

ÉCOLE NATIONALE DES CHARTES
UNIVERSITÉ PARIS, SCIENCES & LETTRES

Joël Féral

licencié ès lettres
diplômé de master

Des *Tables des noms* du Sénat aux données structurées

Expérimentation et évaluation d'une chaîne
de traitement avec des grands modèles de
langue

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

2025

Résumé

Résumé du mémoire en français. Cette page ne doit pas dépasser une page.

Mots-clés : Journal Officiel ; Sénat ; Génération structurée ; LLM ; Mezanno ; séparés par des points-virgules.

Informations bibliographiques : Prénom Nom, *Titre du mémoire. Sous-titre du mémoire*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. [Noms des directeurs.trices], École nationale des chartes, 20245.

Remerciements

M^{ES} remerciements vont tout d'abord à ...

Bibliographie

Sources primaires

Archive nationales, fonds du Premier ministre ; Secrétariat général du Gouvernement ; Direction des Journaux officiels, 1881, URL : https://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?consIr=&frontIr=&optionFullText=&fullText=&defaultResultPerPage=&irId=FRAN_IR_014025&formCaller=GENERALISTE&gotoArchivesNums=false&auSeinIR=false&details=false&page=&udId= (visité le 26/08/2025).

Table annuelle du Journal officiel de la République française Lois et décrets, EN, 1931, URL : <https://gallica.bnf.fr/ark:/12148/bpt6k65430703> (visité le 26/08/2025).

Méthodologie historique

BERT (Jean-François), « L'art de la mise en fiche et ses métamorphoses modernes », *Genesis. Manuscrits – Recherche – Invention*–55 (déc. 2022), ISBN : 9791023107449 Number : 55 Publisher : Sigales, p. 57-70, DOI : 10.4000/genesis.7339.

BOUTIER (Jean), « L'usage historien des archives », dans *Corpus, sources et archives*, Code : Corpus, sources et archives, Tunis, 2001 (Études et travaux de l'IRMC), p. 9-22, DOI : 10.4000/books.irmc.776.

DAUMARD (Adeline) et FURET (François), « Méthodes de l'Histoire sociale : les Archives notariales et la Mécanographie », *Annales*, 14–4 (1959), Company : Persée - Portail des revues scientifiques en SHS Distributor : Persée - Portail des revues scientifiques en SHS Institution : Persée - Portail des revues scientifiques en SHS Label : Persée - Portail des revues scientifiques en SHS Publisher : EHESS, p. 676-693, DOI : 10.3406/ahess.1959.2865.

LAMASSÉ (Stéphane) et RYGIEL (Philippe), « Nouvelles frontières de l'historien », *Revue Sciences/Lettres*–2 (févr. 2014), Publisher : École normale supérieure, DOI : 10.4000/rs1.411.

LEMERCIER (Claire) et ZALC (Claire), *Méthodes quantitatives pour l'historien*, ISSN : 0993-7625, 2008, DOI : 10.3917/dec.lemer.2008.01.

MÜLLER (Bertrand), « De la formation d'un concept à l'invention d'une tradition : les avatars de l'histoire sérielle », dans *Un historien dans ses lendemains : Pierre Chaunu*, dir. Denis Crouzet et Alain Hugon, Code : Un historien dans ses lendemains : Pierre Chaunu, Caen, 2021 (Symposia), DOI : 10.4000/books.puc.15437.

PROST (Antoine), *Douze Leçons sur l'histoire*, Points, Publisher : Le Seuil, Paris, 2010, URL : <https://shs.cairn.info/douze-lecons-sur-l-histoire--9782757820643> (visité le 22/08/2025).

VAN CAMPENHOUDT (Marc), « Une norme de dépouillement terminologique en langue française », *Equivalences*, 21–1 (1992), Publisher : Persée - Portail des revues scientifiques en SHS, p. 121-136, DOI : 10.3406/equiv.1992.1147.

Droit et histoire parlementaires

BARTHÉLEMY (Joseph), *Essai sur le travail parlementaire et le système des commissions*, Country : FR 26 cm., Paris, 1934 (Bibliothèque de l'Institut international de droit public, 5).

Bibliothèque de l'Assemblée nationale (Paris), fr, Page Version ID : 228055770, août 2025, URL : [https://fr.wikipedia.org/w/index.php?title=Biblioth%C3%A8que_de_l%27Assembl%C3%A9e_nationale_\(Paris\)&oldid=228055770#Vers_une_biblioth%C3%A8que_num%C3%A9rique](https://fr.wikipedia.org/w/index.php?title=Biblioth%C3%A8que_de_l%27Assembl%C3%A9e_nationale_(Paris)&oldid=228055770#Vers_une_biblioth%C3%A8que_num%C3%A9rique) (visité le 25/08/2025).

BONNARD (Roger (1878-1944)), *Les règlements des assemblées législatives de la France depuis 1789 (notices historiques et textes)*, fre, Publisher : Paris : Société anonyme du Recueil Sirey, 1926, 1926, URL : <https://www.babordnum.fr/items/show/572> (visité le 26/08/2025).

CONIEZ (Hugo), « L'Invention du compte rendu intégral des débats en France (1789-1848) », *Parlement[s] / Revue d'histoire politique*, 14–2 (déc. 2010), Publisher : L'Harmattan Section : Histoire, p. 146-158, DOI : 10.3917/parl.014.0146.

GARDEY (Delphine), « Scriptes de la démocratie : les sténographes et rédacteurs des débats (1848–2005) », *Sociologie du travail*, 52–2 (juin 2010), Publisher : Association pour le développement de la sociologie du travail, p. 195-211, DOI : 10.4000/sdt.13695.

GARGUILLO (Violaine), *Portails et guides thématiques : Publications officielles France : JORF Débats et documents du Sénat*, fr, URL : <https://bnf.libguides.com/c.php?g=659907&p=4659964> (visité le 25/08/2025).

LEMESLE (Hélène), « Apprendre le travail parlementaire et construire la séparation des pouvoirs dans les années 1870 », *Revue d'histoire du XIXe siècle. Société d'histoire de la révolution de 1848 et des révolutions du XIXe siècle*–35 (déc. 2007), Publisher : Société d'histoire de la révolution de 1848, p. 125-139, DOI : 10.4000/rh19.2132.

Les commissions générales de 1921-1940, fr, URL : <https://www.senat.fr/connaitre-le-senat/lhistoire-du-senat/les-proces-verbaux-des-commissions/les-commissions-generales-de-1921-1940.html> (visité le 27/08/2025).

Les travaux du Sénat de la Troisième République, fr, URL : <https://www.senat.fr/connaitre-le-senat/lhistoire-du-senat/les-travaux-du-senat-de-la-troisieme-republique.html> (visité le 21/08/2025).

MOREL (Benjamin), *Le parlement, temple de la République. De 1789 à nos jours*, Passés composés, Paris, 2024.

PIERRE (Eugene), *Traité de droit politique électoral et parlementaire. Supplément (5e édition complétée par des références au Supplément de 1919) / par Eugène Pierre,...* 1924, URL : <https://gallica.bnf.fr/ark:/12148/bpt6k6228759r> (visité le 26/08/2025).

POUDRA (Jules (1829-1884) Auteur du texte) et PIERRE (Eugène (1848-1925) Auteur du texte), *Traité pratique de droit parlementaire / par Jules Poudra,... et Eugène Pierre,...* 1878, URL : <https://gallica.bnf.fr/ark:/12148/bpt6k58651348> (visité le 26/08/2025).

SAUDRAIS (Hélène), « Aux sources de la loi, les archives parlementaires (XIXe-XXe siècles) », *Revue française de droit constitutionnel*, 101–1 (avr. 2015), Publisher : Presses Universitaires de France Section : Histoire, p. 165-175, DOI : 10.3917/rfdc.101.0165.

Institutions et politiques patrimoniales

BÉQUET (Gaëlle), « La bibliothèque numérique à la Bibliothèque de France : de l'objet-valise à l'objet-frontière (1988-1993) », dans *Trois bibliothèques européennes face à Google : Aux origines de la bibliothèque numérique (1990-2010)*, Code : Trois bibliothèques européennes face à Google : Aux origines de la bibliothèque numérique (1990-2010), Paris, 2015 (Mémoires et documents de l'École des chartes), p. 51-85, DOI : 10.4000/books.enc.14133.

BERMÈS (Emmanuelle), *De l'écran à l'émotion : quand le numérique devient patrimoine*, Country : FR 20 cm. PSL = Université Paris sciences & lettres. Bibliogr. p. 231-234. Index., Paris, 2024, URL : 893734 (visité le 22/08/2025).

BnF-Partenariats, fr, Page Version ID : 222207431, janv. 2025, URL : <https://fr.wikipedia.org/w/index.php?title=BnF-Partenariats&oldid=222207431> (visité le 25/08/2025).

CARON (Bertrand), « Formats de données pour la préservation à long terme : la politique de la BnF » () .

« La numérisation concertée de corpus d'imprimés. Etat des lieux des programmes et des partenaires. Décembre 2014 » (, 2014).

RAUTENBERG (Michel), « Les chemins croisés du patrimoine et de l'imaginaire », dans *L'imaginaire patrimonial : Figures de l'urbanité contemporaine*, Code : L'imaginaire patrimonial : Figures de l'urbanité contemporaine, Rennes, 2024 (Essais), p. 13-52, DOI : 10.4000/13ipp.

RUIZ (Émilien), *Accéder aux numérisations du Journal officiel de la République française, de 1871 à nos jours*, fr-FR, janv. 2013, URL : <https://boiteaoutils.info/2013/01/accéder-aux-numérisations-du-journal/> (visité le 19/08/2025).

WAGNEUR (Jean-Didier), « Gallica : la bibliothèque électronique de la BnF : quel accès pour les personnes handicapées visuelles ? », dans *Bibliothèques et publics handicapés visuels*, Code : Bibliothèques et publics handicapés visuels, Paris, 2002 (Paroles en réseau), DOI : 10.4000/books.bibpompidou.1500.

Humanités numériques

ABADIE (N), BACIOCCHI (S), CARLINET (E), CHAZALON (J), CRISTOFOLI (P), DUMÉNIEU (B), PERRET (J) et TUAL (S), « Approche du projet SoDUCo » () .

CLAVERT (Frédéric), « Une histoire par les données ? Le futur très proche de l'histoire des relations internationales », *Bulletin de l'Institut Pierre Renouvin*, 44–2 (nov. 2016), Publisher : UMR Sirice Section : Histoire, p. 119-130, DOI : 10.3917/bipr1.044.0119.

— *Le goût de l'API / Le goût de l'archive à l'ère numérique*, fr-FR, URL : <https://gout-numerique.net/table-of-contents/archives-nees-numeriques/gout-api> (visité le 22/08/2025).

DENIS (Jérôme) et GOËTA (Samuel), « Les facettes de l'Open Data : émergence, fondements et travail en coulisses », dans *Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus*, dir. Pierre-Michel Menger et Simon Paye, Code : Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus, Paris, 2017 (Conférences), p. 121-138, DOI : 10.4000/books.cdf.5005.

DEROSE (Steven J.), DURAND (David G.), MYLONAS (Elli) et RENEAR (Allen H.), « What is text, really ? », *Journal of Computing in Higher Education*, 1–2 (déc. 1990), p. 3-26, DOI : 10.1007/BF02941632.

DRUCKER (Johanna), *Visualisation. L'interprétation modélisante*, B42, Paris, 2020.

KNUTSEN (Gunnar W.), « Alimenter des bases de données grâce à l'intelligence artificielle », *Histoire & mesure*, XXXIX–2 (déc. 2024), ISBN : 9782713233685 Number : 2 Publisher : Éditions de l'EHESS, p. 99-116, DOI : 10.4000/140kk.

LAMASSÉ (Stéphane) et RYGIEL (Philippe), « Nouvelles frontières de l'historien », *Revue Sciences/Lettres-2* (févr. 2014), Publisher : École normale supérieure, DOI : 10.4000/rs1.411.

LEBRETON (Fanny), « Vers l'ouverture et l'exploration des débats parlementaires : étude d'une méthodologie de structuration et d'enrichissement automatique des données. L'exemple des débats à la Chambre des députés durant la Ve législature de la IIIe République (1889-1893) » (, oct. 2022), p. xv, URL : <https://dumas.ccsd.cnrs.fr/dumas-04538872> (visité le 23/08/2025).

MANUEFIG, *Numériser ce n'est pas éditer (2)*, fr-FR, janv. 2005, URL : <https://figoblog.org/2005/01/18/519/> (visité le 20/08/2025).

POUBLANC (Sébastien) et MARQUÉ (Nicolas), « Introduction au dossier « Historien · nes et numérique : pratiques et expériences vécues » », *Les Cahiers de Framespa. e-STORIA-42* (juill. 2023), Publisher : UMR 5136 – FRAMESPA, DOI : 10.4000/framespa.14370.

PUREN (Marie), *Digital Humanities in the TIME-US Project : Richness and Contribution of Interdisciplinary Methods for Labour History*, arXiv :2410.14222 [cs], oct. 2024, DOI : 10.48550/arXiv.2410.14222.

ROMARY (Laurent) et LOPEZ (Patrice), « GROBID - Information Extraction from Scientific Publications », *ERCIM News*, Scientific Data Sharing and Re-use 100 (janv. 2015), Publisher : ERCIM, URL : <https://inria.hal.science/hal-01673305> (visité le 28/08/2025).

RYGIEL (Philippe), *Historien à l'âge numérique*, Code : Historien à l'âge numérique Publication Title : Historien à l'âge numérique Reporter : Historien à l'âge numérique Series Title : Papiers, Villeurbanne, 2017 (Papiers), URL : <https://books.openedition.org/pressesensib/6303> (visité le 19/08/2025).

TROMPETTE (Pascale) et VINCK (Dominique), « Retour sur la notion d'objet-frontière », *Revue d'anthropologie des connaissances*, 31–1 (juin 2009), Publisher : S.A.C., p. 5-27, DOI : 10.3917/rac.006.0005.

Images et numérisation

BEN SALAH (Ahmed), DUPLOUY (Laurent) et MOREUX (Jean-Philippe), « The digital documents quality control workflow at the BnF (operation, issue, improvement) », *Archiving Conference* (, janv. 2013).

CASEY (R.G.) et LECOLINET (E.), « A survey of methods and strategies in character segmentation », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18–7 (juill. 1996), p. 690-706, DOI : 10.1109/34.506792.

CHIRON (Guillaume), MOREUX (Jean-Philippe), DOUCET (Antoine), COUSTATY (Mickaël) et VISANI (Muriel), « Erreurs OCR et biais d'indexation : impact sur les usages », dans *17ème conférence Extraction et Gestion des Connaissances, Atelier Journalisme Computational*, Grenoble, France, 2017, p. 69-73, URL : <https://hal.science/hal-01455763> (visité le 22/08/2025).

Thierry Claerr et Isabelle Westeel (éd.), *Numériser et mettre en ligne*, Code : Numériser et mettre en ligne Publication Title : Numériser et mettre en ligne Reporter : Numériser et mettre en ligne Series Title : La Boîte à outils, Villeurbanne, 2010 (La Boîte à outils), URL : <https://books.openedition.org/pressesenssib/414> (visité le 20/08/2025).

FLEISCHHACKER (David), GOEDERLE (Wolfgang) et KERN (Roman), *Improving OCR Quality in 19th Century Historical Documents Using a Combined Machine Learning Based Approach*, en, arXiv :2401.07787 [cs], janv. 2024, DOI : 10.48550/arXiv.2401.07787.

KISSLING (Benjamin), « Kraken - A Universal Text Recognizer for the Humanities », dans *Digital Humanities 2019*, Utrecht, Netherlands, 2019, DOI : 10.34894/Z9G2EX.

« La numérisation concertée de corpus d'imprimés. Etat des lieux des programmes et des partenaires. Décembre 2014 » (, 2014).

LEBERT (Marie), [gutenberg.org/cache/epub/27040/pg27040.txt](https://www.gutenberg.org/cache/epub/27040/pg27040.txt), URL : <https://www.gutenberg.org/cache/epub/27040/pg27040.txt> (visité le 17/08/2025).

Mistral OCR / Mistral AI, en, URL : <https://urlr.me/Cqyt9c> (visité le 21/08/2025).

NEUDECKER (Clemens), BAIERER (Konstantin), GERBER (Mike), CLAUSNER (Christian), ANTONACOPOULOS (Apostolos) et PLETSCHACHER (Stefan), « A survey of OCR evaluation tools and metrics », dans *The 6th International Workshop on Historical Document Imaging and Processing*, Lausanne Switzerland, 2021, p. 13-18, DOI : 10.1145/3476887.3476888.

Numérisation de masse : qualité et formats utilisés pour garantir la conservation, fr, URL : <https://www.bnf.fr/fr/numerisation-de-masse-qualite-et-formats-utilises-pour-garantir-la-conservation> (visité le 25/08/2025).

SMITH (R.), « An Overview of the Tesseract OCR Engine », dans *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, ISSN : 1520-5363, Curitiba, Parana, Brazil, 2007, p. 629-633, DOI : 10.1109/ICDAR.2007.4376991.

SMITH (Ray W.), « History of the Tesseract OCR engine : what worked and what didn't », dans ADS Bibcode : 2013SPIE.8658E..02S, 2013, t. 8658, p. 865802, DOI : 10.1117/12.2010051.

Traitements automatiques du langage

BROWN (Tom B.), MANN (Benjamin), RYDER (Nick), SUBBIAH (Melanie), KAPLAN (Jared), DHARIWAL (Prafulla), NEELAKANTAN (Arvind), SHYAM (Pranav), SASTRY (Girish), ASKELL (Amanda), *et al.*, *Language Models are Few-Shot Learners*, arXiv :2005.14165 [cs], juill. 2020, DOI : 10.48550/arXiv.2005.14165.

CHAZALON (Joseph) et CARLINET (Edwin), « Revisiting the Coco Panoptic Metric to Enable Visual and Qualitative Analysis of Historical Map Instance Segmentation », dir. Josep Lladós, Daniel Lopresti et Seiichi Uchida, *Document Analysis and Recognition – ICDAR 2021*, 12824 (2021), Series Title : Lecture Notes in Computer Science, p. 367-382, DOI : 10.1007/978-3-030-86337-1_25.

CHEN (Yunmo), GANTT (William), CHEN (Tongfei), WHITE (Aaron Steven) et DURME (Benjamin Van), *A Unified View of Evaluation Metrics for Structured Prediction*, arXiv :2310.13793 [cs], oct. 2023, DOI : 10.48550/arXiv.2310.13793.

DEVLIN (Jacob), CHANG (Ming-Wei), LEE (Kenton) et TOUTANOVA (Kristina), *BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv :1810.04805 [cs], mai 2019, DOI : 10.48550/arXiv.1810.04805.

FINKEL (Jenny Rose), GREAGER (Trond) et MANNING (Christopher), « Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling », dans *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, dir. Kevin Knight, Hwee Tou Ng et Kemal Oflazer, Ann Arbor, Michigan, 2005, p. 363-370, DOI : 10.3115/1219840.1219885.

GRAHAM (S. Scott), MAJDIK (Zoltan P.) et CLARK (Dave), « Methods for Extracting Relational Data from Unstructured Texts Prior to Network Visualization in Humanities Research », *Journal of Open Humanities Data*, 6–1 (nov. 2020), DOI : 10.5334/johd.21.

GRAVES (Alex), FERNANDEZ (Santiago), GOMEZ (Faustino) et SCHMIDHUBER (Jurgen), « Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks » () .

HUMPHRIES (Mark), LEDDY (Lianne C.), DOWNTON (Quinn), LEGACE (Meredith), McCONNELL (John), MURRAY (Isabella) et SPENCE (Elizabeth), *Unlocking the Archives : Using Large Language Models to Transcribe Handwritten Historical Documents*, arXiv :2411.03340 [cs], nov. 2024, DOI : 10.48550/arXiv.2411.03340.

PEYRÉ (Gabriel) et CUTURI (Marco), *Computational Optimal Transport*, en, arXiv :1803.00567 [stat], mars 2020, DOI : 10.48550/arXiv.1803.00567.

RADFORD (Alec), NARASIMHAN (Karthik), SALIMANS (Tim) et SUTSKEVER (Ilya), « Improving Language Understanding by Generative Pre-Training » () .

VASWANI (Ashish), SHAZER (Noam), PARMAR (Niki), USZKOREIT (Jakob), JONES (Llion), GOMEZ (Aidan N.), KAISER (Lukasz) et POLOSUKHIN (Illia), *Attention Is All You Need*, arXiv :1706.03762 [cs], août 2023, DOI : 10.48550/arXiv.1706.03762.

WEI (Jason), TAY (Yi), BOMMASANI (Rishi), RAFFEL (Colin), ZOPH (Barret), BORGEAUD (Sebastian), YOGATAMA (Dani), BOSMA (Maarten), ZHOU (Denny), METZLER (Donald), *et al.*, *Emergent Abilities of Large Language Models*, arXiv :2206.07682 [cs], oct. 2022, DOI : 10.48550/arXiv.2206.07682.

ZHAO (Wayne Xin), ZHOU (Kun), LI (Junyi), TANG (Tianyi), WANG (Xiaolei), HOU (Yupeng), MIN (Yingqian), ZHANG (Beichen), ZHANG (Junjie), DONG (Zican), *et al.*, *A Survey of Large Language Models*, en, arXiv :2303.18223 [cs], mars 2025, DOI : 10.48550/arXiv.2303.18223.

Philosophie de la technique et médialité

BERT (Jean-François), *Une histoire de la fiche érudite*, Code : Une histoire de la fiche érudite
 Publication Title : Une histoire de la fiche érudite Reporter : Une histoire de la fiche érudite Series Title : Papiers, Villeurbanne, 2017 (Papiers), URL : <https://books.openedition.org/pressesenssib/6211> (visité le 21/08/2025).

BOISSIER (Jean-Louis), *La relation comme forme*, Les Presses du Réel, Dijon, 2009.

BONTEMS (Vincent), « Sur la classification des objets techniques selon Simondon », *Artefact. Techniques, histoire et sciences humaines*–3 (mars 2016), ISBN : 9782271087539 Number : 3 Publisher : Presses universitaires du Midi, p. 183-198, DOI : 10.4000/artefact.8270.
 Jean-Marie Dallet et Bertrand Gervais (éd.), *Architectures de mémoire*, Code : Architectures de mémoire Publication Title : Architectures de mémoire Reporter : Architectures de mémoire Series Title : La Grande Collection ArTeC, Nanterre, 2019 (La Grande Collection ArTeC), URL : <https://books.openedition.org/pupo/25762> (visité le 20/08/2025).

DASTON (Lorraine) et GALISON (Peter), *Objectivité*, Les Presses du Réel, Dijon, 2012.

DEWEY (John), *La formation des valeurs*, La Découverte, Paris, 2011.

LEVI-STRAUSS (Claude), *La pensée sauvage*.

RENON (Anne-Lyse), *Design & sciences*, fr-FR, URL : <https://www.puv-editions.fr/ouvrage/design-sciences/> (visité le 22/08/2025).

SIMODON (Gilbert), *Du mode d'existence des objets techniques*, Aubier, 1958.

*Remuer du papier ne peut pas
être inutile.*

Bruno Latour

Introduction

Pour l'historien, les archives font figure centrale [Jean Boutier]. Pour répondre à une question de recherche, le chercheur va dépouiller, trier, classer, reclasser, recouper les documents ou les données issus de fonds ou de collections recueillies. Cette matière se constitue alors en corpus pour ajuster des questions ou charpenter des réponses. Dans l'épais documentaire s'esquisse des hypothèses, s'élaborent des solutions. L'information enfin extraite des cartons, des rayons ou des documents disponibles sur le Web peut être enfin convoquée, parfois après un laborieux travail de dépouillement. Avec un peu de chance, l'information s'offre presque toute prête : ainsi les lignes budgétaires de tel livre de compte de telle institution ; ainsi telles entrées de tels annuaires professionnels. Tantôt il faudra reproduire manuellement l'information qui a été dénichée feuille à feuille et parfois photographiée à la hâte ; tantôt, avec un peu plus de chance, simplement récupérer des jeux de données quasiment prêts à l'emploi et les traiter selon ce qu'exige les impératifs scientifiques. Une fois les données réunies en lignes et en tableaux, l'enquête ne fait que commencer : c'est qu'il s'agit de supposer des tendances, des points de comparaison ; de dégager aux événements des séries ou des structures explicatives et forger des réponses. Ainsi la "centralité de l'archive" – et des données – dans le travail de l'historien – hier Lucien Febvre, Prost.. – ou du sociologue – Durkheim – ou encore, pourquoi pas, de l'astronome et ses archives voulant restituer un passé aux trajectoires des comètes.

Derrière ces masses de documents sériels, une allure : ce sont des bases de données de papier. Elles contiennent des noms et des nombres, des métiers ou des adresses. Pour le chercheur, ces sources répétitives se prêtent volontiers à une traduction numérique pour faciliter et opérer des traitements plus systématiques en bénéficiant des capacités calculatoires de l'ordinateur. Egalement pour l'archiviste en charge de l'indexation des fonds, elles regorgent de noms qu'il est intéressant d'exposer, par exemple, à son public de généalogistes via les portails de recherche d'archives. Naturellement, il est tenté d'employer l'outil informatique pour déléguer – ou rendre possible – ce travail d'extraction et de structuration de l'information présente dans les masses documentaires. Les grands modèles de langue (les *LLM*) semblent pouvoir faire endosser aux ordinateurs ce labeur de traduction et de structuration

de l'information. Se dessinent alors des enjeux techniques et épistémologiques d'un passage – celui des documents « analogiques » aux données numériques – qu'il convient d'interroger.

Ce mémoire interroge la problématique de la traduction de ces corpus sériels en ensembles de données structurées, dans une perspective d'analyse historienne et archivistique. Historienne d'abord, car il s'agit d'explorer de "nouvelles frontières" disciplinaires impliquant la collaboration entre chercheurs en SHS, informaticiens et institutions patrimoniales en vue d'exploiter à des fins scientifiques des données issus de fonds numérisés ; archivistique ensuite, car la traduction des informations contenues dans les documents en données exploitables par des systèmes informatiques rejoignent les enjeux d'indexation et donc de valorisation d'ensembles documentaires.

Cette problématique de traduction de fonds sériels en données structurées exploitables pour l'analyse quantitative historienne ou pour la valorisation documentaire part d'un constat : nombre de publications administratives ou normatives — annuaires, lois, décrets, tables parlementaires — relèvent d'une production sérielle à forte teneur informationnelle mais échappent aux catégories sensibles habituellement mobilisées dans le rapport aux archives. Leur lecture manuelle est difficile, leur dépouillement peut être décourageant. On s'éloigne ici du "goût de l'archive" d'Arlette Farge qui dépeint une phénoménologie sensible de la source historique pour adopter une approche moins solipsiste de valorisation de documents "sans goût" – mais dont on aura restitué un pluriel. Ces fonds sériels – dont il est difficile de valoriser au même titre que de prestigieuses chartes médiévales ou de précises gravures scientifiques étant donné leur monotone prosaïsme – forment en effet une mémoire institutionnelle précieuse qu'il convient d'interroger dès lors qu'on parvient à les structurer et les croiser avec d'autres sources.

Si on n'arrête pas de louer depuis quelques décennies de nouveaux tournants dans la façon de travailler sur les sources grâce à l'outil numérique, il faut bien avouer que les derniers développements autour des grands modèles de langage accentuent un virage. La fouille de texte, la « lecture à distance » (*distance reading*), et tout ce qui implique l'extraction d'information sémantique, reposent sur une récente synergisation des techniques numériques : tout d'abord, la capacité à restituer un objet physique en image discrète (numérisation) ; de structurer et normaliser l'accès à des images dans des dépôts centralisés via des protocoles HTTP(API) ; de retranscrire de l'image l'information textuelle (OCR) ou des structures de pages (Document Layout Detection) ; et de la structuration sémantique *a posteriori* des entités nommées de façon semi-automatique – moyennement un apprentissage supervisé via l'annotation ou le moissonnage de la production scientifique en ligne, déjà numérisée. Cette synergie des traitements sur l'image numérique et des développement en linguistique computationnelle renouvelle « les frontières de l'historien » en ce sens que, pour peu que les

documents soient numérisées, on puisse automatiser le travail d'extraction de l'information présente dans les sources. La chaîne qui va de la numérisation du document à l'extraction de données sémantiquement structurées et son analyse via l'outil informatique forme un *système technique* (Simondon) et l'historien, le sociologue ou encore le statisticien s'y inscrivent pleinement.

Une précaution : la question technique qui participe de l'administration des nouvelles frontières de l'historien n'est pas seulement un moyen d'automatiser des tâches qui autrefois étaient plus manuelles : elle apporte dans ses rêts de nouveaux enjeux épistémologiques (les données ne se donnent pas, ce sont des *capta*, Drucker), des manières de réactiver des questions ou des réponses (Colingwood), des sources documentaires, des méthodes (impliquant un redision du travail) ou des façons de travailler (Ladurie, l'historien et l'ordinateur). Les questions techniques sont moins une affaire d'automatisation que de sensibilité des systèmes techniques à l'information et leur relation avec le travail humain (Simondon). De même, les instruments ne font pas que servir le travail intellectuel de façon neutre : ils sont aussi de la « théorie réifiée » (Bachelard). Se superposent ainsi le champ énonciatif des systèmes techniques et des disciplines scientifiques qui s'y inscrivent et expriment des connaissances *situées*. Les outils techniques promettent des méthodologies – lesquelles reposent sur de manière d'envisager la division d'un travail pour une tâche déterminée. Les *pipelines* de traitement de données – de l'image au texte structuré – se constituent à la fois comme outil et comme condensation *in silicium* des *a priori* épistémologiques. Les solutions techniques à des problèmes scientifiques sont une affaire de *design* (Anne-Lyse Renon), c'est-à-dire de stratégies épistémiques et valuatives.

Dès lors, comment articuler les enjeux historiographiques propres à ces corpus avec les nouvelles opportunités d'extraction et de structuration automatiques qu'offrent les techniques récentes et leur synergisation ? (OCR, modèles de langage, outils d'annotation) ? Comment construire une chaîne de traitement reproductible, capable de restituer la richesse de ces fonds tout en garantissant la qualité des données produites ? La question de la valuation des méthodes – c'est-à-dire de leur légitimité au regard de ce à quoi on tient savoir –, elle-même, semble dépendre des besoins : un archiviste n'a pas tout à fait les mêmes besoins qu'un historien – et tous les historiens n'ont pas les mêmes besoins. Pour le premier, l'exactitude est de mise ; pour le second qui adopte une approche statistique, des données fiables – c'est-à-dire probablement imparfaites – seront suffisantes pour effectuer des « pesées globales » historiques (Pierre Chaunu).

Ces questionnements auront été les miens durant mon stage à l'EPITA/BnF où j'ai eu l'occasion de travailler à la convergence des nouvelles opportunités de traitements automatiques et de la question de la légitimité épistémique de données produites par des systèmes

techniques, à partir de l'extraction d'informations sous forme structurée du Journal Officiel de 1931, dans une optique de préparation de l'analyse des discours historiques sous la IIIe République. [à développer]

Cette interrogation se décline ainsi selon ces trois axes :

- Que sont les sources sérielles et quels sont leurs apports spécifiques pour répondre à une question de recherche ? En quoi leur structure, leur cumulativité et leur forme normative permettent-elles des lectures nouvelles, notamment en lien avec les pratiques de gouvernement et les dispositifs de publicité du droit sous la IIIe République ? Pour aborder ce point, je me pencherai sur l'analyse des Tables du Sénat de 1931 car, dans une perspective de *reenactement* historien de l'activité parlementaire, elles donneraient un corps à cette problématique de traduction de sources sérielles en sources exploitables pour l'analyse.
- Quelles méthodes et quels outils permettent aujourd'hui d'en automatiser la structuration ? Comment construire un protocole d'extraction cohérent, tenant compte de la matérialité des documents (OCR, mise en page, bruit), des formats de sortie, et des finalités analytiques visées ?
- Comment évaluer la fiabilité des données ainsi produites et garantir leur légitimité scientifique ? Quels critères de qualité, de traçabilité et de transparence permettent de faire de ces résultats des objets d'enquête mobilisables par les historiens ?

A travers ces trois axes, qui constituent chacun une partie de ce mémoire, je veux donc répondre à la problématique du passage de l'information serielle non-structurée présente dans les sources, à la fois sur des aspects documentaires ; techniques (méthodes d'extraction) et valuatives (évaluation et épistémologie de l'évaluation).

Première partie

Des sources sérielles : les Tables
Annuelles du Sénat comme cas
d'usage

Chapitre 1

Le *Journal Officiel* : la parole, la main, la vue et le droit

Reencatement : ensauvagement de l'assemblée nationale. QUalité du débat parlementaire.

Ne pas faire une histoire du *Journal Officiel*, de ses origines ; mais une histoire dans sa position dans un énoncé à travers les façons de le classer.

Les sources sérielles : définition et enjeux

Les sources sérielles désignent des ensembles documentaires produits de manière régulière, cumulative et normalisée, qui permettent d'observer la répétition et l'évolution de phénomènes sociaux ou politiques dans le temps. Pierre Chaunu, qui en a popularisé l'usage dans les années 1960, les définissait comme des matériaux « à série longue », permettant une histoire quantitative de la conjoncture, qu'il s'agisse des prix, des registres paroissiaux ou des trafics maritimes. Dans la lignée de cette approche, Emmanuel Le Roy Ladurie a montré que la régularité et la masse de ces sources ouvraient la voie à une « histoire sérielle » fondée sur l'accumulation et le traitement statistique.

À l'ère numérique, des historiens comme Frédéric Clavert rappellent que ces sources se distinguent moins par leur nature que par leur mode de production : leur caractère répétitif, leur homogénéité formelle et leur relative standardisation les rendent propices à des opérations d'extraction et de structuration automatisées. En ce sens, elles ne livrent pas seulement des contenus, mais offrent un potentiel méthodologique, qui repose sur leur caractère cumulatif.

La sérialité documentaire ne doit cependant pas masquer la dimension éditoriale et institutionnelle de ces objets. Comme l'a montré Alain Desrosières pour les statistiques publiques, la constitution d'une série est toujours le résultat d'un choix social et politique : décider ce qui est compté, comment cela est classé, et pour quels usages. Appliqué aux sources parlementaires, ce constat invite à considérer que les tables du Journal Officiel ne sont pas de simples

instruments techniques : elles orientent la lecture, hiérarchisent les thèmes et construisent une représentation ordonnée de l'activité politique.

I. **Le *Journal Officiel* : entre publicité et promulgation, archives et documentation**

La création du *Journal Officiel de la République française* en 1869 s'inscrit dans une longue histoire des dispositifs de publicité de la loi. Héritier à la fois du *Moniteur universel* (1789–1869), qui transcrivait les débats parlementaires, et du *Bulletin des lois* (1793–1931), garant de la promulgation exécutoire des textes normatifs, le *Journal Officiel* cumule deux fonctions : informer le public des débats parlementaires et conférer force obligatoire aux lois par leur publication. Cette double vocation, attestée dès la Révolution française, fait de ce périodique un instrument essentiel de la transparence parlementaire et de l'effectivité juridique.

Son caractère sériel — une parution quasi quotidienne, au format stable, avec une organisation régulière des rubriques — constitue précisément ce qui en fait une source de premier plan pour l'historien. Sa régularité et sa cumulativité permettent de suivre à la trace l'activité parlementaire, mais elles posent aussi la question de la matérialité du texte publié : transcription fidèle ou reconstitution éditoriale ? Dès lors, le *Journal Officiel* ne doit pas être lu seulement comme un réceptacle de la parole parlementaire, mais comme une œuvre éditoriale, façonnée par les sténographes, rédacteurs et correcteurs qui traduisent l'oralité en texte écrit.

Archives ou documentation ? La place singulière du *Journal Officiel*

La nature archivistique du *Journal Officiel* demeure ambivalente. Juridiquement, il s'agit d'une publication soumise au dépôt légal, conservée à ce titre à la Bibliothèque nationale de France, à la bibliothèque des Assemblées et aux Archives nationales. Pourtant, son intégration dans la **série K des Archives départementales** interroge : peut-on considérer comme archives des documents produits à des fins éditoriales, vendus au public et destinés à être lus ?

L'analyse archivistique rappelle que les archives sont le « collatéral direct » d'une activité administrative, produites sans intention de publication. Or, le *Journal Officiel* est un objet éditorial dont la finalité première est précisément la publicité. Son inscription en série K — aux côtés des lois, ordonnances et arrêtés préfectoraux — illustre néanmoins la manière dont l'État a voulu garantir, dans chaque département, l'accès des citoyens aux textes normatifs. Loin d'être un simple effet du « respect des fonds », cette intégration témoigne d'une stratégie

politique de diffusion centralisée du droit.

Cette tension entre document édité et document d'archives souligne la spécificité des sources sérielles : elles se situent à la frontière entre mémoire administrative et communication publique, entre traces institutionnelles et dispositifs de légitimation.

II. Le *Journal Officiel* : un contexte technique et administratif (1921–1940)

Pour comprendre la matérialité de la source exploitée dans ce mémoire (les tables du Sénat de 1931), il faut rappeler que le *Journal Officiel* n'était pas seulement le produit d'une Chambre parlementaire, mais celui d'une organisation industrielle. Depuis 1880, la publication est assurée par la **Société anonyme coopérative de composition et d'impression des Journaux officiels (SACIJO)**, placée sous tutelle du ministère de l'Intérieur. Linotypistes, rotativistes, correcteurs et personnels administratifs concourent à sa fabrication quotidienne.

La période 1921–1940, bornée par l'acquisition de la *linotype Model 9* et l'interruption de 1940, offre un cadre technique relativement homogène. Elle correspond aussi à une stabilité des pratiques éditoriales, qui permet d'envisager une lecture sérielle. Les archives de la Direction des Journaux officiels (série 19840069 des AN) éclairent cette fabrique administrative : rapports budgétaires, organigrammes, correspondances de service. Elles révèlent un fonctionnement hybride, entre administration d'État et entreprise de presse, qui explique la diffusion massive et régulière de ces volumes.

III. Les « processus métier » de la publicité parlementaire à partir des Tables nominales : analyse des sources

Qualifier la chaîne de production parlementaire en termes de « processus métier » revient à cartographier l'ensemble des opérations qui transforment la parole politique en texte normatif publié. À la IIIe République, ce processus suit plusieurs étapes :

- **Délibérer** : débats oraux au Sénat et à la Chambre des députés, régis par les règlements de 1876, avec une organisation en bureaux et commissions.

Transcrire : sténographes et rédacteurs produisent les *comptes rendus in extenso** , qui

passent par un travail de révision avant impression. **Publier** : les textes sont édités dans le Journal Officiel* et diffusés par abonnement et dépôt légal.

- **Promulguer** : la publication confère force obligatoire aux lois votées, qui ne prennent effet qu'une fois rendues publiques.

Ce continuum — délibérer, transcrire, publier, promulguer — rend manifeste le rôle central du *Journal Officiel*. Il ne s'agit pas seulement d'un témoin documentaire, mais d'un maillon de l'effectivité du droit.

Chapitre 2

Les tables annuelles : des relations documentaires

I. Les tables dans l'environnement du *Journal Officiel*

À côté des livraisons quotidiennes du *Journal Officiel*, le dispositif documentaire de la Troisième République produit un ensemble d'outils de repérage et de cumul : index, tables et recueils annuels. Ces tables, organisées par Chambre et par type de document (séances, questions, interventions, lois, décrets, etc.), constituent un instrument de navigation à travers la masse documentaire accumulée. Elles offrent un second niveau de structuration, indispensable à l'exploitation d'un corpus qui, sans cela, serait pratiquement illisible dans son entier.

Dans ce sens, les tables ne sont pas de simples annexes, mais un élément constitutif du *Journal Officiel*. Leur publication témoigne d'une volonté de rendre praticable la lecture sérielle, en transformant un flot continu de débats en une matière consultable *a posteriori*. Elles permettent aux parlementaires, aux fonctionnaires et aux juristes, mais aussi aux journalistes et au public, de retrouver un débat, une loi ou un orateur dans un ensemble potentiellement infini de pages.

II. Forme et organisation des tables

La table annuelle se présente comme un volume imprimé, distinct des numéros quotidiens mais reprenant la même logique typographique de sobriété. La structuration est généralement alphabétique ou thématique, avec des entrées renvoyant à des numéros de séance ou de page du *Journal Officiel*. Ainsi, le chercheur y trouve à la fois :

- des index de noms (parlementaires, ministres, orateurs) ;

- des index de matières (projets de lois, sujets débattus, thèmes abordés) ;
- des références législatives (dates, intitulés, numéros de lois et décrets).

Cette composition apparemment simple reflète un travail complexe de collecte et de mise en ordre, qui engage des méthodes d'indexation encore largement manuelles dans les années 1930. Les tables matérialisent donc une double médiation : celle de la transcription sténographique, puis celle de la mise en indexation.

III. Informations sémantiques et usages

Les tables ne livrent pas seulement des renvois. Leur organisation alphabétique ou thématique suggère déjà une lecture orientée du corpus. En réordonnant les débats selon les sujets ou les personnes, elles produisent une représentation « secondaire » de l'activité parlementaire :

- **Pour l'historien**, elles permettent de cartographier les thèmes récurrents, d'identifier des trajectoires individuelles de parlementaires, ou encore de suivre la maturation d'une question dans le temps long.
- **Pour les juristes**, elles assurent un repérage efficace des textes normatifs, condition de la sécurité juridique.
- **Pour l'administration**, elles facilitent la réutilisation interne des débats et la circulation de l'information entre services.

En ce sens, les tables possèdent une valeur sémantique propre : elles ne sont pas de simples index, mais des instruments de catégorisation, qui hiérarchisent les contenus du *Journal Officiel* et leur confèrent une visibilité inégale.

Les tables comme « hub » intercorpus

Enfin, les tables établissent des liens entre différents ensembles documentaires. Elles ne se limitent pas aux seuls débats parlementaires, mais relient ceux-ci aux autres publications officielles et à des corpus complémentaires. Elles servent d'articulation entre :

les volumes quotidiens du Journal Officiel ; les recueils législatifs et réglementaires (par exemple le Bulletin des lois*) ;*

- les instruments internes des Chambres (procès-verbaux, rapports de commissions) ;
- les archives départementales (série K), où elles prennent place aux côtés d'autres formes de publicité administrative.

En occupant cette position nodale, les tables fonctionnent comme des « hubs documentaires » : elles permettent de passer d'un corpus à l'autre, et d'inscrire les débats dans l'écosystème plus large des pratiques de gouvernement.

Exemple : Les Tables du Sénat, année 1931

Le volume des *Tables annuelles du Sénat* pour l'année 1931 se présente sous la forme d'un in-octavo relié, composé de plusieurs centaines de pages. La typographie, sobre et régulière, reprend les conventions du *Journal Officiel* : colonnes étroites, numérotation continue, absence d'ornementation. L'ensemble se divise en sections distinctes, qui reflètent les usages concrets des lecteurs.

1. Index des orateurs

On y trouve une **liste alphabétique des sénateurs**, chaque nom suivi de références aux séances où ils sont intervenus. Par exemple :

Tardieu (André) : interventions p. 312, 457, 892.

Cet index permet de retracer rapidement la présence et l'activité d'un parlementaire sur une année complète. Pour l'historien, il offre une base sérielle pour mesurer la visibilité des élus et la fréquence de leur participation aux débats.

2. Table des matières thématiques

La deuxième section regroupe les débats par **matières** :

Finances publiques* : budget, impôts, emprunts. Affaires étrangères* : traités, conventions, mandats. Travail et questions sociales* : assurance chômage, législation ouvrière, retraites.

Chaque entrée renvoie à un numéro de séance du *Journal Officiel*. Ce classement thématique reflète une logique documentaire propre, qui diffère de l'ordre chronologique des séances : il met en valeur la récurrence des thèmes et facilite leur repérage transversal.

3. Références législatives et réglementaires

Enfin, les tables recensent les **lois votées et les décrets publiés** pendant l'année, assortis de leur date et de leur numéro. Ce registre, proche d'un répertoire législatif, assure le lien avec le *Bulletin des lois* et, par extension, avec l'ensemble de la législation nationale.

Analyse

Cet exemple illustre trois dimensions essentielles des tables :

- Leur **fonction instrumentale** : elles servent avant tout de guide, destiné à faciliter la recherche d'une information précise dans un corpus immense.
- Leur **valeur sémantique** : en proposant une catégorisation (par personnes, thèmes, textes), elles produisent une image de l'activité parlementaire qui n'est pas neutre, mais orientée par le mode d'indexation.
- Leur **rôle intercorpus** : en mettant en relation débats, interventions et textes normatifs, elles constituent un point de jonction entre la parole parlementaire et le droit promulgué.

Numérisation du *Journal Officiel* : entre politique documentaire, infrastructures techniques et souveraineté archivistique

Du côté de la BnF, avec le projet Gallica, c'est le « mode image » qui a été choisi : « la bibliothèque se range pour de bon du côté de la reproduction plutôt que de l'édition ».

La numérisation du *Journal Officiel de la République française* (J.O.) constitue un cas exemplaire des tensions de la datafication. Sa mise en ligne sur **Gallica**, la bibliothèque numérique de la BnF, n'est pas seulement le produit d'une opération technique (scanner, OCR, structurer) : elle relève d'une **politique documentaire explicite**, d'une hiérarchie patrimoniale et d'une réflexion sur la transparence démocratique.

Dès le lancement de Gallica dans les années 1990, les corpus officiels et juridiques ont été considérés comme prioritaires dans les programmes de numérisation. La BnF a défini une **politique de numérisation concertée** qui associe ses partenaires institutionnels (bibliothèques, archives, musées, mais aussi administrations parlementaires) et repose sur deux modalités principales :

- la **subvention**, qui permet à des institutions tierces de financer la numérisation de leurs fonds à condition que les résultats soient interopérables et intégrés dans Gallica ;

l'intégration directe dans le marché de numérisation de la BnF, réservée aux corpus volumineux ou emblématiques, comme le Journal Officiel, qui nécessitent une infrastructure robuste et centralisée ([BnF, Numérisation concertée de corpus imprimés, 2018](https://www.bnf.fr/sites/default/files/2018-11/numconcertee_impr_progr_partenaires.pdf?utm_source=chatgpt.com)).*

Le choix du J.O. s'explique à la fois par sa **valeur patrimoniale** (trace officielle de la vie normative de l'État depuis 1870), son **utilité sociale et démocratique** (garantir un accès transparent au droit), et par son **homogénéité formelle** qui facilite les opérations techniques (OCR, segmentation par rubriques, enrichissement par métadonnées). La mise en ligne du J.O. sur Gallica couvre aujourd'hui de larges pans de la Troisième et

de la Quatrième République, même si certaines années restent lacunaires ou peu visibles ([Boîte à Outils, 2013](https://boiteaoutils.info/2013/01/accéder-aux-numérisations-du-journal/?utm_source=chatgpt.com)).

Cette politique s'accompagne d'une **chaîne technique complexe** :

- numérisation en mode image (PDF, JPEG) ;
- reconnaissance optique de caractères (OCR), dont la qualité varie selon la typographie, l'état du papier ou la mise en page ;
- segmentation en unités documentaires (articles, décrets, rubriques) ;
- enrichissement par métadonnées, qui conditionne la découvrabilité dans les moteurs de recherche.

Ces étapes introduisent des biais : erreurs OCR qui faussent la recherche plein texte, segmentation parfois incomplète, normalisation qui gomme des variations de présentation. Le J.O. numérisé n'est donc pas un « miroir » de la source papier, mais un **objet reconfiguré** par la chaîne sociotechnique de la BnF.

Sénat et Assemblée nationale : des acteurs de la numérisation parlementaire

Le processus ne relève pas uniquement de la BnF. Les **deux chambres du Parlement** sont parties prenantes de cette politique de numérisation, en lien étroit avec Gallica.

- Le **Sénat** a engagé la numérisation de ses **Impressions parlementaires** (débats, annexes, rapports) couvrant la Troisième République. Plusieurs campagnes ont permis d'intégrer dans Gallica des volumes allant de 1876 à 1905, puis de 1910 à 1940, avec un travail en cours sur 1906–1909 ([Sénat.fr](https://www.senat.fr/connaitre-le-senat/lhistoire-du-senat/les-travaux-du-senat-de-la-troisieme-republique.html?utm_source=chatgpt.com)). L'**Assemblée nationale** a également engagé la numérisation de ses documents ([BoteOutils](https://boiteaoutils.info/2013/01/accéder-aux-numérisations-du-journal/?utm_source=chatgpt.com)).

Ces initiatives montrent que le Parlement n'est pas un simple producteur de données, mais aussi un **acteur documentaire** qui oriente la sélection, la structuration et la diffusion de ses propres archives. Le croisement entre Gallica, Retronews et les sites institutionnels illustre l'existence d'**écosystèmes documentaires pluriels**, qui médiatisent différemment un même corpus selon les publics visés (chercheurs, citoyens, journalistes).

Archives numérisées et archives nativement numériques : le cas du JORF

Il convient enfin de distinguer deux régimes d'archives :

- Les **archives numérisées**, comme les volumes historiques du J.O. : elles proviennent d'un support papier, transformé par une chaîne technique (numérisation, OCR, indexation). Leur fiabilité est conditionnée par la qualité des scans et des traitements automatiques, et leur diffusion par les choix de formats (PDF image, texte OCRisé, métadonnées).

Les archives nativement numériques, comme le Journal officiel de la République française* contemporain (JORF), désormais produit directement sous forme numérique, structuré en XML, interrogable via des bases de données et accessible via **Légifrance**. Ici, il n'y a pas de passage par l'OCR : les textes sont disponibles en clair, immédiatement exploitables, interopérables et consultables en temps réel.

Cette distinction a des conséquences méthodologiques majeures :

1. **Qualité et fiabilité** : les données du JORF sont plus stables, car issues d'une chaîne de production numérique native.
2. **Temporalité d'accès** : le JORF offre une mise à disposition quasi instantanée, là où les corpus rétro-numérisés accusent des délais et des lacunes.
3. **Structuration** : l'usage du XML et des API ouvre de nouvelles possibilités d'exploitation automatique, d'agrégation et de visualisation.

Ainsi, la numérisation du J.O. historique et la production numérique du JORF contemporain dessinent deux faces d'une même logique : d'un côté, la reconfiguration patrimoniale d'archives imprimées par la BnF ; de l'autre, la fabrique d'archives nativement numériques par l'État. Dans les deux cas, ce sont des **infrastructures sociotechniques** qui conditionnent la circulation, la visibilité et la confiance dans les données.

Deuxième partie

L’enjeu des données structurées : des sources à la base de données

Chapitre 3

Une histoire par les données

Nous venons de voir que les *Tables Annuelles* du Sénat contiennent une véritable mine d'informations pour établir une analyse de l'activité parlementaire. Ces *Tables*, accessibles sur Gallica, avec le jeu des renvois et des index, sont de véritables bases de données de papier. Pour récupérer les informations du *Journal Officiel* de façon automatisée, c'est-à-dire sans reproduire à la main l'ensemble, il faut penser à une chaîne de traitement qui part de ces sources numériques, sous format image, pour pouvoir en capturer l'information. Il s'agit ici de voir comment construire un protocole d'extraction cohérent, en tenant compte de la matérialité des documents eux-mêmes, aussi bien sous leur forme « analogique » que numérique.

Dans ce chapitre, il s'agira de répondre aux problématiques techniques de cette traduction des sources numérisées – c'est-à-dire sous format image – au texte. Ceci imposant de donner un contexte préalable de cette « mise en données »¹ des sources historiques, laquelle est inhérente à la disponibilité de corpus numérisés par les politiques de valorisation des fonds des institutions patrimoniales.

Comment travailler à partir d'une image numérique ? Certes, la représentation photographique et numérique d'un document est lisible pour un oeil humain ; mais du point de vue informationnel, ces images ne sont que des paquets de pixels, une conversion numérique de l'information lumineuse renvoyée par les objets photographiés². Ces pixels ne sont pas, évidemment, les lettres elles-mêmes. Ils sont la traduction sur l'écran de trains d'informations binaires qui, sans le bon décodage, pourrait vouloir dire tout autre chose. Le premier enjeu

¹Frédéric Clavert, « Une histoire par les données ? Le futur très proche de l'histoire des relations internationales », *Bulletin de l'Institut Pierre Renouvin*, 44–2 (nov. 2016), Publisher : UMR Sirice Section : Histoire, p. 119-130, DOI : 10.3917/bipr1.044.0119.

²Thierry Claerr et Isabelle Westeel (éd.), *Numériser et mettre en ligne*, Code : Numériser et mettre en ligne Publication Title : Numériser et mettre en ligne Reporter : Numériser et mettre en ligne Series Title : La Boîte à outils, Villeurbanne, 2010 (La Boîte à outils), URL : <https://books.openedition.org/pressesenssib/414> (visité le 20/08/2025).

pour un travail de capture de l’information est de transformer cette matière matricielle en information textuelle sur laquelle on peut appliquer des traitements. Le texte se présente comme pré-requis pour établir des chaînes de traitement de capture informationnelle. Ce passage de l’image au texte numérique est en fait techniquement une prérogative des tâches de *reconnaissance optique des caractères* – ou « OCR » (*Optical Character Recognition*). Elle butte également sur des problématiques de détection de la mise en page, laquelle fonde un ordre de lecture – et donc un agencement du sens des phrases qu’il faut considérer. [Section 2 : de l’image au texte]

Deuxième problème : une fois ce texte numérique obtenu, comment capturer l’information sémantique qui est présente ? Comment l’ordinateur peut comprendre que tel ensemble des caractères alphanumériques correspond en fait à un sénateur de la Troisième République ? On peut trouver, dans le document, des motifs qui signalent une entité (par exemple, le nom d’un sénateur). Cette approche comme on va le voir, est basée sur la reconnaissance de motifs typographiques. Elle est cependant fragile et dépendante de la qualité de l’OCR – voire des erreurs humaines présentes dans le document d’origine. Elle suppose aussi une forme de connaissance *a priori* synthétique de la représentation de l’information dans le document. Ainsi peut-on se tourner vers des approches extractives qui viennent labelliser l’information textuelle obtenue³ ; ou bien les approches génératives qui « lisent » le texte et restituent l’information comprise et permettent de contourner le problème des exceptions qui forment le corps des documents⁴. Cette chaîne de travail de capture exige une information exploitable par l’ordinateur ; elle a besoin d’une certaine systématité, laquelle est une prérogative des modèles de données – de leur « forme ». La chaîne de traitement commence donc avec l’image, passe par le texte et des méthodes de capture de l’information sémantique qu’elle contient, pour aboutir à une information structurée. L’enjeu n’est pas simple car chaque étape reporte les marges d’erreur des précédentes. [Section 3]

Dans ce chapitre, il s’agira ainsi de dessiner le contexte technique et institutionnel de cette « datafication » des données en vue de leur traitement – et notamment avec les nouvelles opportunités des grands modèles de langage.

³Jenny Rose Finkel, Trond Grenager et Christopher Manning, « Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling », dans *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, dir. Kevin Knight, Hwee Tou Ng et Kemal Oflazer, Ann Arbor, Michigan, 2005, p. 363-370, doi : 10.3115/1219840.1219885.

⁴Alec Radford, Karthik Narasimhan, Tim Salimans et Ilya Sutskever, « Improving Language Understanding by Generative Pre-Training » ().

I. Numériser les sources

1. En « mode texte »

En 1971, un étudiant reproduisait sur un ordinateur *Xerox* la *Déclaration d'indépendance des Etats-Unis*, en caractères alphanumériques **ASCII**. Il s'agissait de Michel Hart, fondateur du **Projet Gutenberg** qui se donnait pour tâche de reproduire et diffuser bénévolement sur le réseau internet des œuvres littéraires du domaine public. Des livres comme *La Bible*, des œuvres de Shakespeare, quelques autres de Lewis Carroll ou de James M. Barrie seront notamment reproduites⁵. Ce travail de « numérisation » est en fait un travail laborieux, sinon de tâcheron : chacune des lettres de chaque livre sera tapée à la main, les unes après les autres. En 1990, avec les débuts du Web, le projet prend un nouvel essor et bénéficie d'une collaboration internationale : les collections s'élèvent à environ 1000 livres en 1997 ; 4000 livres en 2001 ; et 15000 livres en 2005⁶. Entre le livre et la version numérique, il n'y a pas d'images : juste le travail manuel de transcription des caractères. C'est une numérisation des livres « en mode texte »⁷ : l'information textuelle seule, stockée sur disque dur, est reproduite sur l'écran, cela « déstructurant l'objet livre »⁸. En effet, avec cette reproduction en caractères alphanumériques, la structure physique du livre – sa mise en page – est perdue ; mais on peut en revanche rechercher un mot et retrouver un passage plus aisément.

Le texte numérique ne se définit pas seulement comme une reproduction électronique du texte imprimé, mais comme une transformation de l'information en une suite de signes codés. Concrètement, chaque caractère est représenté par une valeur numérique, selon un système de codage – tel que l'**ASCII** (*American Standard Code for Information Interchange*) ou, plus récemment, l'**Unicode**, qui attribue à chaque lettre, chiffre ou symbole une séquence binaire, une suite de *bits*, c'est-à-dire de 0 et de 1. Ce passage de l'écriture alphabétique à la codification binaire permet au texte d'être manipulé comme une donnée discrète : il devient possible de rechercher automatiquement un mot, de compter des occurrences, de structurer des chaînes de caractères.

Cette démarche d'encodage de l'information, qui ne concerne pas ici proprement l'historien, est exemplaire au regard des méthodes de *numérisation* des documents textuels en ce sens qu'elle traduit une forme analogique — physique ou continue — en une forme numérique, discrète. L'opération de transcription manuelle, caractère par caractère, est ici comparable

⁵Marie Lebert, gutenberg.org/cache/epub/27040/pg27040.txt, URL : <https://www.gutenberg.org/cache/epub/27040/pg27040.txt> (visité le 17/08/2025).

⁶*Ibid.*

⁷Emmanuelle Bermès, *De l'écran à l'émotion : quand le numérique devient patrimoine*, Country : FR 20 cm. PSL = Université Paris sciences & lettres. Bibliogr. p. 231-234. Index., Paris, 2024, URL : 893734 (visité le 22/08/2025), 30-33.

⁸*Ibid.*

à celle d'un dépouillement systématique sur archives papier : il s'agit de saisir l'information contenue dans les sources dans un dispositif tabulaire, par exemple un tableau⁹. La similarité n'est toutefois que d'ordre opératoire. Même si les enjeux intellectuels ne sont pas les mêmes, il y a un face-à-face entre un opérateur et la source à restranscrire. Il y a une reproduction qui n'est pas photographique, mais manuelle. Dans le cas de Michel Hart et du Projet Gutenberg, la répétition du texte littéraire reste relativement linéaire et vise une reproduction intégrale. À l'inverse, la transcription historienne – qui n'est pas nécessairement moins laborieuse – suppose *en même temps* une enquête critique : sélectionner, structurer, et souvent synthétiser des données pour les « mettre en table », c'est-à-dire les rendre comparables, quantifiables. La transcription de sources historiques est tournée vers sa structuration, d'une « mise en fiche »¹⁰, même si celle-ci est implicite – par exemple l'ordre des éléments transcrit est une information séquentielle de cheminement de lecture – car c'est une lecture exploratoire. Un chemin documentaire est à inventer, le parcours de lecture n'est pas donné. Ce schème opératoire transcriptif, ce qui est donc commun à la tâche de transcription littérale d'un Michel Hart ou à celle cheminatoire d'un historien à la Pierre Chaunu¹¹, peut être qualifié de travail de *transduction technique* de l'information¹². Cette transduction est un processus par lequel l'information passe d'un support et d'un régime de signification à un autre, selon des contraintes à la fois techniques et des exigences intellectuelles. Dans le cas de la transcription numérique, l'information change de milieu technique et son inscription *in silicium* permet de reconsiderer le texte discrétilisé comme un ensemble d'éléments manipulables, par exemple par des algorithmes de tri. Le texte numérique peut être alors considéré selon différents degrés de structuration : tantôt comme une « répétition linéaire » et brute des sources originales à l'instar des premières éditions du **Projet Gutenberg** ; tantôt comme une information choisie et hiérarchisée, ou du moins tendue vers un raffinement structurel, comme l'implique une mise en tableau.

Dans le cas des historiens, ce passage implique un véritable travail d'individuation des données – c'est-à-dire de leur transformation au regard du contexte technique qui joue ici comme un milieu : il faut découper des flux documentaires continus en unités discrètes (noms, dates, professions, événements par exemple), qui ne préexistent pas à l'opération de transcription mais sont construites, reformulée par elle, avec la contrainte ultérieure de pouvoir

⁹Claire Lemercier et Claire Zalc, *Méthodes quantitatives pour l'historien*, ISSN : 0993-7625, 2008, DOI : 10.3917/dec.1emer.2008.01.

¹⁰Jean-François Bert, « L'art de la mise en fiche et ses métamorphoses modernes », *Genesis. Manuscrits – Recherche – Invention*–55 (déc. 2022), ISBN : 9791023107449 Number : 55 Publisher : Sigales, p. 57-70, DOI : 10.4000/genesis.7339.

¹¹Bertrand Müller, « De la formation d'un concept à l'invention d'une tradition : les avatars de l'histoire sérielle », dans *Un historien dans ses lendemains : Pierre Chaunu*, dir. Denis Crouzet et Alain Hugon, Code : Un historien dans ses lendemains : Pierre Chaunu, Caen, 2021 (Symposia), DOI : 10.4000/books.puc.15437.

¹²chatonsky.

retrouver l’information encodée. Des stratégies d’habillage de l’information saisie sont à envisager du même mouvement, par exemple en dotant le texte d’un « appareil critique » qui permettent un retour au texte original à travers des clés qui en indexent le contenu¹³. La numérisation en mode texte est alors une opération configuratrice : la saisie manuelle se fait l’instrument d’un changement de régime technique du support de l’information car elle implique, à différent degrés, un besoin de mise en structure. Du côté des historiens, si ce travail de transduction informationnelle, plus sophistiqué que la transcription littérale, peut être comparé au travail de terrain du sociologue ou de l’éthnographe – en ce sens qu’elle suscite justement des questions et reconfigure les valuations de l’enquête¹⁴¹⁵ – il s’adosse surtout à l’élaboration de nouvelles données sur les données collectées. Ces « données sur les données » – ces **métadonnées** –, nécessaires par exemple pour mettre en table l’information obtenue, relèveraient d’un *design* ou, autrement dit, d’une stratégie de composition de l’information en vue d’en produire une représentation intellegible pour un traitement ultérieur¹⁶. Ces métadonnées, plus ou moins riches ou explicites mais qui permettent de ranger les données collectées, sont propres à l’historien qui les conçoit et les ajuste selon ce qu’exige sa question de recherche. Si bien que ce processus valutatif de mise en catégorie de l’information capturée accompagne la transcription manuelle et littérale des sources. Elle s’inscrit dans une démarche éditoriale telle que décrite par Steven DeRose, David Durand, Elli Mylonas et Allen Renear, qui se représentent le texte comme une structure hiérarchique ordonnée d’objets et de contenu¹⁷.

Le travail de saisie manuelle, avec ou sans métadonnées, n’est évidemment pas une nouveauté introduite par l’ordinateur. Bien avant l’ère numérique, les historiens s’y adonnaient déjà. Ainsi, Pierre Chaunu qui, en 1947, recopiait à la main, sur papier, les données issues des archives microfilmées et des ouvrages nécessaires à sa thèse, afin de les ordonner et de les exploiter systématiquement¹⁸ ; ou encore Ladurie sous forme de fiches ou de tableaux [Ladurie]. Au-delà du « bricolage »¹⁹ transcriptif, propre à chaque chercheur, l’histoire de la fiche érudite nous indique également des ambitions de systématisation lesquelles conduisent à une véritable ingénierie du glanage documentaire²⁰. Citons par exemple « l’armoire érudite », les

¹³Marc Van Campenhoudt, « Une norme de dépouillement terminologique en langue française », *Equivalences*, 21–1 (1992), Publisher : Persée - Portail des revues scientifiques en SHS, p. 121-136, DOI : 10.3406/equiv.1992.1147.

¹⁴C. Lemercier et C. Zalc, *Méthodes quantitatives pour l’historien...*

¹⁵John Dewey, *La formation des valeurs*, La Découverte, Paris, 2011.

¹⁶Anne-Lyse Renon, *Design & sciences*, fr-FR, URL : <https://www.puv-editions.fr/ouvrage/design-sciences/> (visité le 22/08/2025).

¹⁷Steven J. DeRose, David G. Durand, Elli Mylonas et Allen H. Renear, « What is text, really? », *Journal of Computing in Higher Education*, 1–2 (déc. 1990), p. 3-26, DOI : 10.1007/BF02941632.

¹⁸B. Müller, « De la formation d’un concept à l’invention d’une tradition... ».

¹⁹Claude Levi-Strauss, *La pensée sauvage*.

²⁰J.F. Bert, *Une histoire de la fiche érudite*, Code : Une histoire de la fiche érudite Publication Title : Une

boîtes ou les casiers extensibles de l’entreprise *Borgeaud* permettant de ranger l’information dans des bonnes cases, le casier où la position dans un classeur ayant valeur ici de « métadonnée analogique »²¹. Aujourd’hui encore, malgré l’apparition d’outils de transcription automatique [voir section 2], cette pratique demeure courante : toutes les sources ne sont pas disponibles en version numérique, et le chercheur, tout comme l’étudiant ou le généalogiste, peut être amené à relever lui-même les informations qui l’intéressent, directement en salle d’archives ou lors du dépouillement de fonds imprimés.

Le mode opératoire de la numérisation « en mode texte » constitue en ce sens un cas exemplaire, puisqu’il s’oppose radicalement à la logique de la numérisation photographique, dite « en mode image »²²²³. Il ouvre également des perspectives pour le traitement quantitatif, dans la mesure où il produit une matière directement exploitable : des données susceptibles d’être structurées — par une mise en table ou un encodage hiérarchique tel que la **TEI** — et manipulées — par exemple à travers la recherche de motifs textuels.

2. En « mode image »

À l’opposé du « mode texte », la numérisation en « mode image » repose sur la reproduction photographique des documents, cherchant à restituer leur matérialité visuelle : texte, blancs, marges, typographie, ornements, etc. Tout est fixé dans une matrice de pixels. Héritière des microfilms et des fac-similés, cette pratique connaît une expansion décisive dans les années 1990, avec l’essor du Web et le lancement des premières grandes campagnes institutionnelles de numérisation. Deux projets emblématiques illustrent cette dynamique : Gallica (BnF, 1997) et Google Books (2004)²⁴. Leur ambition est similaire — mettre à disposition le patrimoine imprimé à grande échelle — même si, pour Gallica, il s’agit de s’inscrire dans une mission de service public. Dans les deux cas, la numérisation institutionnelle en mode image suppose des investissements lourds en infrastructures, en personnels et en politiques documentaires²⁵. Elle se distingue ainsi des pratiques de transcription textuelle, souvent issues de gestes individuels ou collaboratifs (historien recopiant ses sources, dépouilements collectifs, corrections d’OCR par **crowdsourcing**). Certes, des chercheurs ou amateurs produisent eux aussi des photographies de documents — parfois propres, parfois très imparfaites —, mais ces fichiers, à défaut d’un soin documentaire, restent isolés, de qualité variable, sans métadonnées

histoire de la fiche érudite Reporter : Une histoire de la fiche érudite Series Title : Papiers, Villeurbanne, 2017 (Papiers), URL : <https://books.openedition.org/pressesensib/6211> (visité le 21/08/2025).

²¹ *Ibid.*

²² E. Bermès, *De l’écran à l’émotion...*

²³ *Numériser et mettre en ligne...*

²⁴ E. Bermès, *De l’écran à l’émotion...*

²⁵ *Numériser et mettre en ligne...,* p. 66-88.

ni garantie de pérennité. Si bien qu'ici, photographier soi-même ses sources – par exemple aux archives – ne revient finalement qu'à remettre à plus tard le travail de dépouillement. Ce qui d'ailleurs n'empêche finalement une laborieuse transcription en mode texte et, cette fois, non plus en face des sources mais de leur reproduction numérique.

Sur le plan technique, ces images sont matricielles : elles fixent l'apparence de la page mais le contenu intellectuel « littéraire » demeure opaques pour l'ordinateur. Une image numérique est un tableau – une *matrice* – d'une largeur et d'une hauteur données, comportant alors largeur × hauteur pixels, pixels qui encode l'information colorimétrique sur trois vecteurs : le paramètre *rouge*, le paramètre *vert*, et le paramètre *bleu*. La combinaison de ces trois paramètres, selon les règles de la synthèse colorimétrique additive, permettent de restituer, pour chaque pixel, l'ensemble des couleurs du spectre visible. Qu'elles proviennent d'une campagne institutionnelle ou d'un smartphone amateur, elles ne permettent pas la recherche plein texte sans OCR ou segmentation. Dans le cas des photographies amateurs, pouvant être floues ou mal cadrées, l'OCR est même impraticable, réduisant ces fichiers à des fac-similés inertes, reportant ainsi la tâche de saisie manuelle non pas à partir des documents matériels, mais à partir de leur représentation numérique. Ce qui permet de souligner que ce n'est pas la technique de numérisation qui « fait » le mode opératoire ; mais bien le rapport de l'opérateur à un système technique plus ou moins sensible à l'information et sa capacité à la transformer²⁶.

En effet, « numériser n'est pas éditer »²⁷. Là où la mise en table implique un travail de curation plus ou moins entendu — sélection, structuration, annotation —, la capture photographique ne livre qu'une matière brute. Sans enrichissement éditorial, sans métadonnées ni transcription automatique, ces images demeurent orphelines de leur contenu textuel : de simples objets visuels, inaccessibles à toute interrogation systématique. Leur valeur scientifique dépend donc entièrement des traitements ultérieurs qui les convertissent en données exploitables. Dans le cas de la numérisation amateur, ces fichiers constituent le plus souvent un point de départ, ou une stratégie mnémone pour différer le travail de transcription lors de l'enquête sur les sources. À l'inverse, pour la numérisation institutionnelle, ils relèvent d'une logique patrimoniale : il s'agit avant tout de restituer des ouvrages de nature variée par la reproduction photographique, solution pragmatique mais qui pose d'emblée la question de l'alternative avec le mode texte. Comme le rappelait Jean-Didier Wagneur à propos de Gallica :

« Nous avions le devoir patrimonial de restituer l'image du document tel qu'il

²⁶Gilbert Simodon, *Du mode d'existence des objets techniques*, Aubier, 1958.

²⁷manuefig, *Numériser ce n'est pas éditer (2)*, fr-FR, janv. 2005, URL : <https://figoblog.org/2005/01/18/519/> (visité le 20/08/2025).

a été déposé et le choix du mode image (fac-similé électronique) s'est imposé. Cette option a été à l'origine de nombreuses questions autour de l'alternative que le mode texte présentait. [...] On voit qu'à terme, la saisie en mode texte aurait débouché sur la nécessité de produire des documents mixtes (texte et image) afin de préserver l'intégrité de tous les documents de nature graphique (illustrations, cartes, reproductions, graphes et expériences scientifiques) figurant dans les ouvrages numérisés. »²⁸

Se dessine une tension durable entre deux régimes de numérisation : d'un côté, la transcription textuelle qui repose sur la couteuse « nécessité de faire saisir les œuvres de plusieurs centaines, voire milliers, d'auteurs » exigeant un « accompagnement scientifique considérable »²⁹ ; de l'autre, la reproduction visuelle, portée par des politiques institutionnelles, qui fabrique des corpus massifs mais souvent réduits à l'état de matière brute.

II. Ce que la numérisation fait aux sources

1. La « datafication »

Cette opposition doit enfin être replacée dans l'horizon plus large de la *datafication*³⁰. La *datafication* est le processus qui vise à quantifier un phénomène de sorte qu'il soit calculable et analysable³¹. Elle est en quelque sorte un « schème opératoire »³² permettant la calculabilité des sources avec un outillage informatique. Cette mise en données « insiste sur la notion de processus » et « se définit par les choix opérés par les organismes qui y procèdent », cela impliquant « les critères d'inclusion [de] corpus à numériser » ; l'élaboration de métadonnées descriptives situées, lesquelles ont un impact sur leur découvrabilité puisque les moteurs de recherche s'y appuient³³.

Si bien que numériser, ce n'est pas seulement reproduire. C'est transformer des artefacts en données, dans un cadre technique, social et institutionnel qui oriente les usages, fixe les normes de conservation et conditionne l'accès même aux sources. Elle met en évidence deux conceptions distinctes de la numérisation : d'un côté, la transcription textuelle, qui construit les données par un travail de sélection et est tendue vers une structuration plus ou moins

²⁸Jean-Didier Wagneur, « Gallica : la bibliothèque électronique de la BnF : quel accès pour les personnes handicapées visuelles ? », dans *Bibliothèques et publics handicapés visuels*, Code : Bibliothèques et publics handicapés visuels, Paris, 2002 (Paroles en réseau), DOI : 10.4000/books.bibpompidou.1500.

²⁹Ibid.

³⁰F. Clavert, « Une histoire par les données ?... ».

³¹Ibid.

³²G. Simodon, *Du mode d'existence des objets techniques...*, p. 236.

³³F. Clavert, « Une histoire par les données ?... », p. 123.

affirmée ; de l'autre, la reproduction visuelle, qui se limite à conserver une représentation plane de la matérialité de l'objet. L'histoire des pratiques documentaires de « mise en données » témoigne de cette tension durable entre deux paradigmes concurrents³⁴ qui sont alors autant techniques qu'institutionnels. En suivant Bruno Latour, on peut dire que « nous ne devrions jamais parler de ‘données’, mais toujours d’‘obtenues’ »³⁵. La saisie manuelle n'est pas une simple transplantation de contenu, mais une construction — un fait mobilisable conditionné par des choix de catégorisation, de format, et d'usage. La reproduction photographique, à l'inverse, ne produit pas directement de données exploitable, mais fige l'apparence visuelle du document, laissant le texte dans un état opaque tant qu'aucune opération d'extraction ou d'annotation n'est réalisée.

Chaque « mise en données » est ainsi moins une instantiation pure de la source dans le giron du binaire qu'une stratégie épistémique de sa présentation au sein d'un réseau socio-technique instituant³⁶. Dans cette perspective, on peut dire avec Cornelius Castoriadis que la numérisation est un geste « instituant » : elle crée de nouvelles manières de faire exister et de rendre visibles les sources, conditionnant les régimes d'intelligibilité qui en découlent³⁷. Or, au sein des politiques contemporaines de numérisation, cette dimension instituante prend une forme discursive et normative particulière : celle du *patrimonial*. Le patrimoine est un « passé-présent » sans cesse réinterprété, qui puise dans l'imaginaire institué et reformé en retour de nouvelles significations. La numérisation est justifiée et orientée par un vocabulaire de démocratisation et de valorisation, dans une fibre tout encyclopédique, tel que défendu par Jean-Didier Wagneur cité précédemment. Autrement dit, ce qui est numérisé n'est pas seulement conservé ou reproduit : il est institué comme patrimoine, doté d'une valeur « émotionnelle » et sociale spécifique³⁸. Les données produites par la numérisation sont donc indissociables d'une politique culturelle qui configure les conditions d'accès, de visibilité et de réutilisation des corpus, ce qui implique nécessairement des biais de sélection ou tout simplement des priorités déterminées par les politiques de conservation de documents fragiles.

³⁴E. Bermès, *De l'écran à l'émotion...*, p. 29.

³⁵Jérôme Denis et Samuel Goëta, « Les facettes de l'Open Data : émergence, fondements et travail en coulisses », dans *Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus*, dir. Pierre-Michel Menger et Simon Paye, Code : Big data et traçabilité numérique : Les sciences sociales face à la quantification massive des individus, Paris, 2017 (Conférences), p. 121-138, DOI : 10.4000/books.cdf.5005.

³⁶A.L. Renon, *Design & sciences...*

³⁷Michel Rautenberg, « Les chemins croisés du patrimoine et de l'imaginaire », dans *L'imaginaire patrimonial : Figures de l'urbanité contemporaine*, Code : L'imaginaire patrimonial : Figures de l'urbanité contemporaine, Rennes, 2024 (Essais), p. 13-52, DOI : 10.4000/13ipp.

³⁸E. Bermès, *De l'écran à l'émotion...*

2. Pratiques historiennes : sphère technique, sphère sociale

La datafication ne consiste donc pas seulement en un enchaînement d'opérations techniques — OCR, structuration, encodage —, mais en un processus situé, au croisement de la sphère technique et de la sphère sociale. Il n'existe pas de « données brutes » : toute donnée est déjà une *capta*, c'est-à-dire une information « prise », construite dans et par un cadre interprétatif qui reflète des choix méthodologiques et institutionnels³⁹. Leur mise à disposition est elle-même également une affaire de visualisation : la mise en ordre des résultats d'une recherche en ligne appartient au giron de la visualisation des données – et comme on l'a sous entendu précédemment, les stratégies épistémiques de représentation de connaissance est une affaire de *design*. La notion de *graphesis*, proposée par Johanna Drucker⁴⁰, insiste justement sur ce point : les visualisations ne se contentent pas de montrer des données, elles constituent un mode de production de savoir en les configurant graphiquement. D'ailleurs certains artistes, comme le duo Vasulka, parlent volontiers de « nouvel espace épistémique », « d'épistémè numérique »⁴¹ ou encore, cette fois du côté de Jean-Louis Boissier, « d'image relation »⁴², ce dernier concept insistant plutôt sur la notion d'interactivité. Pour ces artistes, il est question de faire grincer la dimension interprétative et graphique de l'accès à l'information à travers des interfaces car, si l'habitude peut nous conduire à considérer la présentation des données par des moteurs de recherche comme « objectifs », il faut au contraire montrer qu'il s'agit de dispositifs techniques partiaux, construits à partir de théories, de thèses et d'hypothèses qui ont une historicité⁴³. S'il peut bien y avoir une prétention encyclopédique dans les dispositifs techniques d'accès aux sources⁴⁴, il y a également une dimension esthétique, « jouable », qui est une condition à la lisibilité du dispositif numérique⁴⁵. Dès lors, il faut considérer les données et les métadonnées à la lumière de leur présentation, présentation qui doit faire l'objet de décisions, de stratégies au regard de ce qui importe de montrer ou de cacher – bref : il y a là des *valuations* dans la médialité, en reprenant ce mot dans son sens deweyien⁴⁶, c'est-à-dire une dépendance des schèmes techniques opératoires aux processus de formation des valeurs. Les institutions patrimoniales sont des acteurs exemplaires de cette double dépendance des dispositifs techniques aux attentes sociales ou scientifiques.

³⁹ Johanna Drucker, *Visualisation. L'interprétation modélisante*, B42, Paris, 2020.

⁴⁰ *Ibid.*

⁴¹ Jean-Marie Dallet et Bertrand Gervais (éd.), *Architectures de mémoire*, Code : Architectures de mémoire Publication Title : Architectures de mémoire Reporter : Architectures de mémoire Series Title : La Grande Collection ArTeC, Nanterre, 2019 (La Grande Collection ArTeC), URL : <https://books.openedition.org/pupo/25762> (visité le 20/08/2025).

⁴² Jean-Louis Boissier, *La relation comme forme*, Les Presses du Réel, Dijon, 2009.

⁴³ Lorraine Daston et Peter Galison, *Objectivité*, Les Presses du Réel, Dijon, 2012.

⁴⁴ J.D. Wagneur, « Gallica... ».

⁴⁵ J.L. Boissier, *La relation comme forme...*

⁴⁶ J. Dewey, *La formation des valeurs...*

Un exemple particulièrement riche de ces réflexions se trouve dans le PLAO (pour « Poste de Lecture Assistée par Ordinateur ») imaginé à la BnF dans les années 1990 sous l’impulsion de Bernard Stiegler. Il s’agissait d’une interface destinée à incarner le lecteur « savant » au cœur d’un dispositif instrumenté, permettant non seulement de consulter des textes numérisés, mais aussi de les annoter, de structurer dynamiquement un corpus, de naviguer, d’archiver, de copier ou d’indexer en temps réel. Le PLAO n’est jamais devenu un équipement déployé à large échelle, mais il s’est imposé comme un « objet-frontière » — au sens de Star et Griesemer⁴⁷ — structurant une coalition entre informaticiens, philosophes, historiens et bibliothécaires, engagés dans la conceptualisation d’un nouvel usage de la bibliothèque numérique⁴⁸.

Ces interrogations contemporaines ne surgissent pas *ex-nihilo* : elles prolongent une histoire plus ancienne des rapports entre historiens et outils informatiques. Ces transformations récentes s’inscrivent dans une histoire plus longue des usages de l’informatique par les historiens. Dès les années 1950, des expériences pionnières lient mécanographie et analyse serielle⁴⁹, dans la lignée des « archives quantitatives » chères à l’école des *Annales*. Dans les décennies 1960–1970 voient s’imposer l’ambition d’une histoire totale⁵⁰. La prophétie de Le Roy Ladurie — « l’historien de demain sera programmeur ou ne sera pas » — illustre cet horizon, même si les limites de l’histoire serielle conduisent rapidement à relativiser l’objectivité promise par la machine. Dans les années 1980–1990, l’apparition du micro-ordinateur personnel, puis des premières bases de données relationnelles et de revues comme *Le Médiéviste et l’Ordinateur*, favorise une technicisation diffuse des pratiques, souvent portée par des chercheurs passionnés plutôt que par une politique disciplinaire globale⁵¹. Le Web, à partir de la fin des années 1990, change l’échelle de ces usages : il facilite l’accès et la diffusion des corpus (Gallica, revues.org) et transforme le rapport aux archives grâce aux appareils photo numériques et aux interfaces de recherche. L’émergence des humanités numériques dans les

⁴⁷Pascale Trompette et Dominique Vinck, « Retour sur la notion d’objet-frontière », *Revue d’anthropologie des connaissances*, 31–1 (juin 2009), Publisher : S.A.C., p. 5-27, DOI : 10.3917/rac.006.0005.

⁴⁸Gaëlle Béquet, « La bibliothèque numérique à la Bibliothèque de France : de l’objet-valise à l’objet-frontière (1988-1993) », dans *Trois bibliothèques européennes face à Google : Aux origines de la bibliothèque numérique (1990-2010)*, Code : Trois bibliothèques européennes face à Google : Aux origines de la bibliothèque numérique (1990-2010), Paris, 2015 (Mémoires et documents de l’École des chartes), p. 51-85, DOI : 10.4000/books.enc.14133.

⁴⁹Adeline Daumard et François Furet, « Méthodes de l’Histoire sociale : les Archives notariales et la Mécanographie », *Annales*, 14–4 (1959), Company : Persée - Portail des revues scientifiques en SHS Distributor : Persée - Portail des revues scientifiques en SHS Institution : Persée - Portail des revues scientifiques en SHS Label : Persée - Portail des revues scientifiques en SHS Publisher : EHESS, p. 676-693, DOI : 10.3406/ahess.1959.2865.

⁵⁰Sébastien Poublanc et Nicolas Marqué, « Introduction au dossier « Historien · nes et numérique : pratiques et expériences vécues » », *Les Cahiers de Framespa. e-STORIA-42* (juill. 2023), Publisher : UMR 5136 – FRAMESPA, DOI : 10.4000/framespa.14370.

⁵¹*Ibid.*

années 2000 fait de ces pratiques dispersées un champ identifié, mais aussi un espace de tensions : entre injonctions institutionnelles, résistances disciplinaires et négociations interdisciplinaires. L’histoire des pratiques historiennes « numériques » apparaît ainsi comme un processus long, fait d’allers-retours entre engouements, critiques et réinventions, qui relativise l’idée d’un tournant soudain pour insister sur la continuité d’une adaptation progressive des historiens à leurs outils.

Avec l’essor du Web, la datafication ne se réduit plus à un geste de transcription ou à un choix institutionnel de numérisation : elle s’inscrit dans des infrastructures réticulaires qui conditionnent la circulation et l’usage des corpus. Les documents ne sont pas seulement mis en ligne ; ils sont exposés à travers des protocoles techniques — comme les API — qui déterminent la granularité d’accès, la possibilité de réutilisation et la manière dont les sources sont découvertes. Comme le montre l’exemple du *Goût de l’archive à l’ère numérique* et de sa réflexion sur le « goût de l’API »⁵², l’archive numérisée devient un objet relationnel puisqu’elle n’existe pleinement qu’à travers les réseaux qui l’indexent, la connectent et la rendent interopérable avec d’autres ensembles de données. Cette dimension réticulaire transforme profondément la sphère sociale des archives : les corpus ne sont plus simplement conservés et transmis, mais distribués, exposés, parfois fragmentés, selon des logiques de plateformes et de moteurs de recherche. La découverbarilité des sources dépend ainsi de ces dispositifs techniques saisis dans des processus de *concrétisation*⁵³, qui agissent comme de nouveaux médiateurs documentaires et configurent, en amont, les conditions de possibilité de l’enquête historienne.

Autrement dit, ce que fait la numérisation aux corpus, c’est moins de les rendre « disponibles » que de les reconfigurer : par la sélection de ce qui est numérisé (et de ce qui ne l’est pas), par les formats qui conditionnent l’usage (XML-TEI, bases relationnelles, IIIF), et par les réseaux de diffusion (catalogues, moteurs de recherche, portails institutionnels) qui hiérarchisent leur visibilité. La donnée numérique n’est donc pas un miroir fidèle des sources, mais une construction sociotechnique qui oriente leur appropriation. De ce point de vue, la datafication prolonge les silences de l’archive autant qu’elle ouvre de nouvelles potentialités. Elle produit un double effet : d’un côté, elle consolide des corpus institués par les politiques patrimoniales et documentaires ; de l’autre, elle institue de nouveaux régimes de visibilité et de calcul, rendant possible des analyses sérielles, des croisements de données ou des visualisations inédites. C’est dans cette tension que se joue aujourd’hui la confiance des chercheurs dans les « données » numériques : non comme transparence des sources, mais comme résultat de choix techniques et sociaux qu’il convient de rendre visibles et discutables. La datafica-

⁵²F. Clavert, *Le goût de l’API / Le goût de l’archive à l’ère numérique*, fr-FR, URL : <https://gout-numerique.net/table-of-contents/archives-nees-numeriques/gout-api> (visité le 22/08/2025).

⁵³G. Simondon, *Du mode d’existence des objets techniques...*

tion amplifie ainsi certains silences archivistiques : ce qui n'a pas été consigné, ou ce qui est difficile à transcrire automatiquement, reste hors champ.

Enfin, la datafication est aussi un processus social : elle reflète et prolonge les hiérarchies documentaires héritées. Les corpus numérisés surreprésentent souvent les groupes dominants (élites, institutions, employeurs), au détriment des voix minoritaires. Le danger est alors de tomber dans une réification de la disponibilité où l'historien travaille sur ce qui est disponible, non sur ce qui est historiquement pertinent ou accessible. La question devient dès lors : comment documenter ces biais et construire la confiance dans des données issues de systèmes techniques ?

3. Documents sériels et approches quantitatives

Les documents sériels — registres, recensements, listes nominatives ou annuaires — forment une catégorie d'archives dont la structure répétitive les rend particulièrement compatibles avec les opérations de numérisation et d'encodage. La tradition de l'« histoire sérielle » avait déjà mis en valeur cette potentialité : "l'énormité de la documentation semble avoir paralysé les chercheurs [...] Il semble, cependant, que le moment soit venu pour une nouvelle approche du problème : les techniques modernes, issues des ordinateurs, permettent une véritable révolution historiographique ; elles autorisent le traitement exhaustif d'un très grand nombre de données »⁵⁴. Plus récemment, Mark Crymble a montré que structurer des sources quantitatives reste indissociable des environnements techniques et des savoir-faire mobilisés — rappelant que l'histoire numérique réactive les enjeux — et les limites — de l'histoire sérielle⁵⁵.

Cette transformation se réalise dans ce que Max Kemman décrit comme des *trading zones* de l'histoire numérique, c'est-à-dire des espaces de négociation interdisciplinaire où historiens et acteurs computationnels construisent des langages et des méthodes partagées. Comme le note Kemman, ces interactions montrent que les historiens « construisent différents trading zones par un engagement interdisciplinaire, une négociation des objectifs de recherche et des intérêts individuels » (« *construct different trading zones through cross-disciplinary engagement, negotiation of research goals and individual interests* »)⁵⁶. Autrement dit, la réinvention de l'histoire sérielle se joue dans une interface où se négocient les protocoles, les modèles techniques et les finalités disciplinaires. Dans ce contexte, la logique sérielle n'a pas pour seul point d'appui l'accumulation des données ; elle repose également sur des

⁵⁴Marie Puren, *Digital Humanities in the TIME-US Project : Richness and Contribution of Interdisciplinary Methods for Labour History*, arXiv :2410.14222 [cs], oct. 2024, DOI : 10.48550/arXiv.2410.14222, p. 8.

⁵⁵**crymble**.

⁵⁶**kemman**.

infrastructures techniques qui structurent les effets de comparabilité et d'accès aux corpus. Ce basculement implique pour l'historien une compétence technique forte : il s'agit autant de réaliser une investigation documentaire que de participer activement à la modélisation, au choix des critères, des formats et à la mise en échange des données.

Le projet TIME US en est un excellent exemple : il croise recensements, listes de passagers et registres de naturalisation pour produire des séries comparables, à travers l'usage d'outils de traitement automatique du langage. Comme l'écrit Marie Puren, le défi était de « produire des données quantitatives à partir d'un vaste corpus textuel disparate »⁵⁷. La série serielle y est indissociable de la chaîne technique qui permet d'extraire, annoter et structurer les données.

L'opération serielle est donc dépendante des infrastructures numériques qui en rendent possible la construction et la circulation. Ces couches techniques ne sont pas neutres : elles hiérarchisent les corpus, orientent les comparaisons réalisables et déterminent, en amont, les catégories mobilisables. La numérisation déplace l'histoire serielle dans un régime sociotechnique où la série est inséparable de ses conditions de production et de circulation, où les standards et les interfaces jouent un rôle aussi décisif que les sources elles-mêmes.

⁵⁷ *Ibid.*, p. 5.

Chapitre 4

La reconnaissance optique de caractères

Le choix institutionnel de la BnF, à travers Gallica, de numériser massivement les corpus patrimoniaux en mode image s'inscrit dans une logique de valorisation d'une trace visuelle de l'objet imprimé, indépendamment des évolutions des standards textuels ou des logiciels de lecture. L'image numérique, photographique, motive une certaine expressivité, laquelle est lacuneuse du côté du texte brut. La typographie d'un ouvrage peut parler ; les cartes, dessins et schémas ne sont pas lésés par l'option de cette numérisation photographique. Mais cette stratégie, qui assure une intelligibilité visuelle de la source, laisse son contenu informationnel dans un état opaque pour la machine. C'est précisément pour franchir cette opacité et rendre ces corpus interrogables en plein texte qu'intervient l'OCR (*Optical Character Recognition*), technologie pivot entre la reproduction visuelle et la transformation en données exploitables. Dans ce chapitre, nous reviendrons sur les aspects techniques de l'OCR qui viennent donner à l'image un double textuel, plus prolix que une matrice de pixels.

I. Une vue d'ensemble

La reconnaissance optique de caractères — que l'on désignera désormais sous l'acronyme OCR — désigne l'ensemble des procédés permettant d'extraire automatiquement du texte lisible par machine à partir d'images numérisées. Historiquement, le terme s'applique aux documents imprimés composés avec des caractères typographiques, tandis que l'on réserve plutôt celui de *Handwritten Text Recognition* (HTR) aux documents manuscrits, qui présentent des difficultés spécifiques liées à la variabilité de l'écriture humaine et à l'absence de régularité formelle des lettres. Dans ce qui suit, nous laisserons cet aspect de côté pour nous concentrer sur l'OCR *stricto sensu*.

L’OCR constitue l’opération clef qui rend un document scanné ou photographié interrogeable en plein texte. Sur le plan technique, il s’agit de transformer une matrice de pixels — pouvant être en noir et blanc, en niveaux de gris ou en couleur RVB (rouge, vert, bleu) — en une séquence discrète de symboles numériques. L’unité et l’intégrité des trains d’information binaire étant une prérogative des formats et des algorithmes de compression. Ces couleurs peuvent avoir différents niveaux de finesse : un encodage de la couleur sur un seul bit (1 ou 0), ne laisse que deux possibilités, à l’instar de images bitmap (par exemple, les numérisation de microfilms sur Gallica). Une image RVB, encodée sur 1 octet par canal (soit 8 bits par canal ou 24 bits au total), permet de restituer une gamme plus large de couleurs, avec $(2^8)^3$ possibilités, soit plus de 16 millions de couleurs. Une image en nuance de gris encodée sur un seul octet restitue une gamme de 256 valeurs. Une image RVB d’une profondeur de 8 bits de $h \times l$ contient alors au minimum $h \times l \times 24$ bits d’information, sans compter le processus de compression qui permet de résumer l’information. La quantité d’information est importante : plus elle est haute, plus la réalité continue est échantillonée — donc la photographie du document est plus fidèle — mais elle devient plus lourde et donc couteuse en calcul. Par là, on voit que, du point de vue de l’ordinateur, un caractère n’existe pas comme « lettre » mais comme la traduction visuelle d’une matrice d’informations. La lettre, donc les mots et les phrases, sont comme des agencement de points lumineux dans un espace matriciel — ou plus précisément tensoriel dans le cas des images RVB car chaque pixel est en fait un vecteur en trois dimensions. Certes, un caractère encodé en **Unicode** ou en **ASCII** est également un train binaire. En revanche son décodage découle immédiatement sur une information explicite (un nombre faisant référence à une lettre : par exemple « A », « B » et « C » valent respectivement 65, 66, 67 en ASCII, lesquels valent également 01000001, 01000010 et 01000011 en binaire). Cela n’est pas le cas pour la couleur qui doit être de nouveau interprétée. Un « l » noir placé au milieu d’une page blanche, par exemple, est graphiquement un ensemble de pixels plus ou moins noirs connexes ; mais du point de vue d’une matrice, elle n’est pas vraiment une série de valeurs qui se suivent car il faut parcourir l’ensemble des valeurs pour détecter cette connexité — et cela sans parler bien entendu du bruit pouvant « couper » les ponts entre les pixels. On remarque au passage la difficulté de détecter la ponctuation, où les caractères qui n’ont graphiquement pas de connexité, comme la lettre « i » ou le point d’exclamation¹. Il y a donc ambiguïté entre l’information matricielle et sa référence. Si tous les algorithmes de détection de caractères n’emploient pas des algorithmes de connexité comme il vient d’être vulgarisé, les processus d’OCRisation sont en général sujets aux erreurs² et l’on comprend dès

¹R.G. Casey et E. Lecolinet, « A survey of methods and strategies in character segmentation », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18–7 (juill. 1996), p. 690-706, DOI : 10.1109/34.506792.

²Guillaume Chiron, Jean-Philippe Moreux, Antoine Doucet, Mickaël Coustaty et Muriel Visani, « Er-

lors pourquoi, jusqu’aux années 1990, des initiatives pionnières comme le *projet Gutenberg* préféraient la saisie manuelle du texte³. Aujourd’hui, les documents imprimés en caractères d’imprimerie affichent des scores assez hauts – ce qui n’est d’ailleurs pas le cas des moteurs d’HTR⁴.

La chaîne de traitement qui mène d’une image brute à un texte exploitable peut se décrire en plusieurs volets, si on s’autorise d’abord quelques généralités qu’il conviendra de modérer ensuite. Cette chaîne de traitement, donc, commence par une phase d’acquisition et de normalisation. L’acquisition, c’est l’étape de la prise de vue – on ne prend ici en considération que les documents qui ne sont pas nativement numériques. La qualité de la reconnaissance dépend en effet fortement de l’image initiale : une page légèrement inclinée, gondolée ou affectée par du bruit (comme c’est souvent le cas avec des microfilms granuleux) entraîne une baisse sensible de performance. Pour y remédier, on peut procéder à des corrections géométriques, comme le redressement (*deskew*) ou la compensation des courbures de page (*dewarp*), à un équilibrage des contrastes ou encore à un débruitage. Ces pré-traitements permettent d’optimiser les conditions de reconnaissance mais ne sont pas neutres : appliqués trop agressivement, ils peuvent détruire une partie de l’information utile. C’est la raison pour laquelle, à grande échelle, l’idéal reste une numérisation soignée en amont, qui évite de surcharger une chaîne de traitement déjà coûteuse.

Vient ensuite l’étape de segmentation, où le système doit identifier les unités pertinentes à traiter, de la mise en page aux caractères eux-mêmes – les « glyphs ». Les approches les plus anciennes reposaient sur des heuristiques « simples », telles que la connexité des pixels pour reconnaître des zones d’intérêt à étudier⁵. Les systèmes récents recourent désormais à des méthodes d’apprentissage profond capables de segmenter automatiquement les documents. Mais segmenter ne suffit pas. Faut-il encore mettre en ordre les unités visuelles pour qu’elles deviennent une suite lisible. C’est l’étape de sérialisation, où une ligne d’image est convertie en séquence de caractères qui se charge d’agréger l’ensemble éléments « lus ». Les approches anciennes procédaient caractère par caractère⁶, tandis que les modèles neuronaux modernes lisent directement la ligne entière comme un flux d’information, puis produisent une suite de lettres alignées⁷.

La question de la binarisation – la « mise en noir et blanc » des images – illustre d’ailleurs

reurs OCR et biais d’indexation : impact sur les usages », dans *17ème conférence Extraction et Gestion des Connaissances, Atelier Journalisme Computational*, Grenoble, France, 2017, p. 69-73, URL : <https://hal.science/hal-01455763> (visité le 22/08/2025).

³E. Bermès, *De l’écran à l’émotion...*

⁴G. Chiron, J.P. Moreux, A. Doucet, *et al.*, « Erreurs OCR et biais d’indexation... ».

⁵R. Casey et E. Lecolinet, « A survey of methods and strategies in character segmentation »...

⁶*Ibid.*

⁷Alex Graves, Santiago Fernandez, Faustino Gomez et Jurgen Schmidhuber, « Connectionist Temporal Classification : Labelling Unsegmented Sequence Data with Recurrent Neural Networks » ().

bien l'évolution des méthodes. Jusqu'au début des années 2000, la première étape consistait donc à transformer une image en noir et blanc⁸, notamment pour faciliter la détection de pixels connexes. Pour simplifier : ce qui était noir était de l'encre ; ce qui était blanc du papier. Cette simplification rendait plus facile la détection des lettres par les machines, qui pouvaient alors comparer directement des formes typographiques pré-enregistrées. Mais ce passage forcé au noir et blanc avait un coût : perte d'information sur les nuances, et sensibilité accrue aux défauts d'impression ou aux images bruitées. De même, il n'existe pas d'opération de seuillage universelle. On peut appliquer effectivement un seuil uniforme sur l'image, ou bien effectuer des moyennes locales en faisant glisser une petite matrice — ce sont des méthodes « convolutives », à l'instar du filtre gaussien ou une simple moyenne locale [Daniel Shiffman, @seuillage]. Le choix de la manière de binariser dépend de la qualité de la prise de vue, par exemple de l'uniformité de la lumière sur les pages, laquelle d'ailleurs est difficile à obtenir dans le cas de la prise de vue amateur. Aujourd'hui, cette étape n'est plus incontournable. Elle reste pratique pour traiter des documents très dégradés — comme certains microfilms — mais les systèmes modernes savent travailler directement avec des images en niveaux de gris ou même en couleur. Certains réseaux apprennent automatiquement à « décider » quels contrastes ou seuils sont pertinents, là où les anciennes méthodes appliquaient une règle fixe.

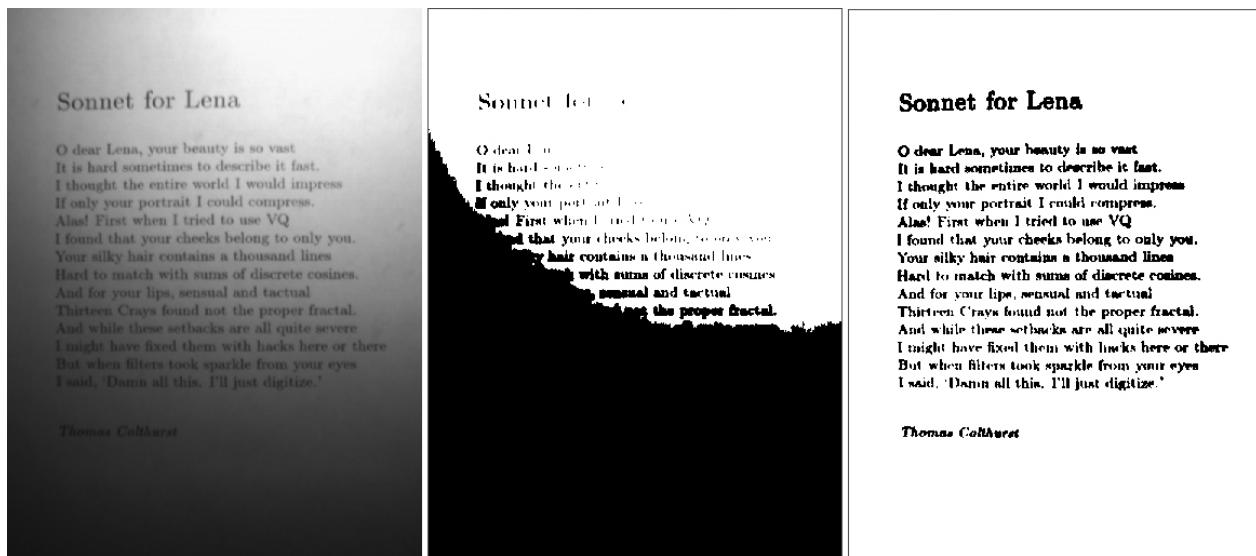


FIG. 4.1 : OfBook

On comprend ici la bascule entre deux époques de l'OCR. Les anciens systèmes fonctionnaient comme une chaîne d'opérations séparées : on nettoyait l'image, on la binarisait, on comparait chaque forme à une base de motifs connus, puis on s'aiderait de dictionnaires pour corriger les erreurs. C'était une sorte de mécanique en plusieurs rouages, où chaque

⁸*Ibid.*

étape pouvait accumuler des approximations. L’OCR, par exemple aux débuts de Tesseract, supposait un document déjà segmenté, ce qui impliquait d’appliquer les traitements sur des entrées standardisées⁹. Les systèmes contemporains, eux, reposent sur des modèles neuro-naux capables d’apprendre l’ensemble du processus de bout en bout : ils prennent une image brute et produisent directement du texte, en intégrant en interne ce qui relevait autrefois de plusieurs étapes distinctes. Cette évolution repose notamment sur les **LSTM** (*Long Short-Term Memory*), des réseaux dits « récurrents » capables de lire une ligne comme une suite d’images et d’en mémoriser le contexte pour prédire la séquence de caractères. Là où un moteur classique devait isoler chaque glyphe pour le comparer à un modèle, un LSTM fonctionne un peu comme un lecteur humain : il garde en mémoire ce qui précède pour interpréter les lettres suivantes, même quand elles sont collées ou abîmées. Ces architectures ne sont évidemment pas infaillibles, mais elles illustrent de manière exemplaire ce que Simondon appelait la *concrétisation technique*¹⁰ : les différentes fonctions du système — nettoyage, segmentation, reconnaissance, correction — deviennent moins séparables et tendent à se fondre dans un dispositif intégré, plus cohérent mais aussi plus opaque. D’un côté, alors, avec le vocabulaire de Gilbert Simondon, on peut dire qu’il y a les dispositifs « abstraits » des méthodes d’OCR ; de l’autre, plus modernes, les dispositifs « concrets »¹¹ ou en voie de concrétisation.

Le schème opératoire « abstrait » – en « rouages » – de la reconnaissance optique des caractères, est décrite dans la littérature des années 1990–2000 comme une architecture en pipeline, où chaque étape de la reconnaissance correspondait à un module distinct : prétraitement, binarisation, segmentation des caractères, classification, puis correction linguistique¹². Dans le cas de Tesseract tel que présenté par Ray Smith, le moteur se contentait de prendre en entrée une image déjà binarisée et supposée contenir une seule colonne de texte¹³. La reconnaissance procédait alors par extraction des composants connectés, regroupés en glyphes, ensuite soumis à un classifieur géométrique, avant qu’une passe linguistique rudimentaire ne privilégie certaines hypothèses sur la base de dictionnaires ou de listes de fréquences¹⁴. Chaque maillon du pipeline pouvait introduire des erreurs, et une approximation en amont — par exemple une mauvaise binarisation ou une segmentation ratée — se répercutait mé-

⁹R. Smith, « An Overview of the Tesseract OCR Engine », dans *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, ISSN : 1520-5363, Curitiba, Parana, Brazil, 2007, p. 629-633, DOI : 10.1109/ICDAR.2007.4376991.

¹⁰G. Simondon, *Du mode d’existence des objets techniques...*

¹¹Vincent Bontems, « Sur la classification des objets techniques selon Simondon », *Artefact. Techniques, histoire et sciences humaines*–3 (mars 2016), ISBN : 9782271087539 Number : 3 Publisher : Presses universitaires du Midi, p. 183-198, DOI : 10.4000/artefact.8270.

¹²R. Casey et E. Lecolinet, « A survey of methods and strategies in character segmentation »...

¹³R. Smith, « An Overview of the Tesseract OCR Engine »...

¹⁴Ray W. Smith, « History of the Tesseract OCR engine : what worked and what didn’t », dans ADS Bibcode : 2013SPIE.8658E..02S, 2013, t. 8658, p. 865802, DOI : 10.1117/12.2010051.

caniquement sur l'ensemble de la chaîne. C'est cette architecture séquentielle, héritée des contraintes matérielles et logicielles de l'époque, que les approches contemporaines cherchent à dépasser en privilégiant des modèles neuronaux capables d'apprendre conjointement plusieurs étapes de traitement. Le tournant marqué en 2018 par l'adoption des réseaux LSTM, qui permettent de traiter une ligne d'image comme une séquence continue de signes, a ouvert la voie à une reconnaissance « end-to-end », de l'image brute jusqu'au texte¹⁵. Mais il serait trompeur de croire que cette évolution a entièrement effacé la logique du *pipeline*. Dans la pratique, même les systèmes modernes conservent une structuration en étapes : prétraitement et normalisation des images, segmentation des zones et des lignes, reconnaissance séquentielle avec LSTM ou Transformers, puis décodage et structuration des données – par exemple avec un export en ALTO/XML – intégration d'un modèle de langue. Les différences résident moins dans l'existence d'un pipeline que dans son degré d'intégration : là où chaque étape était autrefois codée manuellement, elle est désormais en grande partie apprise ou ajustée automatiquement par le modèle¹⁶. Autrement dit, la chaîne ne disparaît pas, mais elle devient plus compacte, moins visible et plus adaptable, ce qui explique à la fois les gains de performance et la difficulté accrue à contrôler ou interpréter chaque rouage.

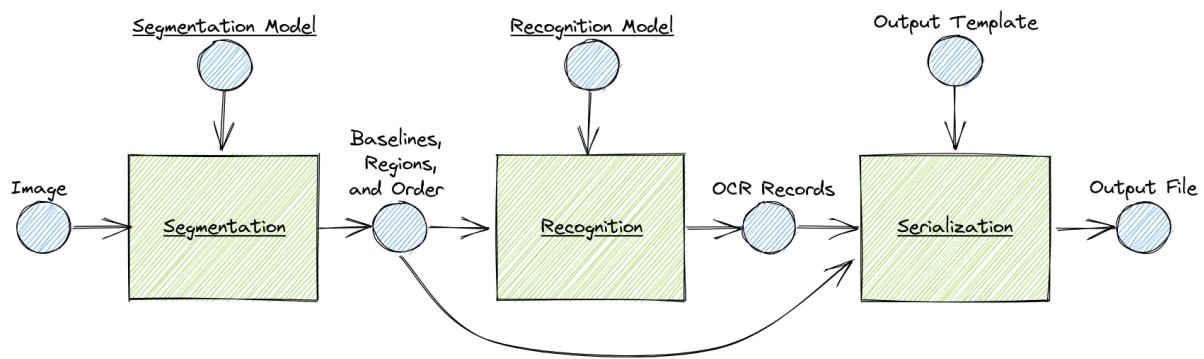


FIG. 4.2 : Schéma de la pipeline Kraken

A noter très récemment une nouvelle évolution des pratiques de transcription automatique avec *Mistral OCR*, présentée en 2025 par *Mistral AI*¹⁷. Contrairement aux OCR classiques centrés sur le texte seul, ce modèle est multimodal – c'est-à-dire qu'il comprend simultanément le texte, les images, les tableaux, voire les équations, les assemblages tex-

¹⁵R. Smith, « An Overview of the Tesseract OCR Engine »...

¹⁶**pipeline-kraken**.

¹⁷*Mistral OCR / Mistral AI*, en, URL : <https://urlr.me/Cqyt9c> (visité le 21/08/2025).

tuels et visuels d'un document, et cela tout en préservant la mise en page et la hiérarchie informationnelle. Il produit une sortie structurée — notamment en Markdown ou JSON — intercalant texte extrait et formes visuelles (images ou équations), avec préservation de l'ordre des éléments, des titres, des tableaux.

II. Les grandes technologies d'OCR

Historiquement, les premiers systèmes d'OCR, développés entre les années 1970 et 1990, étaient conçus pour reconnaître des caractères bien standardisés, tirés d'impressions propres ou de typographies contemporaines. Ils se montraient performants tant que les documents étaient homogènes et en bon état, mais trouvaient rapidement leurs limites face à la diversité des fonds patrimoniaux : imprimés anciens, journaux abîmés, écritures non latines ou manuscrits. C'est cette hétérogénéité croissante qui a conduit au développement de solutions spécialisées, plus flexibles et adaptées aux besoins des bibliothèques et des chercheurs.

Parmi ces moteurs, Tesseract occupe une place singulière. Développé initialement par Hewlett-Packard dans les années 1980 puis repris par Google, il est devenu l'un des logiciels libres d'OCR les plus diffusés au monde, car utilisé notamment dans de grandes bibliothèques numériques comme *Google Books* et publié en *open-source* en 2005¹⁸. Sa quatrième version, publiée en 2018, a marqué une étape importante en intégrant des réseaux de neurones récurrents – les LSTM dont on a brièvement parlé précédemment – ce qui a permis d'améliorer considérablement ses performances sur des typographies variées. Néanmoins, Tesseract reste relativement rigide dès lors qu'il s'agit de traiter des documents abîmés ou des écritures complexes, à l'instar de la typographie vernaculaire ou des écritures manuscrites.

À côté de Tesseract, citons le logiciel propriétaire ABBYY FineReader au coût élevé et au développement fermé limitant les possibilités de personnalisation.

Bien que Tesseract soit devenu un standard libre et largement utilisé pour l'OCR typographique, il n'était pas assez flexible pour répondre aux défis spécifiques des corpus patrimoniaux. C'est dans ce cadre que *Kraken* a vu le jour. Conçu dès l'origine pour pouvoir être entraîné sur des corpus particuliers — imprimés anciens, écritures non latines, manuscrits, gazettes, etc. Benjamin Kiessling, dans *Kraken – a Universal Text Recognizer for the Humanities*¹⁹ décrit comment Kraken élimine les « implicit assumptions on content and layout » propres aux systèmes classiques, tout en permettant l'entraînement d'architectures sur mesure, la reconnaissance de scripts non standards et l'export en formats interopérables tels

¹⁸Id., « An Overview of the Tesseract OCR Engine »...

¹⁹Benjamin Kiessling, « Kraken - A Universal Text Recognizer for the Humanities », dans *Digital Humanities 2019*, Utrecht, Netherlands, 2019, DOI : 10.34894/Z9G2EX.

que PAGE XML ou ALTO XML.

Dans le prolongement de ces initiatives, la plateforme eScriptorium, développée à l'École pratique des hautes études (EPHE/PSL) au sein du projet Huma-Num, combine différents modules de segmentation et de reconnaissance. Elle intègre notamment Pero OCR pour la détection des zones et la segmentation des lignes, et Kraken pour la transcription. eScriptorium offre en outre une interface collaborative qui permet à des équipes de chercheurs, mais aussi à des communautés élargies, d'annoter, d'entraîner et de partager des modèles. Cette dimension participative en fait un outil particulièrement adapté aux humanités numériques, où la correction et l'amélioration collective des résultats constituent un enjeu central.

III. CER, WER : métriques pour évaluer la qualité de l'OCR

Dans ce contexte, la question de l'évaluation des performances devient cruciale. Si la reconnaissance automatique produit un texte exploitable, il est nécessaire de mesurer la qualité de cette transcription afin d'orienter les corrections, de comparer les moteurs ou de juger de l'efficacité d'un entraînement. Deux indicateurs dominent aujourd'hui dans le champ : le **CER** (*Character Error Rate*) et le **WER** (*Word Error Rate*).

Le **CER** mesure le pourcentage d'erreurs au niveau des caractères. Concrètement, il s'agit de comparer la sortie d'un moteur à une transcription de référence (*ground truth* ou « vérité terrain »), puis de calculer le nombre minimal d'opérations nécessaires pour transformer l'un en l'autre : substitutions, insertions ou suppressions. Formellement, ce calcul repose sur la distance de Levenshtein, un algorithme de comparaison de chaînes. Le CER s'exprime comme le rapport entre le nombre d'opérations et le nombre total de caractères de la référence. Ainsi, un CER de 5 % signifie que sur cent caractères transcrits, cinq nécessitent correction.

Le **WER**, de son côté, transpose le même principe au niveau des mots. Il se révèle souvent plus parlant pour l'utilisateur final, car une seule erreur de caractère peut parfois changer entièrement un mot (« loi » devenu « foi »). Mais il est aussi plus sensible aux fautes d'espacement ou de segmentation, ce qui en limite parfois l'usage sur des corpus anciens où les règles typographiques varient fortement.

Dans la pratique, les projets combinent souvent ces deux indicateurs²⁰. Le CER est préféré pour comparer des modèles entre eux, car il est plus fin et moins dépendant des

²⁰David Fleischhacker, Wolfgang Goederle et Roman Kern, *Improving OCR Quality in 19th Century Historical Documents Using a Combined Machine Learning Based Approach*, en, arXiv :2401.07787 [cs], janv. 2024, DOI : 10.48550/arXiv.2401.07787.

conventions d’espacement, tandis que le WER donne une estimation plus intuitive de la lisibilité globale d’un texte. Dans des corpus patrimoniaux difficiles (imprimés du XIX^e siècle, microfilms, manuscrits), atteindre un CER inférieur à 5 % est déjà considéré comme un excellent résultat.

Cependant, il faut rester critique au regard de ces métriques. Comme le montre Neudecker²¹, même les mesures dites « standardisées » – que sont le CER et le WER – ne sont pas directement comparables d’un outil d’évaluation à l’autre. Une source d’écart importante réside dans le calcul des alignements, c’est-à-dire la manière dont on fait correspondre, pas à pas, la transcription automatique produite par la machine et la transcription de référence. Selon l’algorithme d’alignement retenu, par exemple un alignement global de type Levenshtein ou un alignement local optimisé, une même séquence de sorties peut être jugée plus ou moins « erronée ». Cela produit des variations sensibles du CER/WER, surtout dans les cas où l’OCR omet des caractères, inverse des espaces ou fusionne des mots. Dans les humanités numériques, un cas typique est celui des corpus de presse ancienne : un OCR peut confondre les colonnes et inverser l’ordre de lecture, ce qui entraîne de longues séquences décalées. Pour l’algorithme d’alignement, ces erreurs se traduisent en cascades de substitutions et d’insertion/suppressions, gonflant artificiellement le CER, alors que pour l’historien intéressé par la recherche lexicale, la reconnaissance des formes de mots resterait partiellement exploitable. Cette discordance montre que CER et WER n’évaluent pas la pertinence de l’OCR du point de vue des usages, mais reflètent seulement une comparaison mécanique. Elle plaide pour le développement de métriques contextualisées capables d’intégrer la segmentation et l’ordre de lecture, dimensions essentielles pour l’exploitation scientifique des corpus patrimoniaux.

²¹Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos et Stefan Pletschacher, « A survey of OCR evaluation tools and metrics », dans *The 6th International Workshop on Historical Document Imaging and Processing*, Lausanne Switzerland, 2021, p. 13-18, DOI : 10.1145/3476887.3476888.

Point technique

Calcul du CER et du WER

Soit une séquence de référence R (le *ground truth*) et une séquence prédite H (hypothèse). On définit :

- S : le nombre de substitutions,
- D : le nombre de suppressions (*deletions*),
- I : le nombre d'insertions,
- N : le nombre total d'unités dans la référence ($N = |R|$).

La **distance de Levenshtein** correspond à $S + D + I$, soit le nombre minimal d'opérations nécessaires pour transformer R en H .

$$CER = \frac{S + D + I}{N_{\text{caractères}}} \quad WER = \frac{S + D + I}{N_{\text{mots}}}$$

- **CER** : N correspond au nombre total de caractères de la référence.
- **WER** : N correspond au nombre total de mots.

Un score proche de **0 %** indique une reconnaissance quasi-parfaite ; à mesure que l'on progresse vers des valeurs supérieures, le texte devient plus difficilement exploitable.

IV. Numérisation du *Journal Officiel* : une politique documentaire partagée entre la BnF et les institutions parlementaires

Depuis la fin des années 2000, la Bibliothèque nationale de France a développé une politique de numérisation concertée avec ses partenaires, parmi lesquels figurent les bibliothèques parlementaires²². Ce cadre a permis de mettre en place un partenariat durable avec le Sénat et l'Assemblée nationale, en vue de préserver leurs archives et d'en garantir l'accès public. Le Sénat a ainsi confié à la BnF la numérisation de ses *Impressions parlementaires* — débats, rapports et annexes — couvrant la Troisième République : les volumes de 1876 à 1905 et de 1910 à 1940 avaient été traités à la fin de 2021, tandis que la tranche intermédiaire de 1906 à 1909 était encore en cours au début de 2022²³. Pour ma propre recherche centrée sur l'année 1931, cette lacune n'a pas constitué un obstacle, puisque les ressources couvrant la période

²² « La numérisation concertée de corpus d'imprimés. Etat des lieux des programmes et des partenaires. Décembre 2014 », (, 2014).

²³ *Les travaux du Sénat de la Troisième République*, fr, URL : <https://www.senat.fr/connaitre-le-senat/lhistoire-du-senat/les-travaux-du-senat-de-la-troisieme-republique.html> (visité le 21/08/2025).

1910–1940 étaient déjà disponibles en ligne, notamment sur *Gallica* et sur le site du Sénat²⁴. L’Assemblée nationale a engagé une démarche parallèle, rendant accessibles via Gallica ses débats, feuilletons et publications officielles²⁵. Dès 2009, elle avait d’ailleurs conclu un accord avec la BnF pour l’informatisation du catalogue manuscrit (1789–1920) et la numérisation de documents rares, ce qui concernait plutôt son patrimoine ancien que le *Journal officiel* stricto sensu²⁶. Ces campagnes montrent la manière dont les chambres parlementaires ne sont pas de simples « producteurs » de sources, mais de véritables acteurs documentaires : en orientant la sélection, en organisant l’accès et en choisissant les canaux de diffusion (*Gallica*, *Retronews*, portails institutionnels), elles contribuent à instituer de nouvelles modalités de visibilité du droit et du débat démocratique.

Cette organisation illustre également un modèle hybride de gouvernance documentaire : la BnF assure l’infrastructure technique, la numérisation et la diffusion centralisée dans *Gallica*, tandis que les assemblées parlementaires orientent les priorités et conservent une logique éditoriale propre à travers leurs portails²⁷. Ce partage de rôles permet de conjuguer centralisation patrimoniale et valorisation institutionnelle, mais il entraîne aussi une fragmentation des accès : pour reconstituer une série complète, les chercheurs doivent souvent passer de *Gallica* aux sites parlementaires, voire à des plateformes complémentaires comme *Retronews* – proposant une partie des corpus sous accès payant²⁸²⁹.

Du côté de la numérisation proprement dite, les documents sont captés en couleur et produits au format JPEG2000, conformément au référentiel de numérisation des documents « opaques », qui impose une reproduction fidèle de l’original sans retouche automatique et avec une résolution élevée assurant la lisibilité patrimoniale³⁰³¹. L’OCR est systématiquement appliqué et, la qualité de cette conversion est suivie par un « plan d’assurance qualité » qui combine plusieurs méthodes et notamment le CER et le WER que l’on vient d’aborder.

²⁴Violaine Garguilo, *Portails et guides thématiques : Publications officielles France : JORF Débats et documents du Sénat*, fr, URL : <https://bnf.libguides.com/c.php?g=659907&p=4659964> (visité le 25/08/2025).

²⁵Ibid.

²⁶Bibliothèque de l’Assemblée nationale (Paris), fr, Page Version ID : 228055770, août 2025, URL : [https://fr.wikipedia.org/w/index.php?title=Biblioth%C3%A8que_de_l%27Assembl%C3%A9e_nationale_\(Paris\)&oldid=228055770#Vers_une_biblioth%C3%A8que_num%C3%A9rique](https://fr.wikipedia.org/w/index.php?title=Biblioth%C3%A8que_de_l%27Assembl%C3%A9e_nationale_(Paris)&oldid=228055770#Vers_une_biblioth%C3%A8que_num%C3%A9rique) (visité le 25/08/2025).

²⁷Id., *Portails et guides thématiques...*

²⁸Émilien Ruiz, *Accéder aux numérisations du Journal officiel de la République française, de 1871 à nos jours*, fr-FR, janv. 2013, URL : <https://boiteaoutils.info/2013/01/acceder-aux-numerisations-du-journal/> (visité le 19/08/2025).

²⁹BnF-Partenariats, fr, Page Version ID : 222207431, janv. 2025, URL : <https://fr.wikipedia.org/w/index.php?title=BnF-Partenariats&oldid=222207431> (visité le 25/08/2025).

³⁰Numérisation de masse : qualité et formats utilisés pour garantir la conservation, fr, URL : <https://www.bnf.fr/fr/numerisation-de-masse-qualite-et-formats-utilises-pour-garantir-la-conservation> (visité le 25/08/2025).

³¹Bertrand Caron, « Formats de données pour la préservation à long terme : la politique de la BnF » () .



FIG. 4.3

Néanmoins, malgré ces dispositifs de contrôle, l'OCR n'est pas exempt de défauts 4.3 : des éléments de texte peuvent être partiellement tronqués ou effacés, notamment au niveau des marges ou dans les zones de la reliure mal captées, ce qui engendre un certain « bruit » dans les transcriptions, parfois perceptible dans la recherche plein-texte³². Ce phénomène est généralement lié aux contraintes techniques du traitement d'images dans des documents anciens aux reliures serrées ou aux fonds peu perméables à la lumière qui gênent la détection précise des caractères. Dans le cas du *Journal Officiel*, le petit fond est maigre, si bien que les premiers caractères sont avalées par la reliure.

³²Ahmed Ben Salah, Laurent Duplouy et J.P. Moreux, « The digital documents quality control workflow at the BnF (operation, issue, improvement) », *Archiving Conference* (, janv. 2013).

Chapitre 5

Données brutes, données structurées

On l'a vu : la première étape du processus de datafication de sources numérisées et accessibles via les institutions patrimoniales est celle de l'acquisition photographique : un document est scanné ou photographié. On obtient alors une image, fidèle à la source mais muette du point de vue informatique : elle est lisible par l'œil, mais inerte pour un moteur de recherche. La seconde étape, quant à elle, est celle du passage au texte : grâce à la reconnaissance optique de caractères (OCR), l'image est convertie en une suite de signes alphabétiques. On peut alors effectuer des recherches plein texte, mais ce matériau reste linéaire et peu organisé.

La troisième étape, décisive pour l'analyse, est celle de la structuration. Disposer d'un texte est un point de départ, un pré-requis ; mais il s'agit désormais d'identifier dans ce texte des entités, des relations, des attributs — par exemple un nom de sénateur, le sujet de son intervention, le numéro de page correspondant. Ces éléments sont organisés selon un modèle explicite qui rend possible leur comparaison, leur mise en relation et leur analyse. En somme, la structuration est l'opération qui transforme un corpus lisible en données exploitables pour l'analyse. Produire de telles données n'est cependant pas qu'un exercice technique : c'est un travail en trois dimensions — de modélisation (quelle forme donner aux données ?), d'évaluation (comment en juger la qualité ?) et de production (quelles méthodes employer pour transformer un texte en structures fiables ?).

Ce chapitre se concentre sur deux de ces dimensions : comment modéliser des données structurées, et comment les générer à partir de textes dont on veut extraire l'information, par exemple pour des tâches d'indexation. La question de l'évaluation, centrale dans mon stage, quoique abordée ici, fera l'objet d'une partie à part entière.

I. Modéliser des données structurées

La première étape consiste à définir la forme que prendront les données une fois extraites. En effet, la notion de « donnée structurée » recouvre plusieurs types de représentations, plus ou moins adaptées selon les usages et les besoins.

Les *ensembles d'enregistrements* (record sets) correspondent au modèle le plus courant : des collections non ordonnées de tuples, analogues à des tables de base de données. Chaque enregistrement associe une série d'attributs décrivant un objet (nom, fonction, date, etc.). C'est ce format qui est généralement mobilisé dans les tâches d'extraction d'information, telles que la reconnaissance d'entités nommées ou l'extraction de relations. Dans les sciences sociales, on le retrouve par exemple dans les recensements de population, où chaque individu est décrit par un ensemble d'attributs (âge, profession, état matrimonial, lieu de résidence), sans que l'ordre des individus ait d'importance. De même, en histoire politique, la constitution d'un corpus de députés ou sénateurs avec leurs mandats, affiliations et interventions peut être représentée sous forme d'un tel ensemble : une « table » de données où chaque ligne correspond à un parlementaire et chaque colonne à une caractéristique descriptive.

48 Léo Hamon	proposition de loi	création de commissions spécialisées auprès des conseils municipaux
49 Grassard	proposition de résolution	maintenir les parités de change définies pour le franc CFA en décembre 1945
50 Marrane	rapport	aide aux victimes de la Réunion
51 Alain Poher	rapport	réglementation des changes
52 Baron	proposition de résolution	reconsidérer la décision supprimant 5.217 postes dans l'enseignement technique
53 Armengaud	avis	réglementation des changes
54 Charlet	proposition de loi	compléter la loi qui règle les rapports entre locataires et bailleurs de locaux à usage commercial
55 Dorey	rapport	indemniser les viticulteurs de l'Aude victimes de la grêle
56 Henri Buffet	rapport	appliquer à toutes les expéditions de librairie un tarif spécial de transport
57 Durand-Reville	proposition de loi	déterminer le régime fiscal des sociétés coloniales
58 Siaut	rapport	organisation de la production, du transport et de la distribution du gaz
59 Sarrien	rapport	rendre obligatoire le branchement à l'égout dans la ville d'Orléans
60 Caspary	rapport	accorder un congé supplémentaire aux mères de famille exerçant une activité salariée
61 Renaison	rapport	abroger la loi autorisant l'administration des postes et télégraphes à effectuer l'encaissement de
62 Devaud	avis	statut de la formation professionnelle
63 Ott	rapport	enseignement du ski et sur les guides de montagnes
64 Yves Jaouen	proposition de résolution	ajouter deux parlementaires des lieux sinistrés au comité national constitué à cet effet
65 Philippe Gerber	rapport	sinistrés français à l'étranger
66 Devaud	avis	faire verser les allocations familiales entre les mains de la mère de famille

FIG. 5.1 : Fragment d'un tableur produit par une chercheuse à partir du J.O de la Quatrième République, dans le cadre d'un atelier Mezanno. Dans le cadre de l'indexation de la production documentaire des parlementaires, la permutations des lignes n'impacteraient pas le sens des données.

Une variante importante est constituée par les *séquences d'enregistrements* (record sequences), où l'ordre des tuples est significatif. Cette organisation ordonnée peut refléter un critère naturel (par exemple l'ordre temporel) ou une convention éditoriale (comme dans un annuaire ou une table nominative). Dans ces cas, la séquence elle-même porte du sens et facilite certaines analyses (suivi de chronologie, validation croisée, etc.). Ainsi, dans les sciences sociales, l'étude des trajectoires professionnelles ou migratoires repose précisément sur ces séquences : l'enchaînement des postes occupés dans une carrière ou des lieux de résidence

successifs permet de dégager des régularités collectives, de mettre en évidence des bifurcations ou de comparer des parcours types. De même, en histoire parlementaire, les *Tables Nominatives* du *Journal Officiel* ordonnent les interventions des sénateurs selon l'année et la pagination, offrant une séquence exploitable pour analyser l'évolution des prises de parole dans le temps ou la récurrence de certains thèmes.

```
{
  "listes_des_intervenants": [
    {
      "nom_de_famille": "Babin-Chevaye",
      "prenom": "",
      "actions_relatives_a_l_intervenant": [
        {
          "action": {
            "description_action": "Est proclamé secrétaire du Sénat",
            "references_page": [
              8
            ]
          }
        },
        {
          "action": {
            "description_action": "Parle: discuss. d'un projet de loi portant fixation du budget général de l'exer",
            "references_page": [
              582
            ]
          }
        },
        {
          "action": {
            "description_action": "Parle: discuss. d'un projet de loi portant fixation du budget général de l'exer",
            "references_page": [
              719
            ]
          }
        }
      ]
    },
    {
      "nom_de_famille": "Bachelet",
      "prenom": "Alexandre",
      "actions_relatives_a_l_intervenant": "<renvoi d'index>"
    }
  ]
}
```

FIG. 5.2 : Le format JSON restitue ci-dessus une séquence d'enregistrements, ici basée sur l'ordre alphabétique.

D'autres structures plus complexes existent, comme les **arbres**, qui permettent de représenter des hiérarchies ou des relations imbriquées – par exemple en analyse syntaxique –, et les **graphes**, qui offrent une grande flexibilité pour représenter des connaissances consolidées, notamment dans les ontologies ou graphes de connaissances. Ces structures hiérarchiques ou relationnelles sont néanmoins très intéressantes, car elles bénéficient de solides sophistications mathématiques de la théorie des graphes, de l'algèbre linéaire voire de la topologie appliquée aux données. On peut ainsi détecter des communautés (*clustering*), mesurer la centralité ou l'influence de certains nœuds dans un réseau, ou encore identifier des chemins de diffusion ou de dépendance entre entités. En sciences humaines, ces méthodes permettent par exemple

de reconstituer des réseaux de sociabilité, d'analyser la circulation d'idées ou de suivre l'évolution de liens institutionnels. Néanmoins, l'évaluation de ces structures, en particulier les graphes, dépasse généralement le champ de l'extraction d'information proprement dite.

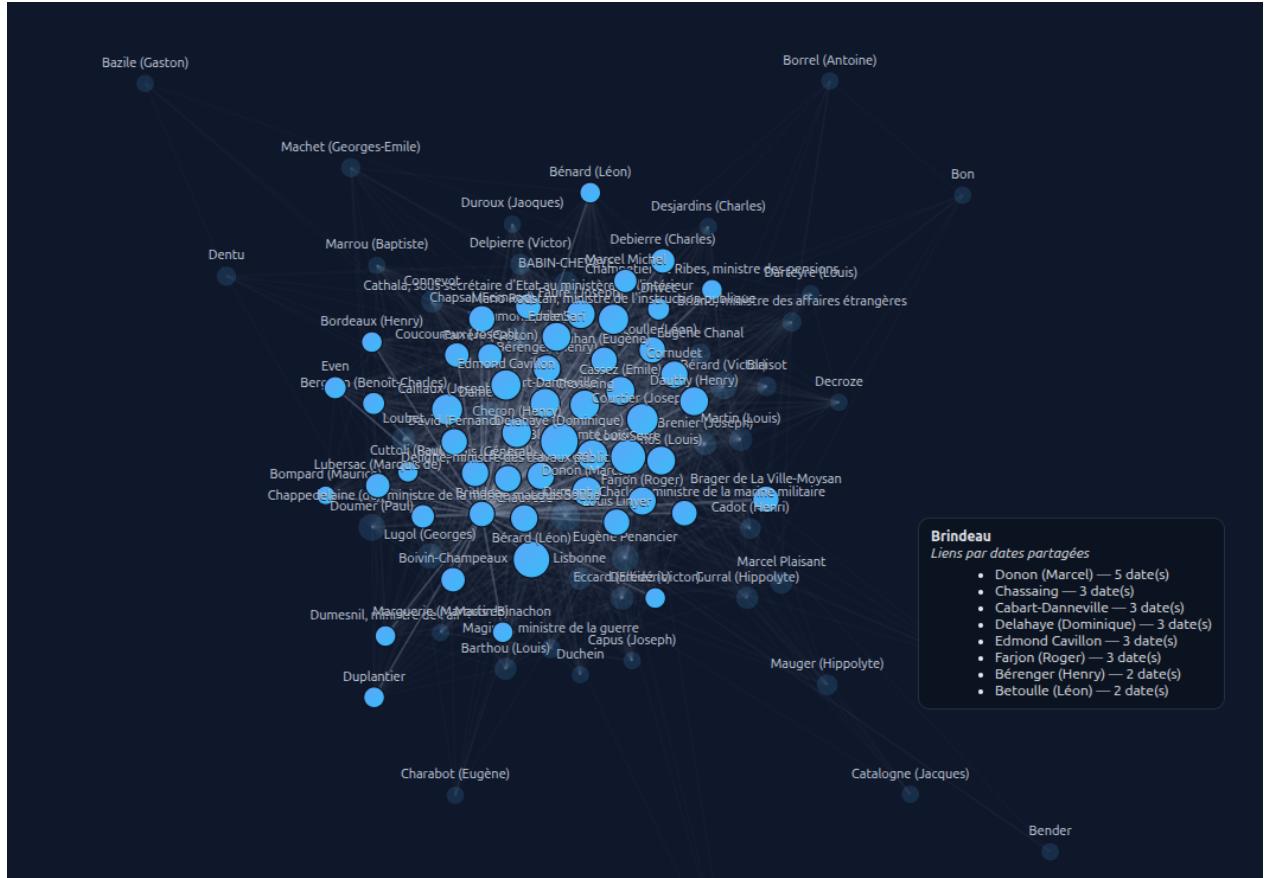


FIG. 5.3 : On peut mettre en relation chaque sénateur (ici représentés par des noeuds) au regard des dates d'interventions partagées.

Dans le cas qui nous occupe, à savoir l'extraction d'informations à partir des documents parlementaires, les *ensembles et séquences d'enregistrements* sont les plus pertinents. Ils correspondent directement à la structure des *Tables nominatives*, où l'on recense des individus (sénateurs), des événements (leurs interventions) et des attributs associés (objets des débats, références de page). C'est donc sur ce type de représentation que nous concentrerons notre attention dans la dernière partie du mémoire. Cependant, la modélisation n'est en pratique pas qu'un exercice intellectuel. L'expérimentation nous renseigne et nous permet d'amender la façon de voir les données, de les mettre en relation. Tout dépend de ce à quoi on tient savoir : si la relation entre les sénateurs importent plus que leur trajectoire intellectuelle, alors un graphe sera plus pertinent qu'une séquence d'enregistrements. Encore une fois, la dimension technique de la modélisation repose sur des enjeux valutatifs de la recherche.

II. Produire des données structurées à partir de texte

Enfin, la question centrale demeure : comment générer des données structurées à partir de texte brut ? Trois grandes familles d’approches dominent : les approches par *pattern matching*, les *approches extractives* et les *approches génératives*¹.

Les approches par *pattern matching* consistent à rechercher dans un texte brut des séquences de caractères ou de mots qui correspondent à des motifs préalablement définis. Elles reposent sur l’idée que certaines informations suivent des régularités formelles ou typographiques que l’on peut capturer par des règles explicites. Les *expressions régulières* – ou *Regex* – en constituent l’exemple le plus connu : elles permettent de détecter des dates, des numéros, ou encore des patronymes en fonction de leur forme.

Les approches extractives reposent sur l’identification directe, dans le texte, des fragments qui correspondent aux champs d’une structure cible. Elles peuvent être mises en œuvre via des règles (expressions régulières, patrons linguistiques) ou via des modèles supervisés d’étiquetage de séquence, comme les *Conditional Random Fields* (CRF)² — des modèles statistiques capables d’apprendre à repérer automatiquement des entités dans un texte à partir d’exemples annotés —, ou encore via des modèles *encodeurs-transformers* tels que **BERT** ou **RoBERTa**³. Ces derniers, pré-entraînés sur de vastes corpus, fournissent des représentations contextuelles des mots : chaque mot est interprété en fonction de son environnement, ce qui permet d’améliorer considérablement la reconnaissance d’entités et la robustesse des extractions. Ces approches présentent l’avantage de limiter le risque d’hallucinations : elles ne « fabriquent » pas d’information absente du texte, puisqu’elles ne font que sélectionner et classer ce qui existe déjà. En revanche, elles nécessitent un entraînement spécifique à la tâche, ce qui implique un corpus annoté, des ressources computationnelles et du temps. Un exemple emblématique de cette méthode est l’outil GROBID, développé à l’INRIA et utilisé par la plateforme HAL⁴, qui mobilise des CRF pour extraire automatiquement les métadonnées des articles scientifiques – par exemple les titres, auteurs et références bibliographiques – à partir de fichiers PDF.

Les approches génératives, en revanche, considèrent l’extraction comme une opération

¹S. Scott Graham, Zoltan P. Majdik et Dave Clark, « Methods for Extracting Relational Data from Unstructured Texts Prior to Network Visualization in Humanities Research », *Journal of Open Humanities Data*, 6–1 (nov. 2020), DOI : 10.5334/johd.21.

²J. R. Finkel, T. Grenager et C. Manning, « Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling »...

³Jacob Devlin, Ming-Wei Chang, Kenton Lee et Kristina Toutanova, *BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv :1810.04805 [cs], mai 2019, DOI : 10.48550/arXiv.1810.04805.

⁴Laurent Romary et Patrice Lopez, « GROBID - Information Extraction from Scientific Publications », *ERCIM News*, Scientific Data Sharing and Re-use 100 (janv. 2015), Publisher : ERCIM, URL : <https://inria.hal.science/hal-01673305> (visité le 28/08/2025).

de traduction : le texte source est « traduit » dans un format cible (JSON, CSV, XML, etc.). Les modèles autoregressifs — c'est-à-dire des modèles qui prédisent chaque mot successif en fonction de tous les mots précédents —, et en particulier les grands modèles de langage (LLM), ont ravivé l'intérêt pour cette voie grâce à leur capacité à généraliser sans entraînement spécifique sur une tâche donnée. Ils peuvent en effet produire des données structurées en *zero-shot*, c'est-à-dire uniquement à partir d'une consigne en langage naturel, ou en *few-shot*, lorsque quelques exemples de sortie attendue suffisent à orienter la génération⁵⁶. Des techniques récentes permettent même de contraindre la génération à respecter des formats prédéfinis, par exemple en filtrant dynamiquement les tokens pour ne retenir que ceux compatibles avec une grammaire donnée⁷. Cette flexibilité rend possible la production de structures complexes et imbriquées, tout en captant des informations implicites. Leur inconvénient majeur reste toutefois le risque d'hallucinations, c'est-à-dire de données hallucinées ou inférées à tort, souvent difficiles à détecter automatiquement.

Dans notre cas, nous avons choisi d'explorer l'efficacité des approches génératives pour traiter les structures répétitives des *Tables nominatives*. Nous mettons à profit les capacités *zero-shot* des LLMs pour évaluer dans quelle mesure ils peuvent produire des données structurées fiables et interopérables à partir de corpus parlementaires. Cependant, il convient de motiver ce choix en abordant les différentes catégories de générations de données que nous venons de donner.

1. Approche à motifs explicites : les ReGex

L'une des premières méthodes mobilisées pour extraire de l'information à partir de textes repose sur l'usage des expressions régulières (Regular Expressions, ou RegEx). Développées dès les années 1950 dans le champ de la théorie des automates, elles se sont imposées comme un outil incontournable pour le traitement automatique de texte. Leur principe est simple : définir un motif formel (un *pattern*) qui décrit la forme que doit prendre une séquence de caractères afin qu'elle soit reconnue par le système. Par exemple, un motif tel que `\d{4}` permet de détecter toutes les occurrences de suites de quatre chiffres, ce qui peut correspondre à des dates dans un texte.

Cette approche présente des avantages certains. Les RegEx sont rapides à exécuter et

⁵Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, *et al.*, *Language Models are Few-Shot Learners*, arXiv :2005.14165 [cs], juill. 2020, DOI : 10.48550/arXiv.2005.14165.

⁶A. Radford, K. Narasimhan, T. Salimans, *et al.*, « Improving Language Understanding by Generative Pre-Training »...

⁷Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, *et al.*, *Emergent Abilities of Large Language Models*, arXiv :2206.07682 [cs], oct. 2022, DOI : 10.48550/arXiv.2206.07682.

extrêmement précises dès lors que l'on connaît à l'avance la forme des données recherchées. Elles sont également peu coûteuses en ressources computationnelles et ne nécessitent aucun entraînement supervisé : il suffit de concevoir le motif pour capturer l'information. Dans le cas de documents semi-structurés ou normés (par exemple la détection de numéros de lois, de dates ou de références bibliographiques), les expressions régulières restent un outil efficace. Elles peuvent aussi servir de composant de prétraitement dans des pipelines plus complexes, en nettoyant ou en normalisant certaines informations avant de les confier à des modèles plus avancés.

Toutefois, leur rigidité limite fortement leur portée. La conception d'une RegEx efficace suppose que l'on ait une connaissance préalable de la forme exacte de ce que l'on cherche — or, dans le cas de sources historiques, cette condition est rarement remplie. Les variations typographiques, les abréviations, les fautes de frappe ou encore les artefacts liés à l'OCR (caractères mal reconnus, coupures de lignes, ligatures) rendent les motifs fragiles. Une expression trop stricte manquera des cas pertinents ; une expression trop souple capturera du bruit et produira des faux positifs. Les approches de recherche floue (*fuzzy matching*) ont été développées pour pallier cette rigidité, mais elles ne permettent que de retrouver ce que l'on connaît déjà à l'avance, et ne s'adaptent pas à des contextes entièrement nouveaux. En d'autres termes, l'approche reste fondamentalement fermée. Elle suppose déjà une connaissance assez fine non pas seulement du « contenu historique » des sources ; mais des suites d'opérations techniques de traduction informationnelle qui commencent à la transcription de l'acte parlementaire au texte OCRisé, en passant par les processus la composition linotypique des imprimeries du *Journal Officiel* et des stratégies d'acquisition numérique.

L'utilisation des RegEx peut être enrichie par des patrons linguistiques plus élaborés (par exemple des grammaires ou des dépendances syntaxiques), par des listes de référence (ou gazetteers, contenant des noms propres, des toponymes, etc.), ou encore par des règles de post-traitement visant à corriger ou normaliser les sorties. Ces ajouts offrent davantage de robustesse et améliorent la précision, mais ils n'échappent pas au principal écueil : la nécessité de concevoir manuellement l'ensemble des règles, ce qui devient rapidement coûteux et peu scalable à grande échelle. L'avantage de ces approches reste toutefois leur explicabilité : contrairement aux modèles neuronaux, les règles définies sont transparentes et permettent de prédire le comportement du système.

Dans le cadre du *Journal Officiel* et des *Tables nominatives*, l'approche par RegEx peut constituer un point de départ utile pour détecter des éléments aux formats relativement standardisés — comme les numéros de page, les années, ou certaines abréviations récurrentes. Cependant, elle montre rapidement ses limites pour capturer la sémantique implicite : qui parle, sur quel sujet, dans quel contexte. C'est précisément ce passage à une structuration

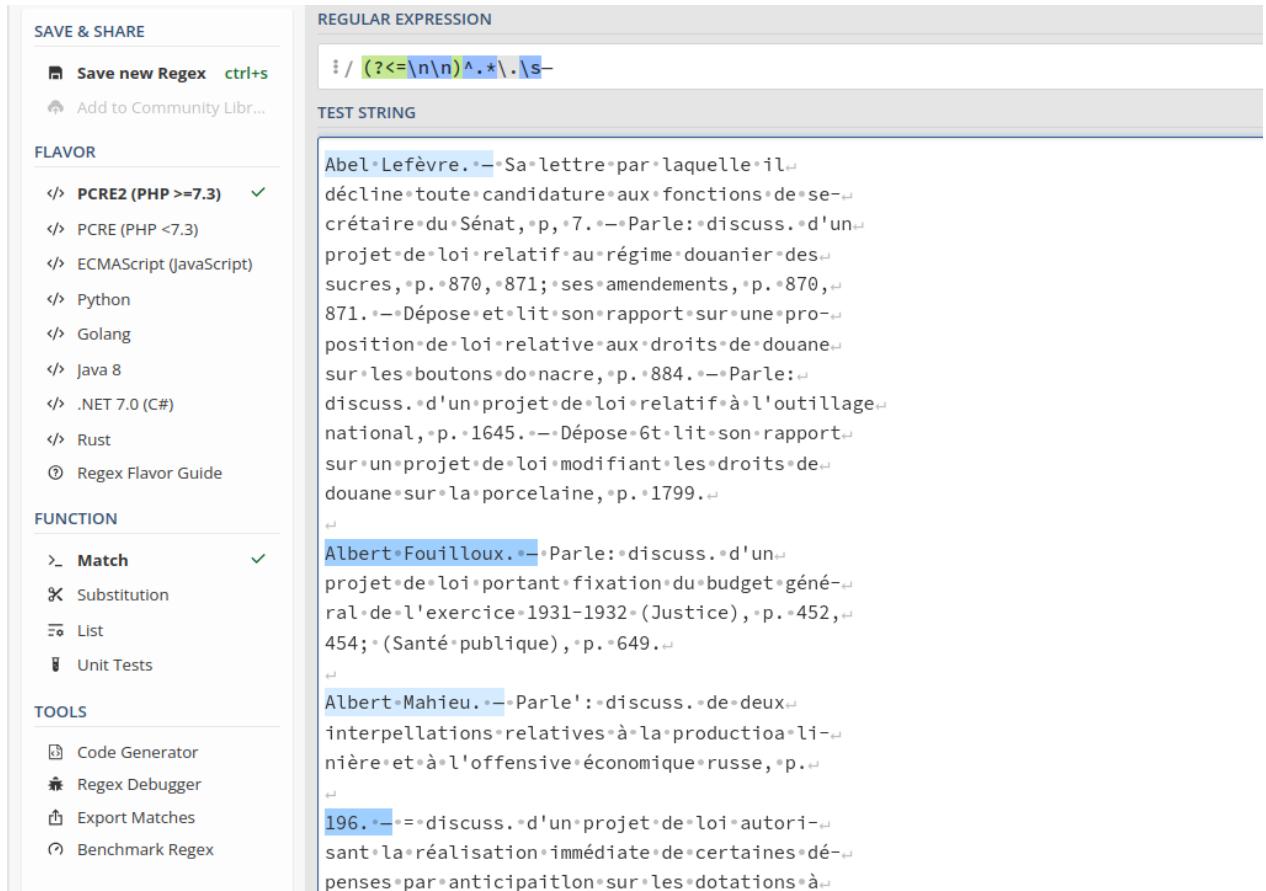


FIG. 5.4 : Exemple de RegEx pour capturer les premiers mots des entrées. La RegEx capture également un numéro de page à cause de l'OCR imparfait qui a doublé des retours chariot. Si on peut avoir une RegEx plus stricte, on prend également le risque d'exclure les noms dont la première lettre aura été mal transcrise – par exemple un « 8 » au lieu d'un « B » ; « 1 » au lieu de « I ». Egaleament, il faudrait considérer toutes les catégories de tirets, ici, seuls cadratins sont considérés. D'autres exceptions sont bien évidement possibles, mais il est difficile de les connaître a priori.

plus riche, allant au-delà de la simple détection de motifs de surface, qui justifie le recours à des approches plus sophistiquées.

2. Approches extractives : BERT

Face aux limites des expressions régulières, la recherche est passée à une nouvelle étape : confier à des modèles statistiques puis neuronaux la tâche de repérer directement, dans un texte, les éléments d'intérêt. Le principe est simple à imaginer : c'est comme si l'on entraînait un étudiant à surligner systématiquement les noms de personnes, les dates ou les thèmes dans un document. Chaque mot ou groupe de mots reçoit ainsi une étiquette (« nom de sénateur », « fonction », « numéro de page », etc.), ce qui permet de transformer le texte en une série d'enregistrements exploitables.

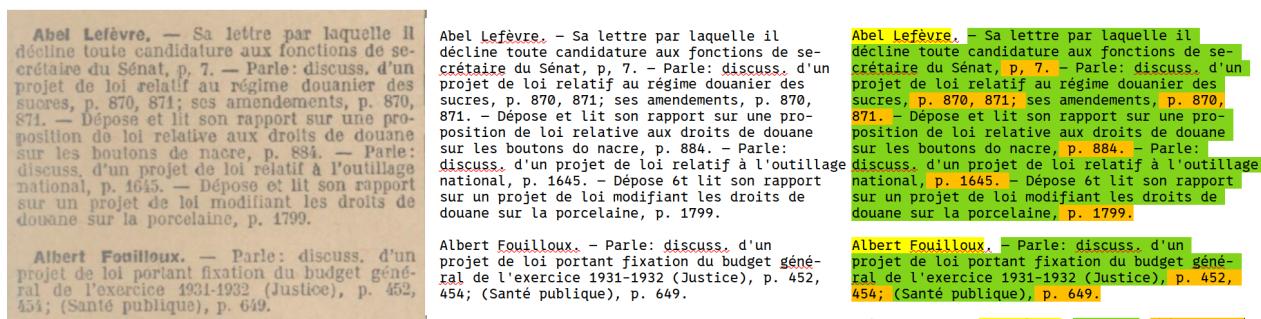


FIG. 5.5 : A gauche, la source (fragment), au milieu le texte OCRIsé, à droite, l'extraction de la sémantique via un Bert qui vient « surligner » l'information.

Au départ, ces systèmes utilisaient des méthodes statistiques assez rigides : on décrivait chaque mot par ses caractéristiques visibles (comme sa terminaison, sa majuscule initiale, son voisinage immédiat) et le modèle apprenait à reconnaître des motifs. Cela marchait, mais seulement si les données ressemblaient beaucoup à celles utilisées pour l'entraînement.

Avec l'apprentissage profond, on est passé à un niveau supérieur. Les réseaux de neurones « séquentiels » ou « récurrents » – par exemple les LSTM dont on parlé précédemment dans les pipelines d'OCR – sont capables de tenir compte de la phrase entière, et donc de mieux comprendre le contexte d'un mot. Ainsi, le système ne se contente pas de voir que « budget » est un substantif, il comprend qu'il apparaît dans un débat sur les finances publiques.

L'arrivée des **Transformers**⁸ en 2017 a marqué une rupture décisive dans le traitement automatique des textes. Contrairement aux approches antérieures fondées sur des réseaux

⁸Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser et Illia Polosukhin, *Attention Is All You Need*, arXiv :1706.03762 [cs], août 2023, DOI : 10.48550/arXiv.1706.03762.

récurrents, qui lisaien t une phrase mot par mot dans un ordre linéaire, les Transformers reposent entièrement sur le mécanisme de *self-attention*. Ce mécanisme permet au modèle de pondérer dynamiquement les relations entre chaque mot et l'ensemble des autres mots de la phrase, quelle que soit leur distance. En d'autres termes, le modèle « décide » à chaque étape quelles parties du texte sont les plus pertinentes pour interpréter un mot donné, et peut ainsi se focaliser sur les éléments décisifs pour comprendre le sens d'une phrase ou même d'un paragraphe entier. Ce changement de paradigme a ouvert la voie à des modèles pré-entraînés massifs, comme **BERT** et ses dérivés, capables de capturer des régularités linguistiques bien au-delà des méthodes séquentielles classiques.

Les modèles **BERT**⁹ – qui implémentent cette architecture conceptuelle – exploitent ainsi pleinement ce paradigme de l'attention pour construire des représentations contextuelles bidirectionnelles. Contrairement aux modèles séquentiels antérieurs – comme les réseaux récurrents ou les LSTM –, qui lisaien t le texte de gauche à droite ou de droite à gauche, BERT intègre simultanément le contexte de gauche et de droite : il ne traite pas les mots un à un, mais les situe dans l'ensemble de la phrase. Cela revient, pour vulgariser, à lire une phrase en « surlignant » dynamiquement chaque mot tout en tenant compte de l'ensemble de la phrase. Entraîné sur des milliards de mots issus de corpus variés, BERT acquiert ainsi une véritable « intuition linguistique » réutilisable dans des domaines spécialisés, comme par exemple l'identification d'entités dans les débats parlementaires.

Les avantages sont clairs : ces modèles reconnaissent mieux les entités malgré les variations de style ou les petites erreurs d'OCR, et ils nécessitent beaucoup moins de travail manuel pour concevoir des règles. Cependant, il y a un revers : ils ne fonctionnent bien que si l'on dispose d'un jeu d'exemples annotés pour les entraîner sur la tâche précise (par exemple, un lot de pages du *Journal Officiel* où les sénateurs et les objets d'intervention sont déjà identifiés). Or, produire ces annotations coûte du temps et de l'expertise. De plus, l'entraînement suppose une certaine puissance de calcul (GPU, logiciels spécialisés), pas toujours disponible dans des projets en sciences humaines.

En résumé : les approches extractives modernes comme BERT sont de véritables surligneurs automatiques très performants, mais ils demandent une phase préparatoire coûteuse (annotations + infrastructure). C'est ce qui explique que les chercheurs en histoire ou en sciences sociales se tournent de plus en plus vers les modèles génératifs (LLMs), qui contournent en partie ce verrou.

⁹ J. Devlin, M.W. Chang, K. Lee, *et al.*, *BERT...*

3. Approches génératives : les LLMs

L'émergence des grands modèles de langage (LLM) – tels que GPT, Claude, LLaMA ou Mistral – a transformé le champ de l'extraction d'information. Par *LLM*, on désigne des modèles d'intelligence artificielle entraînés sur d'immenses corpus textuels (presse, littérature, pages web, documents spécialisés), capables de prédire la suite d'un texte en tenant compte du contexte. Cette propriété, fondée sur l'architecture des Transformers, leur confère une capacité de généralisation : ils peuvent répondre à des questions, résumer, traduire ou extraire des informations sans entraînement spécifique. Contrairement aux approches dites extractives, qui consistent à surligner et classer des portions du texte existant, les LLM reposent sur un principe génératif : ils produisent directement une sortie textuelle, guidée par des instructions. L'extraction d'information n'est donc plus envisagée comme une classification locale, mot par mot, mais comme une reformulation structurée : le modèle lit le texte, en sélectionne les éléments pertinents et les restitue sous une forme organisée, prête à l'analyse.

Le cœur de cette approche réside dans l'instruction donnée au modèle, ce que l'on appelle désormais un *prompt*. Celui-ci peut préciser le type de sortie attendu, qu'il s'agisse d'un tableau, d'un schéma en JSON ou d'une liste hiérarchisée, et inclure des exemples pour orienter la génération. Ainsi, à partir d'un passage du *Journal Officiel* mentionnant « M. Lefèvre intervient au sujet du budget des chemins de fer (p. 234) », le modèle ne se contente pas d'indiquer les entités en jeu ; il est capable de reconstruire une représentation explicite de l'information, par exemple : « nom » : « Dupont », « sujet » : « budget des chemins de fer », « page » : 234¹. Cette capacité à produire une structure à la volée marque un changement de paradigme. L'utilisateur ne construit plus un modèle spécialisé via un entraînement complexe, mais formule une consigne linguistique à laquelle le modèle s'adapte immédiatement.

Cette approche présente plusieurs atouts. Elle se distingue d'abord par sa flexibilité : un même modèle peut accomplir une grande variété de tâches, qu'il s'agisse d'extraire des entités, de résumer des documents ou de convertir un texte en tableau, sans nécessiter de jeu de données annoté ni d'entraînement spécifique. Elle se caractérise également par sa rapidité de mise en œuvre : quelques *prompts* bien conçus suffisent à obtenir des résultats exploitables, ce qui réduit considérablement les coûts et les délais d'un projet. À cela s'ajoute une robustesse linguistique liée au pré-entraînement massif de ces modèles sur d'immenses corpus : ils reconnaissent des formulations inhabituelles, des variations lexicales ou syntaxiques, et produisent malgré tout une sortie cohérente. Enfin, leur capacité à générer directement dans des formats structurés tels que le JSON, le CSV ou l'XML rend leurs productions immédiatement intégrables dans des bases de données ou des environnements analytiques.

Mais ces avantages ne doivent pas masquer certaines limites. Les sorties générées ne sont pas toujours stables : un même prompt appliqué à des passages similaires peut donner

lieu à des structures divergentes, ce qui complique l’automatisation complète du processus. Les modèles sont en outre sujets aux hallucinations, c'est-à-dire qu'ils peuvent inventer des données absentes du texte original, une dérive particulièrement problématique du point de vue scientifique. La qualité des résultats dépend aussi de la formulation du prompt, qui devient un paramètre méthodologique central et parfois délicat à maîtriser. Enfin, si l'usage ponctuel d'un LLM est peu coûteux, le traitement de corpus volumineux peut rapidement représenter un investissement financier ou technique non négligeable, selon que l'on mobilise des API commerciales ou des modèles libres installés localement.

Dans le domaine des sciences humaines et sociales, ces modèles ouvrent néanmoins des perspectives inédites. Ils permettent d'extraire automatiquement des informations structurées à partir de corpus historiques, en évitant l'étape fastidieuse et coûteuse de l'annotation manuelle. Des expérimentations récentes ont déjà montré leur potentiel pour la transcription, l'annotation ou l'extraction d'entités dans des fonds patrimoniaux. Pour le *Journal Officiel* et ses *Tables nominatives*, ils offrent la possibilité de transformer directement des listes semi-structurées en tableaux exploitables, en capturant les noms des sénateurs, les objets d'intervention et les références paginées. Une tâche autrefois fastidieuse, nécessitant un dépouillement patient et souvent incomplet, devient ainsi automatisable à grande échelle.

En définitive, les approches génératives déplacent le centre de gravité de l'extraction d'information. Là où l'on cherchait auparavant à construire des modèles spécialisés et rigides, il s'agit désormais d'orchestrer une génération encadrée par des consignes précises. Cette plasticité ouvre un champ nouveau pour la capture sémantique des corpus textuels, mais elle suppose d'accompagner leur usage de garde-fous méthodologiques, afin de garantir la fiabilité et la traçabilité des données produites.

III. La sortie structurée via LLM pour le *Journal Officiel*

Il est intéressant de revenir sur l'approche générative car c'est celle-ci qui a été retenue. En effet : l'application des approches génératives par grands modèles de langage, dans le cas du *Journal Officiel*, et plus spécifiquement des *Tables nominatives* du Sénat, a l'avantage d'être simple et rapide à mettre en place. Ces tables, qui fonctionnent comme des index annuels de l'activité parlementaire, présentent une organisation qui semble régulière : elles listent les noms des sénateurs, précisent le sujet de leurs interventions et renvoient aux pages correspondantes. Cette apparente homogénéité masque cependant une série de difficultés. Cette partie ne présentera pas les détails techniques, mais quelques enjeux techniques et

avantages à considérer, lesquelles seront développés dans la partie suivante. La mise en page, par exemple, peut varier d'une année à l'autre ; les noms sont parfois abrégés ou tronqués ; et les textes portent les stigmates d'un OCR imparfait, introduisant du bruit ou des confusions typographiques. À cela s'ajoute la diversité des formulations, qui rend incertaine toute tentative de repérage par motifs fixes. Capturer la sémantique de ces tables suppose donc de dépasser la simple reconnaissance de structures formelles et d'accéder à une représentation organisée et exploitable de l'information.

1. Simplicité

Les LLMs offrent un compromis assez intéressant. Contrairement aux approches extractives classiques, qui nécessitent la constitution d'un corpus annoté et l'entraînement d'un modèle spécialisé, les LLMs peuvent être mobilisés rapidement avec un simple appel d'API. Il suffit de leur fournir quelques exemples de sortie attendue au sein d'un prompt pour orienter leur génération. Cette souplesse en fait une solution pragmatique, adaptée à des projets exploratoires en sciences humaines où le temps et les ressources d'annotation peuvent être limités. Une campagne d'annotation peut être longue voire approximative, sans accord inter-annotateur qui permet de désambiguifier l'étiquetage des données. De plus, le recours à ces API accessibles ou à des modèles libres déployés localement permet de réduire les coûts techniques initiaux, tout en laissant aux chercheurs la possibilité de se concentrer sur les choix méthodologiques plutôt que sur les contraintes d'infrastructure.

2. Sortie structurée, génération structurée

C'est ici qu'entre en jeu une distinction essentielle : celle entre **sortie structurée** et **génération structurée**. Dans le premier cas, dit de sortie structurée, on exploite la capacité du LLM à imiter des formats qu'il a déjà rencontrés lors de son entraînement. Au lieu de laisser parler le LLM avec la prolixité qu'on lui connaît, on le « guide » par le prompt en lui imposant un format attendu — par exemple un dictionnaire JSON comportant les champs ‘« nom » : « ... »’, ‘« intervention » : « ... »’, ‘« page » : « ... »’. Le modèle, plutôt que de produire une réponse discursive en langage naturel, ajuste alors son texte pour respecter ce format. C'est une manière pragmatique et peu coûteuse d'obtenir des données directement exploitables. La génération structurée va un pas plus loin : elle ne se contente pas de guider la sortie a posteriori, mais constraint en amont les chemins de génération possibles. Techniquement, il s'agit d'agir sur les probabilités de production des tokens, en forçant le modèle à n'émettre que des suites de symboles compatibles avec une grammaire donnée, comme JSON ou XML. Cette mise « sur rails » de la génération, qu'on peut comparer à un automate fini qui valide

chaque token au moment où il est choisi, permet d'éviter les écarts ou les incohérences de format.

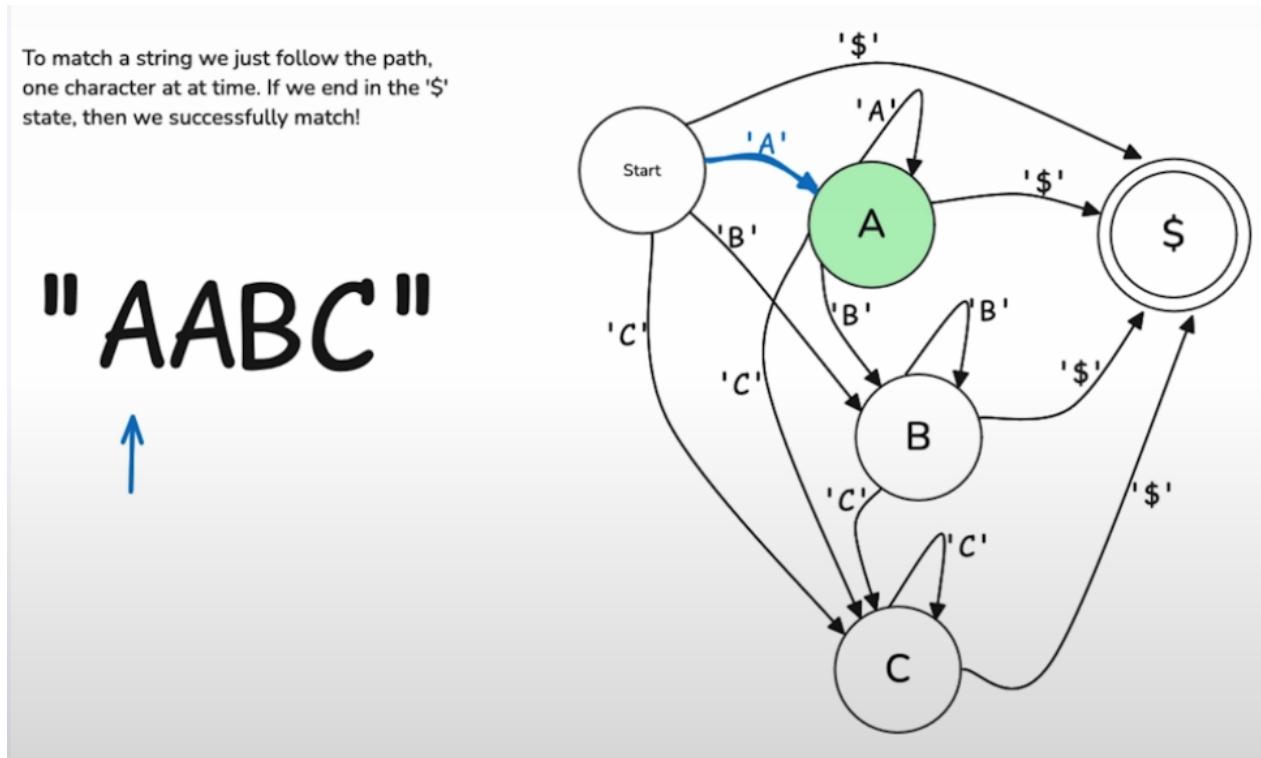


FIG. 5.6 : Automate fini. Source : Outlines (Open AI)

Dans notre cas, comme on le verra dans la partie suivante, la première approche aura été choisie : une sortie structurée imposée par le prompt, sans recours à une génération contrainte au niveau des probabilités. L'avantage de ces deux méthodes, et particulièrement de la sortie structurée, est de rendre possible une intégration directe dans des pipelines d'analyse. Le texte brut devient immédiatement une table exploitable ou un ensemble d'enregistrements importables dans une base relationnelle. En outre, cette structuration assure la traçabilité : chaque donnée extraite peut être reliée au fragment de texte d'où elle provient, ce qui rend possible une vérification par l'historien.

Il reste néanmoins essentiel de souligner les limites et les précautions que cette approche implique. Les sorties produites par les LLMs ne sont pas toujours stables : un même prompt appliqué à des passages semblables peut générer des structures divergentes, incomplètes ou incohérentes. Les modèles peuvent aussi souffrir d'hallucinations, inventant des informations absentes du texte original. Enfin, la qualité des résultats dépend fortement de la formulation du prompt, dont de petites variations peuvent entraîner de grandes différences dans les réponses. Pour ces raisons, la mise en place de protocoles de validation est indispensable. Cela peut passer par un contrôle humain sur des échantillons, par l'application de règles automa-

tiques de cohérence (par exemple vérifier qu'un numéro de page est bien un entier valide), ou encore par l'alignement systématique des sorties avec des référentiels externes, tels que les listes officielles des sénateurs d'une année donnée. Ces garde-fous garantissent que les données produites conservent leur fiabilité scientifique et puissent être mobilisées dans le cadre d'une recherche historique rigoureuse. Une fois obtenues, ces données structurées ouvrent un large éventail de possibilités : suivre longitudinalement l'activité des sénateurs, mesurer la fréquence et la thématique de leurs interventions, ou encore croiser ces informations avec d'autres corpus, tels que les débats parlementaires complets ou les archives législatives. Là où le dépouillement manuel demeurait fastidieux et fragmentaire, la structuration par LLM transforme le *Journal Officiel* en un matériau interrogable, apte à nourrir des analyses sérielles, des visualisations de réseaux d'intervenants ou des comparaisons thématiques.

Cependant, si l'utilisation des LLMs pour la capture sémantique du *Journal Officiel* constitue une approche pragmatique, combinant rapidité de mise en œuvre et faible coût – notamment parce qu'elle évite la lourde étape d'annotation –, elle comporte des risques majeurs. Les modèles peuvent produire des hallucinations, c'est-à-dire des sorties erronées, incomplètes ou inventées, qui fragilisent leur usage scientifique. Dès lors, l'enjeu central n'est plus seulement de produire des données, mais de mesurer le degré de confiance que l'on peut accorder aux résultats. Cette évaluation doit être envisagée à l'échelle de toute la chaîne de traitement : depuis l'image OCRisée jusqu'aux données structurées produites par le modèle.

3. L'outil « Corpusense » du projet Mezanno : une pipeline de l'image numérisée à la donnée structurée

Nous avons décrit, jusqu'ici, les différents passages d'une chaîne de traitement : de la source à sa numérisation en mode image ; de l'image à sa transcription textuelle par OCR ; et enfin du texte brut à une structuration exploitable. Chacune de ces briques comprend elle-même des sous-chaînes, parfois complexes, comme dans le cas de l'OCR. Mais une fois les choix techniques arrêtés — souvent contraints par un cahier des charges où comptent la simplicité de mise en œuvre et les coûts —, il devient possible de stabiliser une pipeline cohérente, adaptée à une typologie documentaire précise.

Or, toutes les configurations documentaires n'appellent pas les mêmes traitements. La reconnaissance d'imprimés homogènes ne relève pas des mêmes techniques que celle des manuscrits. Un corpus comme le *Journal Officiel*, avec ses colonnes régulières et ses tables nominatives, offre une structure répétitive qui se prête bien à une automatisation de bout en bout. À l'inverse, la presse politique du XIX^e siècle, mêlant rubriques, illustrations et typographies variées, déjoue les chaînes trop uniformes et exige des ajustements manuels plus

lourds. C'est pourquoi il paraît irréaliste d'espérer un outil unique, « universel », capable de traiter indifféremment toutes les configurations documentaires. Une approche pragmatique et utile pour les chercheurs en sciences humaines et sociales consiste à développer des pipelines spécialisées, pensées pour une catégorie de documents donnée et pour un usage scientifique identifié. La concrétisation de pipelines vise la solidarisation des rouages, de sorte à former un outil cohérent bénéficiant d'une certaine unicité et d'une intégration à la réticularité technique documentaire. La mise en oeuvre d'une application Web, n'exigeant pas d'installation de logiciels spécifiques ou d'expertise informatique précise, permettrait de répondre à la fois à la concrétisation du schème opératoire transcriptif et structurant tout en s'insérant à l'écologie des sources disponibles notamment via IIIF.

C'est précisément ce choix qu'assume le projet Mezanno, dans la continuité d'AGODA¹⁰ et de SoDuCo¹¹. En ciblant les documents sériels, il s'appuie sur leur régularité formelle — colonnes, entrées répétitives, enregistrements standardisés — pour proposer un outil qui va au-delà de la seule transcription. Contrairement à *eScriptorium*, qui se concentre sur la reconnaissance et l'annotation collaboratives de textes manuscrits ou imprimés, l'application Web *Corpusense* vise une sortie structurée dès la fin du processus. Autrement dit, le résultat n'est pas seulement un texte OCRisé, mais directement une table exploitable, où les entités et attributs pertinents (noms, sujets, pages, dates, etc.) sont alignés et prêts à être interrogés. Un des atouts d'une focalisation sur les documents sériels tient aussi à leur caractère répétitif, plus difficile à valoriser que des estampes scientifiques ou des revues artistiques d'avant-garde. Là où ces sources peuvent sembler fastidieuses à dépouiller manuellement — précisément parce qu'elles sont redondantes, uniformes — elles se prêtent particulièrement bien à une structuration automatique. La régularité qui en faisait jadis la « monotonie » pour l'historien devient ici un avantage méthodologique, puisque la machine excelle à identifier et organiser des motifs récurrents. Ce renversement illustre combien les corpus jugés les plus arides, tels que les annuaires, registres ou tables, constituent un terrain privilégié pour expérimenter des pipelines de traitement : ils transforment la répétition en données analysables, ouvrant ainsi la voie à des analyses sérielles et comparatives à grande échelle.

Le pari est donc double : d'un côté, capitaliser sur la « facilité relative » qu'offrent les documents sériels pour tester et fiabiliser la pipeline ; de l'autre, répondre à un besoin concret des historiens, celui de disposer rapidement de données comparables, exportables et réutilisables, plutôt que de simples images ou transcriptions linéaires. Cette question de

¹⁰Fanny Lebreton, « Vers l'ouverture et l'exploration des débats parlementaires : étude d'une méthodologie de structuration et d'enrichissement automatique des données. L'exemple des débats à la Chambre des députés durant la Ve législature de la IIIe République (1889-1893) » (, oct. 2022), p. xv, URL : <https://dumas.ccsd.cnrs.fr/dumas-04538872> (visité le 23/08/2025).

¹¹N Abadie, S Baciocchi, E Carlinet, J Chazalon, P Cristofoli, B Duménieu, J Perret et S Tual, « Approche du projet SoDUCo » () .

l'usage de ces données est, bien évidemment, solidaire à une exigence de scientificité : peut-on faire confiance aux données produites, lesquelles sont produites le long d'une chaîne de traitement. Cette chaîne de traitement, on l'a vu, compose avec l'erreur. La question de l'évaluation de ces données est donc cruciale et constitue un enjeu à part entière, enjeu qui a pris une place centrale pendant mon stage.

Troisième partie

**Expérimenter et évaluer pour
comprendre : une démarche
historienne outillée**

Chapitre 6

L'outil *Corpusense* : une chaîne de traitement pour les sources historiques

L'outil *Corpusense* du projet Mezanno a pour ambition de proposer aux chercheurs et chercheuses en sciences humaines et sociales un dispositif dédié à l'exploitation de corpus d'archives sérielles. Là où la partie précédente du mémoire a montré la diversité des enjeux liés à la numérisation et à la mise à disposition de ces sources — qualité imparfaite de l'OCR, absence de structuration exploitable dans les textes bruts, masse de données qui rend toute saisie manuelle irréaliste — *Corpusense* vise à apporter une réponse technique réaliste. « Réaliste », car il s'appuie pour cela sur des briques logicielles déjà disponibles et rapides à mettre en place, qu'il agence en une chaîne cohérente correspondant au schème opératoire transcriptif défini précédemment, allant de la source numérisée à la donnée structurée.

Le fonctionnement de *Corpusense* repose principalement sur l'usage d'**API**, c'est-à-dire d'interfaces de programmation qui permettent à des logiciels hétérogènes de communiquer entre eux en suivant un protocole défini. Les API fonctionnent comme des guichets d'information qui délivrent, selon les requêtes, des données demandées, par exemple une série d'images et leurs métadonnées dans le cas de **IIIF**. Plutôt que de réimplémenter des modules complexes, l'outil tire parti de services spécialisés en les appelant directement via leurs API. Trois d'entre elles structurent le dispositif : IIIF, donc, qui est le standard largement adopté par les institutions patrimoniales, qui permet de charger et manipuler des images numérisées de manière normalisée. La constitution de corpus se fait donc à partir des dépôts d'archives numérisées, en général par des institutions patrimoniales. Il faut compter également une « APIsation » du moteur **Pero OCR** par l'EPITA ; qui fournit comme on l'a vu une reconnaissance optique des caractères adaptée aux corpus historiques et multilingues. Et, enfin, l'API de **Mistral**, un modèle de langage, mobilisée pour transformer les textes OCRisés en sorties structurées adaptées aux besoins des chercheurs (voir 6.1). Ainsi, il s'agit

I. Une instance de pipeline « classique »

de reconstituer une chaîne opératoire, composé de différentes briques agencées et constituées en outil.

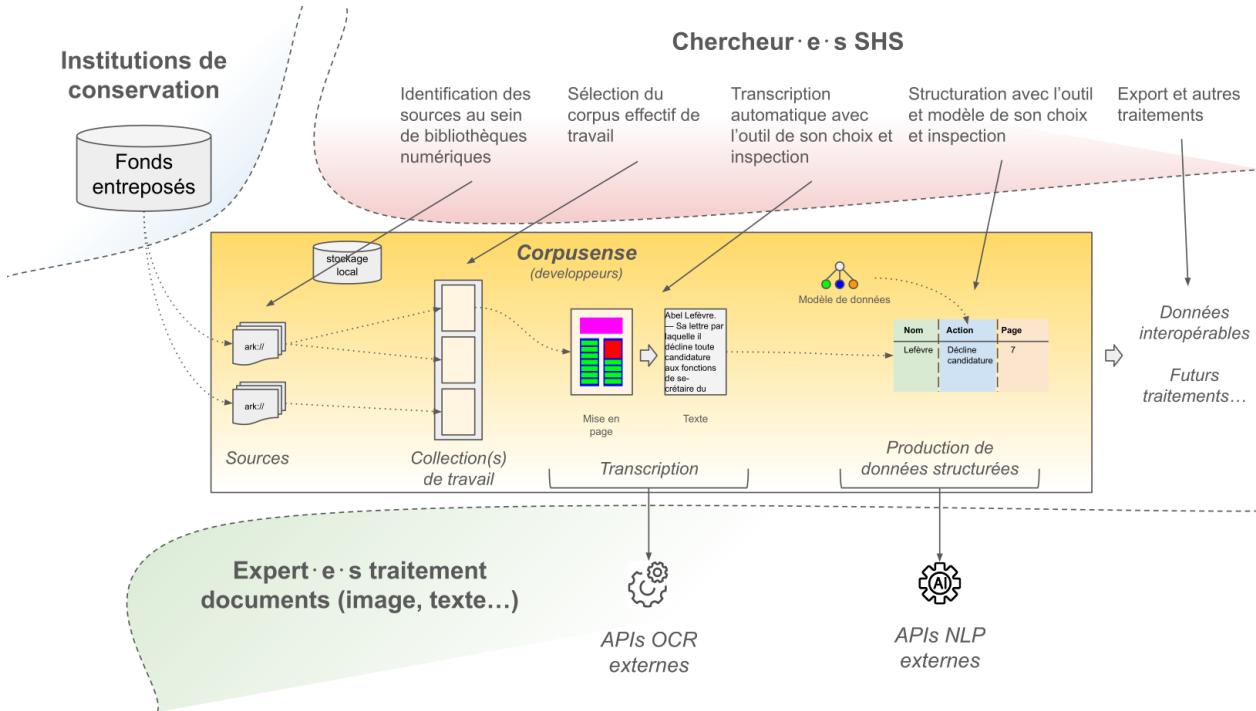


FIG. 6.1

L'outil motive donc une double orientation. D'une part, il repose sur une chaîne de traitement unifiée, pensée pour transformer des documents d'archives sous forme image en données structurées, interopérables et prêtes à être analysées. D'autre part, il se veut flexible et accessible, de façon à laisser aux chercheurs SHS une autonomie réelle dans la conduite de leurs travaux. Concrètement, une application Web comme *Corpusense* permet de constituer des corpus documentaires à partir de dépôts variés d'archives numérisées et d'y appliquer, sans compétences techniques avancées, des traitements qui aboutissent à des données exploitables.

I. Une instance de pipeline « classique »

L'objectif de *Corpusense* veut fournir une infrastructure technique robuste pour solidariser les différents rouages (à savoir la constitution des corpus, OCR, extraction) 6.2.

La chaîne de traitement se déploie en plusieurs étapes successives. Tout d'abord, la sélection et organisation des sources. Le point de départ réside dans la constitution du corpus. Celui-ci peut provenir de fonds institutionnels (par exemple Gallica, la BnF ou des archives universitaires), ou de collections numérisées indépendantes. Les documents, le plus souvent disponibles sous forme d'images, sont alors recensés et organisés dans un format exploitable.



FIG. 6.2

Dans ce contexte, *Corpusense* s’appuie sur le protocole IIIF, largement adopté dans le domaine patrimonial. IIIF permet non seulement d’accéder aux images numérisées de manière normalisée, mais aussi de les manipuler (zoomer, rogner, annoter) et de les intégrer de façon homogène, quelle que soit l’institution d’origine. Ce recours à un standard interopérable assure la portabilité des corpus et facilite leur exploitation au-delà du cadre spécifique de ce projet.

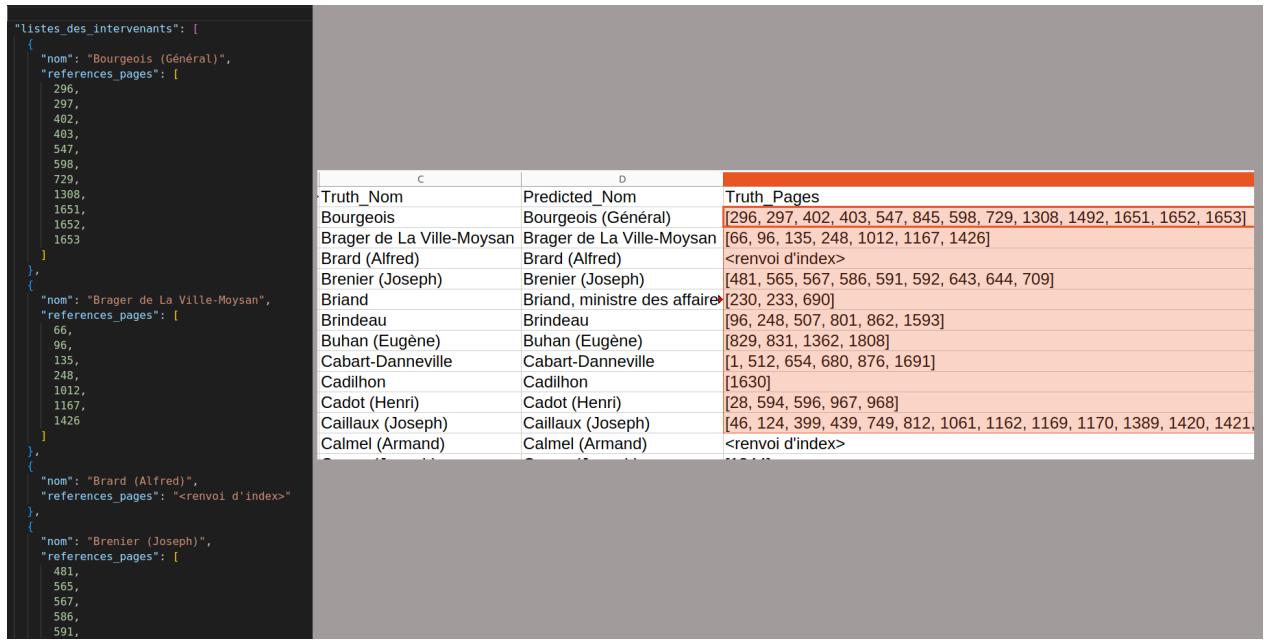
Ensuite, la transcription par OCR. La deuxième étape consiste à convertir ces images en texte grâce à un moteur de reconnaissance optique de caractères (OCR). *Corpusense* s’appuie principalement sur le moteur PERO OCR.

Enfin, la sortie structurée (avec l’API Mistral), laquelle produit un document en JSON, qui peut d’ailleurs être converti aisément en format CSV.

Le choix du format JSON pour représenter les données structurées ne relève pas seulement de la contrainte imposée par l’API Mistral même si cette disponibilité oriente un tel choix. Ce format présente en fait plusieurs avantages décisifs dans le cadre d’un outil comme *Corpusense*. D’un point de vue technique, JSON est un standard commun pour l’échange de données : léger, lisible par l’humain, directement exploitable par la plupart des langages de programmation et facilement convertible en d’autres formats, qu’il s’agisse de tables ou de bases de données relationnelles ou documentaires. Cette plasticité garantit la réutilisation des résultats, quel que soit l’environnement de recherche dans lequel ils sont ensuite mobilisés. L’intérêt du JSON réside dans l’implémentation de la dialectique clé/valeur, laquelle se prête particulièrement bien à des tâches « indexatoires », davantage convenante à la logique tabulaire, plus rigide. Là où un tableau, par exemple au format CSV, oblige à « aplatisir » l’information, JSON permet de conserver les relations entre entités (par exemple entre un intervenant et les différentes pages où il est cité) et d’accueillir des variations de granularité sans perdre la cohérence du tout. Par exemple, chaque sénateur peut intervenir un certain nombre de fois : en JSON, chaque référence d’intervention est un élément manipulable, distinct des autres ; et l’ensemble de ces références est une liste de taille variable. Dans un tableau, on pourrait soit agréger ces différentes références dans une unique colonne séparés avec un séparateur arbitraire 6.3 – impliquant alors de parser *a posteriori* ces séries de nombres – ou bien, mais c’est ici une option assez malheureuse, de constituer autant de colonnes que d’interventions. La séparation des valeurs en JSON fait partie de sa grammaire

I. Une inscription dans l'outil Corpusense : une chaîne de traitement pour les sources historiques

quand, côté tables, la séparation de valeurs numériques est en fait une chaînes de caractères qu'il faut « spliter » pour retomber sur une certaine modularité.



C	D	
Truth_Nom	Predicted_Nom	Truth_Pages
Bourgeois	Bourgeois (Général)	[296, 297, 402, 403, 547, 845, 598, 729, 1308, 1492, 1651, 1652, 1653]
Brager de La Ville-Moysan	Brager de La Ville-Moysan	[66, 96, 135, 248, 1012, 1167, 1426]
Brard (Alfred)	Brard (Alfred)	<renvoi d'index>
Brenier (Joseph)	Brenier (Joseph)	[481, 565, 567, 586, 591, 592, 643, 644, 709]
Briand	Briand, ministre des affaires	[230, 233, 690]
Brindeau	Brindeau	[96, 248, 507, 801, 862, 1593]
Buhan (Eugène)	Buhan (Eugène)	[829, 831, 1362, 1808]
Cabart-Danneville	Cabart-Danneville	[1, 512, 654, 680, 876, 1691]
Cadilhon	Cadilhon	[1630]
Cadot (Henri)	Cadot (Henri)	[28, 594, 596, 967, 968]
Caillaux (Joseph)	Caillaux (Joseph)	[46, 124, 399, 439, 749, 812, 1061, 1162, 1169, 1170, 1389, 1420, 1421]
Calmel (Armand)	Calmel (Armand)	<renvoi d'index>

FIG. 6.3 : Comparaison

Il est ainsi possible de représenter des cas simples ou complexes au sein d'un même corpus, ce qui correspond mieux à la réalité hétérogène des archives numérisées. A ce stade, on remarque une difficulté pour l'historien qui mobilise des données structurées. Dans une démarche outillée, par exemple avec l'outil *Corpusense*, elle résiderait peut-être moins dans la capacité à programmer qu'à envisager la forme des données, qu'à modéliser un problème ou un ensemble de faits. Cet exercice n'est pas trivial : veut-on décrire des entités et leurs attributs, suivre leur activité dans le temps, ou cartographier leurs relations ? Chacun de ces choix renvoie à des modèles de données distincts, et donc à des manières différentes de faire parler les sources. L'outil ne fait pas disparaître les problématiques de modélisation des données. On retrouve ici la notion de *valuation* au sens deweyien : les moyens de l'enquête — ici, la forme des données et les dispositifs techniques qui les produisent — dépendent des fins poursuivies, mais ces fins elles-mêmes ne sont jamais figées. Elles peuvent être révisées, ajustées ou enrichies au fil des expérimentations, en fonction de ce que les données rendent possible ou non. En ce sens, modéliser les données revient parfois à réactiver, au présent, des gestes interprétatifs analogues à ceux que Collingwood décrivait sous le terme de *reenactment* : l'historien ne fait pas que collecter des informations, il rejoue l'acte de pensée, reconstruit les problèmes tels qu'ils se posaient aux acteurs du passé, mais à travers des médiations techniques. L'historien n'a donc pas affaire à une simple « conversion numérique » de

sa pratique, mais bien à une reconfiguration de ses schèmes interprétatifs sous l’effet des opérations techniques. Dans cette perspective, parler de « numérisation du métier d’historien »¹ ne désigne pas seulement une facilitation instrumentale par les outils numériques : cela renvoie à l’intégration de nouveaux schèmes techniques dans l’enquête elle-même, qui orientent la manière de modéliser les mots et les faits. L’historien ne saurait être l’aliéné des dispositifs techniques — au sens où Simondon entend l’aliénation comme l’usage aveugle d’outils méconnus — mais participe au contraire à leur individuation, en les inscrivant consciemment dans son milieu de recherche et dans ses gestes interprétatifs.

De plus, cette solidarisation des différentes briques techniques — de cette médiation technique pour constituer des questions et des réponses — ne fait pas disparaître la question de l’évaluation des données produites. Au contraire, elle la rend plus pressante. Car exploiter scientifiquement des données issues d’un assemblage *abstrait* de techniques *concrètes* — pour reprendre le vocabulaire de Gilbert Simondon — suppose de fonder une confiance raisonnée : confiance dans les outils choisis, dans la cohérence de la chaîne opératoire, mais aussi dans la capacité des chercheurs à expliciter les conditions de production des données qui alimentent leur analyse. Dans le cas d’une pipeline comme *Corpusense*, on a affaire à un tel assemblage abstrait : une juxtaposition de fonctions spécialisées (OCR, segmentation, structuration), qui ne forment pas encore un objet intégré mais dont la coopération doit être évaluée comme un tout. Ce qui permet de fonder cette confiance est donc l’*évaluation* de la chaîne de traitement, c’est-à-dire à la fois la performance de chaque module et la cohérence globale de l’ensemble.

II. Le travail sur *Corpusense*

Au moment du stage, *Corpusense* était encore en cours de développement ; toutes les fonctionnalités n’étaient pas encore disponibles. Je l’ai donc utilisé principalement pour la constitution du corpus, l’OCRisation et le téléchargement du texte issu de la *Table nominale* de 1931. En revanche, pour la génération de la sortie structurée à l’aide de l’API Mistral, je travaillais directement depuis mon ordinateur, en effectuant les appels manuellement. Cette fonctionnalité est aujourd’hui intégrée dans *Corpusense*, mais elle ne l’était pas encore au moment de mes expérimentations.

Ce décalage n’affecte pas la validité des travaux réalisés : les modules sollicités — notamment Mistral pour la production JSON — sont identiques, seul le mode d’appel diffère. Les expérimentations sur la structuration des données ont donc été conduites à partir du texte brut, en dehors de l’application, mais elles prolongent directement le schème opératoire mis en place par *Corpusense*.

¹S. Poublanc et N. Marqué, « Introduction au dossier « Historien · nes et numérique... ».

III. Ateliers à l'EHESS et à la BnF

Avant de passer à l'évaluation proprement dite, arrêtons-nous sur un cas concret qui illustre bien les difficultés rencontrées. Lors d'un atelier mené à l'EHESS par les développeurs de *Corpusense*, une chercheuse — sans formation particulière en informatique — a expérimenté l'outil sur un corpus de la Quatrième République. En une demi-heure, elle a pu obtenir un jeu de données massif, comprenant plus de 1200 entités liées à l'activité parlementaire, plus précisément sur la production de documents parlementaires. L'exercice montre à quel point la chaîne de traitement peut être efficace et accessible : un travail qui aurait pris des semaines en dépouillement manuel est désormais réalisable en un temps réduit.

Cependant cette réussite apparente masque plusieurs écueils. La chercheuse, tâche peu facile oblige, a eu du mal à définir précisément son modèle de données : quelles entités retenir ? quelles relations considérer comme pertinentes ? quelle granularité adopter ? Faut-il penser en terme d'acteurs ? De documents produits ? Il a fallu essayer différents modèles pour obtenir un résultat satisfaisant ; redéfinir clairement ce que l'on tenait à savoir sur la période étudiée. On pourrait, en reprenant le mot de Leroy Ladurie, l'historien a tout intérêt à être *designer* – plutôt d'ailleurs que « programmeur » – car l'enjeu est d'être capable de modéliser des données. De plus, même si la structuration a fonctionné, la question centrale reste ouverte : les données ainsi produites sont-elles fiables ? Dans quelle mesure peut-on leur faire confiance pour alimenter une enquête historique, et non seulement une démonstration technique ? A l'issue de cet atelier, qui visait avant tout à expérimenter l'outil qu'à produire de véritables données pour une question de recherche déterminée, le besoin de savoir si les données étaient fiables était urgent : ainsi, par ce cas exemplaire qui ne fait qu'illustrer une problématique qui a fondé les tenants et aboutissants du stage, la grande question de l'*évaluation* et de la confiance que l'on peut porter aux données.

Chapitre 7

La sortie structurée via LLM appliquée à la Table des Noms du Sénat : une approche empirique

La pertinence scientifique des analyses repose sur l'évaluation qui ne se réduit pas à un simple contrôle technique. Elle engage une véritable réflexion méthodologique. Dans le cadre du stage, cette réflexion méthodologique s'est adossée à l'expérimentation sur la façon d'évaluer des données, générées par la pipeline. Evaluation et expérimentation peut sembler antinomiques. Par évaluation, on entend une dimension scientifique, protocolaire. La notion de « créativité » que motive au fond celle d'expérimentation et sa dialectique de l'essai-erreur, ne semble pas de mise. Pourtant, la question de la métrique à laquelle s'adosse l'évaluation ne va pas de soi : car il faut évaluer à la fois l'indexation du contenu par le système technique ; mais également sa structure de façon conjointe. Il faut également pouvoir apprécier les absences, les hallucinations ; éventuellement, les différentes façons d'OCRiser le texte, tester si un OCR réputé parfait conduit à des résultats finalement très proches de données très imparfaites. De plus, l'évaluation s'adosse à des *vérités terrain* qui, comme on l'a vu pour l'OCR, forment un domaine de référence sur laquelle quantifier les écarts avec la génération du LLM. Ces vérités terrain, qui permettent de fonder une analyse permettant d'émuler l'objectivité, sont pourtant le fruit de décisions qui ne sont pas évidentes : comment nommer ses métadonnées, quelle structure – ou *schéma* – conviendrait le mieux à nos données ? La modélisation, comme on vient de le voir également est une véritable affaire de design. Il y a donc un arsenal de paramètres à prendre en compte. Pour arrêter un protocole d'évaluation, il faut dégager des critères de façon empirique – et donc expérimenter et réduire les paramètres selon ce qu'il semble le plus pertinent. Le mot de Gaston Bachelard selon lequel les instruments seraient de la « théorie réifiée » est particulièrement approprié à notre cas,

car les moyens pour mesurer la qualité des données produites par la pipeline dépend de nos valuations. Cette évaluation prend sens dans un processus itératif où techniques et de valeurs amendées par l’expérimentation et par les objectifs fixés par une question de recherche.

Bien évidemment, cette expérimentation-évaluation s’adosse, non pas à une problématique technique pure, mais à une question de recherche en histoire – bien qu’elle soit un motif pour guider l’exploration documentaire et computationnelle –, à savoir : cartographier l’activité parlementaire pour l’année 1931, conformément au *reenactement* collingwoodien suggéré en première partie de ce mémoire sur la qualité du débat parlementaire. Cette motivation historienne, en quatre mois seulement, ne saurait être mené de bout en bout. L’enjeu est donc de maîtriser et évaluer le protocole d’extraction ; de tester sa complexité et vérifier qu’il est satisfaisant à ce stade, avant de mener des projets plus ambitieux.

I. Expérimentations

1. Prise en main intuitive du problème de la génération de données

L’expérimentation menée s’est organisée autour de deux volets complémentaires : d’une part, la génération de données à partir des *Tables* parlementaires proprement dite ; et, d’autre part, leur évaluation à l’aune d’une vérité terrain soigneusement construite.

Dans un premier temps, le travail a consisté en une phase d’exploration technique visant à s’approprier la tâche d’évaluation des sorties structurées produites par un modèle génératif. L’objectif principal était de vérifier la faisabilité d’une extraction fiable des informations essentielles : les noms des intervenants au Sénat (qu’il s’agisse de sénateurs ou de ministres interpellés) et les dates de leurs interventions.

Ces dates ne figurent pas explicitement dans les *Tables* ; elles doivent être déduites par le biais des références de page, lesquelles constituent des indicateurs temporels indirects. En d’autres termes, la pagination continue du *Journal Officiel* rend possible un fléchage des interventions dans le temps, en reliant chaque entrée des tables à la séance correspondante comme on le verra en temps venu.

Pour produire les données structurées, plusieurs pistes ont été explorées. La plus immédiate a été de recourir à des prompts « naturels » en utilisant directement l’interface Web de Mistral, laissant au modèle une certaine liberté interprétative dans sa réponse.

L’extraction d’une réponse dans l’environnement d’un chat résiste à l’extraction systématique, notamment parce qu’elle est entourée de politesses qui se surajoutent à l’information qui nous intéresse – ici la liste des sénateurs 7.1 . Mais également parce que cette réponse est émise dans un environnement – en ligne – qu’il est difficile à intégrer dans une pipeline.

A screenshot of a Mistral API response. It shows a list of names: Abel Lefèvre, Albert Fouilloux, Albert Mahieu, Albert Meunier, and Alexandre Bachelet. Below the list is a note: "Si tu veux des détails supplémentaires sur leurs interventions ou d'autres informations, fais-moi signe !". At the bottom left is a timestamp "16:12". At the bottom right are several small icons: a thumbs up, a magnifying glass, a refresh symbol, and a square.

Voici la liste des sénateurs présents dans le texte que tu m'as partagé :

- Abel Lefèvre
- Albert Fouilloux
- Albert Mahieu
- Albert Meunier
- Alexandre Bachelet

Si tu veux des détails supplémentaires sur leurs interventions ou d'autres informations, fais-moi signe !

16:12

▲ 🔍 ⌂ ⏪

FIG. 7.1 : Illustration du résultat d'un prompt avec Mistral (en simulant le modèle disponible en avril 2025) où je soumets un extrait des Tables OCRisées. Comme on le voit, c'est un résultat qui mélange une liste de sénateurs et politesses qui sont ici inconvenantes à une extraction sémantique systématique.

Les appels à l'API Mistral permettent donc à la fois d'interroger un LLM comme on l'a fait, mais depuis un script qui peut récupérer les données, les sauvegarder en local, d'appliquer divers traitements.

Ce qui nous permet désormais de justifier l'usage de Mistral : l'appel à l'API était gratuit, sans restrictions, et supportait la génération par sortie structurée – c'est-à-dire la capacité de contraindre le modèle à suivre un schéma formel représentant nos données. Autrement dit, on peut utiliser, sans aucune installation ou dépense, un LLM pouvant être « constraint » à formuler du JSON – et rien que du JSON – par un schéma formel afin de produire des objets comparables. De fait, ce schéma 7.2 était produit avec la librairie Pydantic qui permet de convertir des classes Python qui représentent le schéma de données, en modèle JSON.

Mais un schéma, à lui seul, ne suffit pas : il doit être combiné à un prompt qui guide le modèle. Le tout premier essai reposait sur des instructions très simples :

« Extrayez les informations du texte fourni. Je veux la liste des noms et prénoms de toutes les personnes mentionnées (des sénateurs). » « Attention au bruit : tout ce qu'il y a dans le texte n'est pas forcément un sénateur. » « Voici mon texte : »

Ces instructions rudimentaires avaient pour objectif principal de vérifier le bon fonctionnement de l'appel à l'API et le retour des données dans le format JSON. En pratique, le pipeline consistait en un script Python qui, à partir du texte OCRisé, soumettait au modèle Mistral

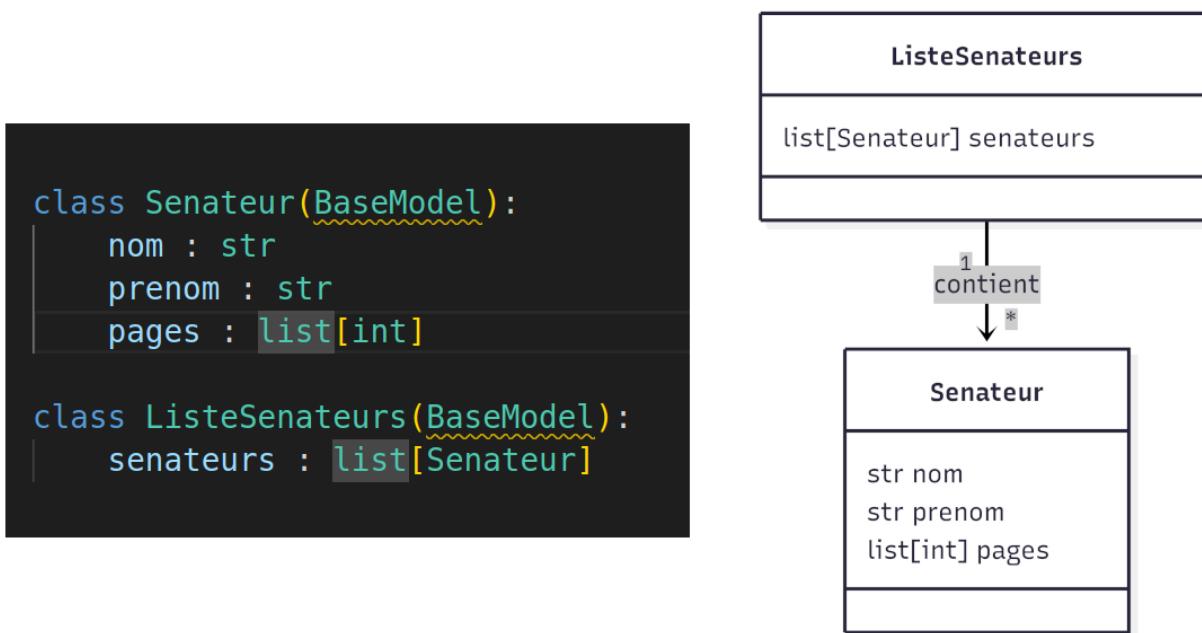


FIG. 7.2 : A gauche, l’implémentation en Pydantic via Python d’un des premiers modèles de données imaginé pour la génération guidée (à droite).

un prompt et un schéma, puis récupérait les données produites¹.

Toutefois, dès les premiers tests, des limites sont apparues : les clés choisies dans le modèle, comme *ListeSenateur* ou *Senateur*, étaient trop restrictives. Or, les *Tables nominatives* ne mentionnent pas seulement des sénateurs, mais également des ministres interpellés ou intervenants extérieurs. Un schéma trop strict risquait donc d’induire des omissions. Pour contourner ce problème, il a été décidé d’adopter un vocabulaire plus fonctionnel, centré sur les *intervenants* plutôt que sur une fonction institutionnelle définie de façon explicite. Une fois cette chaîne de traitement permettant de produire des données essayée, il était l’heure de consolider un protocole d’évaluation de cette méthode de génération de métadonnées.

2. Design, prompt et vérité terrain : trouver le bon modèle de données

Un second volet des expérimentations a porté sur la construction de la vérité terrain, élément indispensable à toute évaluation. En théorie, elle sert de référence absolue pour mesurer les performances du modèle. En pratique, elle résulte d’une série de décisions éditoriales et méthodologiques qui influencent directement les résultats. L’exemple de Louis Barthou, sénateur et ministre de la Guerre en 1931, illustre ce point : fallait-il regrouper toutes ses

¹`pipeline_mezz`.

interventions sous une seule entrée ou distinguer ses apparitions selon ses fonctions 7.3 ? Le choix de la granularité maximale – séparer chaque occurrence – prévient les confusions d’homonymes, mais complique l’évaluation en raison de la tendance du modèle à fusionner les mentions.

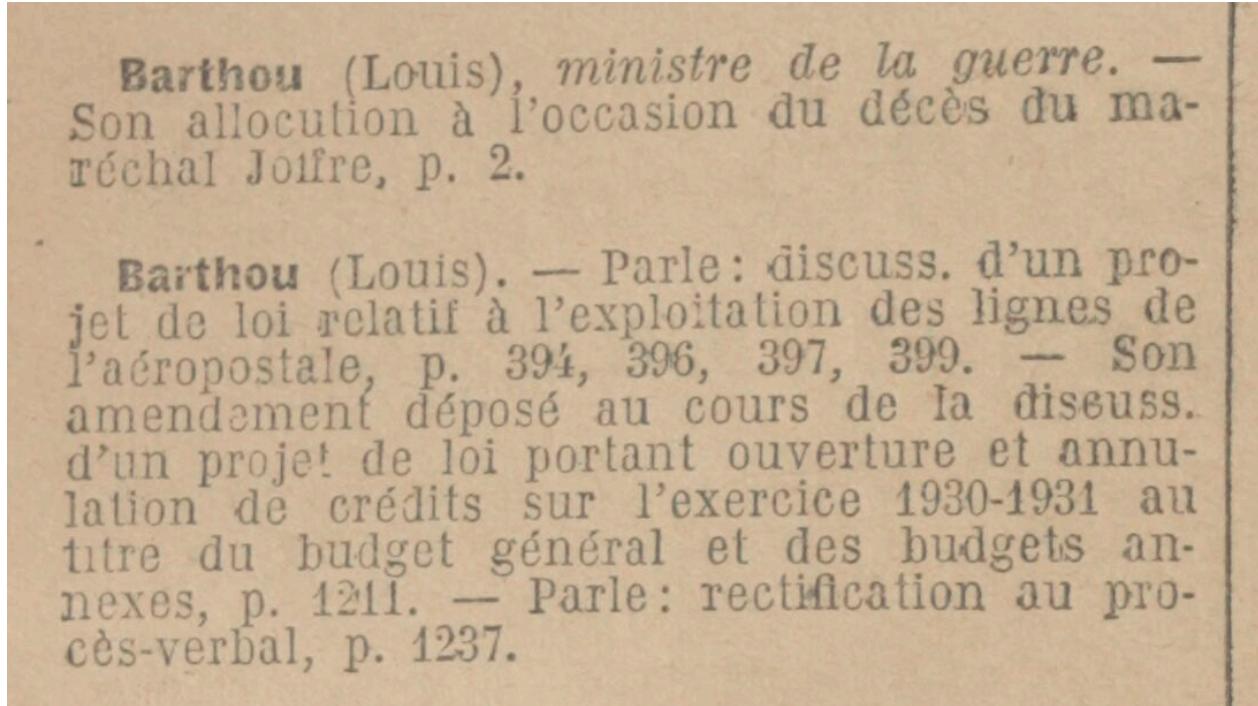


FIG. 7.3

Sur le plan technique, la vérité terrain est lié au prompt et au schéma de données (??). Le schéma de données fait office de « carte » qui vient guider la production textuelle du LLM. Il indique « où » mettre les bons mots dans les bonnes boîtes, les bons intitulés avec les bonnes clés. Pour éviter de démultiplier les paramètres à tester, le prompt (voir : ANNEXE 31) et le schéma (ANNEXE 32) ont été stabilisés à l’issu d’un processus itératif pour évaluer la granularité adéquat. Dans le cadre de l’évaluation, seul le nom du sénateur et les pages de références ont été retenues. Pour guider au mieux la génération du JSON par le LLM, on ajoute à notre schéma des descriptions pour qu’il puisse être plus attentif pour qu’il puisse plus facilement mettre les données dans les bonnes boîtes 7.4.

Le schéma (7.4) joue donc ici le rôle de grammaire opérationnelle : il décrit précisément la structure attendue des sorties (nom, rôle et références paginaires des intervenants), et constraint le modèle à produire un JSON validable. Les premières clés (« Senator », « Liste-Senator ») ont rapidement montré leurs limites : elles risquaient l’invisibilité d’autres acteurs – comme les ministres – et biaisaient l’extraction. La version finale de la vérité terrain (ANNEXE 33) adopte des catégories plus neutres (« Intervenant »), révélant que la modélisation

```
class Intervenant(BaseModel):
    nom: str = Field(..., description="Nom (et prénom s'il existe) de
    l'intervenant")
    references_pages: Union[List[int], str] = Field(...,
                                                    description="Liste des
                                                    numéros de page où
                                                    l'intervenant est référencé
                                                    ou sinon <index
                                                    cross-reference>")

class IntervenantAuSenat(BaseModel):
    listes_des_intervenants: List[Intervenant] = Field(...,
                                                       description="Liste de
                                                       tous les intervenants
                                                       avec leurs références de
                                                       pages respectives")
```

FIG. 7.4

n'est jamais un simple exercice technique mais un choix interprétatif.

Quant aux données produites automatiquement par cette chaîne de traitement stabilisées et appliquées sur l'ensemble des pages des *Tables* du Sénat de 1931 (extraits en annexe : ANNEXE 34), elles semblent être cohérentes. Mais, à défaut d'évaluation, on ne peut tirer de conclusion décisives.

3. Préparer l'évaluation : comparer, apparier

L'évaluation des sorties structurées générées par un LLM reposant sur un schéma, un prompt repose donc sur une comparaison systématique entre deux ensembles : la vérité terrain construite manuellement comme référence, et les données produites automatiquement par le modèle. L'objectif est maintenant de mesurer objectivement la fidélité de la génération, d'identifier les omissions, duplications ou erreurs, et d'évaluer la robustesse globale du système. Lors de la phase exploratoire autour de l'évaluation, plusieurs approches méthodologiques et outils de comparaison ont été expérimentés afin de prendre en main les méthodes classiques de mesure de similarité et identifier des indicateurs adaptés à l'évaluation qualitative de données produites par un modèle génératif.

Comparer une vérité terrain soigneusement construite avec des données générées par un modèle de sortie structurée ne se résume pas à vérifier la présence ou l'absence d'éléments identiques. Plusieurs sources de divergences compliquent l'évaluation : des noeuds peuvent être omis, inventés ou dupliqués ; des erreurs d'OCR altèrent le texte ; des éléments sont

regroupés ou mal hiérarchisés dans la structure JSON. La comparaison doit donc dépasser une logique purement textuelle pour capturer ces écarts dans toute leur complexité, tout en restant robuste et interprétable. La comparaison n'est pas un simple « face à face » entre des données réputées parfaites et des données générées. L'expérimentation a permis de mettre en lumière certaines méthodes – et d'en rejeter d'autres comme la *Tree Edit Distance*, davantage basée sur la structure des données que leur contenu.

La première étape a consisté à expérimenter la comparaison chaque élément textuel de la vérité terrain avec chaque élément généré, ce qui conduit à aplatis temporairement la structure JSON pour se concentrer sur le contenu. La distance de Levenshtein a été privilégiée, car elle mesure simplement le coût minimal d'édition (insertion, suppression, substitution) entre deux chaînes de caractères ; elle est donc adaptée pour évaluer des données issues d'un pipeline OCR + LLM, où les erreurs sont souvent lexicales plutôt que conceptuelles. Concrètement, ces distances ont été organisées en une *matrice de similarité* (7.5), dans laquelle chaque cellule représente le coût de transformation d'un élément de la vérité terrain en un élément généré.

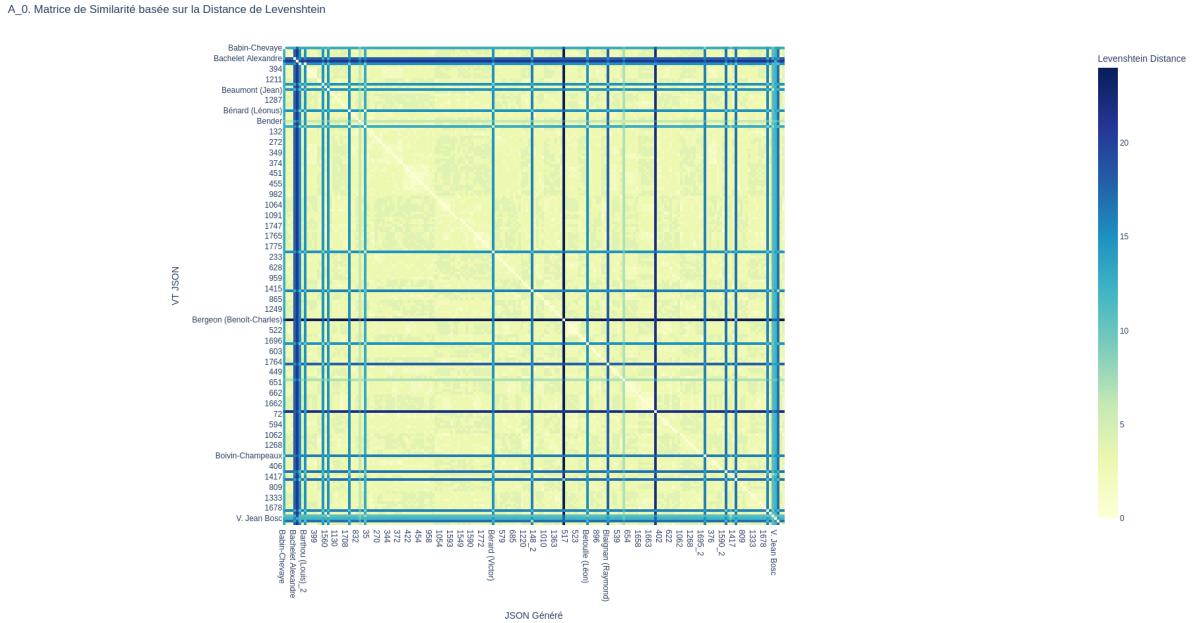


FIG. 7.5 : Matrice de similarité générée avec Python. Les points clairs représentent des distances faibles, c'est-à-dire des mots très similaires.

Ce dispositif visuel met en évidence des phénomènes difficiles à saisir autrement : comme les décalages – des zones sombres signalent des omissions ou des décalages d'alignement, le modèle ayant « perdu le fil » avant de se resynchroniser – et les regroupements – des correspondances multiples sur une même ligne indiquent que plusieurs éléments ont été agrégés dans une seule entrée.

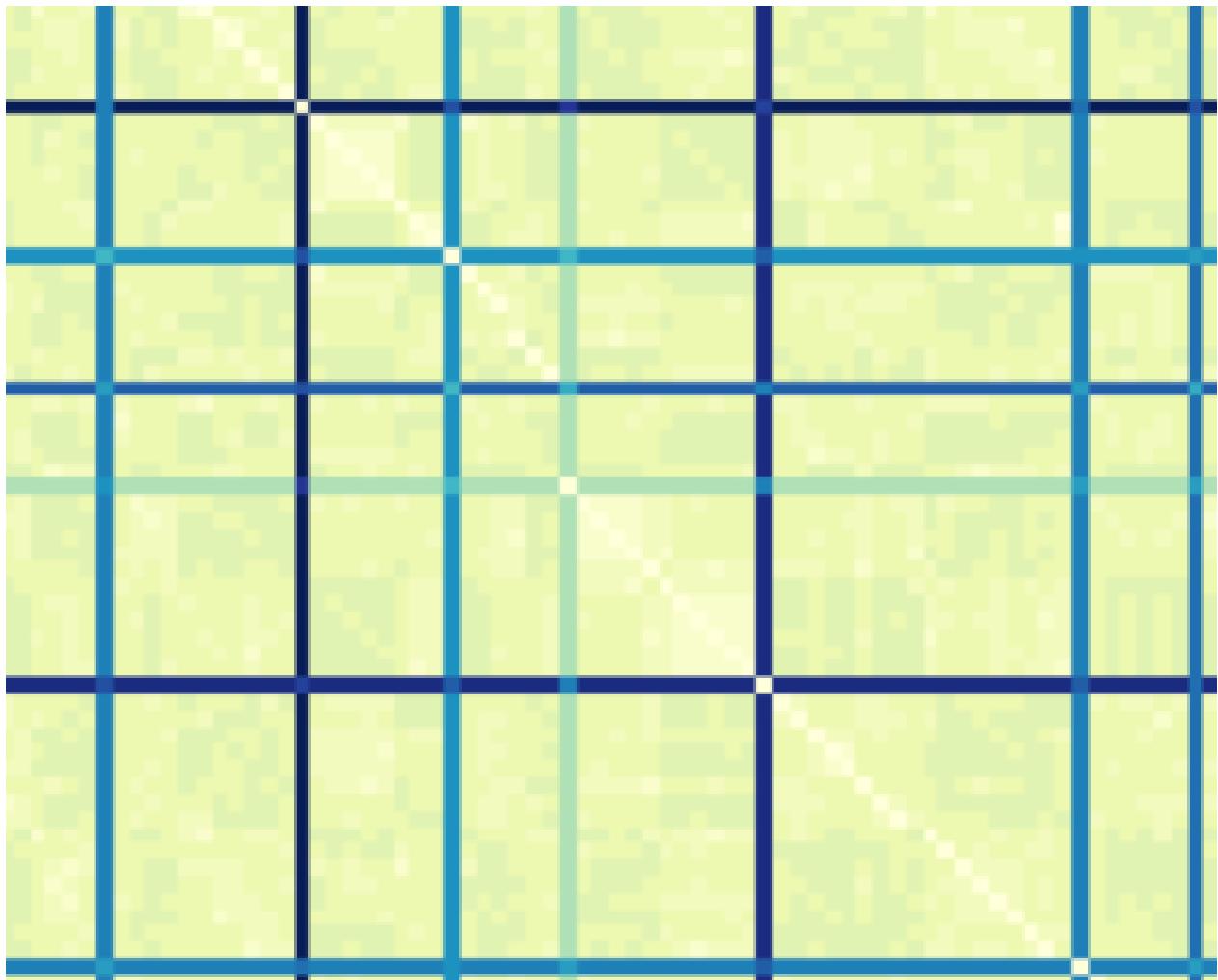


FIG. 7.6 : Zoom sur la matrice de similarité : un décalage de la diagonale peut signifier une omission.

L’observation de ces matrices ont d’ailleurs conduit, à la marge, à une réflexion plus conceptuelle. En effet, ces matrices ne forment pas toujours un espace métrique au sens strict, car des propriétés comme l’inégalité triangulaire peuvent être bafouées – autrement dit, l’ensemble des distances peuvent dessiner un paysage impossible où il n’y a pas de cohérence avec les distances. Pourtant, on observe des zones de cohérence locale qui respectent suffisamment ces propriétés pour fonctionner comme des espaces métriques restreints. Là où il y a émergence de propriétés métriques, il y a l’hypothèse d’une correspondance entre la vérité terrain et les données générées. Deux ensembles identiques comparés forment un espace métrique : on pourrait alors calculer « l’effort » d’une matricité de similarité pour devenir un véritable espace métrique. Encore une fois, il s’agit d’une réflexion marginale qui n’a pas été développée ensuite. Cependant, cela permet de mettre en avant certains garde-fous dans le cadre de l’évaluation, à savoir le respect de certaines propriétés mathématiques liées à la mesure.

Ensuite reste l’étape de la comparaison proprement dite. Comment mettre correspondance les bonnes paires ? Comment comparer toutes les entrées de la vérité terrain avec celles générées ? Si la matrice de similarité représente toutes les relations entre les vérités terrain et les données générées, c’est une cartographie relationnelle qui ne nous permet pas d’apparier les ensembles en vue d’évaluer la qualité des prédictions ; elle ne dit pas comment assembler nos entrées. De même, à cause des désynchronisations, il faut pouvoir aller comparer des éléments qui ne sont pas nécessairement « en-face », c’est-à-dire qui ne partagent pas le même rang.

Pour ce travail de mise en correspondance, intervient le transport optimal. Introduit par Gaspard Monge en 1781 – dans son *Mémoire sur la théorie des déblais et remblais* – et reformulé par l’économiste soviétique Leonid Kantorovitch en 1942, le transport optimal vise à trouver la correspondance la plus économique entre deux ensembles : ici, les éléments de la vérité terrain et ceux générés. Plutôt que de tester toutes les permutations possibles, dont le nombre croît de manière factorielle, le transport optimal représente le problème sous forme d’une matrice de coût – qui ici peut nous être donnée à partir de la matrice de similarité – et cherche à minimiser la somme des coûts d’appariement. Cette méthode permet donc d’obtenir un alignement global optimal, même lorsque les ensembles ont des tailles différentes ; de détecter les correspondances « justifiées » au sens du coût minimal, tout en identifiant les zones où la correspondance est faible ou artificielle ; de dépasser une logique purement séquentielle pour analyser la distribution globale des erreurs et omissions. Notre coût, ici, ce sont les « coûts d’édition » autrement dit nos distances de Levenshtein, exprimée par la matrice de similarité.

Cette approche est intéressante pour documenter l’état des sorties générées : elle ne

prétend pas résoudre les problèmes d'ordre ou de sémantique, par exemple, apparier « sénateur » et « parlementaire » demanderait des distances sémantiques – bien que c'est une piste envisageable avec les *embeddings* dans le cas d'une analyse thématique des interventions parlementaires. En revanche, elle fournit une mesure robuste de dissemblance structurelle que nous offre justement le paysage des distances de Levenshtein. En pratique, les vérifications manuelles réalisées sur de petits ensembles de données ont confirmé que les appariements identifiaient correctement les correspondances : chaque élément de la vérité terrain était relié à son équivalent prédit par le modèle. Ces tests ont permis d'esquisser un protocole d'évaluation fiable, en ancrant l'intuition dans des outils mathématiques et informatiques éprouvés.

Il ne s'agit cependant pas d'une démarche inédite : elle transpose à l'analyse des textes structurés un principe ancien et solide de l'évaluation de segmentation en vision par ordinateur. Comme l'explique Chazalon et Carlinet² à propos de la segmentation panoptique de cartes historiques, il est à la fois possible et pertinent de reformuler l'appariement des régions segmentées — contenus visuels — via un cadre bipartite, d'utiliser des métriques classiques (précision, rappel, F-score)³. Appliquée à nos données textuelles, cette approche inspire l'idée que l'appariement entre vérité terrain et données générées ne repose pas seulement sur le calcul statistique, mais peut aussi être qualifié, visualisé et interprété selon des logiques proches de celles de l'image numérique. En tout cas, une phase de l'expérimentation a consisté à comprendre la nature du transport optimal – notamment avec les travaux de Gabriel Peyré⁴ – afin d'en saisir les limites – l'appariement ne regarde pas le sens des mots et ne prend pas compte de l'aspect séquentielle des entrées – et les avantages – permet justement de surmonter le problème des oubliés et de faire correspondre des données qui ne sont pas « en face ». Cela impliquant ainsi d'intégrer *a posteriori*, dans les métriques qui rendent compte de la qualité des données générées également la qualité de l'appariement.

Toutefois, ces expérimentations mettent également en évidence la dimension heuristique de cette démarche : les résultats dépendent autant du choix des mesures de similarité (distance de Levenshtein) que de la stratégie d'appariement retenue (transport optimal), ce qui engage une réflexion critique sur le design même de l'évaluation. Ainsi, même si la méthode hérite de protocoles éprouvés de l'imagerie, son adaptation à un contexte textuel demande des choix méthodologiques soigneux et explicités.

²Joseph Chazalon et Edwin Carlinet, « Revisiting the Coco Panoptic Metric to Enable Visual and Qualitative Analysis of Historical Map Instance Segmentation », dir. Josep Lladós, Daniel Lopresti et Seiichi Uchida, *Document Analysis and Recognition – ICDAR 2021*, 12824 (2021), Series Title : Lecture Notes in Computer Science, p. 367-382, DOI : 10.1007/978-3-030-86337-1_25.

³Ibid.

⁴Gabriel Peyré et Marco Cuturi, *Computational Optimal Transport*, en, arXiv :1803.00567 [stat], mars 2020, DOI : 10.48550/arXiv.1803.00567.

II. Recouper des données pour l’analyse historienne

Dans une perspective historienne, la capacité à naviguer dans les volumes du *Journal officiel* ne repose pas uniquement sur la qualité des extractions textuelles, mais sur le recouplement systématique des données. L’enjeu est de relier des références issues des tables — ici limitées à des numéros de pages et aux occurrences nominales — à des dates précises de publication, condition essentielle pour contextualiser les débats ou décisions mentionnés. Pour ce faire, nous avons exploité les manifests IIIF fournis par l’API Gallica, qui décrivent chaque volume sous forme d’une structure hiérarchisée : chaque page est identifiée, ordonnée et associée à une URL stable. Ce cadre technique permet de reconstituer un tableau complet « page → date », en combinant l’ordre des pages du manifeste avec les métadonnées des volumes et les dates présentes sur les pages de titre. Cette opération ne vise pas à créer un modèle sophistiqué de reconnaissance ou d’indexation, mais plutôt à bâtir un pivot de référence déterministe pour toute analyse ultérieure : il devient possible de passer d’une citation brute (une page) à son contexte temporel exact.

Il est important de souligner qu’à ce stade, cette correspondance se fonde sur les pages extraites par le système et rien n’indique que cette extraction repose sur des données fiables. Si ces données, à défaut d’évaluation, ne sont pas exploitables scientifiquement pour répondre à une question de recherche, il est cependant convenable de vérifier qu’on peut, à partir d’elles, avoir des informations plus intéressantes que des références de pages, comme des informations chronologiques.

1. Des pages aux dates : utilisation de l’API Gallica et des métadonnées des manifestes

La constitution d’un tableau « page → date » s’appuie sur les services IIIF proposés par Gallica, qui fournissent pour chaque document numérisé un manifest JSON décrivant l’ensemble de ses pages. Ce *manifest* joue le rôle de table des matières numérique : il contient, pour chaque page, son ordre dans le volume, son libellé (numéro de page ou folio) et des identifiants pérennes vers l’image et le contenu OCRisé. La première étape consiste à interroger automatiquement cette ressource – via des requêtes HTTP en lignes de commande, et notamment avec « curl » – pour extraire la séquence complète des pages.

Ensuite, les pages sont indexées : chaque entrée du manifest est associée à son URL lisible et à son numéro de séquence. Cette structure séquentielle est cruciale, car elle fournit un pivot neutre pour relier des références issues d’autres sources (tables imprimées, index, etc.).

La seconde étape exploite les pages de titre qui portent la date d’édition. Ces infor-

mations, accessibles via l’API Gallica sont extraites puis assignées à des plages de pages contiguës : ainsi, toutes les pages comprises entre deux dates identifiées héritent de la date correspondante.

C	D	E	F	pages
1	jour	mois	url_apres_redirection	url_manifest
2	13	1	https://gallica.bnf.fr/ark:/12148/bpt6k220610p.item	https://gallica.bnf.fr/ark:/12148/bpt6k220610p/manifest.json
3	15	1	https://gallica.bnf.fr/ark:/12148/bpt6k2206113.item	https://gallica.bnf.fr/ark:/12148/bpt6k2206113/manifest.json
4	20	1	https://gallica.bnf.fr/ark:/12148/bpt6k220612h.item	https://gallica.bnf.fr/ark:/12148/bpt6k220612h/manifest.json
5	23	1	https://gallica.bnf.fr/ark:/12148/bpt6k220613x.item	https://gallica.bnf.fr/ark:/12148/bpt6k220613x/manifest.json
6	29	1	https://gallica.bnf.fr/ark:/12148/bpt6k220614b.item	https://gallica.bnf.fr/ark:/12148/bpt6k220614b/manifest.json
7	30	1	https://gallica.bnf.fr/ark:/12148/bpt6k220615r.item	https://gallica.bnf.fr/ark:/12148/bpt6k220615r/manifest.json
8	3	2	https://gallica.bnf.fr/ark:/12148/bpt6k2206165.item	https://gallica.bnf.fr/ark:/12148/bpt6k2206165/manifest.json
9	5	2	https://gallica.bnf.fr/ark:/12148/bpt6k220617k.item	https://gallica.bnf.fr/ark:/12148/bpt6k220617k/manifest.json
10	10	2	https://gallica.bnf.fr/ark:/12148/bpt6k2206180.item	https://gallica.bnf.fr/ark:/12148/bpt6k2206180/manifest.json
11	12	2	https://gallica.bnf.fr/ark:/12148/bpt6k220619d.item	https://gallica.bnf.fr/ark:/12148/bpt6k220619d/manifest.json
12	17	2	https://gallica.bnf.fr/ark:/12148/bpt6k2206202.item	https://gallica.bnf.fr/ark:/12148/bpt6k2206202/manifest.json
13	19	2	https://gallica.bnf.fr/ark:/12148/bpt6k220621g.item	https://gallica.bnf.fr/ark:/12148/bpt6k220621g/manifest.json
14	24	2	https://gallica.bnf.fr/ark:/12148/bpt6k220622w.item	https://gallica.bnf.fr/ark:/12148/bpt6k220622w/manifest.json
15	26	2	https://gallica.bnf.fr/ark:/12148/bpt6k2206239.item	https://gallica.bnf.fr/ark:/12148/bpt6k2206239/manifest.json
16	3	3	https://gallica.bnf.fr/ark:/12148/bpt6k220624q.item	https://gallica.bnf.fr/ark:/12148/bpt6k220624q/manifest.json
17	5	3	https://gallica.bnf.fr/ark:/12148/bpt6k2206254.item	https://gallica.bnf.fr/ark:/12148/bpt6k2206254/manifest.json
18	6	3	https://gallica.bnf.fr/ark:/12148/bpt6k220626g.item	https://gallica.bnf.fr/ark:/12148/bpt6k220626g/manifest.json
19	10	3	https://gallica.bnf.fr/ark:/12148/bpt6k220627z.item	https://gallica.bnf.fr/ark:/12148/bpt6k220627z/manifest.json
20	12	3	https://gallica.bnf.fr/ark:/12148/bpt6k220628c.item	https://gallica.bnf.fr/ark:/12148/bpt6k220628c/manifest.json
21	13	3	https://gallica.bnf.fr/ark:/12148/bpt6k220629s.item	https://gallica.bnf.fr/ark:/12148/bpt6k220629s/manifest.json
22	17	3	https://gallica.bnf.fr/ark:/12148/bpt6k220630f.item	https://gallica.bnf.fr/ark:/12148/bpt6k220630f/manifest.json
23	19	3	https://gallica.bnf.fr/ark:/12148/bpt6k220631v.item	https://gallica.bnf.fr/ark:/12148/bpt6k220631v/manifest.json
24	20	3	https://gallica.bnf.fr/ark:/12148/bpt6k220632b.item	https://gallica.bnf.fr/ark:/12148/bpt6k220632b/manifest.json
25	21	3	https://gallica.bnf.fr/ark:/12148/bpt6k220633p.item	https://gallica.bnf.fr/ark:/12148/bpt6k220633p/manifest.json
26	23	3	https://gallica.bnf.fr/ark:/12148/bpt6k2206343.item	https://gallica.bnf.fr/ark:/12148/bpt6k2206343/manifest.json
27	24	3	https://gallica.bnf.fr/ark:/12148/bpt6k220635n.item	https://gallica.bnf.fr/ark:/12148/bpt6k220635n/manifest.json
28	25	3	https://gallica.bnf.fr/ark:/12148/bpt6k220636x.item	https://gallica.bnf.fr/ark:/12148/bpt6k220636x/manifest.json
29	26	3	https://gallica.bnf.fr/ark:/12148/bpt6k220637b.item	https://gallica.bnf.fr/ark:/12148/bpt6k220637b/manifest.json
30	27	3	https://gallica.bnf.fr/ark:/12148/bpt6k220638r.item	https://gallica.bnf.fr/ark:/12148/bpt6k220638r/manifest.json
31	28	3	https://gallica.bnf.fr/ark:/12148/bpt6k2206395.item	https://gallica.bnf.fr/ark:/12148/bpt6k2206395/manifest.json
32	29	3	https://gallica.bnf.fr/ark:/12148/bpt6k220640t.item	https://gallica.bnf.fr/ark:/12148/bpt6k220640t/manifest.json

FIG. 7.7 : Fragment du tableau de correspondance page/date en .csv

En combinant ces deux sources (manifest + extraction de dates), on construit une table de correspondance exhaustive reliant chaque page d’un volume à sa date exacte de parution. Cette table, exportée en CSV, sert ensuite de base à des analyses historiennes : une référence de page extraite d’un index thématique peut être immédiatement replacée dans son contexte temporel. L’approche, entièrement automatisable, repose sur des standards ouverts (IIIF, JSON, HTTP) et sur des outils simples de parsing et de jointure de données, ce qui garantit sa robustesse et sa réutilisabilité. (ANNEXE 34)

Chapitre 8

Livrables

I. Jeux de données pour le protocole d'évaluation de la sortie structurée

1. Protocole d'évaluation de la sortie structurée

L'objectif de l'évaluation est de comparer les sorties produites par le LLM – notées P, pour *prédiction* – au *ground truth* – la vérité terrain, notée G. Chaque instance de donnée correspond à une liste d'entrées, chacune constituée d'un nom d'orateur et d'une liste de pages référencées.

Un défi majeur, on l'a vu, tient au fait que le modèle peut produire le bon ensemble d'entrées, mais dans un ordre différent, ou avec de légères variations structurelles. Pour résoudre cela, nous adoptons une stratégie d'alignement flexible inspirée utilisant le transport optimal pour établir une correspondance *one-to-one* entre prédictions et référence. Cette approche neutralise la contrainte d'ordre et tolère des divergences mineures, permettant une évaluation robuste.

L'OCRisation est indépendante et en amont de la comparaison :

1. Chaque page image est d'abord traitée indépendamment par le moteur OCR PERO pour détecter et transcrire le texte.
2. Compte tenu des difficultés persistantes de segmentation de mise en page, nous produisons trois variantes de transcription par page afin de refléter la variabilité de ce type de pipeline : une OCRisation « brute », sans segmentation ; une OCRisataion qui repose sur une segmentation manuelle ; et une émulation d'une OCRisation réputée parfaite, fait à la main, qui correspond à la vérité terrain.

Dans nos expérimentations, pour des raisons pratiques, chaque page est traitée indépendamment, mais la méthode peut être étendue à des contextes multi-pages.

Bien que les LLMs puissent être incités à générer des sorties structurées, il est essentiel de contraindre leurs générations au format attendu. Parmi les méthodes existantes, la plus efficace repose sur le filtrage des tokens valides au moment de l'inférence à l'aide d'un validateur externe (par exemple un automate fini). Cette capacité est disponible dans certaines API commerciales étendues.

Notre processus d'extraction s'appuie donc sur :

- un texte prétraité par OCR,
- un schéma prédéfini (JSON),
- un prompt naturel,
- et une API supportant les sorties contraintes.

Pour cette étude, nous avons retenu l'API Mistral, avec le modèle *Minstral 8B Instruct v2410*, en raison de ses bonnes performances *zero-shot*, de son coût modéré et de la disponibilité publique de ses poids à des fins de recherche.

2. Vérité terrain, prompt et schéma : un échantillon des Tables

A l'issue de la démarche d'expérimentation en vue de l'évaluation a été produit le design de vérité terrain, accompagné de son schéma et du prompt, qui sera comparées à l'information extraite par le LLM. La tâche d'extraction repose sur l'identification des entrées individuelles dans les *Tables nominatives* — c'est-à-dire les noms des sénateurs et leurs références de pages. (ANNEXE 33, 34)

Un échantillon aléatoire de cinq pages consécutives a été choisi pour transcription et annotation manuelles. Comme les entrées peuvent s'étendre sur plusieurs pages, certaines se retrouvent tronquées. Dans ces cas, le LLM a été explicitement instruit d'ignorer les éléments incomplets. Ce choix reflète les défis réalistes de l'extraction, où une prise en compte multi-pages (ou en flux continu) serait nécessaire en production, mais dépasse le cadre de cette étude.

L'évaluation porte sur 109 entrées réparties sur cinq pages, chacune testée dans les trois conditions OCR.

3. Prédictions obtenues avec le modèle Mistral 8b avec l'extraction guidée par schéma

Pour l'année 1931, les *Tables des noms* du Sénat couvrent 14 pages et comptent environ 300 entrées, chacune correspondant à une intervention en assemblée. Chaque entrée est associée à un orateur et détaille différents types d'actions (demandes d'interpellation, discussions de projets de loi, lecture de rapports de commission, dépôt d'amendements, etc.), accompagnées d'une référence de page renvoyant à la transcription complète de l'intervention.

Ces transcriptions sont publiées dans les *Débats parlementaires* du Sénat. Les tables sont donc liées fonctionnellement aux transcriptions par la pagination. Comme la numérotation des pages est continue sur toute l'année, chaque référence permet de dater précisément l'intervention correspondante.

II. Une métrique pour l'évaluation des données générées par la sortie structurée via LLM

1. Mise en correspondance des prédictions et de la vérité terrain avec le transport optimal

Pour comparer les entrées prédites et de référence, nous définissons une distance normalisée $d_e(g_i, p_j)$ qui combine deux composantes :

- **Nom de l'orateur (texte)** : distance de Ratcliff/Obershelp (basée sur la plus longue sous-chaîne commune), après minuscule et suppression des espaces. La distance normalisée $d_n(g_i, p_j) \in [0, 1]$ vaut 0 pour un match exact et 1 pour une dissimilarité totale.
- **Pages référencées (ensembles)** : distance Intersection-over-Union (IoU) :

$$d_p(g_i, p_j) = 1 - \frac{|ref_pages(g_i) \cap ref_pages(p_j)|}{|ref_pages(g_i) \cup ref_pages(p_j)|}$$

- **Distance d'entrée :**

$$d_e(g_i, p_j) = d_n(g_i, p_j) \times d_p(g_i, p_j)$$

Un appariement *one-to-one* entre prédictions et vérité terrain est alors établi par transport optimal [19], minimisant la distance totale et fournissant une base rigoureuse pour l'évaluation.

2. Limites des métriques classiques : précision, rappel et F1

Les métriques usuelles (précision, rappel, F1) sont couramment utilisées pour évaluer les tâches d'extraction. Cependant, dans notre protocole, où les entrées sont alignées par transport optimal, elles deviennent trompeuses :

- l'appariement injectif force une correspondance complète, ce qui maximise artificiellement la précision,
- le rappel ne reflète pas les entrées manquantes ou ajoutées,
- la F1 hérite de ces biais et surestime les performances.

3. Integrated Matching Quality (IMQ)

Pour dépasser ces limites, nous exploitons directement la distance d_e pour finir un score de qualité $q_i = 1 - d_e(g_i, p_i)$, qui reflète la proximité entre une entrée prédite et sa référence.

Plutôt que de fixer un seuil arbitraire pour décider du « bon » appariement, nous calculons la proportion de correspondances de qualité supérieure à un seuil t , puis intégrons sur tout l'intervalle $[0, 1]$:

$$IMQ = \int_0^1 F(t) dt$$

où $F(t)$ est la fraction des correspondances de qualité $q_i \geq t$.

L'IMQ résume ainsi la qualité globale des appariements, récompensant à la fois leur nombre et leur proximité. Un score de 1 indique un alignement parfait. Cette métrique continue, indépendante de seuils arbitraires, est particulièrement adaptée aux sorties LLM, où de légères divergences sont fréquentes même sous fortes contraintes structurelles.

4. Résultats et analyse

Nous avons appliqué notre méthode d'appariement sur cinq pages distinctes (109 entrées), chacune traitée indépendamment et présentant des qualités OCR variables. Le tableau ci-dessous présente les résultats pour chaque page OCRisée sans segmentation ni correction, avec les tailles des ensembles de référence et prédits, ainsi que le nombre de correspondances retenues par transport optimal :

Source Précision (biaisée) Rappel (biaisé) IMQ Entrées de référence Entrées prédites Correspondances						
				page 02 1.0000 0.9565 0.9059 23 22 22		page 03
				1.0000 1.0000 0.8928 25 25 25		page 04 1.0000 1.0000 0.9591 19 19 19

page 05 | 1.0000 | 1.0000 | 0.8636 | 19 | 19 | 19 | | page 10 | 1.0000 | 1.0000 | 0.8193 | 23 | 23 | 23 |

Toutes les pages affichent une précision et un rappel « biaisés » parfaits ; mais comme discuté en section précédente, ces métriques sont limitées car elles découlent directement de l'appariement injectif. Elles ne reflètent pas la qualité réelle des alignements.

L'**IMQ**, en revanche, fournit une évaluation plus fine, en capturant la distribution des qualités de correspondances. Pour toutes les pages traitées, les scores IMQ restent élevés (entre 0.8193 et 0.9591), montrant une homogénéité forte entre correspondances. L'IMQ évalue donc à la fois la complétude et la proximité sémantico-syntaxique des appariements, jouant un rôle hybride entre rappel qualitatif et précision pondérée.

Variations entre pages

Pages 5 et 10 : IMQ plus bas, lié à des incohérences typographiques. De nombreux prénoms n'y sont pas mis entre parenthèses après le nom, contrairement à l'attendu dans le ground truth* (21 % des entrées sur la page 5, 39 % sur la page 10). Cela augmente artificiellement la distance textuelle et dégrade la qualité perçue des correspondances.

- **Page 3** : malgré une précision/rappel parfaits, IMQ plus faible (0.8928), dû à des problèmes OCR causés par un pli dans la reliure, générant du bruit visuel.
- **Page 2** : IMQ élevé (0.9059) malgré un rappel imparfait. Cela s'explique par un biais d'échantillonnage : le prompt avait été calibré sur cette page, ce qui améliore artificiellement la performance. Toutefois, les résultats solides sur la page 4 (IMQ = 0.9591) confirment la robustesse du dispositif.

Un cas particulier : sur la page 2, une personne est mentionnée deux fois (comme sénateur et comme ministre). Le *ground truth* distingue ces deux entrées, tandis que le LLM les fusionne. Cela réduit artificiellement le rappel mais correspond à une rationalisation fonctionnelle du modèle.

Comparaison avec OCR « parfait »

Lorsque l'on compare avec l'OCR jugé « parfait » (corrigé manuellement), les résultats s'améliorent globalement :

| Source | Précision (biaisée) | Rappel (biaisé) | IMQ | Entrées de référence | Entrées prédictes | Correspondances | | —— | ————— | ——— | — | ————— | ————— | ————— | ————— | | page 02 | 1.0000 | 1.0000 | 0.9513 | 23 | 23 | 23 | | page 03 | 1.0000 | 1.0000 | 0.9430 | 25 | 25 | 25 | | page 04 | 1.0000 | 1.0000 | 0.9821 | 19 | 19 | 19 | | page 05 | 1.0000 | 1.0000 | 0.8778 | 19 | 19 | 19 | | page 10 | 1.0000 | 1.0000 | 0.8966 | 23 | 23 | 23 |

Dans certains cas, les versions OCR bruitées donnent des résultats paradoxalement meilleurs. Par exemple, les en-têtes courants capturés par l'OCR bruité fournissent un contexte utile pour les entrées tronquées en début de page. Ainsi, sur la page 2, le LLM a correctement reproduit la double mention (sénateur/ministre), alors que l'OCR corrigé ne l'a pas permis.

Cela montre que la performance dépend non seulement du LLM, mais aussi de l'adéquation entre ses comportements et la conception du *ground truth*.

- Le prompt apparaît comme un paramètre critique : certains écarts ne sont pas liés au modèle, mais aux instructions données.
- Un schéma de granularité raisonnable, couplé à un prompt générique, permet d'obtenir des résultats fiables sans nécessiter une connaissance « atomique » des spécificités documentaires.
- L'analyse statistique page par page révèle des indices sur les exceptions structurelles internes aux documents (choix typographiques ou institutionnels), qui peuvent être significatives pour l'historien.

III. Conclusion

Ce travail a exploré l'utilisation des grands modèles de langage pour la génération de données structurées à partir de sources historiques, à travers une étude de cas centrée sur les *Tables nominatives* du Sénat français de 1931. L'approche — combinant OCR, structuration guidée par schéma et génération contrainte via LLM — a produit des résultats évalués grâce à une métrique plus adaptée, l'**IMQ**, intégrée dans un protocole d'alignement optimal reliant données de référence et données prédites.

L'introduction de la métrique IMQ s'est révélée essentielle : elle permet d'évaluer la qualité de structuration au-delà des scores classiques de précision/rappel, inadéquats dans ce contexte.

Plusieurs pistes s'ouvrent pour renforcer la robustesse et la généralisation de l'approche :

- **Relier plus directement données extraites et questions de recherche** : il s'agit d'assurer que les hypothèses de réponse formulées à partir des données générées restent robustes dans le temps.
- **Évaluer le prompt lui-même** : cette étape reste à formaliser pour parvenir à un protocole d'évaluation véritablement complet.

- **Repenser la structuration de données** : elle ne doit pas être considérée comme un simple prétraitement neutre, mais comme un choix déterminant pour les analyses historiques possibles.

De ce point de vue, le prompt et le schéma de données apparaissent comme des **méta-paramètres** du système de production de données historiques. Leur génération et leur ajustement doivent être conçus comme faisant partie intégrante de la chaîne de traitement.

Une voie prometteuse consiste à **systématiser et automatiser ce processus de météo-optimisation**, afin de rendre ces approches reproductibles, transparentes et accessibles à des utilisateurs non spécialistes.

- Brown et al. → ‘¹’

- 2 3 4 567 89101112
, , , ,

¹T. B. Brown, B. Mann, N. Ryder, *et al.*, *Language Models are Few-Shot Learners...*

²M. Puren, *Digital Humanities in the TIME-US Project...*

³Jules (1829-1884) Auteur du texte Poudra et Eugène (1848-1925) Auteur du texte Pierre, *Traité pratique de droit parlementaire / par Jules Poudra,... et Eugène Pierre,...* 1878, URL : <https://gallica.bnf.fr/ark:/12148/bpt6k58651348> (visité le 26/08/2025).

⁴Stéphane Lamassé et Philippe Rygiel, « Nouvelles frontières de l'historien », *Revue Sciences/Lettres*–2 (févr. 2014), Publisher : École normale supérieure, DOI : 10.4000/rsl.411.

⁵Eugène Pierre, *Traité de droit politique électoral et parlementaire. Supplément (5e édition complétée par des références au Supplément de 1919) / par Eugène Pierre,...* 1924, URL : <https://gallica.bnf.fr/ark:/12148/bpt6k6228759r> (visité le 26/08/2025).

⁶Hélène Saudrais, « Aux sources de la loi, les archives parlementaires (XIXe-XXe siècles) », *Revue française de droit constitutionnel*, 101–1 (avr. 2015), Publisher : Presses Universitaires de France Section : Histoire, p. 165-175, DOI : 10.3917/rfdc.101.0165.

⁷Hélène Lemesle, « Apprendre le travail parlementaire et construire la séparation des pouvoirs dans les années 1870 », *Revue d'histoire du XIXe siècle. Société d'histoire de la révolution de 1848 et des révolutions du XIXe siècle*–35 (déc. 2007), Publisher : Société d'histoire de la révolution de 1848, p. 125-139, DOI : 10.4000/rh19.2132.

⁸Roger (1878-1944) Bonnard, *Les règlements des assemblées législatives de la France depuis 1789 (notices historiques et textes)*, fre, Publisher : Paris : Société anonyme du Recueil Sirey, 1926, 1926, URL : <https://www.babordnum.fr/items/show/572> (visité le 26/08/2025).

⁹Hugo Coniez, « L'Invention du compte rendu intégral des débats en France (1789-1848) », *Parlement/s, Revue d'histoire politique*, 14–2 (déc. 2010), Publisher : L'Harmattan Section : Histoire, p. 146-158, DOI : 10.3917/parl.014.0146.

¹⁰Benjamin Morel, *Le parlement, temple de la République. De 1789 à nos jours*, Passés composés, Paris, 2024.

¹¹Antoine Prost, *Douze Leçons sur l'histoire*, Points, Publisher : Le Seuil, Paris, 2010, URL : <https://shs.cairn.info/douze-lecons-sur-l-histoire--9782757820643> (visité le 22/08/2025).

¹²H. Coniez, « L'Invention du compte rendu intégral des débats en France (1789-1848) »...

- 131415161718192021
- Chen et al. → ²²
- Clavert Muller → ²³
- Devlin et al. → ²⁴
- Finkel, Grenager Manning → ²⁵
- Humphries et al. → ²⁶
- Kirillov et al. → ²⁷
- Kišš, Beneš Hradiš → ²⁸

¹³ *Archive nationales, fonds du Premier ministre; Secrétariat général du Gouvernement; Direction des Journaux officiels*, 1881, URL : [https://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?consIr=&frontIr=&optionFullText=&fullText=&defaultResultPerPage=&irId=FRAN_IR_014025&formCaller=GENERALISTE&gotoArchivesNums=false&auSeinIR=false&details=false&page=&udId="](https://www.siv.archives-nationales.culture.gouv.fr/siv/rechercheconsultation/consultation/ir/consultationIR.action?consIr=&frontIr=&optionFullText=&fullText=&defaultResultPerPage=&irId=FRAN_IR_014025&formCaller=GENERALISTE&gotoArchivesNums=false&auSeinIR=false&details=false&page=&udId=) (visité le 26/08/2025).

¹⁴ *Table annuelle du Journal officiel de la République française Lois et décrets*, EN, 1931, URL : <https://gallica.bnf.fr/ark:/12148/bpt6k65430703> (visité le 26/08/2025).

¹⁵ P. Rygiel, *Historien à l'âge numérique*, Code : Historien à l'âge numérique Publication Title : Historien à l'âge numérique Reporter : Historien à l'âge numérique Series Title : Papiers, Villeurbanne, 2017 (Papiers), URL : <https://books.openedition.org/pressesenssib/6303> (visité le 19/08/2025).

¹⁶ Delphine Gardey, « Scriptes de la démocratie : les sténographes et rédacteurs des débats (1848–2005) », *Sociologie du travail*, 52–2 (juin 2010), Publisher : Association pour le développement de la sociologie du travail, p. 195–211, DOI : 10.4000/sdt.13695.

¹⁷ Jean Boutier, « L’usage historien des archives », dans *Corpus, sources et archives*, Code : Corpus, sources et archives, Tunis, 2001 (Études et travaux de l’IRMC), p. 9–22, DOI : 10.4000/books.irmc.776.

¹⁸ Ichiro Hasegawa, *Orbits of Ancient and Medieval Comets*, URL : https://adsabs.harvard.edu/full/1979PASJ...31..257H?utm_source=chatgpt.com (visité le 25/08/2025).

¹⁹ J. Chazalon et E. Carlinet, « Revisiting the Coco Panoptic Metric to Enable Visual and Qualitative Analysis of Historical Map Instance Segmentation »...

²⁰ Joseph Barthélémy, *Essai sur le travail parlementaire et le système des commissions*, Country : FR 26 cm., Paris, 1934 (Bibliothèque de l’Institut international de droit public, 5).

²¹ *Les commissions générales de 1921-1940*, fr, URL : <https://www.senat.fr/connaitre-le-senat/lhistoire-du-senat/les-proces-verbaux-des-commissions/les-commissions-generales-de-1921-1940.html> (visité le 27/08/2025).

²² Yunmo Chen, William Gantt, Tongfei Chen, Aaron Steven White et Benjamin Van Durme, *A Unified View of Evaluation Metrics for Structured Prediction*, arXiv :2310.13793 [cs], oct. 2023, DOI : 10.48550/arXiv.2310.13793.

²³ clavertmuller.

²⁴ J. Devlin, M.W. Chang, K. Lee, et al., *BERT*...

²⁵ finkelmanning.

²⁶ Mark Humphries, Lianne C. Leddy, Quinn Downton, Meredith Legace, John McConnell, Isabella Murray et Elizabeth Spence, *Unlocking the Archives : Using Large Language Models to Transcribe Handwritten Historical Documents*, arXiv :2411.03340 [cs], nov. 2024, DOI : 10.48550/arXiv.2411.03340.

²⁷ kirillov.

²⁸ kisshradis.

- Knutsen → ²⁹
- Kodym Hradis → ³⁰
- Kohút Hradis → ³¹
- Kojima et al. → ³²
- Liu et al. → ³³
- de Marneffe et al. → ³⁴
- Mintz et al. → ³⁵
- Mistral AI → ³⁶
- Morel → ³⁷
- Nadeau Sekine → ³⁸
- Peyré Cuturi → “
- Radford et al. (2018) → ³⁹
- Radford et al. (2019) → ⁴⁰
- Wei et al. → ⁴¹
- Willard Louf → ⁴²
- Yuan Sester → ⁴³

²⁹Gunnar W. Knutsen, « Alimenter des bases de données grâce à l'intelligence artificielle », *Histoire & mesure*, XXXIX–2 (déc. 2024), ISBN : 9782713233685 Number : 2 Publisher : Éditions de l'EHESS, p. 99–116, DOI : 10.4000/140kk.

³⁰kodymhradis.

³¹kohuthradis.

³²kojima.

³³liu.

³⁴marneffe.

³⁵mintz.

³⁶mistrala.

³⁷B. Morel, *Le parlement, temple de la République. De 1789 à nos jours...*

³⁸nadeausekine.

³⁹radford2018.

⁴⁰radford2019.

⁴¹J. Wei, Y. Tay, R. Bommasani, *et al.*, *Emergent Abilities of Large Language Models...*

⁴²willardlouf.

⁴³yuansester.

- Zhang Shasha → ^{‘44’}
- Zhao et al. → ^{‘45’}

⁴⁴**zhangshasha.**

⁴⁵Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, *et al.*, *A Survey of Large Language Models*, en, arXiv :2303.18223 [cs], mars 2025, DOI : 10.48550/arXiv.2303.18223.

Quatrième partie

Conclusion

Annexe A

Le titre très long de la première annexe

Table des matières

Résumé	i
Remerciements	iii
Bibliographie	v
Sources primaires	v
Méthodologie historique	v
Droit et histoire parlementaires	vi
Institutions et politiques patrimoniales	vii
Humanités numériques	viii
Images et numérisation	ix
Traitement automatique du langage	xi
Philosophie de la technique et médialité	xii
	xiii
Introduction	xv
I Des sources sérielles : les Tables Annuelles du Sénat comme cas d'usage	1
1 Le <i>Journal Officiel</i> : la parole, la main, la vue et le droit	3
I. Le <i>Journal Officiel</i> : entre publicité et promulgation, archives et documentation	4
II. Le <i>Journal Officiel</i> : un contexte technique et administratif (1921–1940)	5
III. Les « processus métier » de la publicité parlementaire à partir des Tables nominales : analyse des sources	5
2 Les tables annuelles : des relations documentaires	7
I. Les tables dans l'environnement du <i>Journal Officiel</i>	7

II.	Forme et organisation des tables	7
III.	Informations sémantiques et usages	8
1.	Index des orateurs	9
2.	Table des matières thématiques	9
3.	Références législatives et réglementaires	9
II	L'enjeu des données structurées : des sources à la base de données	13
3	Une histoire par les données	15
I.	Numériser les sources	17
1.	En « mode texte »	17
2.	En « mode image »	20
II.	Ce que la numérisation fait aux sources	22
1.	La « datafication »	22
2.	Pratiques historiennes : sphère technique, sphère sociale	24
3.	Documents sériels et approches quantitatives	27
4	La reconnaissance optique de caractères	29
I.	Une vue d'ensemble	29
II.	Les grandes technologies d'OCR	35
III.	CER, WER : métriques pour évaluer la qualité de l'OCR	36
IV.	Numérisation du <i>Journal Officiel</i> : une politique documentaire partagée entre la BnF et les institutions parlementaires	38
5	Données brutes, données structurées	41
I.	Modéliser des données structurées	42
II.	Produire des données structurées à partir de texte	45
1.	Approche à motifs explicites : les ReGex	46
2.	Approches extractives : BERT	49
3.	Approches génératives : les LLMs	51
III.	La sortie structurée via LLM pour le <i>Journal Officiel</i>	52
1.	Simplicité	53
2.	Sortie structurée, génération structurée	53
3.	L'outil « Corpusense » du projet Mezanno : une pipeline de l'image numérisée à la donnée structurée	55

III Expérimenter et évaluer pour comprendre : une démarche historienne outillée	59
6 L'outil Corpusense : une chaîne de traitement pour les sources historiques	61
I. Une instance de pipeline « classique »	62
II. Le travail sur Corpusense	65
III. Ateliers à l'EHESS et à la BnF	66
7 La sortie structurée via LLM appliquée à la Table des Noms du Sénat : une approche empirique	67
I. Expérimentations	68
1. Prise en main intuitive du problème de la génération de données	68
2. Design, prompt et vérité terrain : trouver le bon modèle de données .	70
3. Préparer l'évaluation : comparer, apparier	72
II. Recouper des données pour l'analyse historienne	77
1. Des pages aux dates : utilisation de l'API Gallica et des métadonnées des manifestes	77
8 Livrables	79
I. Jeux de données pour le protocole d'évaluation de la sortie structurée	79
1. Protocole d'évaluation de la sortie structurée	79
2. Vérité terrain, prompt et schéma : un échantillon des Tables	80
3. Prédictions obtenues avec le modèle Mistral 8b avec l'extraction guidée par schéma	81
II. Une métrique pour l'évaluation des données générées par la sortie structurée via LLM	81
1. Mise en correspondance des prédictions et de la vérité terrain avec le transport optimal	81
2. Limites des métriques classiques : précision, rappel et F1	82
3. Integrated Matching Quality (IMQ)	82
4. Résultats et analyse	82
III. Conclusion	84
IV Conclusion	89
A Titre court	91