>
[71] <https://gitlab.com>

centaines ou des milliers de pages pour trouver peut-être une unique mention de l'élément qui nous intéresse tout particulièrement. Enfin, grâce à la richesse des données encodées et possiblement liées à des bases de données, il est possible de créer de l'interactivité avec le lecteur qui pourra créer des visualisations au gré de ses requêtes : réseaux de personnes, arbres de relations, frises chronologiques, etc. (figure 11).

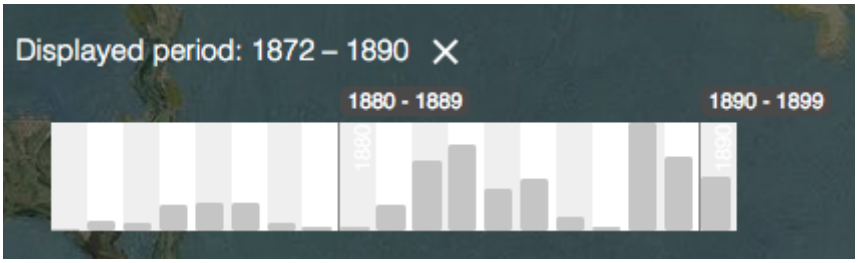


Figure 11: Visualisation de la chronologie des lettres de Vincent Van Gogh

2. Comment choisir un outil de publication

Afin de s'assurer du bon déroulé de la mise en ligne du corpus, il est important de bien choisir son outil dès le début du projet pour éviter de se retrouver face à des limitations techniques, notamment l'impossibilité d'afficher un ou plusieurs éléments de son corpus, ou de s'imposer trop de maintenance, grâce à l'offre d'un *back-end*[72] préétabli. Il faudra également évaluer les besoins réels. Si une visualisation simple sans interrogation dynamique d'une base de données suffit, il n'est pas nécessaire de surdimensionner les moyens pour la mise en ligne. La mise en place d'une solution *ad hoc* pour des productions limitées dans le temps ou sans soutien institutionnel peut s'avérer complexe et compromettre la réalisation d'un projet en raison d'un manque de moyens ou d'une expertise technique adéquate. Pour faciliter la mise en ligne des corpus, plusieurs solutions, de l'outil de génération d'une page GitHub aux CMS, sont disponibles. Les CMS ou *Content Management System* (« système de gestion de contenu » en français) sont des programmes informatiques qui permettent d'assister la création d'un site internet, afin d'éviter l'utilisation d'une solution maison, plus compliquée et généralement bien moins pratique. Comme le dit Elena Pierazzo, utiliser une solution *prêt-à-porter* plutôt qu'une création maison *Haute Couture*, peut aider à diffuser plus amplement les éditions numériques et offre des environnements plus durables et pérennes[73]. Avoir recourt à des outils ou CMS pour la mise en ligne du corpus est également mieux adapté à des néophytes, car cela demande moins de maintenance et de connaissances en programmation, tout en requérant toutefois un minimum de connaissances dans des langages tels que HTML, CSS ou XPath, afin de faire fonctionner au mieux l'outil. Enfin, ces solutions sont habituellement documentées, ce qui

[72] Le *back end* désigne l'ensemble des composants et des technologies qui fonctionnent en arrière-plan pour gérer et stocker les données, exécuter les processus nécessaires, et fournir les informations au *front end* (la partie visible pour les utilisateurs). Dans le contexte d'un corpus numérique, le *back end* inclut des éléments tels que les bases de données où les textes et métadonnées sont stockés, les serveurs qui hébergent le site web, et les applications qui traitent les requêtes des utilisateurs.

[73] Pierazzo, E. What future for digital scholarly editions? From Haute Couture to Prêt-à-Porter, *International Journal of Digital Humanities*, 2019, https://doi.org/10.1007/s42803-019-00019-3

permet de les prendre plus aisément en main et d'aider au déploiement en ligne, quand cela est possible. Cependant, dans le choix de son outil, il est important de prendre en considération leur affinité avec la gestion de fichiers TEI, leurs fonctionnalités (visualisation, fouille de texte, etc.), et surtout les ressources nécessaires pour assurer la pérennité du site et des fichiers.

3. Quelques outils de publication en ligne

Les différents outils dédiés à la mise en ligne de corpus numériques présentent chacun des caractéristiques uniques qui peuvent être utiles en fonction des objectifs scientifiques et des types de documents à afficher. En plus de permettre la visualisation des textes, ces outils sont souvent conçus pour afficher les enrichissements apportés par l'encodage en TEI, telles que la reconnaissance des entités nommées (voir II.C.2) ou la comparaison des différents témoins d'une édition critique. Bien que l'installation locale de ces outils et l'affichage des documents puissent être relativement simples, la mise en ligne d'un serveur pour rendre ces contenus accessibles pose des défis techniques. Dans ce contexte, il peut être nécessaire de faire appel à un·e ingénieur·e ou de recourir à des versions en ligne de ces outils, lorsque disponibles. Alternativement, des structures comme Huma-Num peuvent offrir des solutions d'hébergement web et un soutien pour l'installation des instances.

OMEKA

OMEKA[74] est un logiciel libre de gestion de bibliothèque numérique. Certains projets, tels que sur la plateforme EMAN (figure 12) ou le projet EHRI (figure 13), l'utilisent pour l'affichage de fichiers XML dans le cas de certains projets. Ce logiciel nécessite des manipulations spécifiques qui peuvent impliquer beaucoup de maintenance et entraîner des complications. Par conséquent, s'il est l'outil de référence pour la mise en ligne de bibliothèques numériques et permet d'afficher les images accompagnant les fichiers TEI, il n'est propice pas à une navigation fine dans des fichiers textes ni des plus adaptés pour la visualisation de fichiers TEI enrichis.

Figure 12: Plateforme EMAN utilisant OMEKA

Figure 13: Site Diplomatic Reports utilisant OMEKA

[74] https://omeka.org/

CETEIcean

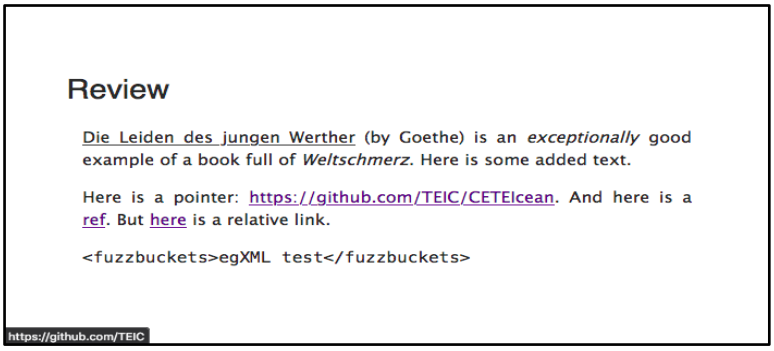


Figure 14 : Exemple de CETEIcean

Souvent présenté comme le successeur de *TEI Boilerplate*[75], *CETEIcean*[76] est une bibliothèque Javascript, qui permet l'affichage de fichiers TEI sous la forme de HTML dans le navigateur. Le site dédié au projet[77] propose alors plusieurs options de déploiement en ligne, dont une solution peu coûteuse via la création de pages GitHub. L'outil est pertinent, quand on souhaite une simple présentation du texte sans besoin d'affichage dynamique ou d'interrogation d'une base de données. L'outil n'offre que très peu d'interactivité et est principalement utile pour la production de sites statiques qui ont l'avantage de demander peu de maintenance.

MaX

Le *Moteur d'affichage XML*, ou *MaX*[78], est un outil ouvert développé par la MRSH de Caen[79]. Il offre la possibilité de visualiser un corpus à l'aide d'une base de données baseX avec un client graphique, ce qui facilite les requêtes et l'exploration des données. Le projet *Estrades*[80] s'appuie sur ces outils pour proposer une infrastructure afin d'offrir une logique de plateforme et de support à la recherche à plus large échelle. Outre les outils de structuration, d'annotation, il permet également de lier aux corpus XML TEI des bases de données, de type *Heurist*, regroupant des entités, des index pour permettre une utilisation et une interrogation optimales, à l'aide de XQuery, des données connexes aux documents textuels grâce au pipeline *Heimdall*[81].

[75] https://dcl.luddy.indiana.edu/teibp/
[76] https://teic.github.io/CETEIcean/
[77] https://github.com/TEIC/CETEIcean
[78] Une documentation extensive de l'outil est disponible au lien suivant : https://pdn-certic.pages.unicaen.fr/max-documentation/.
[79] « MaX · MRSH · Maison de la Recherche en Sciences Humaines », <https://mrsh.unicaen.fr/max>. Max est téléchargeable à l'adresse suivante : <https://outils.biblissima.fr/en/xml-editing-tools/#publication>.
[80] https://estrades.huma-num.fr/
[81] https://gitlab.huma-num.fr/datasphere/heimdall

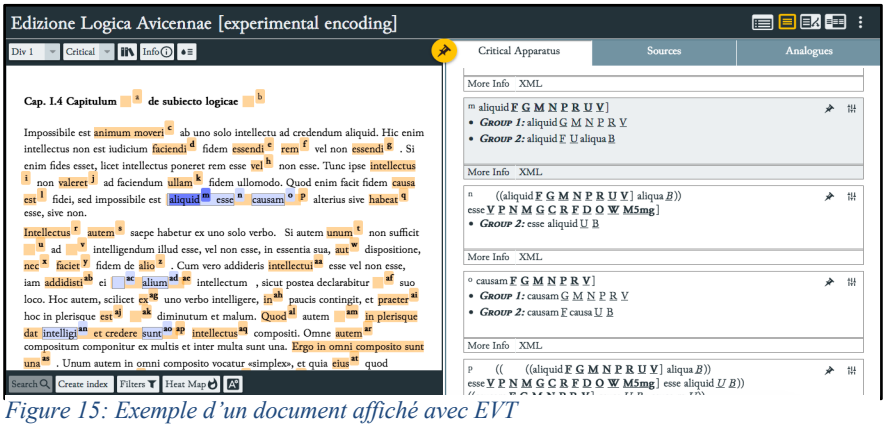


Figure 15: Exemple d'un document affiché avec EVT

Edition Visualisation Technology (EVT)[82] est un outil léger, librement accessible et libre accès, dont le code est disponible sur GitHub[83]. Présenté lors des conférences DH 2015[84] et 2016[85], il a été créé pour l'édition critique numérique de fichiers XML dans le cadre du projet *Digital Vercelli Book project*[86]. Il possède une interface utilisateur et ne requiert pas de connaissance en programmation de la part de son utilisateur. C'est un outil qui a été conçu pour des éditions numériques avec des apparats critiques, mais aussi pour la parallélisation des différents témoins d'un même texte. Pour répondre aux besoins croissants de la communauté TEI, de nouveaux développements d'EVT sont prévus pour en faire un outil plus général pour la publication web de documents basés sur la TEI, et proposer un service équivalent à celui de *TEI Publisher*.

TEI Publisher

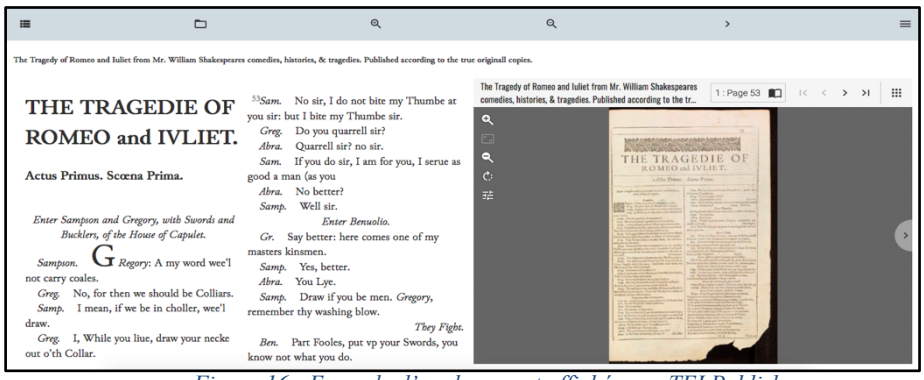


Figure 16 : Exemple d'un document affiché avec TEI Publisher

[82] http://evt.labcd.unipi.it/
[83] https://github.com/evt-project/evt-viewer/
[84] http://evt.labcd.unipi.it/abstracts/DHBenelux2015-Antwerp_[abstract].pdf
[85] http://evt.labcd.unipi.it/posters/DH2016-Krakow.pdf
[86] http://vbd.humnet.unipi.it/beta2/

TEI Publisher[87] est un outil pour la publication pour des fichiers TEI. Descendant de TEI Simple[88], ce CMS, uniquement dédié à des éditions XML, fonctionne avec des *templates* déjà créés et adaptées à toute sorte de visualisation (facsimilé ou non, index d'entités, informations liées, etc.), ainsi que des fichiers de transformations, sous la forme d'ODD, qui permettent de déclarer comment on veut afficher chaque élément de son fichier XML. Il offre ainsi la possibilité de profiler la visualisation du corpus directement sur la modélisation des données en TEI. Afin de découvrir ces diverses possibilités, *TEI Publisher* permet de tester les différentes options de visualisation dans un « terrain de jeux »[89]. L'outil est également pourvu d'une documentation exhaustive et détaillée[90].

Quelques références pour aller plus loin :
– Bridget Almas, Hugh Cayless, Thibault Clérice, [et al.], « Distributed Text Services (DTS): A Community-Built API to Publish and Consume Text Collections as Linked Data », *Journal of the Text Encoding Initiative*, Text Encoding Initiative Consortium, janvier 2023.
– Amit Kumar, Susan Schreibman, Stewart Arneil, [et al.], « <teiPublisher>: A Repository Management System for TEI Documents », *Literary and Linguistic Computing*, vol. 20 / 1, mars 2005, p. 117-132.
– Elena Pierazzo, "What future for digital scholarly editions? From Haute Couture to Prêt-à-Porter", *International Journal of Digital Humanities*, 2019, https://doi.org/10.1007/s42803-019-00019-3

Constituer et transcrire son corpus, modéliser les données, mettre en ligne les fichiers, telles sont les étapes essentielles d'une chaîne basique pour la production et la diffusion en ligne des textes. Chaque phase, qu'il s'agisse de la transcription, de la modélisation en TEI ou de la structuration du contenu, joue un rôle crucial dans la création de corpus numériques exploitables et accessibles. Ces étapes permettent non seulement de garantir la pérennité et la réutilisation des données, mais aussi d'enrichir le texte de manière à en faciliter l'analyse et la diffusion scientifique. Toutefois, bien que ces étapes soient indispensables, elles peuvent se révéler insuffisantes face à l'évolution des outils numériques et aux volumes croissants de données textuelles à traiter. Dans ce contexte, les avancées récentes en matière d'automatisation offrent des solutions pour accélérer et optimiser ces processus. La transcription automatique, l'analyse de la mise en page ou encore l'enrichissement des textes par l'annotation linguistique et la reconnaissance d'entités nommées sont autant d'outils qui, bien que perfectibles, ouvrent de nouvelles perspectives pour la gestion des corpus. C'est pourquoi, dans la deuxième partie de cet exposé, nous nous pencherons sur ces technologies d'automatisation et leur intégration dans la chaîne d'acquisition textuelle, en mettant en lumière les opportunités qu'elles offrent ainsi que les défis techniques qu'elles posent.

[87] https://teipublisher.com/exist/apps/tei-publisher-home/index.html
[88] https://www.balisage.net/Proceedings/vol15/html/Wicentowski01/BalisageVol15-Wicentowski01.html
[89] https://teipublisher.com/exist/apps/tei-publisher/index.html
[90] https://teipublisher.com/exist/apps/tei-publisher/documentation

II-Automatisation et enrichissement de la chaîne d'acquisition de texte

Dans cette section, nous examinerons comment les technologies actuelles permettent d'automatiser certaines étapes d'une chaîne d'acquisition de texte, et comment des processus comme l'acquisition automatique et l'enrichissement via l'annotation linguistique ou la reconnaissance d'entités nommées (NER, *Name Entities Recognition*) peuvent être intégrés. Bien que les outils d'annotation soient aujourd'hui performants pour des langues modernes bien dotées comme le français ou l'anglais, il reste essentiel de prévoir des étapes de contrôle régulières et adaptées, selon les exigences de qualité nécessaires à l'exploitation future du corpus. Par ailleurs, plus les étapes d'automatisation font appel à des technologies complexes telles que la vision par ordinateur, le traitement automatique des langues (TAL) et les techniques d'apprentissage profond (*deep learning*), plus la gestion du projet devient complexe, surtout pour de grands corpus nécessitant l'enchaînement fluide des différentes étapes sans intervention humaine. L'intégration de ces technologies requiert donc, au minimum, une maîtrise des bases de la programmation, en particulier en *Python*, afin de développer et ajuster les protocoles de traitement.

A. Acquisition automatique du corpus

La transcription automatique des textes (ATR, pour *Automatic Text Recognition*)[91] est une méthode de plus en plus utilisée dans les sciences humaines pour acquérir des corpus à grande échelle. Cette technologie a le potentiel de transformer de manière significative l'étude des textes en permettant de travailler à des échelles inédites, susceptible de remettre en question des canons historiques et littéraires[92]. Elle favorise également une recherche en contexte, grâce à l'accessibilité accrue d'un plus grand nombre de textes. Cette section présente les différentes facettes de la transcription automatique, ses défis, et les spécificités des différents types de documents. La phase d'acquisition du texte, centrale dans tout projet, conditionne en grande partie les usages et exploitations futurs. Il est donc essentiel de bien comprendre les enjeux méthodologiques de l'ATR et leurs conséquences.

L'ATR repose sur des logiciels souvent basés sur des technologies d'apprentissage automatique, capables de convertir des images de textes en données numériques exploitables.

[91] Parmi les logiciels les plus connus : Clemens Neudecker, Konstantin Baierer, Maria Federbusch, [et al.], « OCR-D: An end-to-end open source OCR framework for historical printed documents », *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, New York, Association for Computing Machinery, 2019, (« DATeCH2019 »), p. 53-58, Thomas M. Breuel, « Ocropy: Python-based tools for document analysis and OCR », 2014, Thomas M. Breuel, « The OCRopus open source OCR system », *Document Recognition and Retrieval XV*, vol. 6815, SPIE, 2008, p. 120-134., Saurabh Dome et Asha P Sathe, « Optical Charater Recognition using Tesseract and Classification », *International Conference on Emerging Smart Computing and Informatics (ESCI)*, 2021, p. 153-158.

[92] Ariane Pinche et Peter Stokes, « Historical Documents and Automatic Text Recognition: Introduction », *Journal of Data Mining & Digital Humanities*, Historical Documents and automatic text recognition, Episciences.org, mars 2024.

Si des solutions « clé en main » existent pour la reconnaissance optique de caractères (OCR) sur des documents imprimés, leur efficacité et leur complexité varient en fonction des caractéristiques des documents. Bien que l'acronyme « OCR » suggère une lecture caractère par caractère, les systèmes d'ATR contemporains tendent à analyser des lignes entières, plutôt que des lettres isolées, permettant de retranscrire des écritures manuscrites avec des lettres liées. Contrairement à une transcription manuelle, qui repose sur la compréhension sémantique, l'ATR traite des ensembles de pixels pour générer des caractères numériques. Cette approche machine présente des limites, notamment face à des mises en page complexes (notes marginales, insertions interlinéaires) ou des styles de texte inhabituels, là où un humain déchiffrerait facilement. Cela signifie aussi qu'il est difficile de demander à un logiciel d'ATR de transcrire tout en résolvant des abréviations ambiguës, où le sens de la phrase et le contexte sont essentiels pour trancher entre plusieurs choix de résolution quand la machine ne voit les lignes qu'une à une.

1. Usage et limites de la transcription automatique

L'OCR est une technologie qui existe depuis plusieurs années de manière très fonctionnelle, toutefois certains types de document résistent encore aux logiciels clef en main. C'est le cas des incunables, des imprimés anciens ou encore des journaux qui présentent de nombreux défis : accidents d'impression et effacement de caractères, complexité des fontes, qualité du papier[93], densité de la mise en page qui conduit les lignes de plusieurs colonnes de texte à être mal distinguées les unes des autres, etc. Les langues non alphabétiques présentent aussi un défi important pour l'automatisation de leur transcription puisqu'au lieu de devoir apprendre un peu moins d'une centaine de caractères, les systèmes de transcription doivent apprendre à distinguer entre des milliers de signes possibles.

Les documents manuscrits, quant à eux, sont encore plus difficiles à traiter pour les logiciels d'ATR en raison de plusieurs phénomènes. En tout premier lieu, les variations de forme entre des textes écrits par des mains différentes, mais aussi celles pour une même lettre au sein d'un texte écrit par une même personne, multipliant ainsi le nombre de tracés possibles pour une même lettre[94]. Ce sont autant de variations[95] que le système doit apprendre à reconnaître. D'autres facteurs de variations formelles s'ajoutent à cela : une variation géographique et diachronique importante, bien connue des paléographes, des variations formelles causées par le support et l'outil d'écriture employés, ou encore les différents styles d'écriture[96]. Les abréviations, qui sont fréquentes dans les documents manuscrits, constituent également un point d'achoppement méthodologique important.

D'autres types de documents présentent des difficultés : par exemple, les documents tapuscrits rédigés à la machine à écrire. Bien que dépourvus de cursivité, ils présentent des

[93] Lorsqu'ils sont trop fins et laissent voir l'encre du verso par transparence.
[94] Par exemple, dans le mot « filleul », les deux « ll » sont rarement tracés de la même manière que le « l » final, contrairement à ce que l'on pourrait observer dans le cas d'une écriture mécanique.
[95] On verra plus bas que la notion même de variation doit être considérée avec attention, dans la mesure où ce sont les instructions données par l'utilisateur qui vont définir quels signes sont des variantes d'un même caractère.
[96] Les écritures cursives rendent quasi impossible la distinction des lettres au sein d'un mot, tandis que la lecture d'une écriture livresque sera plus aisée.

qualités d'encre inégales, des corrections manuelles ou des ratures qui entraînent des jeux de superpositions. Ces documents mêlent parfois écriture manuscrite et tapuscrite, en particulier lorsqu'il s'agit de documents annotés. La mixité des écritures, dont les différentes catégories supposent parfois d'employer différentes stratégies d'annotation et invitent à la création de modèles de transcription « multi-modaux » (dans le sens où ils doivent pouvoir traiter plusieurs modes d'écriture), est aussi l'occasion d'un questionnement méthodologique. Cette mixité est d'autant plus fréquente dans les formulaires administratifs pré-imprimés.

2. Comprendre l'apprentissage automatique pour mieux utiliser l'ATR

L'ATR, nous l'avons dit, est une technologie qui relève de l'apprentissage automatique. Bien qu'elle permette de transcrire plus rapidement une plus grande quantité de pages, sa prise en main peut sembler une aventure de longue haleine. Outre la maîtrise des logiciels, l'ATR peut nécessiter de créer ses propres modèles de transcription et donc des données d'entraînement, mais aussi des données de tests qui facilitent et rendent plus objectif le contrôle de la qualité de la transcription. Tout ceci suppose de se familiariser avec quelques concepts et méthodes de l'apprentissage automatique, même si les logiciels facilitent autant que possible l'entraînement des algorithmes[97].

L'ATR repose sur des cycles d'apprentissage qui permettent de générer des modèles de transcription entraînés à partir de données annotées. Durant son entraînement, le modèle apprend à reproduire les données d'entraînement, dites vérité de terrain, à partir desquelles il doit élaborer des règles générales. Il est important de distinguer la production d'un jeu de données d'entraînement, de l'application d'un modèle. Un modèle de transcription peut être envisagé comme la partie d'un algorithme qui définit ce que le système est capable de reconnaître à partir d'une image (notamment les caractères disponibles). Certains logiciels, comme *Kraken*[98], permettent d'exporter cette partie de l'algorithme sous la forme d'un fichier, tandis que d'autres, comme *Transkribus*[99], ne permettent d'y accéder que par l'intermédiaire d'un logiciel et de sa base de données.

Il est crucial de faire la distinction entre deux phases de l'utilisation de l'ATR : la phase de production d'un modèle de transcription et la phase d'application du modèle. Avant toute chose, il est utile d'identifier les modèles de transcription publics associés au logiciel choisi[100]. Tester les modèles existants permet de définir s'il est nécessaire ou non d'entraîner son propre modèle de transcription, et donc, dans les cas les plus chanceux, de pouvoir passer directement à la phase d'application de la transcription sur tout le corpus. Dans de nombreux cas, entraîner un modèle de transcription sur mesure permet de gagner en précision et d'obtenir une meilleure

[97] Nous proposons ici des définitions générales qui ne remplacent pas une lecture approfondie de la documentation des outils utilisés.
[98] Benjamin Kiessling, « Kraken - an Universal Text Recognizer for the Humanities », Utrecht, CLARIAH, 2019.
[99] Philip Kahle, Sebastian Colutto, Günter Hackl, [et al.], « Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents », *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 04, 2017, p. 19-24.
[100] Pour Transkribus, il est possible de consulter une liste des modèles publics et même, de tester ces modèles sur quelques fichiers (https://www.transkribus.org/ai-text-recognition). Pour Kraken, il existe un répertoire de modèles prêts à être téléchargés (https://zenodo.org/communities/ocr_models/).

transcription. Pour entraîner un modèle, il est nécessaire de fournir au logiciel des exemples de transcription des sources, c'est-à-dire des données d'entraînement. Face à un corpus très hétérogène, il peut être utile d'entraîner plusieurs modèles de transcription, il faut donc créer autant de jeux de données d'entraînement. Ces derniers sont idéalement composés de pages représentatives des sources, sélectionnées de manière aléatoire. La préparation de ces données doit être faite avec soin, car la qualité du modèle en dépend.

La production d'un modèle de transcription suppose souvent des allers-retours entre production des données d'entraînement, entraînement du modèle, test du modèle, et reprise de la production de données d'entraînement. Inutile de chercher à atteindre l'objectif d'une transcription parfaite à 100 %, car, si l'objectif est de produire un texte sans erreur, il faudra prévoir de relire le résultat de la transcription automatique (ne serait-ce que pour vérifier que des lignes n'ont pas été omises), ou trouver des stratégies pour corriger le résultat à grande échelle. Enfin, certains problèmes sont inhérents aux sources et ne sont pas corrigibles au niveau du modèle, comme le problème de la segmentation des mots qui n'est pas toujours claire dans les documents manuscrits ou les césures en fin de ligne (voir II.C.1). Comment évaluer la qualité d'un modèle de transcription ? La précision des modèles peut être déterminée à l'aide d'un corpus de test, composé d'images annotées correspondant à la transcription que l'on attend du modèle. Idéalement, les données de test n'ont pas été utilisées pour entraîner le modèle afin d'éviter un biais de « surapprentissage » sur des données issues de l'entraînement. Lors de l'évaluation du modèle, le texte généré (la prédiction) est comparé au texte attendu (la référence) présent dans les données de test. La différence entre les deux versions est exprimée sous la forme de deux scores. (1) Le CER, ou *character error rate* pour taux d'erreur par caractère, est le plus communément utilisé[101]. (2) Un autre score est parfois utilisé en complément : le WER, pour *word error rate* ou taux d'erreur au mot[102]. Le WER ne donne pas une idée claire de la précision d'un modèle, il sert plutôt à évaluer si les erreurs sont réparties sur tous les mots du texte ou bien sur une moindre portion[103]. En général, la phase de production d'un modèle de transcription devrait viser à atteindre un modèle avec un score inférieur à 10 %, voire 5 % d'erreurs au caractère, soit une précision entre 90 % et 95 %. Pour les documents les plus lisibles, il est même possible d'atteindre un taux d'erreur en dessous de 2 %[104].

Le succès de l'application d'un modèle de transcription à l'ensemble du corpus est dépendant de deux critères : (1) le modèle est bien adapté aux sources à traiter et (2) la bonne détection en amont des lignes sur la page lors de la phase d'analyse de la mise en page (II.B.1)[105]. Si le modèle de transcription a besoin d'être amélioré, il est tout à fait possible

[101] Voir <https://harmoniseatr.hypotheses.org/glossary-atr#CERID>
[102] Voir <https://harmoniseatr.hypotheses.org/glossary-atr#WordErrorRateID>
[103] Les CER et le WER ne sont que des scores et ne remplacent pas une évaluation qualitative de la transcription permettant également d'évaluer de la lisibilité du texte en dépit des erreurs. Il existe des logiciels pour évaluer les modèles en manipulant le CER pour ignorer les erreurs portant sur les questions de casse, sur les chiffres ou encore sur les accents, voir Terriel, L., & Chagué, A. KaMI-lib (Version 0.1.3) [Computer software]. https://doi.org/10.5281/zenodo.1234.
[104] Tobias Hodel, David Schoch, Christa Schneider, [et al.], « General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example », *Journal of Open Humanities Data*, vol. 7 / 0, Ubiquity Press, juillet 2021, p. 13.
[105] En effet, une segmentation fautive peut signifier que des lignes sont en partie ou entièrement ignorées durant la phase de transcription, voire que le logiciel tente de transcrire des lignes qui n'en sont pas. Si on applique

d'utiliser une partie des données transcrites automatiquement, de les corriger précautionneusement, puis de les utiliser comme nouvelles données d'entraînement afin d'accélérer le processus. Enfin, il existe également un catalogue de jeux de données d'entraînement prêtes au réemploi, couvrant différentes langues, périodes et systèmes d'écriture : *HTR-United*[106]. Ce catalogue liste des jeux de données soigneusement préparées et décrites par d'autres utilisateurs de l'ATR et qui peuvent être réutilisés[107]. Une fois les données d'entraînement produites, et si les conditions de réutilisation des images l'autorisent, elles peuvent à leur tour alimenter la banque de données du catalogue et permettre l'entraînement de modèles génériques capables de traiter de vastes collections de documents[108] soit directement à l'aide d'un affinage du modèle[109]. En outre, procéder de la sorte permet de réduire collectivement les coûts d'entraînement, ainsi que l'impact écologique de l'utilisation de telle technologie.

3. Normalisation des pratiques de transcription

Afin d'optimiser les résultats des modèles en garantissant l'homogénéité des prédictions, l'établissement de règles de transcription est crucial dans le cas de la transcription automatique. Elles permettent (1) de fournir au modèle des exemples de transcription qui ne sont pas contradictoires et qui ne vont donc pas l'empêcher d'apprendre efficacement et (2) d'évaluer le modèle selon les mêmes règles que celles de l'entraînement. Pour établir des recommandations de transcription, il est essentiel de considérer que les données seront utilisées dans le cadre d'un apprentissage automatique. Plus on fournit de classes à apprendre (soit de caractères), plus le modèle sera long à entraîner, plus il sera lourd, et plus il nécessitera de données. Il est également crucial d'assurer l'homogénéité de la représentation des caractères : un même signe doit toujours être représenté par le même caractère Unicode pour assurer la cohérence des prédictions. Enfin, il est important de minimiser les situations où le modèle doit deviner des signes non écrits, comme dans le cas du développement des abréviations, afin d'éviter des résultats totalement aberrants. Ainsi, les transcriptions nécessaires pour la phase d'ATR ne répondent pas toujours aux règles de transcription attendues pour un projet d'édition.

---

directement la transcription sans vérifier que les lignes sont dans l'ensemble correctement détectées, on peut avoir du mal à comprendre pourquoi un modèle présentant de bons scores produit ensuite des transcriptions incorrectes.

[106] < https://htr-united.github.io/>.

[107] Alix Chagué, Thibault Clérice, Laurent Romary. HTR-United : Mutualisons la vérité de terrain ! *DHNord2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux*, MESHS, Nov 2021, Lille, France. ⟨hal-03398740⟩

[108] Ariane Pinche, Thibault Clérice, Alix Chagué, Jean-Baptiste Camps, Malamatenia Vlachou-Efstathiou, et al.. CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts: A generalized set of guidelines and models for Latin scripts from Middle Ages (8th-16th century). *Digital Humanities - DH2024*, Aug 2024, Washington DC, ⟨hal-04346939⟩.

[109] Reul, C., Tomasek, S., Langhanki, F., Springmann, U., « Open-Source Handwritten Text Recognition on Medieval Manuscripts Using Mixed Models and Document-Specific Finetuning », Uchida, S., Barney, E., Eglin, V. (eds) *Document Analysis Systems*, 2022. Lecture Notes in Computer Science, vol 13237. Springer, Cham. https://doi.org/10.1007/978-3-031-06555-2_28, Gille Levenson, M. « Towards a general open dataset and model for late medieval Castilian text recognition (HTR/OCR) », *JDMDH*, 16 octobre 2023, Documents historiques et reconnaissance automatique de texte - https://doi.org/10.46298/jdmdh.10416, Pinche, A. « Generic HTR Models for Medieval Manuscripts. The CREMMALab Project », *JDMDH*, 2023, Documents historiques et reconnaissance automatique de texte - https://doi.org/10.46298/jdmdh.10252.

Les résultats devront être révisés, le cas échéant, si l'on désire s'éloigner de la représentation fidèle du texte tel qu'il apparaît sur son support. Dans la mesure où le logiciel de transcription automatique reproduit les exemples qui lui sont donnés durant l'apprentissage, il est important de comprendre que ce sont aussi ces exemples qui vont définir les règles à suivre. Pour une transcription élaborée pour un entraînement dans laquelle tous les <r> sont tantôt transcrits <r>, tantôt <ꝛ> (r rotunda) et où les <s> sont tantôt transcrits <s> tantôt <ſ> (s long), le modèle apprendra à distinguer quatre lettres distinctes : <s>, <ſ>, <r> et <ꝛ>, là où d'autres modèles, dont les données d'entraînement proposeraient une transcription qui ne fait pas la distinction entre les allographes, pourraient n'en voir que deux (<r> et <s>). Dans le premier cas, on considère qu'une variation graphique constitue un caractère à part entière, tandis que dans le second, on considère qu'il ne s'agit que de tracés différents pour un seul et même caractère. Autre cas de figure, si l'on apprend au modèle à transcrire une abréviation comme <duḍ> (pour dudit) avec ces trois lettres en particulier, mais que dans les données de test, cette même abréviation est transcrite par <dud> ou encore sous une forme développée <dudit>, alors la forme gardant l'abréviation modèle sera nécessairement considérée comme fautive.

Il existe plusieurs approches méthodologiques pour l'établissement de règles de transcription des documents historiques avec des approches plus ou moins imitatives de la source[110], leur pleine compréhension demande d'avoir des bases en paléographie et de savoir lire parfaitement ses sources. Récemment, l'initiative CATMuS[111], pour *Consistent Approaches to Transcribing ManuScripts*, a proposé l'établissement de règles pour les textes en langue romane de la période médiévale à aujourd'hui proposant une approche graphématique, c'est-à-dire qui respecte les graphies originales du document (abréviations, graphies anciennes, ponctuation, etc.), mais qui ne retranscrit pas les variations de forme des lettres (par exemple entre le s rond et le s long)[112]. S'inscrire dans une approche compatible avec ces règles de transcription permet de rendre ses données facilement réutilisables par d'autres.

Il existe de nombreux logiciels pour appliquer ou entraîner des modèles de transcriptions. Les principaux ont été listés par le projet *Harmonise ATR Workflow*[113], ce paysage logiciel est en constante évolution. Le choix du logiciel est très important et peut dépendre des besoins spécifiques d'un projet. Les ressources financières et technologiques peuvent également se révéler déterminantes dans le choix d'une solution payante ou d'une solution gratuite. La question des formats d'export disponibles dans un logiciel et du degré de fermeture de l'accès aux données n'est pas à négliger non plus. En effet, certaines plateformes ne permettent pas aux utilisateurs de garder la main sur les modèles entraînés. Les conditions de réutilisation des numérisations des sources sont à prendre à compte. Si elles ne permettent pas une réutilisation

[110] Jean-Baptiste Camps, *La Chanson d'Otinel : édition complète du corpus manuscrit et prolégomènes à l'édition critique*, thèse de doctorat, Université Paris-Sorbonne - Paris IV, 2016, Peter Robinson et Elizabeth Solopova, « Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue », juillet 1993, 10.5281/zenodo.4050360, Dominique Stutzmann, « Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? », *Codicology and Palaeography in the Digital Age*, 2:34, 2010.
[111] <https://catmus-guidelines.github.io>.
[112] Ariane Pinche, Thibault Clérice, Alix Chagué, [et al.], « CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts », *Digital Humanities Conference*, 2024.
[113] Voir <https://harmoniseatr.hypotheses.org/literature-tools-links-and-help>.

commerciale, alors il faudra s'assurer que le logiciel choisi ne les réutilise pas à votre insu. À l'inverse, les solutions gratuites peuvent nécessiter une plus grosse prise en charge des aspects infrastructurels, avec une installation qui peut demander plus de connaissances techniques que n'en dispose l'équipe.

Pour conclure, les aspects méthodologiques de la transcription offre des perspectives passionnantes pour l'étude des textes, quoique complexes et demandant un vaste éventail de connaissances. Fort heureusement, il existe de plus en plus de ressources permettant de comprendre les technologies sous-jacentes et les bonnes pratiques qu'elles supposent. La transcription automatique suppose bien souvent de déconstruire des étapes qui nous paraissent évidentes lorsque nous transcrivons manuellement : ne pas transcrire, développer et normaliser le texte d'un même geste. En suivant des règles rigoureuses et en utilisant les outils appropriés, en prenant le temps de comprendre comment fonctionnent ces outils et les différentes étapes qui doivent être mises en place, il est finalement possible d'obtenir des transcriptions de haute qualité qui pourront contribuer à l'avancée de la recherche académique et à la préservation du patrimoine historique.

Quelques références pour aller plus loin :
– Anne Baillot et Marieke Koenig, « Automatic Text Recognition : Harmonizing ATR Workflows », [En ligne : https://harmoniseatr.hypotheses.org]. Consulté le 8 juillet 2024.
– Tobias Hodel, David Schoch, Christa Schneider, [et al.], « General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example », *Journal of Open Humanities Data*, vol. 7 / 0, Ubiquity Press, juillet 2021, p. 13.
– Ariane Pinche, Thibault Clérice, Alix Chagué, [et al.], « CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts », *Digital Humanities Conference*, 2024.

B. Analyse de mise en page et automatisation de la structuration du corpus

Avant de pouvoir transcrire automatiquement un corpus, il est nécessaire de reconnaître les lignes de texte et l'espace occupé par les caractères, notamment les hampes, hastes et signes diacritiques afin de localiser précisément le texte sur l'image et, par conséquent, les signes à transcrire. Aujourd'hui, en vision par ordinateur, certains logiciels de reconnaissance automatique d'écriture, comme Kraken ou les solutions proposées par Transkribus, permettent également d'identifier les différents éléments qui composent la page d'un document, ce qui correspond à la segmentation de l'image en zones. Ces informations, comme nous le verrons plus précisément dans la suite de cet exposé, peuvent par la suite être réutilisées afin d'automatiser pour partie la structuration du texte en TEI.

1. Segmentation du texte : définition et état de l'art

La segmentation de texte est non seulement la première étape, mais aussi une tâche fondamentale dans l'acquisition automatique de contenus textuels. Elle permet de localiser les

lignes de texte et d'identifier la structure d'un document. Dans le premier cas, cette étape est obligatoire pour bon nombre d'outils de transcription automatique qui reposent sur le résultat de la segmentation pour extraire du texte[114], tandis que le second reste facultatif, mais permet de sauvegarder la mise en page d'un document et fournit donc les éléments nécessaires pour reconstruire les paragraphes, identifier des colonnes de texte, etc. La segmentation est parfois appelée zonage, analyse de documents, analyse optique de mise en page, ou analyse de la mise en page. Cette étape est subdivisée en deux processus distincts sur lesquels nous reviendrons : la segmentation des lignes de texte, et la segmentation des régions de texte.

La segmentation des lignes de texte peut être réalisée de deux manières différentes : (1) les lignes de texte sont annotées manuellement ou (2) elles sont annotées automatiquement à l'aide d'outils d'apprentissage machine où les modèles ont été entraînés à partir d'annotations manuelles (ou vérité terrain). Dans le cas de l'utilisation et de l'entraînement de modèles pour réaliser ces tâches, il est possible de contrôler la qualité de réalisation de ces tâches grâce à des évaluations quantitatives[115]. Plusieurs méthodes et outils se distinguent dans le domaine de la segmentation. Les outils les plus répandus sont basés sur l'apprentissage machine (*machine learning*) et l'apprentissage profond (*deep learning*). Les techniques multimodales, telles que *layoutLMv3*[116], combinent les caractéristiques (ou *features*) visuelles et textuelles des images pour détecter les régions de texte. Cependant, l'utilisation exclusive des caractéristiques visuelles reste la plus courante. Cette méthode est employée par le moteur Kraken pour détecter les lignes et les régions de texte, ainsi que par le modèle de détection d'objet YOLO[117] pour la détection des seules régions de texte. Il est important de noter que ces outils combinent souvent deux tâches en une seule : la détection et la classification, ou l'attribution d'une étiquette, des lignes et/ou régions. La segmentation des lignes de texte et la segmentation des régions ne sont pas des processus mutuellement exclusifs et n'ont pas besoin d'être effectuées en même temps. L'utilisation combinée des deux précédents outils est aujourd'hui répandue au sein de la communauté des humanités numériques[118].

Définition des lignes et des zones

Lorsqu'une ligne de texte est détectée par la plupart des outils de segmentation ou tracée manuellement dans une image, deux éléments sont produits automatiquement en même temps :

---

[114] Benjamin Kiessling, « Kraken - a Universal Text Recognizer for the Humanities », *DataverseNL*, 2019.

[115] Deux métriques sont généralement utilisées : l'IoU, ou intersection over union qui mesure le chevauchement entre la prédiction et la vérité terrain ; et le mAP, ou *mean average precision*, qui évalue la précision moyenne des prédictions sur différentes classes.

[116] Yupan Huang, Tengchao Lv, Lei Cui, [et al.], « LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking », arXiv, 2022.

[117] Glenn Jocher, Ayush Chaurasia et Jing Qiu, « Ultralytics YOLO », 2023, < https://github.com/ultralytics/ultralytics>.

[118] Thibault Clérice, « You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine », *JDMDH*, décembre 2023, Najem-Meyer Sven et Romanello Matteo, « Page Layout Analysis of Text-heavy Historical Documents: a Comparison of Textual and Visual Approaches », *arXiv*, 2022, Thibault Clérice, Juliette Janes, Hugo Scheithauer, [et al.], « Layout Analysis Dataset with SegmOnto », *DH2024 - Annual conference of the Alliance of Digital Humanities Organizations*, Washington DC, 2024.

– Une *baseline* : ligne virtuelle qui passe par au moins deux points avec des coordonnées de type (x, y) sur laquelle le texte est accroché. Il est possible d'avoir à la place une *topline*. Choisir l'une ou l'autre dépend du système graphique. Ainsi, pour l'alphabet arabe, par exemple, le choix se portera sur une *baseline*, tandis que pour l'alphabet hébraïque, on préférera une *topline*.
– Un masque : polygone défini par au moins trois points, chacun défini par un jeu de coordonnées qui délimite la zone de pixels contenant le texte de la ligne. Le masque est généralement calculé automatiquement à partir de la *baseline* ou de la *topline*.

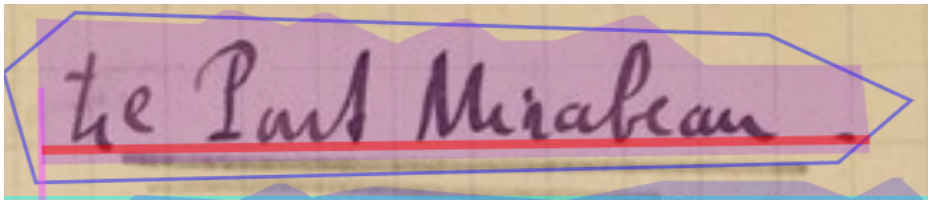


Figure 17: Une ligne de texte annotée avec une baseline et un masque. Guillaume Apollinaire, Le Pont Mirabeau, Bibliothèque nationale de France (BnF). <https://gallica.bnf.fr/ark:/12148/btv1b525056707/f33/>

Ces deux éléments constituent la ligne de texte qui sera transcrite[119]. Enfin, le dernier élément pour compléter la ligne est la transcription du texte même. Dans l'illustration précédente, la ligne de texte serait donc :

- la *baseline*,
- le masque,
- la transcription : « Le Pont Mirabeau ».

Il est généralement possible d'attribuer une étiquette aux *baselines* et *toplines* de sorte à pouvoir les traiter ultérieurement (section II.B.2). Cela reste facultatif et n'impacte pas la transcription.

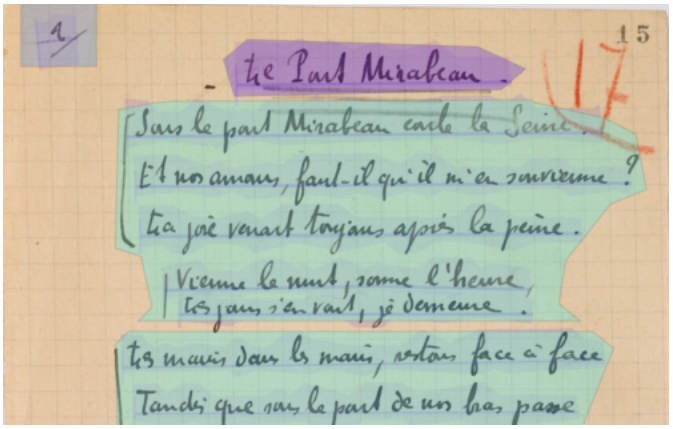


Figure 18: Plusieurs régions de texte annotées. Guillaume Apollinaire, Le Pont Mirabeau, BnF, < https://gallica.bnf.fr/ark:/12148/btv1b525056707/f33/>

[119] Certains outils, comme *Calfa Vision*, proposent des systèmes sans segmentation (DAN), Chahan Vidal-Gorène, Boris Dupin, Aliénor Decours-Perez, [et al.], « A Modular and Automated Annotation Platform for Handwritings: Evaluation on Under-Resourced Languages », *Document Analysis and Recognition – ICDAR 2021*, éds. Josep Lladós, Daniel Lopresti et Seiichi Uchida, Cham, Springer International Publishing, 2021, p. 507-522.

Une région ou zone de texte rassemble des lignes qui constituent un ensemble logique au niveau de la structure du document, par exemple un paragraphe, un titre, ou des éléments plus complexes comme des tableaux, des colonnes de texte, des images, etc. Le terme *bounding box* est employé pour décrire un polygone défini par au moins trois couples de coordonnées (x, y) qui englobent une ou plusieurs lignes de texte. Les outils de segmentation permettent d'attribuer une étiquette correspondant à la fonction de la *bounding box* dans la structure du document (section II.B.2). Dans l'exemple (fig. 18), des *bounding boxes* ont été tracées autour du numéro de page (en bleu), du titre (en violet) et des strophes (en vert). Ainsi, toutes les lignes de texte présentes dans ces régions de texte héritent elles-mêmes du label des zones auxquelles elles appartiennent.

La détection des régions de texte n'a pas d'incidence technique sur la transcription automatique de texte. Autrement dit, les outils d'ATR n'ont besoin que de la détection des lignes pour transcrire. En revanche, la détection des régions est cruciale du point de vue de la structuration du texte. Sans information spatiale et logique de la mise en page, le contenu textuel n'est plus qu'un texte ininterrompu sans hiérarchie, d'autant plus lorsqu'il s'agit de contenus hautement structurés tels les dictionnaires, les textes glosés, etc. Dans le cas d'un texte à plusieurs colonnes par exemple, sans information de mise en page, les différentes colonnes seraient confondues entre elles et l'ordre de lecture des lignes perturbé. Enfin, un document structuré à partir de la détection des régions de texte rend possible certaines tâches en aval telles que la réduction de la portée des requêtes ou la mise en place de recherches à facettes, une navigation plus aisée à l'intérieur d'un document transcrit, l'exclusion de certaines sections (par exemple les titres courants), etc.

Pour classifier les lignes et les régions, nous recommandons l'utilisation de SegmOnto[120], conçu pour décrire le contenu des sources textuelles. Ce projet propose un vocabulaire contrôlé pour le nommage du contenu des sources textuelles, homogénéisant les données de segmentation des outils d'ATR, tant pour les données d'entraînement que pour les données prédites. Ce dernier s'appuie sur les termes de la codicologie pour proposer un vocabulaire simple et générique afin de décrire les sources du Moyen Âge à aujourd'hui avec pour limite que sa conception s'appuie essentiellement sur des documents occidentaux. Il permet de classer les zones présentes sur une page et les lignes contenues dans ces zones selon une syntaxe simple (en anglais) basée sur trois paramètres :

1. Le type de zone/ligne, qui est obligatoire et dont la valeur est contrôlée ;
2. Le sous-type de zone/ligne, qui est facultatif et dont la valeur est recommandée, mais ouverte ;
3. Un suffixe, qui est aussi facultatif.

Les composants suivent une syntaxe simple : zone:subtype#suffixe, donnant par exemple, MainZone:column#1. Ce vocabulaire a permis la création de jeux de données généralistes

[120] Simon Gabay, Ariane Pinche, Kelly Christensen, [et al.], « SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles », décembre 2023.

pour entraîner des modèles d'analyse de la mise en page. Ainsi, le projet *LADaS*[121] a personnalisé *SegmOnto* pour une annotation plus fine des images, compatible avec la TEI. Ainsi, on peut avoir des sous-types comme paragraphes, strophes, ou répliques de théâtre[122]. Toutefois, attention, plus le vocabulaire est précis, plus le nombre de classes du modèle augmente, nécessitant un jeu de données important comprenant tous les types à annoter. Cette initiative, en s'appuyant sur un vocabulaire contrôlé documenté, a permis non seulement de proposer une modélisation fine de l'annotation pour la mise en page des données, mais aussi la création d'un jeu de données partageables et *open source* permettant d'entraîner de nouveaux modèles de segmentation[123]. Ainsi, SegmOnto permet de garantir la création de données cohérentes grâce à l'utilisation de termes standardisés pour décrire les lignes et régions. Le partage des données est facilité, et les jeux de données peuvent être mis en commun. Autrement dit, un vocabulaire contrôlé contribue à la création de données respectant les principes *FAIR*[124].

Exporter les données

Les résultats de la segmentation des lignes de texte et des régions de texte peuvent être sauvegardés à l'aide de deux standards XML dédiés : ALTO XML et PAGE XML[125]. Ils sont aujourd'hui les formats les plus répandus et sont utilisés pour stocker les détails de la mise en page des documents textuels segmentés en association avec le texte acquis lors de la transcription automatique ou non. Un fichier XML ALTO ou PAGE se situe au niveau de l'image, et non du document entier. La transformation des résultats de la segmentation dans ces formats est effectuée automatiquement par les outils de segmentation. Il est également possible de transformer ces exports en XML TEI. Ce standard dédié à l'édition de texte prévoit la sauvegarde de ces informations[126]. Ces trois formats peuvent être utilisés comme format pivot pour structurer semi-automatiquement du texte acquis depuis une image.

---

[121] Thibault Clérice, Juliette Janès, Hugo Scheithauer, Sarah Bénière, Laurent Romary, et al., « Layout Analysis Dataset with SegmOnto », *DH2024 - Annual conference of the Alliance of Digital Humanities Organizations*, Aug 2024, Washington DC, ⟨hal-04513725⟩

[122] Voir les guidelines du projet : <https://github.com/DEFI-COLaF/LADaS/blob/main/AnnotationGuide.md>

[123] Clérice, T., Janès, J., Scheithauer, H., Bénière, S., Langlais, P., Romary, L., Sagot, B., & Bougrelle, R. « Layout Analysis Dataset with SegmOnto (LADaS) [Data set] », https://github.com/DEFI-COLaF/LADaS

[124] Voir https://www.go-fair.org/fair-principles/. Consulté le 22/07/2024.

[125] Voir les spécifications ALTO XML, https://www.loc.gov/standards/alto/ et PAGE XML, https://github.com/PRImA-Research-Lab/PAGE-XML. Consulté le 22/07/2024.

[126] Hugo Scheithauer, Alix Chagué et Laurent Romary, « Which TEI representation for the output of automatic transcriptions and their metadata? An illustrated proposition », 2022.

2. Automatisation de la structuration à partir de la segmentation

Figure 19: Exemple de segmentation réalisée avec eScriptorium et le vocabulaire contrôlé SegmOnto.

Ainsi, à partir d'une segmentation automatique des différentes zones d'une page, il est possible d'automatiser partiellement la production d'un fichier XML TEI. Cette automatisation permet d'organiser les données textuelles dans un format standard. Dans le cas de la TEI, la balise permet de lister, à l'aide de l'élément `<surface>`, les différentes zones d'une image en les associant à leurs coordonnées. Une telle automatisation s'appuie sur des scripts — en XSLT, en Python, ou en combinant les deux — pour convertir les fichiers XML ALTO ou PAGE en XML TEI[127].

La production du fichier TEI peut varier en finesse selon la complexité des scripts et le modèle de segmentation utilisé en amont. La méthode la plus simple consiste à récupérer le texte et à l'organiser en grandes unités, correspondant aux différentes zones du document (chaque page correspondant à un fichier PAGE ou ALTO)[128]. Cependant, une telle démarche demande un travail approfondi sur le modèle de segmentation et de modélisation des données pour structurer les informations textuelles dans le fichier TEI au sein de la balise `<body>` de la manière la plus pertinente possible, initiant une phase de pré-éditorialisation du texte. Cela nécessite de mettre en place un *mapping* entre les zones identifiées lors de la phase de segmentation et la structuration du fichier TEI attendu, incluant des zones de corps de texte, de titres, de notes, etc.

[127] Voir, par exemple, Alix Chagué et Hugo Scheithauer, « Page2tei, an XSL Transformation to transform PAGE XML into TEI XML », 2021, <https://github.com/TEI4HTR/page2tei>. Floriane Chiffoleau, Alix Chagué et Hugo Scheithauer, « Alto2tei, an XSL Transformation to transform XML ALTO into TEI XML », 2022, <https://github.com/TEI4HTR/alto2tei>, pour une liste plus exhaustive des scripts existants voir <https://github.com/cneud/ocr-conversion>.

[128] Voir le projet LEPIDEMO : LECTAUREP PIPELINE DEMONSTRATOR qui permet de créer des fichiers TEI à partir de transcriptions générées avec Kraken et *eScriptorium*. Voir également Alix Chagué et Hugo Scheithauer, « LEPIDEMO, a Pipeline Demonstrator for LECTAUREP to go from eScriptorium to TEI-Publisher », 2021, <https://github.com/lectaurep/lepidemo>.

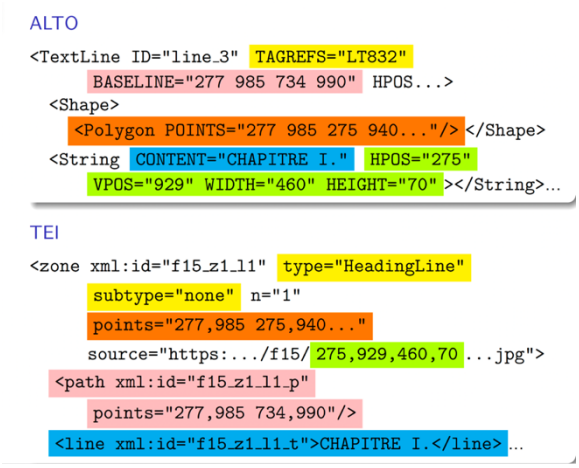


Figure 21: Exemple de conversion possible depuis XML ALTO vers XML TEI, A. Pinche, K. Christensen, et S. Gabay, « Between automatic and manual encoding, » TEI Conference, 2022.

Pour rendre ce *mapping* possible, un système de description et de nommage des différentes zones de la page est nécessaire pour faciliter leur identification et leur transformation. Ce système de nommage peut être *ad hoc*, nécessitant des scripts de transformation personnalisés, ou s'appuyer sur le vocabulaire d'un projet existant, comme SegmOnto[129], si celui-ci répond aux objectifs de recherche du projet en cours. Cette approche présente plusieurs avantages : (1) gagner du temps dans l'élaboration d'un vocabulaire adapté, (2) réutiliser les scripts du projet modèle, et (3) créer des données compatibles avec le projet modèle, permettant ainsi de mutualiser les données afin d'améliorer les modèles de segmentation.

SegmOnto	TEI
NumberingZone	<fw type="NumberingZone">
QuireMarksZone	<fw type="QuireMarksZone">
RunningTitleZone	<fw type="RunningTitleZone">
MarginTextZone	<note type="MarginTextZone">
MainZone	<ab type="MainZone">
DefaultLine	<lb>
HeadingLine	<hi rend="HeadingLine">
DropCapitalLine	<hi rend="DropCapitalLine">

Figure 20: Exemple de mapping des zones SegmOnto avec un encodage TEI, extrait du projet Gallic(opor)a.

Enfin, la structuration du fichier TEI final peut être complétée par l'ajout de métadonnées dans le <teiheader>, récupérées à partir des informations disponibles, soit via le manifeste IIIF des images numérisées, soit à partir des catalogues en ligne des bibliothèques comme la BnF grâce à l'identifiant ARK du document numérisé.

[129] Simon Gabay, Ariane Pinche, Kelly Christensen, [et al.], « SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles », décembre 2023.

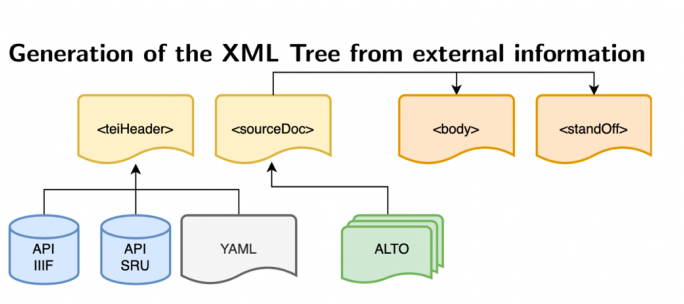


Figure 22: Pipeline du projet Gallic(orpor)a pour la génération d'un fichier TEI à partir des informations des XML ALTO et des métadonnées issues du manifeste IIIF et du catalogue général de la BnF.

Quelques références pour aller plus loin :
– Thibault Clérice, « You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine », JDMDH, décembre 2023.
– Simon Gabay, Ariane Pinche, Kelly Christensen, [et al.], « SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles », décembre 2023, <https://hal.science/hal-04343404>.
– Najem-Meyer Sven et Romanello Matteo, « Page Layout Analysis of Text-heavy Historical Documents: a Comparison of Textual and Visual Approaches », *arXiv,* 2022.

C. Enrichissement du texte : annotations linguistiques et entités nommées

Les outils d'acquisition automatique de texte permettent aujourd'hui de constituer à moindre coût de vastes corpus, devenus impossibles à parcourir ou à étudier manuellement. L'exploitation du texte brut ou du plein texte, bien que potentiellement utile, ne permet pas de tirer pleinement parti du corpus. Il peut se révéler utile d'enrichir le texte d'informations linguistiques ou contextuelles. Par exemple, lorsqu'on s'intéresse au lexique, il est souvent nécessaire de retrouver toutes les formes d'un verbe ou d'un mot en une seule requête, quelle que soit sa flexion, surtout face à un corpus présentant une forte variation graphique, comme c'est le cas pour les langues vernaculaires médiévales ou les langues à déclinaison comme le latin. L'ajout d'informations linguistiques, telles que la nature ou la fonction des mots, permet également de réaliser des études quantitatives sur la syntaxe ou de faciliter des analyses stylométriques, où l'on peut, par exemple, se concentrer sur les mots-outils pour identifier la signature stylistique d'un auteur. Par ailleurs, l'annotation des entités nommées est aujourd'hui très prisée pour constituer des bases de données à partir de ces corpus, permettant ainsi de cartographier les lieux cités ou d'analyser les réseaux de relations autour des personnes mentionnées. Toutefois, avant de pouvoir automatiser ces tâches, il est nécessaire d'établir en amont un texte où les mots sont clairement délimités dans le flux textuel, c'est-à-dire où les espaces typographiques entre les mots sont normalisées (ce qui n'est pas toujours le cas dans les sources manuscrites) et où les césures en fin de ligne sont résolues en réassociant les parties du mot séparées par un saut de ligne.

1.Segmentation linguistique et identification de la césure à la ligne

Segmentation linguistique

Nous entendons, ici, par segmentation linguistique, la segmentation des lignes du texte en « mots et signes de ponctuation », appelés « token ». Le cas d'usage est celui de la sortie d'ATR qui mène soit à des erreurs de segmentation de la part du copiste soit de l'outil de transcription[130]. Ainsi, dans l'exemple ci-dessous (figure 23), on note une erreur de segmentation linguistique double : l'omission d'un espace entre l'article et le nom de ville (*de barna*), et l'omission d'un espace présent au sein du nom propre (*bar na*).

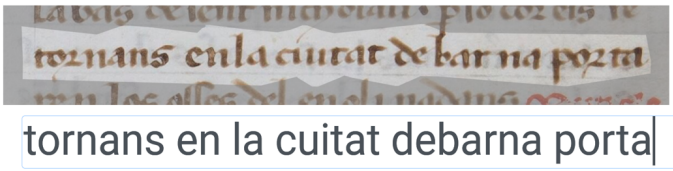


Figure 23: Transcription automatisée sur l'interface eScriptorium (modèle CATMuS medieval 1.0.0) ; Legenda aurea, BnF Espagnol 44 (catalan en réalité), fol. 11r.

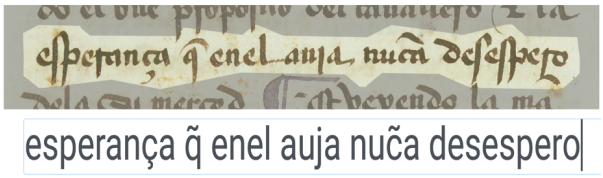


Figure 24: Exemple illustrant la méthode de segmentation de Boudams

À cela s'ajoute parfois, pour favoriser une automatisation optimale de l'annotation, la nécessité de mettre en conformité la segmentation des mots avec les normes du dictionnaire de référence. Dans l'exemple ci-dessus (figure 24), on observe une agglutination de la préposition et de l'article « enel » (en + el) qu'il faudra normaliser en désagglutinant les deux termes pour faciliter l'annotation.

Pour ce qui est des méthodes de segmentation, il existe différents outils. Le plus développé est celui de Thibault Clérice, qui propose un outil destiné originellement à des sources en *scripta continua* pour lesquelles la segmentation est fondamentale[131]. L'outil, nommé *Boudams*, fondé sur l'apprentissage supervisé, propose une classification de chacun des caractères de la ligne selon deux classes : intérieur de mot (x) ou fin de mot (S).

	Sample
Input String	`Ladamehaitees'enparti`
Mask String	`xSxxxSxxxxxSxxxSxxxxS`
Output String	`La dame haitee s'en parti`

Table 1: Input, mask and human-readable output generated by the model. x are WC and S are WB

Figure 25: Exemple de segmentation linguistique médiévale (interface eScriptorium): ». Libro del caballero Zifar, BnF Espagnol 36, fol. 3 v.

[130] Les espaces représentent une grande proportion des caractères mal identifiés (par ajout ou omission) par les modèles d'ATR: Ariane Pinche, « Generic HTR Models for Medieval Manuscripts. The CREMMALab Project », *JDMDH*, octobre 2023.

[131] Thibault Clérice, « Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin », *JDMDH*, avril 2020.

Les résultats attendus atteignent les 98 %[132], ce qui est suffisant pour de la lecture distante[133]. Pour une visée éditoriale, le texte devra être corrigé plus attentivement et manuellement, mais le temps gagné peut-être important sur des corpus dont la segmentation linguistique diffère fortement des normes éditoriales actuelles.

Identification de la césure à la ligne

Outre la segmentation linguistique, afin de rétablir un continuum textuel, il est nécessaire d'identifier les césures de fin de ligne. La tâche est alors simplifiée si la césure est notée dans le document original, d'autant plus si les transcriptions utilisent un signe unique normalisé pour la signaler (voir I.A.2). Toutefois, dans les documents historiques, dans de nombreux cas, la césure n'est signalée par aucun signe.

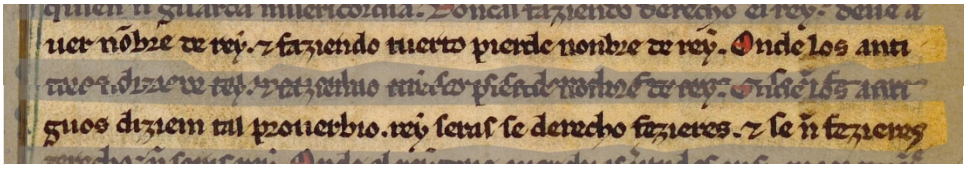


Figure 26: Exemple de césure à la ligne à identifier : « […]Onde los anti/guos dixième tal prouerbio […] ». Fuero Juzgo, BnF Espagnol 256, fol. 6r. On remarque un possible signe de césure à la fin de la première ligne, très subtil.

Leur identification est chronophage, mais peut être raisonnablement automatisée. Plusieurs méthodes peuvent être envisagées :

1. Une méthode à base de règles, qui viendrait comparer les caractères en début et fin de ligne à un dictionnaire de formes existantes, ce qui suppose de posséder un tel dictionnaire, et de pouvoir gérer efficacement les cas ambigus ; pour les langues pré-orthographiques, la grande variabilité graphique rend la constitution d'un tel dictionnaire très difficile, voire impossible ;
2. L'utilisation d'un outil fondé sur l'apprentissage machine qui puisse gérer les cas ambigus et qui se fonde sur un corpus d'exemples complets et non pas sur un dictionnaire de formes.

En réalité, l'identification de la césure à la ligne peut être assimilée à une sous-tâche de la segmentation linguistique puisqu'il s'agit de déterminer de façon ponctuelle – une fois par ligne – si un caractère en particulier (le caractère qui joint la fin d'une phrase et le début d'une autre, soit le dernier caractère de chaque ligne) correspond à un milieu ou à une fin de mot. En d'autres termes, en fusionnant chaque ligne deux à deux et en identifiant le caractère qui fait le lien (arbitrairement, le dernier caractère de la première ligne), on pourra identifier si la fin de ligne correspond à une fin de mot, ou à un intérieur de mot.

[132] Selon l'auteur, et confirmé dans M. Gille Levenson sur corpus castillan, voir Gille Levenson, « Towards a general open dataset and model for late medieval Castilian text recognition (HTR/OCR) », *JDMDH*, octobre 2023.
[133] Maciej Eder, « Mind your corpus: systematic errors in authorship attribution », *Literary and Linguistic Computing*, vol. 28 / 4, décembre 2013, p. 603-614.

Enfin, il est possible de traiter et de représenter ces informations au format XML-TEI: l'élément utilisé sera <lb> (pour marquer le début d'une ligne: *line beginning*) associé à l'attribut *@break* pour indiquer la présence ou non d'une césure à l'aide des valeurs *yes* et *no* (figure 27).

```
<text>
    <body>
        <div type="livre">
            <div type="partie">
                <div type="chapitre" n="1.2.2">
                    <ab>
                        <lb break="yes" xml:id="elem_eSc_line_5505eb5d"/>uł in irascibili uł in ꝯcupiscibili. accipiēdo g
                        <lb break="no" xml:id="elem_eSc_line_e1b0418b"/>útutē moralē large ꝓut ipā prudētia dr q̄dā
                        <lb break="yes" xml:id="elem_eSc_line_8b795cfa"/>útus moral̄ dicē possum ꝯ fm has uíii. potē
                        <lb break="yes" xml:id="elem_eSc_line_964cf1db"/>cias aīe inq⸗ ht ēē uirt. sūpte sūt.4. útu
                        <lb break="no" xml:id="elem_eSc_line_77537134"/>tes cardinales .s. prudētia. iustīa. fortitudo
                        <lb break="yes" xml:id="elem_eSc_line_3d414633"/>ꝛ tempantia. Na⸗ prudētia ē in intllu. iusti
                        <lb break="no" xml:id="elem_eSc_line_aab973cc"/>cia i uolūtate. fortitudo in irascibili. tēpan
                        <lb break="no" xml:id="elem_eSc_line_80698dac"/>tia i ꝯcupiscibili</ab>
                </div>
            </div>
        </div>
    </body>
</text>
```

Figure 27: Exemple d'identification automatisée et d'annotation en TEI de la césure à la ligne sur le De Regimine Principum de Gilles de Rome (transcription automatisée du manuscrit BNF Lat 6477, norme graphématique).

2. Annotations linguistiques

Par annotation linguistique, nous entendrons principalement l'annotation lexicale et grammaticale. Les traitements de syntaxe ne seront pas abordés dans ce chapitre. La tâche d'annotation lexico-grammaticale consiste en l'enrichissement d'un corpus textuel d'informations de type lexical avec l'ajout d'un lemme, ainsi que d'information sur la nature (POS pour *Part Of Speech*) et la morphologie (nombre, genre, temps, personnes, etc.). Ces annotations peuvent faciliter la consultation et le requêtage des textes, mais aussi l'édition des certains textes en permettant une normalisation avancée des graphies du texte.

Tokens et annotations linguistiques

La division d'un corpus en *token* est une tâche complexe. Par exemple, dans l'expression « pomme de terre », combien peut-on compter de mots/tokens ? En général, on tendra à le considérer comme un seul mot, mais cette expression pourrait aussi relever de ce que l'on nomme les expressions multimots (*Multiword Expressions* ou MWE en anglais) dont l'identification est encore difficile aujourd'hui. De la même manière, pour les verbes au passé composé comme « j'ai mangé », combien avons-nous de mots ? On aura tendance actuellement à distinguer deux tokens : l'auxiliaire et le participe passé, qui pourront éventuellement être reliés lors d'une phase ultérieure d'analyse de dépendances (analyse syntaxique du discours)[134]. Pour des tâches de traitement computationnel et d'analyses linguistiques simples, nous pouvons considérer que le mot/token se distingue par un séparateur de type espace typographique ou ponctuation.

[134] Du point de vue technique, il peut être intéressant d'opérer une tokenisation, à l'aide de règles XSLT ou de scripts Python, des sources pré-structurée en XML-TEI du moment où cette structuration n'est pas encore trop complexe –, afin de pouvoir croiser les informations structurelles et les informations linguistiques.

En ce qui concerne les langues à forte variation graphique, dont l'orthographe n'est pas fixée (par exemple, les langues romanes d'époque médiévale et pré-moderne), l'annotation lexico-grammaticale constitue une phase de normalisation du corpus. Ainsi, dans l'extrait suivant : « *Silènes estoyent iadis petites boites telles que voyons de present es bouticqs des apothecaires* […]. » (François Rabelais, *Gargantua*, Prologue, édition de F. Juste, Lyon, 1534). Les termes suivants : *estoyent, iadis*, *bouticqs*, *apothecaires*, *pinctes* montrent une grande variabilité graphique. Pour pouvoir comptabiliser ces formes, il est primordial de les ramener à une forme standard : leur entrée dans le dictionnaire, c'est-à-dire le *lemme*, et pour les analyser de leur adjoindre une annotation linguistique en POS et morphologie[135]. La consultation et le requêtage de grandes bases de données linguistiques de langues à forte variation graphique sont alors grandement facilités. Grâce au lemme, nul besoin de prendre en compte dans la requête toutes les variantes graphiques pour récupérer toutes les occurrences. Produire des requêtes pour identifier des usages linguistiques peut se révéler extrêmement complexe si l'on ne s'appuie que sur les formes d'un corpus, car il faut identifier toutes les variations possibles d'un même mot et toutes les combinaisons possibles de ces variations. L'utilisation d'outils et de patrons (comme les expressions régulières) peut permettre dans certains cas de contourner cette difficulté, mais ne l'annule pas complètement. Dans le cas d'un corpus lemmatisé et annoté en parties du discours et en morphologie, cet obstacle disparaît, il suffit dès lors de produire des requêtes sur les analyses (un lemme, une classe grammaticale, etc.) sans s'intéresser à leur réalisation graphique[136].

Jeux d'étiquettes et référentiels de lemmes

La linguistique de corpus s'intéresse à l'étude statistique des corpus textuels massifs et a vu la création d'un grand nombre de jeux d'étiquettes qui ont permis l'annotation linguistique de vastes corpus dont Cattex, Eagles, UD[137]. Il est important ici de différencier le jeu d'étiquettes et le format de données. La dernière expression désigne la présentation globale des données linguistiques, quand la première désigne le code utilisé pour les informations linguistiques (classe grammaticale, morphologie). Le format de données le plus utilisé est le format CONLL-U qui présente les données de la forme suivante : Index dans la phrase, forme, lemme, parties du discours, morphologie, analyse syntaxique. Certains jeux d'étiquettes ne suivent pas ce format et proposent des données tabulaires comprenant la forme, le lemme, les parties du discours et la morphologie. Nous présentons ici trois jeux de données utilisés aujourd'hui.

---

[135] Notons qu'aujourd'hui un certain nombre d'outils de traitement du langage naturel ne passent pas par une phase intermédiaire ou préalable d'annotation lexico-grammaticale, mais par ce que l'on appelle des plongements de mots ou de phrase, c'est à dire des représentations informatisées du sémantisme d'un mot ou d'une phrase. C'est le cas des outils à base de *transformers* par exemple.

[136] Un des langages de requêtes linguistiques les plus utilisés aujourd'hui est CQL (*Corpus Query Language*). Il permet de produire des requêtes complexes sur les formes, les parties du discours, les lemmes ou la morphologie, sur un ou plusieurs mots suivis : voi Oliver, Christ, « A modular and flexible architecture for an integrated corpus query system »n *COMPLEX'94*, 1994, vol. 15. Par exemple, si l'on considère le jeu d'étiquettes de parties du discours de Universal Dependencies (voir plus bas), la requête [pos!="DET"][pos="NOUN"] permet de faciliter l'identification des syntagmes nominaux sans déterminants comme on en voit dans le texte de Rabelais.

[137] Nous pouvons aussi citer d'autres jeux d'étiquettes historiques et moins répandus, tel que Multext.

EAGLES

Eagles[138] est un standard de traitement naturel du langage créé au début des années 1990. Il a la particularité de proposer un jeu d'étiquettes qui fusionne les parties du discours et la morphologie. Sa seconde particularité est de désigner chaque type d'analyse par sa position. Ainsi, dans l'étiquette NCFP000, chacune des lettres désigne successivement la partie du discours, puis la personne, le nombre ; les trois chiffres 0 finaux servent de valeurs de remplissage. Voici un exemple d'annotation de l'extrait de Rabelais proposé plus haut :

Index	**Forme**	**Lemme**	**POS+MORPH**
1	Silènes	silène	NCFP000
2	estoyent	être	VMII3P0
3	iadis	jadis	RG
4	petites	petit	AQ0FP00
5	boites	boîte	NCFP000
6	telles	tel	AQ0FP00
7	que	que	PR00000
8	voyons	voir	VMIP1P0
9	de	de	SP
10	present	présent	NCMS000
11	es	en+le	SP+DA0CS0
12	bouticqs	boutique	NCFP000
13	des	de+les	SP+DA0CP0
14	apothecaires	apothicaire	NCMP000

CATTEX

CATTEX2009 est un projet français d'annotation morphologique[139], utilisé dans divers projets en linguistique, dont le corpus BFM (Base de Français Médiéval)[140]. Le principe de l'annotation est de proposer des couples d'analyse type=valeur, dont les champs sont séparés par le signe < | >. Les labels des étiquettes sont en français. Les analyses de parties du discours sont séparées par un point <.> en cas d'agglutination des étiquettes.

Token	**Lemma**	**PoS**	**Morph**
Silènes	silène	NOMcom	MODE=imp|NOMB.=s|GENRE=m
estoyent	être	VERppa	NOMB.=p
iadis	iadis	ADVgen	MORPH=empty
petites	petit	ADJqua	NOMB.=p|GENRE=f
boites	boite	NOMcom	NOMB.=p|GENRE=f
telles	tel	ADJind	NOMB.=p|GENRE=f
que	que	CONsub	MORPH=empty
voyons	voir	VERcjg	MODE=ind|TEMPS=pst|PERS.=1|NOMB.=p
de	de	PRE	MORPH=empty

[138] Geoffrey Leech et Andrew Wilson, « Standards for Tagsets », Hans Van Halteren, (éd.). *Syntactic Wordclass Tagging*, Dordrecht, Springer Netherlands, 1999, p. 55-80.
[139] Céline Guillot, Sophie Prévost et Alexei Lavrentiev, « Principes d'annotation Cattex09 », <http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_principes_2.0.pdf>.
[140] Céline Guillot, Serge Heiden et Alexei Lavrentiev, « Base de français médiéval : une base de référence de sources médiévales ouverte et libre au service de la communauté scientifique », *Diachroniques. Revue de Linguistique française diachronique*, 2018, p. 168.

present	présent	NOMcom	NOMB.=p
es	en+le	VERcjg	MODE=ind|TEMPS=pst|PERS.=2|NOMB.=s
bouticqs	boutique	VERppe	NOMB.=s
des	de_le	PRE.DETdef	NOMB.=p
apothecaires	apothicaire	NOMcom	NOMB.=p|GENRE=f

Universal Dependencies

Universal Dependencies (UD) est un ensemble de jeux d'étiquettes orienté vers l'annotation syntaxique (*treebanks*), mais qui permet également l'annotation des lemmes, des PoS et de la morphologie. Encore relativement récent, il est aujourd'hui adopté par un grand nombre de projets d'annotation, qui en déclinent les principes selon leurs usages.

Index	**Forme**	**Lemme**	**PoS**	**Morph**
1	Silènes	silène	NOUN	Gender=Fem|Number=Plur
2	estoyent	être	VERB	Mood=Ind|Number=Plur|Person=3|Tense=Imp|VerbForm=Fin
3	iadis	jadis	ADV	_
4	petites	petit	ADJ	Gender=Fem|Number=Plur
5	boites	boîte	NOUN	Gender=Fem|Number=Plur
6	telles	tel	ADJ	Gender=Fem|Number=Plur
7	que	que	SCONJ	_
8	voyons	voir	VERB	Mood=Imp|Number=Plur|Person=1|Tense=Pres|VerbForm=Fin
9	de	de	ADP	_
10	present	présent	NOUN	Gender=Masc|Number=Sing
11	es	en+le	PREP+DET	Definite=Ind|Number=Plur|PronType=Art
12	bouticqs	boutique	NOUN	Gender=Fem|Number=Plur
13-14	des	_	_	_[141]
13	de	de	ADP	_
14	les	le	DET	Definite=Def|Number=Plur|PronType=Art
15	apothecaires	apothicaire	NOUN	Gender=Masc|Number=Plur

La liste présentée ci-dessus montre les diverses façons et l'équivocité de l'annotation linguistique. Cette diversité peut représenter un risque pour l'interopérabilité des données et la constitution de corpus larges par accumulation successive. Aujourd'hui, le jeu d'étiquettes le plus utilisé et qui semble acquérir *de facto* une valeur de standard est *Universal Dependencies*. Nous en recommandons l'utilisation, afin de faciliter l'échange des données et, notamment, la constitution de corpus linguistiques interrogeables de façon croisée, favorisant ainsi la diachronie ou la création de jeux de données multilingues.

[141] Les agglutinations sont d'abord identifiées dans une ligne indépendante (en indiquant l'index du premier et du dernier mot un fois décomposé), puis la forme décomposée en fonction des éléments qui la composent.

Documenter ses pratiques d'annotation

L'annotation lexico-grammaticale n'est pas une tâche univoque, il n'y a pas de manière unique d'annoter un corpus. Elle peut être menée selon des principes scientifiques et linguistiques ainsi que des objectifs qui peuvent induire des pratiques différentes. Pour les langues romanes se pose de manière quasi-systématique la question de l'adjectivation des participes. Dans le cas de « luisant », doit-on le considérer comme un verbe (participe présent utilisé comme adjectif) ou bien comme un adjectif[142] ? Cet exemple a vocation avant tout à illustrer la difficulté inhérente à la tâche d'annotation linguistique d'un corpus ; nous ne pouvons qu'insister sur l'importance de la documentation des réflexions qui ont mené aux choix d'annotation, qui importent autant que les choix en eux-mêmes. Par ailleurs, il est fréquent qu'un projet s'écarte quelque peu du standard choisi en raison de ses objectifs scientifiques. Il est alors important de documenter cet écart, afin de pouvoir rendre les données les plus interopérables possibles à terme. Par exemple, le projet d'annotation du castillan médiéval e-CAM[143] a décidé de ne pas suivre la règle commune en matière d'agglutination et de conserver telles quelles certaines de ces formes pour des raisons de simplicité, ainsi qu'en raison des phénomènes de métathèse qui apparaissent souvent dans les formes agglutinées en castillan[144]. La documentation des pratiques est alors indispensable pour rendre compte des écarts par rapport à la norme en vigueur.

Outre la question des jeux d'étiquettes, un point fondamental est la connaissance et la documentation du ou des référentiels de lemmes utilisés. Quelles sont les règles de lemmatisation en vigueur dans le corpus (approche conservatrice *versus* approche normalisatrice)[145] ? Quels sont les dictionnaires de référence utilisés ? Comment est gérée l'homographie des lemmes ? Deux lemmes peuvent être homographes, pour des raisons d'évolution phonétique, par exemple. Dans ce cas, le choix de leur désambiguïsation est souvent arbitraire et s'appuiera sur un dictionnaire déjà constitué[146]. L'important, encore une fois, réside dans l'explicitation des choix d'annotation.

Quels outils pour l'annotation linguistique

Il existe un grand nombre d'outils qui permettent l'annotation grammaticale. Cette variété se réduit plus ou moins en fonction de l'état de langue choisi. On peut citer l'outil historique *TreeTagger*[147], mais aussi le lemmatiseur à base de règles développé à l'ATILF, LGERM[148].

---

[142] Nous recommandons de considérer ces formes comme des verbes et non pas comme des adjectifs, afin de faciliter des études en diachronie.

[143] Pour plus de détails, voir <https://agorantic.univ-avignon.fr/wp-content/uploads/sites/13/2024/01/P2_eCAM-AAP-blc-Agorantic-2024.pdf>.

[144] Par exemple, la forme "*tomaldo*", pour "*tomadlo*" ("prenez-le"), rend impossible l'identification du point de scission exact entre le verbe et le pronom enclitique.

[145] Par exemple, en Castillan, la forme *vergüeña* (vergogne, honte) pourra être lemmatisée en *vergüenza* dans le cas d'une lemmatisation normalisatrice, ou comme *vergüeña* selon des règles plus conservatrices.

[146] Pour le français médiéval : *Tobler-Lommatzch*,; pour le castillan: le *Dictionnaire de l'Académie Espagnole*, ou le *Dictionnaire Historique de la Langue Espagnole*, ou le DEMEL.

[147] Helmut Schmid, *Probabilistic Part-of-Speech Tagging Using Decision Trees*, 1994.

[148] Gilles Souvay, « LGeRM: un outil d'aide à lemmatisation du français médiéval », *18th International Conference on Historical Linguistics ICHL 2007*, 2007.

Citons, de même, la bibliothèque *NLTK* (*Natural Language Tool Kit*), très utilisée dans le domaine du traitement du langage naturel, et sa version adaptée aux langues classiques, *CLTK* (*Classical Language Tool Kit*)[149]. Dans le domaine des langues romanes, et plus particulièrement ibériques, l'outil *Freeling*[150] est utile. Il s'agit d'un système de traitement automatique des langues modulaire (possibilités d'annotations : lemmes, parties du discours, morphologie, entre autres) qui adapte le standard *EAGLES* pour ses annotations. En ce qui concerne le castillan médiéval, il s'agit actuellement du seul outil en accès ouvert qui permette l'annotation, bien que des projets de production de corpus annotés et de modèles soient en cours[151]. L'usage croissant de méthodes fondées sur l'apprentissage automatique a contribué à la grande diversification des outils disponibles, comme les projets *Stanza*[152] qui utilise *UD*, et *Spacy*[153] qui possède divers jeux d'étiquettes en fonction des modèles (toutefois UD est aujourd'hui prédominant). D'autres outils ont été pensés pour les langues rares et peu dotées, comme *Pie*[154]. Ce lemmatiseur qui se fonde sur l'apprentissage machine a donné lieu à un grand nombre de corpus et de modèles d'annotation[155]. Il est aujourd'hui utilisé à l'École nationale des chartes avec le projet *Deucalion*[156], système d'annotation en ligne proposant divers modèles en accès libre (ancien français, français moderne, latin, grec, hollandais). Ces outils fonctionnent avec les langages de programmation les plus communément utilisés, comme Python, offrent des API faciles d'utilisation et sont aisément intégrables à des chaînes de traitement complètes. Ils demandent tout de même quelques compétences en ingénierie pour une bonne prise en main. Enfin, les outils d'annotation sont comparés à l'aide de métrique et de scores d'évaluation, en particulier l'exactitude (*accuracy*), qui désigne le taux de prédictions correctes par rapport au volume total du corpus. On aura tendance à préciser ces scores en distinguant la qualité de l'annotation sur des *tokens* connus, ambigus et inconnus[157], ce qui

[149] Kyle P. Johnson, Patrick J. Burns, John Stewart, [et al.], « The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages », *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, éds. Heng Ji, Jong C. Park et Rui Xia, Online, Association for Computational Linguistics, 2021, p. 20-29.

[150] Lluís Padró et Evgeny Stanilovsky, « FreeLing 3.0: Towards Wider Multilinguality », *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, ELRA, 2012.

[151] Matthias Gille Levenson, Olivier Brisville-Fertin, Maria Díez Yáñez, [et al.], « Construcción de un corpus de evaluación de la anotación léxico-gramatical del castellano medieval (siglos 13-15) », Santiago de Compostela, 2021.

[152] Peng Qi, Yuhao Zhang, Yuhui Zhang, [et al.], « Stanza: A Python Natural Language Processing Toolkit for Many Human Languages », *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, éds. Asli Celikyilmaz et Tsung-Hsien Wen, Online, Association for Computational Linguistics, 2020, p. 101-108.

[153] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, [et al.], « spaCy: Industrial-strength natural language processing in python », 2020.

[154] Enrique Manjavacas, Ákos Kádár et Mike Kestemont, « Improving Lemmatization of Non-Standard Languages with Joint Learning », *Proceedings of the 2019 Conference of the North*, Minneapolis, Minnesota, 2019, p. 1493-1503.

[155] Jean-Baptiste Camps, Thibault Clérice, Frédéric Duval, [et al.], « Corpus and Models for Lemmatisation and POS-tagging of Old French », *JDMDH*, 2021

[156] Thibault Clérice, « Pie Extended, an extension for Pie with pre-processing and post-processing », *Zenodo*, 2020, voir en ligne : <https://dh.chartes.psl.eu/deucalion/>.

[157] Uniquement pour les outils fondés sur l'apprentissage supervisé, les outils à base de règle ne s'appuyant pas sur des corpus d'entraînement. Pour être précis, on pourra distinguer les tokens et les cibles (*targets*) ambiguës,

permet d'évaluer la capacité du modèle à annoter en contexte et à produire des analyses sur des données nouvelles. Il est difficile de donner des scores idéaux en ce qui concerne l'annotation lexico-grammaticale. Pour les outils d'annotation d'ancien français, les modèles les plus récents obtiennent des scores 97 % pour les lemmes et les PoS sur un corpus dit « in-domain », et de 91 % pour les lemmes et environ 95 % pour les PoS sur un corpus dit « out-of-domain », c'est-à-dire de nature distincte du corpus d'entraînement[158]. Cette métrique est censée représenter le fonctionnement du modèle dans un contexte de recherche « normal ». En ce qui concerne le castillan médiéval, une évaluation a été produite en 2021 qui montre que *Freeling* prédit correctement 91,6 % des lemmes, 93 % des parties du discours, et 90,6 % de la morphologie, en prenant en compte les entités nommées[159].

```
<text>
    <body>
        <div type="livre" n="3">
            <div type="partie" n="3">
                <div type="chapitre" n="22">
                    <div type="traduction">
                        <p n="JijdHudsajJ">
                            <w lemma="et" pos="CC">Et</w>
                            <w lemma="el" pos="DA0MS0">el</w>
                            <w lemma="3" pos="A00MS0">iii°</w>
                            <w lemma="ser" pos="VSIP3S0">es</w>
                            <w lemma="hacer" pos="VMG0000">faziendo</w>
                            <w lemma="saeta" pos="NCFP000">saetas</w>
                            <w lemma="que" pos="PR0CN000">que</w>
                            <w lemma="llamar" pos="VMIP3P0">llaman</w>
                            <w lemma="rueca" pos="NCFP000">rruecas</w>
                            <pc lemma="," pos="Fp">.</pc>
                            <w lemma="et" pos="CC">Et</w>
                            <w lemma="este" pos="DD0FP0">éstas</w>
                            <w lemma="ser" pos="VSIP3P0">son</w>
                            <w lemma="en" pos="SPS00">en</w>
                            <w lemma="medio" pos="NCMS000">medio</w>
                            <w lemma="hueco" pos="AQ0FP0">huecas</w>
                            <pc lemma="," pos="Fc">,</pc>
```

Figure 28: Texte castillan annoté via Freeling et structuré en TEI

Intégration et réutilisation des données dans la chaîne d'acquisition

La TEI dispose d'un module permettant la représentation des informations linguistiques d'un corpus[160]. Une des façons les plus simples de le faire est d'utiliser l'élément <w> (word) et ses attributs spécifiques @lemma, @pos, @msd. Si les formats les plus courants pour conserver des annotations linguistiques sont le CSV ou le TSV, dans le cadre de la publication d'un corpus textuel, il peut s'avérer pertinent de tout conserver dans le format standard qu'est le XML TEI, voire d'injecter ses informations dans le corps de son texte structuré pour permettre des requêtes précises sur le corpus ou encore l'appel à des glossaires ou dictionnaires.

---

ces dernières désignant les lemmes jamais vus lors de l'entraînement, quand les premières désignent les formes uniquement

[158] Voir Frédéric Duval, Lucence Ing, Jean-Baptiste Camps, [et al.], « Lemmatisation de l'ancien français : Présentation du modèle et des outils de l'École des chartes », *XXXe Congrès International de Linguistique et de Philologie Romanes*, Société de linguistique romane, Jul 2022, La Laguna, Tenerife, Espagne. pp.1001-1012, <10.46277/SLR.18.2023.1001-1012>, <hal-04013381>.

[159] En ce qui concerne ce dernier outil, il est à savoir qu'il est perfectible, étant donné que plus de la moitié des erreurs de lemmatisation sont dues à de l'homographie, pour une étude plus approfondie voir : Matthias Gille Levenson, Olivier Brisville-Fertin, Maria Díez Yáñez, [et al.], « Construcción de un corpus de evaluación de la anotación léxico-gramatical del castellano medieval (siglos 13-15) », Santiago de Compostela, 2021.

[160] Voir la section https://www.tei-c.org/release/doc/tei-p5-doc/en/html/AI.html#AILA des règles de la TEI.

Une fois l'annotation linguistique réalisée, il est possible de combiner les informations linguistiques et les informations structurelles, codicologiques, historiques ou ecdotiques d'un corpus. L'utilisation des API d'outils d'annotation présentés plus haut peut permettre d'annoter directement un corpus en XML-TEI, pour peu que celui-ci ait été tokénisé en amont – ce qui n'est pas toujours une tâche aisée, en fonction de la langue[161]. On peut dès lors envisager la création de chaînes de traitement global. *FALCON* (*For Alignment, Lemmatization and Collation,* voir figure 29)[162] et *TEICollator*[163] par exemple, se servent des annotations pour la collation des variantes dans la tradition textuelle[164]. Le principe de ces outils est de pouvoir croiser les informations grammaticales et lexicales avec les informations structurelles, voire codicologiques, pour faciliter l'étude du corpus par la suite en produisant automatiquement des corpus collationnés en XML-TEI.

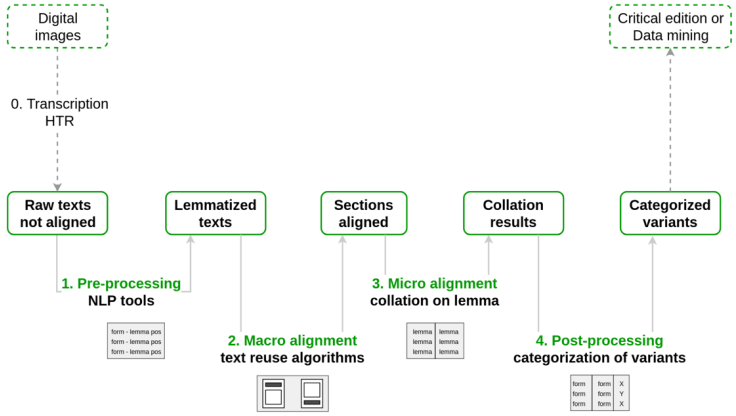


Figure 29: Pipeline de FALCON, voir Jean-Baptiste Camps, Elena Spadini et Lucence Ing, « Collating Medieval Vernacular Texts. Aligning Witnesses, Classifying Variants », DH2019 Digital Humanities Conference 2019, Utrecht, Netherlands, 2019.)

3. La reconnaissance d'entités nommées : l'extraction d'informations au service de l'enrichissement des textes

En plus de l'annotation linguistique, il est désormais possible d'annoter automatiquement les noms de lieux, de personnes ou d'organisations. Cette pratique permet de mieux appréhender l'information textuelle et de créer, à partir de ces annotations, des bases de données, des cartes, ou encore des analyses de réseaux, enrichissant ainsi l'exploitation des corpus.

[161] Une fois le corpus tokénisé, on peut extraire les formes contenues dans les balises <w>, produire des annotations, et enfin les réinjecter dans le corpus. Pour ce faire, l'outil d'annotation ne doit pas opérer de re-tokénisation du corpus, comme cela peut parfois arriver, avec Freeling par exemple.
[162] Jean-Baptiste Camps, Lucence Ing et Elena Spadini, « Flacon : A processing workflow for automated collation », 2021.
[163] Matthias Gille Levenson, « TeiCollator: a TEI to TEI workflow », *TEI2022: Text as data*, Newcastle, United Kingdom, 2022.
[164] En effet, deux formes alignées qui sont distinctes, mais qui partagent le même lemme, la même partie du discours et la même morphologie constituent une variante graphique. Il est également possible d'identifier un ensemble de classes de variantes (morphologiques, lexicales, etc), afin de faciliter l'étude de la tradition textuelle.

La reconnaissance d'entités nommées : définition

La reconnaissance d'entités nommées (*Named Entity Recognition*, NER) est une tâche de traitement automatique des langues qui consiste à identifier et classer des éléments spécifiques, tels que des noms de lieux, de personnes ou d'organisations, dans des données non structurées. En attribuant à ces entités des classes sémantiques, la NER permet de structurer ces informations pour les traiter automatiquement dans des processus en aval de la chaîne d'acquisition et traitement des données[165]. L'extraction d'information inclut également l'extraction de relations, qui permet de détecter les relations sémantiques présentes dans un texte, par exemple entre deux entités comme des personnes. Elle permet d'enrichir le texte, et, à grande échelle, ouvre la voie à la fouille de données, facilitant ainsi l'analyse approfondie des relations au sein des corpus. Les linguistes définissent une entité nommée comme une « unité linguistique de nature référentielle »[166]. Une entité nommée peut prendre la forme d'un mot ou de plusieurs mots, et fait référence à une entité unique d'un système donné (noms de lieux, noms de personnes, planètes, etc.). Cependant, les entités nommées ne sont pas forcément des noms propres, et leur forme linguistique change en fonction des corpus d'application. Pour les corpus biomédicaux, par exemple, l'extraction d'entités nommées portera davantage sur les noms de molécules, de médicaments, de pathologies, etc., plutôt que des noms de lieux. Par ailleurs, dans notre cas, nous nous intéressons à une extraction réalisée au sein de données textuelles, mais il existe également des outils de NER pour les données audios.

Ainsi, la NER permet d'enrichir les sources, et leur donne un nouveau point d'entrée, en autorisant par exemple la lecture distante. Les entités nommées peuvent être indexées et facilitent l'exploration de grandes collections de textes grâce, par exemple, à l'utilisation de moteurs de recherche à facettes. La NER est également le socle d'autres tâches, comme la désambiguïsation et l'*entity linking*, que nous décrirons dans la suite de cet exposé.

État de l'art

Pour extraire les entités nommées, il existe plusieurs méthodes. Leur annotation peut être réalisée à la main, tandis que leur extraction semi-automatique peut être exécutée par des systèmes de reconnaissance par règles. Ces deux solutions ne sont cependant pas les plus répandues du fait de leur nature chronophage. Les modèles d'apprentissage machine et d'apprentissage profond leur sont aujourd'hui préférés. Actuellement, il est possible de réutiliser soit des modèles multilingues, soit des modèles dédiés à des langues spécifiques. Un certain nombre d'outils et de bibliothèques Python (souvent les mêmes que pour la lemmatisation) sont disponibles gratuitement pour accomplir cette tâche, tels que *Stanza* et *SpaCy*[167]. Ces bibliothèques proposent des dizaines de modèles de NER couvrant plusieurs

[165] Rosa Stern, *Identification automatique d'entités pour l'enrichissement de contenus textuels*, thèse de doctorat, Université Paris-Diderot - Paris VII, 2013.
[166] Yoann Dupont, *La structuration dans les entités nommées*, Thèse de doctorat, Sorbonne Paris Cité, 2017.
[167] Peng Qi, Yuhao Zhang, Yuhui Zhang, [et al.], « Stanza: A Python Natural Language Processing Toolkit for Many Human Languages », *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, éds. Asli Celikyilmaz et Tsung-Hsien Wen, Online, Association for

langues, comme l'anglais, le français, l'arabe ou le japonais. Elles présentent l'avantage d'être faciles d'accès et d'utilisation, même avec une connaissance basique de Python, tout en offrant une grande efficacité. Cependant, ces modèles restent « génériques » et peu adaptés à des usages spécifiques, dans la mesure où ils sont souvent entraînés à extraire uniquement quatre grands types d'entités : les noms de personnes, les lieux, les organisations et une catégorie « divers » (*miscellaneous*), qui regroupe les entités nommées qui n'appartiennent pas aux catégories citées précédemment. En outre, ces modèles sont généralement conçus pour fonctionner sur des variantes modernes et bien représentées des langues. Ainsi, dans certains cas, il pourra être nécessaire d'entraîner de nouveaux modèles de NER en annotant soi-même au préalable les entités nommées spécifiques nécessaires à un projet donné.

Annoter les entités nommées

L'annotation des entités nommées et la constitution de jeux de données ne sont pas tout à fait comparables à l'annotation d'entités nommées dans un fichier TEI. Les fichiers TEI annotés et enrichis avec des entités nommées le sont principalement à des fins d'édition, utilisant des balises telles que <persName> pour les noms de personnes ou <placeName> pour les noms de lieux, que l'on peut lier à des déclarations uniques contenant des informations complémentaires sur l'entité tout au long du texte. Ces annotations sont conçues pour structurer un texte à éditer et incluent toutes les informations nécessaires à la création d'index (voir I.B.2). Cependant, ces annotations en TEI, si elles existent déjà, peuvent être réutilisées pour créer des corpus d'entraînement au format IOB grâce à des transformations XSLT ou Python.

Les annotations NER servent à créer des corpus de vérité terrain pour l'entraînement de modèles et fournir des annotations plus légères et souples que la TEI. Les données sont généralement annotées au format IOB (pour *Inside-Outside-Beginning)*[168] qui permet d'annoter un token et de lui attribuer une étiquette correspondant à une classe d'entité nommée selon sa position (au début ou à l'intérieur de l'expression qui désigne l'entité). Le format IOB est répandu dans le monde de la recherche et de l'industrie, et rend interopérables les jeux de données. Par exemple, dans le cas d'une phrase simple telle que « Sorbonne Université est située à Paris », le résultat serait selon le format IOB :

Sorbonne	B-ORG
Université	I-ORG
est	O
située	O
à	O
Paris	B-LOC

« Sorbonne Université » est une entité nommée composée de plusieurs mots, dont « Sorbonne » est le mot qui commence l'entité, et « Université » un mot composant l'entité. D'autres mots comme « est » ou « située » ne correspondent pas à des entités nommées, et

Computational Linguistics, 2020, p. 101-108, Matthew Honnibal, Ines Montani, Sofie Van Landeghem, [et al.], « spaCy: Industrial-strength natural language processing in python », 2020.

[168] Erik F. Tjong Kim Sang et Fien De Meulder, « Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition », *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, p. 142-147.

reçoivent donc l'étiquette « *outside* ». Les quatre classes d'entités nommées les plus répandues et citées précédemment reçoivent généralement les étiquettes suivantes : ORG pour un nom d'organisation ; LOC pour un nom de lieu ; PER pour un nom de personne ; et MISC pour divers. Il existe également des formats d'annotation plus complexes comme le BIOES (*Beginning-Inside-Outside-Ending-Single Element*)[169]. Ces fichiers d'annotation sont présentés sous la forme de textes bruts organisés en ligne et colonnes (TSV).

Une fois les données annotées, dans le cadre d'une chaîne d'acquisition, il est essentiel de pouvoir réintégrer ces données dans la structure TEI du document final. Les défis techniques à surmonter sont similaires à ceux de la phase d'annotation linguistique, notamment garantir une bonne tokenisation des termes et associer à chaque token un identifiant permettant de réaligner les données annotées avec leur équivalent dans le fichier TEI (voir II.C.1). Ce réalignement peut se faire directement dans le corps du texte ou via la balise <standOff>. Toutefois, pour ceux qui cherchent des solutions nécessitant moins d'investissement technique, il existe aujourd'hui des interfaces et logiciels tels que TEI Publisher qui permettent l'annotation d'entités nommées et l'utilisation de modèles SpaCy directement dans des fichiers TEI via une interface graphique[170].

Désambiguïsation des entités nommées

Une fois les entités nommées extraites, un défi supplémentaire se pose : celui de la désambiguïsation des entités nommées et de les lier à une notice d'autorité en ligne, pour capturer précisément à quelles entités ces mentions se réfèrent, malgré les potentielles homonymies : par exemple pour Paris, avoir la capacité de préciser, si nous faisons référence à Paris, capitale de la France (Q90 dans Wikidata) ou bien à Paris, Texas (Q830149). La désambiguïsation des entités nommées consiste également à résoudre les ambiguïtés qui peuvent exister sur le type d'entités nommées (nom de lieu, nom de personne, etc.), notamment dans le cas d'expressions composées de plusieurs mots. Dans un texte comportant une mention du « parvis Rosa-Parks à Paris » par exemple, il est important de déterminer que « parvis Rosa-Parks » fait référence au lieu et doit être annoté en tant que tel, et non en tant que nom de personne. Ce processus repose souvent sur le contexte dans lequel l'entité est mentionnée, et est généralement compris dans l'extraction des entités nommées effectuée par les modèles de NER des librairies mentionnées précédemment.

[169] Erik F. Tjong Kim Sang et Fien De Meulder, « Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition », *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, p. 142-147.

[170] Voir le site de TEI Publisher, https://teipublisher.com/exist/apps/tei-publisher-home/index.html, et plus particulièrement la page de documentation sur son interface de documentation https://teipublisher.com/exist/apps/tei-publisher/documentation/web-annotations. Consultés le 24/07/2024.

Le *linking* des entités nommées, ou *entity linking,* va au-delà de la désambiguïsation en associant à chaque mention une entrée spécifique dans une base de connaissance, comme Wikidata ou DBpedia. Par exemple, dans le cas d'un fichier TEI, un identifiant vers une notice d'autorité peut être ajouté à une entité dans sa notice (voir figure 30). Le *linking* contribue à l'enrichissement d'un texte et à l'amélioration de la recherche d'information pour en faciliter l'analyse. Plusieurs outils proposent une automatisation du linking, tels que Stanza, SpaCy ou encore Entity Fishing[171].

```
<person xml:id="p0002">
    <persName>
        <forename>Fitzwilliam</forename>
        <surname>Darcy</surname>
    </persName>
    <idno type="wikidata">Q2207092</idno>
</person>
```

Figure 30: Exemple de notice en TEI, dans un élément <standOff>. Le linking est effectué à l'aide de l'élément <idno>, qui qui permet de pointer vers une base de connaissance.

L'enrichissement d'un texte par l'annotation linguistique et la reconnaissance d'entités nommées (NER) constitue une étape essentielle dans l'exploitation des corpus numériques. L'annotation linguistique apporte des informations précieuses sur la nature et la fonction des mots, permettant ainsi des études stylistiques, syntaxiques ou quantitatives. Quant à la NER, elle facilite la structuration et la mise en relation d'éléments clés du texte, tels que les noms de lieux, de personnes ou d'organisations, ouvrant la voie à des analyses plus complexes, comme la cartographie des lieux ou l'analyse de réseaux de relations. Ces processus d'enrichissement, bien qu'automatisés, requièrent une vigilance particulière pour assurer leur qualité, notamment en termes de normalisation, de précision des annotations et de segmentation des unités. Sans une segmentation correcte, il devient impossible de réconcilier et réintégrer les annotations dans la chaîne d'acquisition du texte, contraignant à travailler sur des fichiers séparés et risquant de perdre l'organisation hiérarchique du texte. Dans le cadre d'une automatisation partielle, il peut être judicieux d'annoter d'abord le texte, puis de le structurer dans un second temps pour éviter les décalages entre les différentes versions du texte. Malgré ces défis, ces technologies représentent un levier puissant pour transformer des textes bruts en données exploitables et structurées, ouvrant la voie de recherches de plus grande ampleur et plus approfondies en linguistique, histoire et littérature.

[171] Peng Qi, Yuhao Zhang, Yuhui Zhang, [et al.], « Stanza: A Python Natural Language Processing Toolkit for Many Human Languages », *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, éds. Asli Celikyilmaz et Tsung-Hsien Wen, Online, Association for Computational Linguistics, 2020, p. 101-108, Matthew Honnibal, Ines Montani, Sofie Van Landeghem, [et al.], « spaCy: Industrial-strength natural language processing in python », 2020, Luca Foppiano et Laurent Romary, « Entity-fishing: a DARIAH entity recognition and disambiguation service », *Journal of the Japanese Association for Digital Humanities*, vol. 5 / 1, novembre 2020, p. 22.

Quelques références pour aller plus loin :

– Jean-Baptiste Camps, Thibault Clérice, Frédéric Duval, [et al.], « Corpus and Models for Lemmatisation and POS-tagging of Old French, *JDMDH*, 2021.
– Yoann Dupont, *La structuration dans les entités nommées*, Thèse de doctorat, Sorbonne Paris Cité, 2017.
– Geoffrey Leech et Andrew Wilson, « Standards for Tagsets », in Hans Van Halteren, (éd.). *Syntactic Wordclass Tagging*, éd. Hans Van Halteren, Dordrecht, Springer Netherlands, 1999, p. 55-80.
– Rosa Stern, *Identification automatique d'entités pour l'enrichissement de contenus textuels*, thèse de doctorat, Université Paris-Diderot - Paris VII, 2013.

En conclusion, cet article a exploré l'ensemble du processus de production d'une chaîne d'acquisition numérique du texte, depuis la constitution du corpus jusqu'à l'automatisation des tâches d'acquisition et d'enrichissement. L'objectif de cette présentation n'est pas d'imposer une méthode rigide, mais de fournir un aperçu des nombreuses options disponibles, en s'appuyant sur nos propres expériences. Si certaines étapes sont incontournables pour la publication en ligne, telles que la constitution, la transcription, la structuration et la mise en ligne du corpus, d'autres, comme l'ajout d'entités nommées ou l'annotation linguistique, sont facultatives et peuvent être adaptées en fonction des besoins spécifiques du projet. Nous avons également tenu à souligner qu'il n'est pas nécessaire d'automatiser toutes les étapes de chaque projet. Ainsi, la transcription peut être faite à la main ou récupérée à partir d'une édition nativement numérique, de même pour l'ajout des entités, ainsi que leur « linking » à des notices d'autorité[172]. En revanche, si les annotations linguistiques peuvent en théorie être ajoutées manuellement, dans les faits, la tâche se révèle bien trop chronophage.

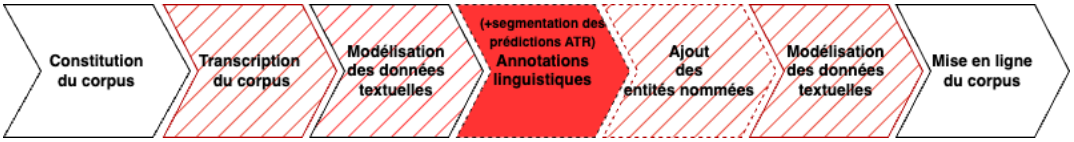


Figure 31: Exemple de chaîne d'acquisition, en pointillés les étapes facultatives, en hachurés les étapes qui peuvent être automatisées, en rouge, les étapes à automatiser. La modélisation des données apparaît à deux moments clés de la chaîne : elle peut être réalisée à la fois avant et/ou après l'enrichissement du corpus, puis affinée avant sa mise en ligne.

Ainsi, la première partie a détaillé les étapes essentielles d'une chaîne de production basique de la constitution du corpus à sa mise en ligne. Nous espérons avoir présenté les bases nécessaires pour la modélisation en XML TEI, tout en insistant sur l'importance de structurer et d'enrichir ces données pour garantir leur exploitation future. Enfin, la mise en ligne, dernière étape de la chaîne, est cruciale pour la diffusion des travaux, et le choix des outils doit être mûrement réfléchi en tenant compte des implications techniques et financières, comme la maintenance des serveurs. Dans certains cas, un site statique, simple à maintenir, accompagné de données brutes déposées sur des plateformes comme Zenodo, GitLab ou GitHub, peut s'avérer une solution plus durable qu'un site dynamique, difficile à maintenir et sujet à l'obsolescence.

Dans un second temps, nous avons exploré les avancées récentes en matière d'automatisation et d'intelligence artificielle, en démontrant comment des technologies telles que la reconnaissance automatique de texte (ATR), la lemmatisation et la reconnaissance d'entités nommées (NER) peuvent optimiser les processus de transcription et d'annotation. Toutefois, intégrer ces technologies dans les chaînes d'acquisition traditionnelles reste complexe et soulève plusieurs défis techniques. Bien que ces outils puissent accomplir des tâches spécifiques, il n'existe pas encore de chaîne complète regroupant toutes les étapes mentionnées. De plus, les outils d'automatisation, bien qu'efficaces, ne sont pas infaillibles, en particulier pour les langues peu dotées ou les états anciens des langues. Travailler sur des documents complexes exige souvent un investissement supplémentaire pour créer et partager

[172] Ces annotations peuvent être assisté avec l'utilisation de *Leaf Writter* ou TEI Publisher avec l'ajout de lien vers des notices wikiData ou du VIAF, ainsi que l'ajout de RDF dans les fichiers TEI.

des données d'entraînement adaptées, afin d'entraîner des modèles plus performants et de garantir des résultats de qualité, tout en contribuant à l'avancée des recherches futures.

Dans cet exposé, nous nous sommes concentrés sur la publication de sources uniques, sans aborder la question de la publication de sources multiples ni la collation semi-automatisée des traditions manuscrites. Bien que cette partie de la chaîne d'acquisition ne soit pas totalement inexplorée, elle demeure encore à un stade expérimental et nécessite des approfondissements avant de pouvoir être pleinement intégrée dans des pipelines plus complexes. Quelques travaux préliminaires, tels que CollateX[173], FALCON[174] ou teiCollator[175], méritent toutefois d'être mentionnés. Les recherches futures menées dans le cadre de l'ERC Prima par E. Pierazzo, qui se concentre sur les traditions manuscrites complexes, pourraient également apporter des avancées significatives dans ce domaine.

Enfin, l'utilisation de ces technologies ne relève pas uniquement d'une expertise d'ingénierie. Au contraire, les résultats les plus fructueux émergent souvent de la combinaison d'une expertise technique et d'une connaissance approfondie des documents. Ce dialogue entre compétences permet de structurer les données de manière optimale. Ainsi, la réussite d'un projet repose bien souvent sur une synergie entre ces deux pôles de compétences, qu'elle soit incarnée par une personne maîtrisant ces deux domaines ou par une équipe pluridisciplinaire où les expertises sont partagées[176].

[173] R. Haentjens Dekker et G. Middell, « Computer-Supported Collation with CollateX: Managing Textual Variance in an Environment with Varying Requirements », in B. Dr. Maegaard, (éd.). *Supporting Digital Humanities 2011: Answering the unaskable*, éd. B. Dr. Maegaard, Copenhagen, Denmark, 2011.

[174] Jean-Baptiste Camps, Lucence Ing et Elena Spadini, « Flacon : A processing workflow for automated collation », 2021.

[175] Matthias Gille Levenson, « TeiCollator : une chaîne de traitement ecdotique semi-automatisée », *XXXe Congrès International de Linguistique et Philologie Romane*, 2022.

[176] Pour la rédaction de cet article, la répartition des rôles a été la suivante : Alix Chagué a pris en charge les sections relatives à la transcription et à l'acquisition automatique du corpus. Floriane Chiffoleau s'est occupée des sections portant sur la modélisation du corpus en XML TEI ainsi que sur la mise en ligne. Hugo Scheithauer a rédigé les parties consacrées à l'analyse automatique de la mise en page et à la reconnaissance des entités nommées. Matthias Gille Levenson s'est concentré sur les sections traitant de la segmentation linguistique et de l'annotation linguistique. Ariane Pinche a rédigé l'introduction, la partie sur la constitution du corpus, l'automatisation de la structuration à partir de la segmentation, tout en assurant la coordination des différentes étapes de rédaction et en harmonisant la version finale du texte.

Bibliographie sélective

I-Acquisition du texte

ANDREWS, Tara L., « The Third Way: Philology and Critical Edition in the Digital Age », *Variants*, vol. 10, 2013, p. 61-76, [En ligne : https://boris.unibe.ch/43071/].

BAILLOT, Anne et KOENIG, Marieke, « Automatic Text Recognition: Harmonizing ATR Workflows », [En ligne : https://harmoniseatr.hypotheses.org]. Consulté le 8 juillet 2024.

CHAGUÉ, Alix, CLÉRICE, Thibault et ROMARY, Laurent, « HTR-United : Mutualisons la vérité de terrain ! », *DHNord2021 - Publier, partager, réutiliser les données de la recherche : les data papers et leurs enjeux,* Lille, 2021, [En ligne : https://hal.science/hal-03398740].

CLÉRICE, Thibault, « You Actually Look Twice At it (YALTAi): using an object detection approach instead of region segmentation within the Kraken engine », *Journal of Data Mining & Digital Humanities* (Historical Documents and ATR), décembre 2023, [En ligne : http://arxiv.org/abs/2207.11230].

CLÉRICE, Thibault, JANÈS, Juliette, SCHEITHAUER, Hugo, [et al.], « Layout Analysis Dataset with SegmOnto (LADaS) », [En ligne : https://github.com/DEFI-COLaF/LADaS].

CLÉRICE, Thibault, JANES, Juliette, SCHEITHAUER, Hugo, [et al.], « Layout Analysis Dataset with SegmOnto », *DH2024 - Annual conference of the Alliance of Digital Humanities Organizations*, Washington DC, 2024, [En ligne : https://inria.hal.science/hal-04513725].

CLÉRICE, Thibault, JANÈS, Juliette, SCHEITHAUER, Hugo, [et al.], « DEFI-COLaF/LADaS: 2024-02-20 - First release », Zenodo, 2024, [En ligne : https://zenodo.org/doi/10.5281/zenodo.10682623].

CLÉRICE, Thibault, PINCHE, Ariane, VLACHOU-EFSTATHIOU, Malamatenia, [et al.], « CATMuS Medieval: A multilingual large-scale cross-century dataset in Latin script for handwritten text recognition and beyond », 2024, [En ligne : https://inria.hal.science/hal-04453952].

GABAY, Simon, PINCHE, Ariane, CHRISTENSEN, Kelly, [et al.], « SegmOnto: A Controlled Vocabulary to Describe and Process Digital Facsimiles », décembre 2023, [En ligne : https://hal.science/hal-04343404].

HODEL, Tobias, SCHOCH, David, SCHNEIDER, Christa, [et al.], « General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example », *Journal of Open Humanities Data*, vol. 7 / 0, Ubiquity Press, juillet 2021, p. 13, [En ligne : http://openhumanitiesdata.metajnl.com/articles/10.5334/johd.46/].

KAHLE, Philip, COLUTTO, Sebastian, HACKL, Günter, [et al.], « Transkribus - A Service Platform for Transcription, Recognition and Retrieval of Historical Documents », *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 04, 2017, p. 19-24.

KIESSLING, B., TISSOT, R., STOKES, P., [et al.], « eScriptorium: An Open Source Platform for Historical Document Analysis », *International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2, 2019, p. 19-19.

KIESSLING, Benjamin, « Kraken - a Universal Text Recognizer for the Humanities », DataverseNL, 2019, [En ligne : https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/Z9G2EX].

LEVENSON, Gille, « Towards a general open dataset and model for late medieval Castilian text recognition (HTR/OCR) », *Journal of Data Mining & Digital Humanities* (Historical Documents and automatic text recognition), octobre 2023, [En ligne : https://jdmdh.episciences.org/12264].

PINCHE, Ariane, « Generic HTR Models for Medieval Manuscripts. The CREMMALab Project », *Journal of Data Mining & Digital Humanities* (Historical Documents and automatic text recognition), octobre 2023, [En ligne : https://jdmdh.episciences.org/11592].
PINCHE, Ariane, CHRISTENSEN, Kelly et GABAY, Simon, « Between automatic and manual encoding », *TEI 2022 conference : Text as data*, Newcastle, 2022, [En ligne : https://hal.science/hal-03780302].
PINCHE, Ariane, CLÉRICE, Thibault, CHAGUÉ, Alix, [et al.], « CATMuS-Medieval: Consistent Approaches to Transcribing ManuScripts », *Digital Humanities Conference*, 2024, [En ligne : https://inria.hal.science/hal-04346939].
REUL, Christian, TOMASEK, Stefan, LANGHANKI, Florian, [et al.], « Open Source Handwritten Text Recognition on Medieval Manuscripts Using Mixed Models and Document-Specific Finetuning », *Document Analysis Systems*, éds. Seiichi Uchida, Elisa Barney et Véronique Eglin, Cham, Springer International Publishing, 2022, p. 414-428.
SCHEITHAUER, Hugo, CHAGUÉ, Alix et ROMARY, Laurent, « Which TEI representation for the output of automatic transcriptions and their metadata? An illustrated proposition », 2022, [En ligne : https://inria.hal.science/hal-04001303].
SOUSA NETO, Arthur Flor DE, BEZERRA, Byron Leite Dantas, TOSELLI, Alejandro Héctor, [et al.], « HTR-Flor++: A Handwritten Text Recognition System Based on a Pipeline of Optical and Language Models », *Proceedings of the ACM Symposium on Document Engineering 2020*, New York, , 2020, p. 1-4, [En ligne : https://doi.org/10.1145/3395027.3419603].
SPRINGMANN, Uwe, REUL, Christian, DIPPER, Stefanie, [et al.], « Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin », *arXiv*, 2018, [En ligne : http://arxiv.org/abs/1809.05501].
STUTZMANN, Dominique, « Ontologie des formes et encodage des textes manuscrits médiévaux », *Document numérique*, Vol. 16, 2013, p. 81-95, [En ligne : https://www.cairn.info/revue-document-numerique-2013-3-page-81.htm].
STUTZMANN, Dominique, « Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? », Franz Fischer, Christiane Fritze, Georg Vogeler, (éds.). *Kodikologie und Paläographie im digitalen Zeitalter 2 = Codicology and Palaeography in the Digital Age 2*, 2011, (« Schriften des Instituts für Dokumentologie und Editorik »), p. 247-277, [En ligne : https://shs.hal.science/halshs-00596970].
SVEN, Najem-Meyer et MATTEO, Romanello, « Page Layout Analysis of Text-heavy Historical Documents: a Comparison of Textual and Visual Approaches », arXiv, 2022, [En ligne : http://arxiv.org/abs/2212.13924].
VITALI-ROSATI, Marcello, « Les chercheurs en SHS savent-ils écrire ? », *The Conversation*, 2018, [En ligne : http://theconversation.com/les-chercheurs-en-shs-savent-ils-ecrire-93024].

II-Modalisation des données en TEI

ALMAS, Bridget, CAYLESS, Hugh, CLÉRICE, Thibault, [et al.], « Distributed Text Services (DTS): A Community-Built API to Publish and Consume Text Collections as Linked Data », *Journal of the Text Encoding Initiative*, janvier 2023, [En ligne : https://journals.openedition.org/jtei/4352].
BAŃSKI, Piotr, « Why TEI stand-off annotation doesn't quite work », 2010, [En ligne : https://www.balisage.net/Proceedings/vol5/print/Banski01/BalisageVol5-Banski01.html].
BUARD, Pierre-Yves, *Modélisation des sources anciennes et édition numérique*, thèse de doctorat, Université de Caen, 2015, [En ligne : https://hal.science/tel-01279385].

BURNARD, Lou, « What is TEI Conformance, and Why Should You Care? », *Journal of the Text Encoding Initiative*, 2020, [En ligne : https://journals.openedition.org/jtei/1777].
BURNARD, Lou, SCHÖCH, Christof et ODEBRECHT, Carolin, « In search of comity: TEI for distant reading », *Journal of the Text Encoding Initiative*, mars 2021, [En ligne : https://journals.openedition.org/jtei/3500].
DEROSE, Steven J., « Markup Overlap: A Review and a Horse », *Proceedings of the Extreme Markup Languages® Conference*, Montréal, 2004, [En ligne : http://www.mulberrytech.com/Extreme/Proceedings/html/2004/DeRose01/EML2004DeRose01.html].
DEROSE, Steven J., DURAND, David G., MYLONAS, Elli, [et al.], « What is text, really? », *Journal of Computing in Higher Education,* vol. 1 / 2, décembre 1990, p. 3-26, [En ligne : https://doi.org/10.1007/BF02941632].
DUVAL, Frédéric, « Pour des éditions numériques critiques. L'exemple des textes français », *Le texte à l'épreuve du numérique*, vol. 73, Médiévales, Saint-Denis, France, Presses universitaires de Vincennes, Université Paris VIII, 2017, (« Médiévales »), p. 13-30.
GALLERON, Ioana, DEMONET, Marie-Luce, MEYNARD, Cécile, [et al.], « Consortium Cahier – Corpus d'auteurs pour les humanités : informatication, édition, recherche », 2022, [En ligne : https://cahier.hypotheses.org/files/2018/12/guide_edition_EVENT_2018_1.pdf].
« Guidelines for Editors of Scholarly Editions » [En ligne : https://www.mla.org/Resources/Guidelines-and-Data/Reports-and-Professional-Guidelines/Guidelines-for-Editors-of-Scholarly-Editions], consulté le 22 décembre 2023.
HENNY, Jean-Michel, « Politique numérique (questions 33-42) », *L'édition scientifique institutionnelle en France: État des lieux, matière à réflexions, recommandation*, Association des Editeurs De la Recherche et de l'Enseignement Supérieur, 2015, p. 85-92.
IDMHAND, Fatiha et GALLERON, Ioana, « Guide pour la FAIRisation des données des corpus d'auteurs [Groupe de travail Date-Cahier] », 2020, [En ligne : https://cahier.hypotheses.org/guides/guide-pour-la-fairisation-des-donnees-des-corpus-dauteurs].
KUMAR, Amit, SCHREIBMAN, Susan, ARNEIL, Stewart, [et al.], « <teiPublisher>: A Repository Management System for TEI Documents », Literary and Linguistic Computing, vol. 20 / 1, mars 2005, p. 117-132, [En ligne : https://doi.org/10.1093/llc/fqh047].
LAVRENTIEV, Alexei et GUILLOT-BARBANCE, Céline, « La BFM 2022 : un corpus pour les recherches diachroniques en français médiéval et au-delà », *Corpus, Bases, corpus et langage - UMR 6039*, janvier 2024, [En ligne : https://journals.openedition.org/corpus/8601].
MCGANN, Jerome, « Dante Gabriel Rossetti and the Betrayal of Truth », *Victorian Poetry*, vol. 26 / 4, West Virginia University Press, 1988, p. 339-361, [En ligne : https://www.jstor.org/stable/40002201].
PIERAZZO, Elena, « What future for digital scholarly editions? From Haute Couture to Prêt-à-Porter », *International Journal of Digital Humanities,* vol. 1 / 2, juillet 2019, p. 209-220, [En ligne : https://doi.org/10.1007/s42803-019-00019-3].
RENEAR, Allen H., « Text Encoding », Susan Schreibman, Raymond Georges Siemens, John M. Unsworth, (éds.). *A companion to digital humanities*, Malden, MA, Blackwell Publishing, 2004, p. 218-270.
RENEAR, Allen, MYLONAS, Elli et DURAND, David G., « Refining Our Notion of What Text Really Is: The Problem of Overlapping Hierarchies », *Research in Humanities Computing*, 1996.
ROBERTSON, Michael, « The Walt Whitman Archive », *Journal of American History*, vol. 99 / 3, décembre 2012, p. 1019-1020, [En ligne : https://doi.org/10.1093/jahist/jas486].

SINCLAIR, Stéfan et ROCKWELL, Geoffrey, « Chapitre 12. Les potentialités du texte numérique », in Marcello Vitali-Rosati, Michael E. Sinatra, (éds.). *Pratiques de l'édition numérique*, Presses de l'Université de Montréal, 2014, (« Parcours numérique »), p. 191-204, [En ligne : https://books.openedition.org/pum/337].
TURSKA, Magdalena, CUMMINGS, James et RAHTZ, Sebastian, « Challenging the Myth of Presentation in Digital Editions », *Journal of the Text Encoding Initiative*, septembre 2016, [En ligne : https://journals.openedition.org/jtei/1453].

III-Enrichissement du texte

CAMPS, Jean-Baptiste, CLÉRICE, Thibault, DUVAL, Frédéric, [et al.], « Corpus and Models for Lemmatisation and POS-tagging of Old French », *Journal of Data Mining and Digital Humanities*, 2021, [En ligne : http://arxiv.org/abs/2109.11442].
CAMPS, Jean-Baptiste, ING, Lucence et SPADINI, Elena, « Collating Medieval Vernacular Texts: Aligning Witnesses, Classifying Variants », *DH2019 Digital Humanities Conference*, Utrecht, 2019, [En ligne : https://dh-abstracts.library.cmu.edu/works/10074].
CLÉRICE, Thibault, « Deucalion Latin Lemmatizer », *Zenodo*, 2019, [En ligne : https://zenodo.org/record/2707476#.XO14Ay3pNTZ].
CLÉRICE, Thibault, « Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin *», Journal of Data Mining & Digital Humanities*, vol. 2020, Episciences.org, avril 2020, [En ligne : https://jdmdh.episciences.org/6264/pdf].
CLÉRICE, Thibault, « Pie Extended, an extension for Pie with pre-processing and post-processing », Zenodo, 2020, [En ligne : https://doi.org/10.5281/zenodo.3883589].
CLÉRICE, Thibault, JOLIVET, Vincent et PILLA, Julien, « Building infrastructure for annotating medieval, classical and pre-orthographic languages: the Pyrrha ecosystem », *Digital Humanities 2022*, Jul 2022, Tokyo, Japan.
DUPONT, Yoann, *La structuration dans les entités nommées*, thèse de doctorat, Sorbonne Paris Cité, 2017, [En ligne : https://theses.fr/2017USPCA100].
EDER, Maciej, « Mind your corpus: systematic errors in authorship attribution », Literary and Linguistic Computing, vol. 28 / 4, décembre 2013, p. 603-614, [En ligne : https://doi.org/10.1093/llc/fqt039].
ENRIQUE MANJAVACAS, THIBAULT CLÉRICE et MIKE KESTEMONT, « Emanjavacas/pie v0.2.3 », Zenodo, 2019, [En ligne : https://zenodo.org/record/2654987#.XP-nvC3M0fM].
FOPPIANO, Luca et ROMARY, Laurent, « Entity-fishing: a DARIAH entity recognition and disambiguation service », *Journal of the Japanese Association for Digital Humanities*, vol. 5 / 1, novembre 2020, p. 22, [En ligne : https://inria.hal.science/hal-01812100].
GILLE LEVENSON, Matthias*, Le Regimiento de los prínçipes et sa glose : étude et édition numérique de la partie sur le gouvernement de la cité en temps de guerre (III, 3)*, thèse de doctorat, École Normale Supérieure de Lyon, 2023.
GILLE LEVENSON, Matthias, « TeiCollator : une chaîne de traitement ecdotique semi-automatisée », *XXX^e Congrès International de Linguistique et Philologie Romane*, 2022, [En ligne : https://hal.science/hal-03715059].
GILLE LEVENSON, Matthias, BRISVILLE-FERTIN, Olivier, DÍEZ YÁÑEZ, Maria, [et al.], « Construcción de un corpus de evaluación de la anotación léxico-gramatical del castellano medieval (siglos 13-15) », Santiago de Compostela, 2021.
GUILLOT, Céline, PRÉVOST, Sophie et LAVRENTIEV, Alexei, « Principes d'annotation Cattex09 », [En ligne : http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_principes_2.0.pdf].

HEIDEN, Serge, MAGUÉ, Jean-Philippe et PINCEMIN, Bénédicte, « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement », vol. 2 / 3, Edizioni Universitarie di Lettere Economia Diritto, 2010, p. 1021, [En ligne : https://halshs.archives-ouvertes.fr/halshs-00549779].

HONNIBAL, Matthew, MONTANI, Ines, VAN LANDEGHEM, Sofie, [et al.], « spaCy: Industrial-strength Natural Language Processing in Python », 2020.

JOHNSON, Kyle P., BURNS, Patrick J., STEWART, John, [et al.], « The Classical Language Toolkit: An NLP Framework for Pre-Modern Languages », *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, éds. Heng Ji, Jong C. Park et Rui Xia, Online, Association for Computational Linguistics, 2021, p. 20-29, [En ligne : https://aclanthology.org/2021.acl-demo.3].

LEECH, Geoffrey et WILSON, Andrew, « Standards for Tagsets », in Hans Van Halteren, (éd.). *Syntactic Wordclass Tagging*, Springer Netherlands, 1999, p. 55-80, [En ligne : https://doi.org/10.1007/978-94-015-9273-4_5].

MANJAVACAS, Enrique, KÁDÁR, Ákos et KESTEMONT, Mike, « Improving Lemmatization of Non-Standard Languages with Joint Learning *», Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota, Association for Computational Linguistics*, 2019, p. 1493-1503, [En ligne : http://aclweb.org/anthology/N19-1153].

PADRÓ, Lluís et STANILOVSKY, Evgeny, « FreeLing 3.0: Towards Wider Multilinguality », *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, ELRA, 2012.

PRÉVOST, Sophie, GUILLOT, Céline, LAVRENTIEV, Alexei, [et al.], « Jeu d'étiquettes morphosyntaxiques CATTEX2009 », *Technical report*, École normale supérieure de Lyon, Lyon, 2013, [En ligne : https://palafra.github.io/assets/palafrafro_Cattex2009_2.0.pdf].

QI, Peng, ZHANG, Yuhao, ZHANG, Yuhui, [et al.], « Stanza: A Python Natural Language Processing Toolkit for Many Human Languages *», Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, éds. Asli Celikyilmaz et Tsung-Hsien Wen, Online, Association for Computational Linguistics, 2020, p. 101-108, [En ligne : https://aclanthology.org/2020.acl-demos.14].

SOUVAY, Gilles, « LGeRM: un outil d'aide à lemmatisation du français médiéval *», 18th International Conference on Historical Linguistics*, 2007.

STERN, Rosa, *Identification automatique d'entités pour l'enrichissement de contenus textuels*, thèse de doctorat, Université Paris-Diderot - Paris VII, 2013, [En ligne : https://theses.hal.science/tel-00939420].

TJONG KIM SANG, Erik F. et DE MEULDER, Fien, « Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition », *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL*, 2003, p. 142-147, [En ligne : https://aclanthology.org/W03-0419].

IV-Chaîne d'acquisition

BÉNIÈRE, Sarah, CHIFFOLEAU, Floriane et SCHEITHAUER, Hugo, « Streamlining the Creation of Holocaust-related Digital Editions with Automatic Tools », 2024, [En ligne : https://inria.hal.science/hal-04594190].

BURNS, Patrick J., « Building a Text Analysis Pipeline for Classical Languages », *Building a Text Analysis Pipeline for Classical Languages*, De Gruyter Saur, 2019, p. 159-176, [En ligne : https://www.degruyter.com/document/doi/10.1515/9783110599572-010/html].

CHAGUÉ, Alix et CHIFFOLEAU, Floriane, « An accessible and transparent pipeline for publishing historical egodocuments », *WPIP21 - What's Past is Prologue: The NewsEye International Conference,* Virtual, Austria, 2021, [En ligne : https://hal.science/hal-03173038].

CHAGUÉ, Alix et SCHEITHAUER, Hugo, « LEPIDEMO, a Pipeline Demonstrator for LECTAUREP to go from eScriptorium to TEI-Publisher », 2021, [En ligne : https://github.com/lectaurep/lepidemo].

CHIFFOLEAU, Floriane, « Keeping it open: a TEI-based publication pipeline for historical documents », 2021, [En ligne : https://hal.science/hal-04357295].

CHIFFOLEAU, Floriane et BAILLOT, Anne, « Le projet DAHN : une pipeline pour l'édition numérique de documents d'archives », 2022, [En ligne : https://hal.science/hal-03628094].

CHIFFOLEAU, Floriane, BAILLOT, Anne et OVIDE, Manon, « A TEI-based publication pipeline for historical egodocuments - the DAHN project », Next Gen TEI, 2021 - TEI Conference and Members' Meeting, Virtual, United States, 2021, [En ligne : https://hal.science/hal-03451421].

CHIFFOLEAU, Floriane et SCHEITHAUER, Hugo, « Leveraging EHRI Online Editions for training automated edition tools », 2024, [En ligne : https://inria.hal.science/hal-04594084].

CHRIST, Oliver, « A modular and flexible architecture for an integrated corpus query system », COMPLEX'94, vol. 15, 1994, [En ligne : ftp://ftp.ims.uni-stuttgart.de/pub/projekte/tc/christ:complex-submission.ps.gz].

« MaX · MRSH · Maison de la Recherche en Sciences humaines » [En ligne : https://mrsh.unicaen.fr/max/]. Consulté le11 mars 2024.

PINCHE, Ariane, GABAY, Simon, CHRISTENSEN, Kelly, [et al.], « Gallic(orpor)a: Extraction, annotation et diffusion de l'information textuelle et visuelle en diachronie longue », Bibliothèque nationale de France, 2022, [En ligne : https://hal.science/hal-04024750].

« Pipeline for digital scholarly editions », *DiScholEd,* 2023, [En ligne : https://github.com/DiScholEd/pipeline-digital-scholarly-editions].

« PluCo · MRSH · Maison de la Recherche en Sciences humaines » [En ligne : https://mrsh.unicaen.fr/pluco/]. Consulté le 11 mars 2024.

PORTE, Guillaume, ROGER, Julia, ALLAIN, Anne-Laure, [et al.], « « Diverses chaînes de production éditoriale » (janvier 2022). [Vidéo]. Canal-U. », [En ligne : https://e-diffusion.uha.fr/video/4628-diverses-chaines-de-production-editoriale-janvier-2022/]. Consulté le 11 mars 2024.

SAFARYAN, Anahit, ANDREWS, Tara L. et ATAYAN, Tatevik, « Continuous Integration Systems for Critical Edition: The Chronicle of Matthew of Edessa », Zenodo, 2019, [En ligne : https://zenodo.org/records/5498352].

SERETAN, violetta, « Sharing the Experience: Workflows for the Digital Humanities », *Digital Critical Edition of Apocryphal Literature: Sharing the Pipeline,* DARIAH-Campus, juin 2020, [En ligne : https://campus.dariah.eu/en/resource/events/sharing-the-experience-workflows-for-the-digital-humanities].