ÉCOLE NATIONALE DES CHARTES
UNIVERSITÉ PARIS, SCIENCES & LETTRES

Joël Féral
licencié ès lettres
diplômé de master

L'outil Mezanno pour les approches quantitatives en sciences humaines et sociales

De l'élaboration d'un corpus de documents sériels à l'extraction automatisée de données structurées : les tables annuelles du Journal Officiel (1931 - 1935) comme cas d'usage

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

Mes remerciements vont tout d'abord à Jean-Philippe M., Joseph C., Marie P. et Sébastien C. qui m'ont fait confiance pour participer à l'élaboration du projet Mezanno.

Je remercie également mes parents ET SURTOUT DGETTO !!!

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'épars documentaire s'esquissent 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 tentent 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 redivision 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 sérielle 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

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

qsd

1. 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.

2. 1.1. Le Journal Officiel : entre publicité et promulgation

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 recomposition 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.

3. 1.2. 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.

4. 1.3. Le Journal Officiel dans son 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.

5. 1.4. Les « processus métier » de la publicité parlementaire

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 1

Chapitre 2 — Les tables annuelles : des relations entre corpus

1. 2.1. 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.

2. 2.2. 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.

3. 2.3. 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.

4. 2.4. 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.

5. 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 régu-

lière, 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.

6. 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.

Deuxième partie

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

Chapitre 2

Une histoire par les données

Les *Tables Annuelles* du Sénat, comme on vient de le voir, 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 numérisées. 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 technique 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 » des sources historiques, laquelle est inhérente à la disponibilité de corpus numérisés par les politiques de valorisation des fonds des institutions patrimoniales. [Section 1 : Datafication des corpus : « numériser »]

Ensuite, premier problème : 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. 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 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, un sénateur inaugure chaque paragraphe). 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 approche extractive (les Bert) ou bien les approches génératives qui « lisent » le texte et restitue l'information comprise et permettent de contourner le problème des exceptions qui forment le corps des documents. Cette chaîne de travail de capture est tourné vers un format de cette information exploitable par l'ordinateur. 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.

I. Datafication des corpus : *numériser*

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 oeuvres littéraires du domaine public. La Bible, les oeuvres de Shakespeare, quelques autres de Lewis Carroll ou de James M. Barrie seront notamment reproduites [Marie Lebert]. Ce travail de « numérisation » est en fait un travail laborieux : chacune des lettres de chaque livre sera tapée à la main, les unes après les autres. En 1990, de façon contemporaine à la jeunesse 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 2005 [Marie Lebert]. Entre le livre et la version numérique,

il n'y a pas d'image : juste le travail de transcription manuel des caractères. C'est une numérisation des livres « en mode texte » [Bermès, 30-33] : l'information textuelle seule, stockée sur disque dur, est reproduite sur l'écran, cela « destructurant l'objet livre » [Bermès]. 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 – ainsi 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 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 tableur [Claire Lemercier, Claire Zalc]. La similarité n'est toutefois que d'ordre opératoire. Il y a un face-à-face entre l'opérateur et la source à restranscrire. 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 suppose une enquête critique : sélectionner, structurer, et souvent synthétiser des données pour les « mettre en table », c'est-à-dire les rendre comparables et cumulables. Cet schème opératoire transcriptif peut être qualifié, avec Simondon, de travail de *transduction technique* : 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 matérielles et intellectuelles. L'information change de milieu et, en contexte numérique, son inscription *in silicium*, permet de reconsidérer le texte discrétisé comme un ensemble d'éléments manipulables. 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, 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 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 « apparat critique » qui permettent un retour au texte original à travers des clés qui en indexe le contenu [Marc Van Campenhoudt]. 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'enquête [Dewey ; Claire Lemercier, Claire Zalc] – elle 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 [Anne-Lyse Renon]. Elle est propre à l'historien qui ajuste ces catégories à ses besoins. Si bien que ce processus valuatif de mise en catégorie de l'information capturée accompagne la transcription manuelle et littérale des sources. Elle est ce proche de la 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 contenu [Steven Derose].

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 systématiquement [Bertrand Müller] ; ou encore Laduriesous forme de fiches ou de tableaux [Ladurie]. 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 » [Bermès]. 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). Leur ambition est similaire — mettre à disposition le patrimoine imprimé à grande échelle — à noter que leur logique diverge : lorsque Gallica s'inscrit dans une mission de service public, Google privilégie la puissance de l'indexation et la recherche plein texte, au détriment de l'intégrité patrimoniale des objets – et, à ses débuts, au détriment du droit d'auteur français.

La numérisation en mode image suppose des investissements lourds en infrastructures, en personnels et en politiques documentaires. Elle se distingue ainsi des pratiques de transcription textuelle, souvent issues de gestes individuels ou collaboratifs (historien recopiant ses sources, dépouillements collectifs, corrections d'OCR par crowdsourcing). Certes, des chercheurs ou amateurs produisent eux aussi des photographies de documents — parfois propres, parfois « à l'arrache » —, mais ces fichiers restent isolés, de qualité variable, sans métadonnées ni garantie de pérennité. Ils ne deviennent pas des corpus mais des archives personnelles. Les institutions, au contraire, inscrivent leurs campagnes dans des stratégies de patrimonialisation, produisant des ensembles cohérents et diffusables à large échelle.

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, souvent 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.

Il importe ainsi de rappeler que numériser n'est pas éditer [Bermès, Poupeau]. Là où la mise en table implique un véritable travail de curation — 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 amateure, ces fichiers constituent le plus souvent un point de départ, ou une stratégie mnémonique 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 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. » [Jean-Didier Wagneur]

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 » ; 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.

3. La « datafication »

Cette opposition doit enfin être replacée dans l'horizon plus large de la *datafication* (Clavert, Rygiel). La *datafication* est le processus qui vise à quantifier un phénomène de sorte qu'il soit calculable et analysable [Frédéric Clavert]. Elle est en quelque sorte un « schème opératoire » [Simondon] 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 appuient [Frédéric Clavert, 123].

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 de structuration ; 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 concurrents [Bermès, 29] 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' » [Datactivist]. 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 exploitables, 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 instanciation pure de la source dans le giron du binaire qu'une stratégie épistémique de sa présentation au sein d'un réseau sociotechnique instituant [Anne-Lyse Renon]. 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 découlent. Or, au sein des politiques contemporaines de numérisation, cette dimension instituante prend une forme discursive et normative particulière : celle du *patrimonial*. La numérisation est justifiée et orientée par un vocabulaire de conservation, de démocratisation et de valorisation. Autrement dit, ce qui est numérisé n'est pas seulement conservé ou reproduit : il est institué comme patrimoine, doté d'une valeur symbolique « émotionnelle » et sociale spécifique [Bermès]. 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.

4. 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 institutionnels [Johanna Drucker]. 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*. D'ailleurs certains artistes, parlant volontiers de « nouvel espace épistémique » ou « d'*épistémè numérique* », ont justement mis au centre de leur travail cette question « interprétative » et graphique de l'accès à l'information à travers des interfaces [Architecture de Mémoire, Jean-Marie Dallet].

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 sérielle [Adeline Daumard, François Furet], dans la lignée des « archives quantitatives » chères à l'école des Annales. Les décennies 1960–1970 voient s'imposer un véritable culte du chiffre, nourri par l'accès aux centres de calcul et par l'ambition d'une histoire totale [Sébastien Poublanc, Nicolas Marqué]. 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 sérielle 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 globale [Sébastien Poublanc, Nicolas Marqué]. 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 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 [Bernard Stiegler] 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 », l'archive numérisée devient un objet relationnel : 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 découvrabilité des sources dépend ainsi de ces dispositifs techniques saisis dans des processus de concrétisation [Simondon], 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 datafication 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 ?

5. 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 bilbliothèque 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/11/num$_c$*oncertee*$_i$*mpr*$_p$*rogr*$_p$*artenaires.pdf?utm$_s$*ource = 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/acceder-aux-numerisations-du-journal/ ?utm$_s$*ource = 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.

6. 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$_s ource = chatgpt.com)).$*L'$**Assemblée n** $//bnf.libguides.com/c.php?g = 659907p = 4659962utm_source = chatgpt.com)$; $[BoteOutils](https : //boiteaoutils.info/2013/01/acceder - aux - numerisations - du - journal/?utm_source = chatgpt.com))$.

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).

7. 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, interrogeable 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.

Chapitre 3

De l'image au texte : la reconnaissance optique de caractère (OCR)

Doubler les images du texte.

Le choix institutionnel de numériser massivement les corpus patrimoniaux en mode image tient à la volonté de restituer au plus près la matérialité visuelle des documents. Dès les années 1990, les grandes bibliothèques nationales — à commencer par la BnF avec Gallica — ont privilégié la photographie ou le scan intégral des pages afin de garantir une reproduction fidèle : typographie, mise en page, ornements, blancs, marques de lecture ou d'usage. Cette approche s'inscrit dans une logique de patrimonialisation : conserver une trace stable et exploitable de l'objet imprimé, indépendamment des évolutions des standards textuels ou des logiciels de lecture. L'image constitue en effet un témoin pérenne, qui permet aussi bien la consultation visuelle du document par le lecteur que la réédition de fac-similés numériques. Mais cette stratégie, qui assure l'intégrité visuelle et symbolique de la source, laisse le texte dans un état opaque pour la machine. C'est précisément pour franchir cette opacité et rendre ces corpus interrogeables 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.

1. Extraire automatiquement du texte

La reconnaissance optique de caractères (OCR, *Optical Character Recognition*) désigne l'ensemble des procédés permettant d'extraire automatiquement du texte lisible par machine à partir d'images numérisées. Elle constitue l'opération clef qui rend un document photographié ou scanné interrogeable en plein texte. Concrètement, l'OCR vise à transformer une matrice de pixels (mode image) en une suite de signes discrets (mode texte), selon la logique de la

datafication déjà évoquée : passage d'un continu visuel à un discret alphabétique.

Un document numérisé est d'abord une matrice de pixels codant des intensités lumineuses. Pour l'ordinateur, un caractère imprimé n'existe pas en tant que lettre, mais comme une forme graphique composée de zones plus ou moins sombres. L'OCR consiste à :

1. **Prétraiter l'image** (binarisation, redressement, suppression du bruit visuel, segmentation en zones de texte, lignes et mots) afin d'obtenir des silhouettes de caractères.

1. **Reconnaître les caractères** par comparaison avec des modèles préexistants. Deux approches coexistent :

la reconnaissance par patterns* (méthodes classiques, basées sur la correspondance visuelle de formes) ; * la reconnaissance statistique et neuronale (méthodes modernes, utilisant apprentissage profond et réseaux de neurones convolutifs pour apprendre les formes typographiques).

1. **Restituer un texte** sous forme encodée (UTF-8 ou Unicode), éventuellement enrichi de métadonnées de position (*ALTO XML*, *hOCR*) permettant de conserver l'ancrage spatial des mots dans l'image.

Le résultat dépend de nombreux facteurs : qualité du scan, état matériel du document, typographie, langue, mais aussi du degré d'adaptation du modèle de reconnaissance aux sources traitées.

2. Les grandes technologies d'OCR

Historiquement, les premiers systèmes d'OCR (années 1970–1990) étaient conçus pour reconnaître des caractères bien standardisés (typographies contemporaines, impressions propres). Avec l'essor de la numérisation patrimoniale et la variété des fonds (imprimés anciens, manuscrits, journaux abîmés), les limites de ces approches ont conduit au développement de solutions spécialisées.

* **Tesseract OCR** : développé par HP dans les années 1980 puis repris par Google, Tesseract est l'un des moteurs les plus utilisés, notamment dans les grandes bibliothèques numériques (Google Books, Internet Archive). Depuis sa version 4 (2018), il intègre des réseaux de neurones récurrents (LSTM), améliorant sa performance sur des typographies variées. Cependant, il reste peu flexible sur des écritures complexes ou des documents abîmés.

* **ABBYY FineReader** : solution propriétaire largement utilisée par les institutions pour sa robustesse industrielle. Performante sur les imprimés modernes, elle demeure coûteuse et peu ouverte, limitant sa personnalisation.

***Kraken** : développé à partir de l'expérience d'*OCRopus* (projet open-source de Google), Kraken est un moteur open-source basé sur l'apprentissage profond. Sa force réside dans la possibilité d'entraîner des modèles sur des corpus spécifiques (par exemple, typographies anciennes, textes en alphabets non latins). Il est aujourd'hui largement utilisé pour les fonds patrimoniaux et manuscrits.

***eScriptorium / Pero OCR** : environnement développé à l'École pratique des hautes études (EPHE/PSL) et au sein du projet* Huma-Num*. Il combine des outils de segmentation (basés sur les réseaux de neurones convolutifs de Pero OCR) et de transcription (via Kraken). eScriptorium fournit une interface collaborative pour annoter, entraîner des modèles et produire des transcriptions à grande échelle. Cette approche est particulièrement adaptée aux humanités numériques, car elle permet d'impliquer chercheurs et communautés dans l'amélioration des modèles.

* **HTR (Handwritten Text Recognition)** : pour les manuscrits, l'OCR cède la place à la HTR, qui repose sur des principes analogues mais adaptés aux écritures manuscrites. La plateforme **Transkribus**, par exemple, est aujourd'hui la référence pour les fonds manuscrits européens, permettant d'entraîner des modèles spécifiques sur des écritures d'archives.

3. Enjeux et limites

Les performances de l'OCR sont hétérogènes : sur des imprimés du XX siècle en bon état, la reconnaissance atteint souvent plus de 95

*des corrections manuelles ou collaboratives (*crowdsourcing*), * l'entraînement de modèles spécialisés, * l'élaboration de pipelines de traitement intégrant segmentation, normalisation linguistique et encodage structuré (TEI, ALTO XML).

L'OCR ne produit donc jamais une « donnée brute », mais un texte *obtenu*, dont la fiabilité dépend du corpus, des choix techniques et du travail éditorial qui l'accompagne.

Comme on l'a vu dans le chapitre précédent, les *Tables Annuelles* du Sénat sont disponibles sur Gallica. D'un point de vue technique, ces *Tables* sont des documents numérisés, c'est-à-dire des images dont on aura discrétisé l'information.

Chapitre 4

Données brutes, données structurées : quelques enjeux de l'interopérabilité.

En prolongement de notre problématique initiale, nous structurons notre revue autour de trois questions clés :

1. Comment modéliser efficacement des données structurées ?

1. Comment évaluer la qualité des données structurées produites par ces approches ?

1. Comment générer des données structurées à partir de texte ?

1. Modéliser des données structurées

Les données structurées peuvent prendre plusieurs formes, parmi lesquelles :

- **Ensembles d'enregistrements (Record Sets)** : ce sont des ensembles non ordonnés de tuples, analogues à des tables de base de données où les colonnes représentent les attributs des objets. Cette structure est couramment utilisée dans les tâches d'extraction d'information, comme la reconnaissance d'entités nommées [18] ou l'extraction de relations [15].
- **Séquences d'enregistrements (Record Sequences)** : ce sont des versions ordonnées des ensembles d'enregistrements, où l'ordre des éléments a un sens. La séquence peut refléter un critère (par exemple temporel) ou faciliter certaines tâches comme la validation croisée, à l'image des annuaires.
- **Arbres** : ces structures hiérarchiques sont souvent utilisées pour représenter des relations imbriquées ou des dépendances, comme en analyse syntaxique de dépendances [14].

- **Graphes** : structures flexibles utilisées pour représenter des connaissances consolidées, comme les ontologies ou les graphes de connaissances. Bien qu'ils soient largement étudiés, leur évaluation sort en général du champ de l'extraction d'information, et dépasse donc le cadre de ce travail.

Dans cet article, nous nous concentrons sur les ensembles et séquences d'enregistrements, qui sont les plus pertinents pour notre étude de cas : l'extraction de données structurées à partir de documents parlementaires. Les différents modèles que nous avons expérimentés sont décrits en section 3.

2. Évaluer la qualité des données structurées

L'évaluation de la qualité des données structurées peut être globalement classée en deux familles de métriques : les métriques de type distance d'édition et les métriques d'appariement, telles que définies dans [2].

- **Métriques de distance d'édition** : elles reposent sur des processus d'optimisation complexes et sont souvent coûteuses en calcul. Leur interprétabilité est limitée, car elles ne fournissent pas de comparaison directe entre données produites et attendues. Exemples : la distance de Levenshtein (pour les comparaisons caractère par caractère), ou la distance d'édition d'arbres [25] pour des structures hiérarchiques. Des métriques générales existent aussi pour les graphes, mais leur complexité les rend souvent impraticables.

- **Métriques d'appariement** : plus interprétables, elles identifient explicitement les éléments qui correspondent entre les données produites et attendues. La plupart reposent sur un appariement biparti entre les ensembles prédits et de référence, et calculent des scores à partir du nombre d'éléments appariés [2]. Les plus courantes incluent la mesure F1 (précision + rappel) et l'indice de Jaccard (similarité d'ensembles). Peu d'études cependant se concentrent sur les cas de données structurées ou d'appariement partiel, où les données produites ne coïncident pas parfaitement avec les attendues.

Il est intéressant de noter que la communauté de la vision par ordinateur rencontre exactement le même problème. Par exemple, le *COCO Panoptic Segmentation Challenge* [7] propose un cadre similaire, où l'évaluation combine détection, segmentation et classification via un appariement optimal entre les surfaces prédites et de référence. Une démarche analogue est adoptée par Chen et al. [2].

Dans nos travaux, nous adoptons une métrique d'appariement basée sur un appariement optimal entre ensembles de données structurées. Cette approche généralise l'appariement biparti tout en intégrant les correspondances partielles. Elle fournit à la fois une évaluation quantitative de la qualité des données et une identification des éléments manquants ou « hallucinés », offrant ainsi des pistes d'amélioration concrètes.

3. Produire des données structurées à partir de texte

Les approches de génération de données structurées à partir de texte se divisent en deux grandes catégories : **détection (extractive)** et **génération (abstractive ou générative)**.

- **Approches extractives** : elles identifient dans le texte les fragments correspondant à des champs ou éléments de la donnée structurée. Les méthodes traditionnelles reposaient sur des règles (expressions régulières, heuristiques). Plus récemment, des modèles d'apprentissage, comme les modèles de séquence étiquetée (par ex. CRF [5]) ou les modèles encodeurs-transformers (par ex. BERT [4]), dominent. Leur conception impose un alignement strict entre texte d'entrée et étiquettes de sortie, réduisant le risque d'hallucinations (c'est-à-dire d'inventions de données). Mais ces approches nécessitent un entraînement spécifique à la tâche, donc des données annotées, des ressources computationnelles et du temps.

- **Approches génératives** : elles exploitent des modèles autoregressifs pour « traduire » le texte en un format structuré cible. L'essor des LLMs a ravivé l'intérêt pour cette voie, en raison de leurs capacités de généralisation impressionnantes, même sans entraînement spécifique [1, 20, 21]. Les sorties peuvent être contraintes à des formats précis (JSON, schémas complexes) grâce au filtrage dynamique de tokens valides [23]. Ces modèles peuvent produire des structures complexes et imbriquées, et inférer des éléments implicites. Leur inconvénient majeur est cependant leur propension aux hallucinations, difficiles à détecter.

Dans cet article, nous explorons l'efficacité des approches génératives pour gérer les structures répétitives présentes dans les index parlementaires ou *Tables*. Nous mettons à profit les capacités *zero-shot* des LLMs et évaluons la viabilité de cette approche pour générer des données structurées dans ce contexte spécifique.

Chapitre 5

Du texte à la donnée structurée : capturer la sémantique

I. Approche à motifs explicites : les ReGex

Une première approche naïve d'extraction de l'information du texte : les regex. Puissants, rapides. Mais rigide et implique de connaître à l'avance la forme de ce qu'on cherche, ce qui n'est pas trivial ! Il faut aussi partir du principe que l'on a pas une connaissance synthétique a priori de l'information. Il y a toujours un « hic ». Fragile face au bruit ocr, aux fautes typographiques inattendues ; et avoir une regex plus souple, c'est aussi prendre le risque de capter du bruit.

> La recherche floue Un moyen de diluer la rigidité des motifs ; mais ne permet que de trouver ce que l'on connaît à l'avance. Dans un optique d'extraction massive, on veut tout sortir automatiquement.

> Automates finis !

La contrainte forte des regex Intéressant à coupler avec d'autres approches plus souple comme on le verra.

- **Patrons linguistiques** (grammaires, dépendances syntaxiques)
- **Listes de référence / gazetteers**
- **Règles de post-traitement**

=> Avantage : explicable, prévisible => Limite : peu robustes aux variations inattendues

II. Approches extractives : l'approche Bert (one-to-one)

L'approche du surlignage 1 to 1.
Principe : le modèle apprend à repérer les entités dans un texte via des annotations.

- **Modèles supervisés classiques** : CRF, SVM, MaxEnt
- **Neuraux séquentiels** : BiLSTM-CRF, CNN-LSTM
- **Transformers extractifs** : BERT, RoBERTa, CamemBERT en mode NER

=> Avantage : généralise mieux, bonne précision => Limite : nécessite des données annotées et un entraînement

III. Approches génératives : les LLMs

L'usage croissant de l'intelligence artificielle par les historiens

3]*multiplielespossibilitsdeproductiondejeuxdedonneshistoriques.L'avnementdesgrandsmodlesdelangage*

L'utilisation des LLMs ouvre de nouvelles perspectives pour l'extraction de données structurées

13]*partirdedocumentshistoriques.Danscecontexte,undficentralrsidedanslaproductiondesortiesstructure*

Deux questions fondamentales demeurent cependant :

1. comment passer d'un texte brut à une représentation structurée exploitable, telle qu'un tableau ou un fichier CSV ;
2. comment évaluer la qualité et la fiabilité des données extraites.

Cet article aborde ces deux aspects à travers une étude de cas concrète : l'extraction d'informations structurées à partir des *Tables nominatives* (ou *Tables des noms*) du Sénat français de 1931, qui constituent un index de l'activité parlementaire classé par nom. Nous explorons une approche de génération faiblement contrainte à l'aide d'un LLM, et proposons une méthode pour représenter les données cibles, guider le processus d'extraction et évaluer les performances du système. Au-delà de ce cas spécifique, l'étude vise à contribuer à une réflexion plus large sur la faisabilité et les limites des modèles génératifs pour la structuration de données historiques.

Les *Tables des noms* du Sénat français furent publiées durant la Troisième République (1870–1940)[2]. Dans l'écosystème documentaire plus large du *Journal Officiel* — qui vise à reconstituer l'activité parlementaire et ses issues juridiques ou réglementaires en France —, les *Tables nominatives* du Sénat offrent un relevé concis et systématique des interventions des sénateurs en séance publique. Ces index étaient conçus pour accompagner la transcription des débats[3] et en faciliter la consultation. Compilés manuellement une fois par an, ils recensent chaque intervention d'un sénateur ou d'un membre du gouvernement, précisent l'objet de son discours et indiquent la pagination correspondante. Si ces tables étaient particulièrement utiles à une époque où la recherche plein texte dans les débats parlementaires numérisés n'était pas possible, elles conservent aujourd'hui encore une forte valeur pour les historiens. Une extraction systématique des données permettrait de suivre l'activité parlementaire sur le long terme, de quantifier les interventions de sénateurs affiliés à certains mouvements politiques, ou encore de soutenir la validation croisée des entités nommées extraites des débats eux-mêmes.

Notre objectif est d'extraire des données structurées de ces tables ; pour nos premières expérimentations, nous nous concentrons sur une seule table nominative, celle de 1931. Le début des années 1930 marque en effet l'entrée du parlementarisme français dans une phase de déclin, qui culmine avec la chute de la Troisième République en 1940

17].*L'analysedelatablede*1931*permetdeposerlesbasesd'unetudelargiel'ensembledeladcennie*, *afindemieux*

Après un état de l'art des approches existantes pour l'extraction et l'évaluation de données structurées (Travaux connexes), nous présentons trois contributions principales :

1. nous concevons et mettons en œuvre une chaîne de traitement guidée par un schéma, pour extraire des informations structurées à partir des *Tables nominatives* du Sénat, en combinant OCR et génération à base de prompt avec un LLM (Schéma et pipeline) ;

2. nous introduisons un protocole d'évaluation adapté à cette tâche, incluant une méthode d'alignement par appariement optimal et une métrique continue qui prend en compte les sorties partielles et bruitées (Protocole d'évaluation) ;

3. nous fournissons une évaluation empirique de l'extraction par LLM dans ce contexte historique, montrant que les performances du modèle sont fortement influencées par la conception du prompt et du schéma de données — deux éléments que nous proposons de considérer comme des paramètres critiques du processus de modélisation global (Résultats).

Principe : le modèle produit directement le résultat structuré à partir du texte, sur la base d'une consigne en langage naturel.

- **LLMs** (GPT, Claude, Mistral) en extraction via prompt
- **Fine-tuning génératif** (T5, GPT-4 en mode extraction JSON)

=> Avantage : très flexible, pas besoin de jeu d'entraînement spécialisé => Limite : variabilité, hallucinations, besoin de validation

IV. Approches hybrides

Principe : combiner plusieurs catégories dans un flux de traitement.

- Exemple : Gazetteer pour repérer des entités connues + BERT pour les autres + Regex pour les formats normés + validation humaine

=> Avantage : maximiser précision et rappel => Limite : complexité d'intégration

V. La sortie structurée via LLM pour la capture sémantique du Journal Officiel

Expliquer le choix des LLMs (rapide et pas cher + pas d'entrainement ou de spécialisation).

Troisième partie

Expérimenter et évaluer pour comprendre : une démarche historienne outillée

Chapitre 6

Evaluer le protocole de capture sémantique

De l'importance de Valuation, évaluation. La question de la métrique. Article co-écrit.

Chapitre 7

Expérimentations

Un travail d'exploration technique Une fois la tâche définie (éval de la sortie structurée). TED, donut, Métriqe de l'inégalité triangulaire.

Chapitre 8

Livrable

I. Extraction guidée par schéma des interventions parlementaires à partir des index historiques

1. Les Tables nominatives du Sénat de 1931

Il existe une édition des *Tables du Journal Officiel* pour chaque année, comprenant typiquement environ 450 pages dans les années 1930. La section intitulée *Tables des noms* — qui inclut à la fois le Sénat et la Chambre des députés — s'étend sur une quarantaine de pages, dont environ quinze pour la partie Sénat.

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.

2. Chaîne de traitement pour l'extraction structurée guidée par schéma

Malgré les avancées récentes des grands modèles de vision et langage (LVLMs), leurs performances *end-to-end* en zéro-shot restent insuffisantes pour des tâches d'OCR à haute précision. Pour contourner cette limite, nous adoptons une chaîne de traitement simple,

exploitant les forces de composants spécialisés afin de maximiser la précision globale de l'extraction.

1. Chaque page image est d'abord traitée indépendamment par le moteur OCR **PERO**

$8, 10, 11]pourdtecterettranscrireletexte.Comptetenudesdifficultspersistantesdesegmentationdemi$

2. Une fois les transcriptions obtenues, nous concaténons le texte issu de toutes les pages pertinentes pour former un flux unique. Celui-ci est fourni en entrée à un LLM, chargé de produire les données structurées cibles.

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)

$23].CettecapacitestdisponibledanscertainesAPIcommercialestendues.$

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 *Ministral 8B Instruct v2410*

$16], enraisondesesbonnesperformanceszero\text{-}shot, desoncotmodretdeladisponibilitpubliquedesespoidsdesf$

3. Modélisation des données et définition du schéma

L'objectif principal est d'extraire des informations structurées sur l'activité parlementaire du Sénat français en 1931, dans la perspective de construire une frise interactive représentant la densité d'interventions au cours du temps. Pour garantir la fiabilité et l'interprétabilité de cette visualisation, il est nécessaire de fournir des indicateurs clairs de qualité d'extraction.

Le lien direct entre fiabilité de l'extraction et confiance dans les réponses à des questions de recherche demeure un défi. Nous concentrons donc notre évaluation sur des métriques bien définies, qui quantifient la similarité entre structure prédite et structure de référence (*ground*

truth). Ces métriques ne rendent pas encore compte de l'impact sémantique des erreurs, mais constituent une base transparente d'évaluation.

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. Ces numéros de pages servent d'indicateurs temporels indirects, puisque la pagination est continue sur l'année. L'information extraite est représentée en JSON, format à la fois structuré et interopérable, permettant un traitement ultérieur (CSV, analyses).

Le schéma de données utilisé pour guider le LLM est défini au niveau des noms d'orateurs et de leurs références de pages, ce qui suffit à notre objectif. Nous utilisons la bibliothèque **Pydantic** pour formaliser ce schéma de manière compatible JSON, avec validation stricte des types et descriptions intégrées. Ces descriptions jouent le rôle d'étiquettes sémantiques, améliorant la clarté du prompt et l'orientation du modèle.

4. Construction du prompt pour l'extraction

Le prompt fourni au LLM (cf. Annexe B) vise à guider l'extraction des participants (sénateurs, ministres) et de leurs références de pages à partir du texte. Sa structure est la suivante :

* **Définition de la tâche** : préciser que chaque entrée correspond à une personne ayant participé aux activités du Sénat (sénateurs, ministres, etc.), dont il faut extraire le nom et les références de pages. * **Clarification des termes** : donner des définitions pour des termes clés comme « entrée » ou « action », afin d'éviter toute ambiguïté. * **Cas particuliers** : indiquer comment traiter les cas de renvois d'index (sans numéros de pages, mais renvoyant vers une autre entrée) et les entrées fragmentées sur plusieurs pages (ignorées pour simplifier cette première étude). * **Instructions de formatage** : spécifier les règles de présentation des noms (prénom entre parenthèses après le nom de famille) et des références de pages.

Le prompt pourrait être encore amélioré en y intégrant davantage de contexte historique ou d'exemples représentatifs (*few-shot prompting*).

5. Raffinement itératif du schéma et du prompt

Au cours de l'étude, le schéma et le prompt ont été ajustés de manière itérative. En effet, le LLM proposait parfois spontanément des manières alternatives — et parfois plus efficaces — de structurer l'information extraite. Par exemple, il avait tendance à dédupliquer les références de pages répétées pour un même orateur, simplifiant ainsi le résultat au-delà du schéma initial.

La construction du *ground truth* a donc dû trouver un équilibre entre respect strict du schéma formel et prise en compte des tendances de structuration du LLM.

Afin d'éviter tout biais, tous les ajustements de schéma et de prompt ont été réalisés uniquement sur une page de développement (dite « page 02 »). Les autres pages ont servi exclusivement aux tests finaux, suivant les bonnes pratiques en apprentissage automatique.

À l'avenir, ce processus pourrait être automatisé ou enrichi par des stratégies systématiques d'ingénierie de prompts.

II. Expériences

La configuration expérimentale a été mise en place à travers une phase de développement initiale, durant laquelle le schéma de données, les instructions du prompt et les données de référence structurées pour une page de développement ont été affinés de manière itérative. Une fois cette phase finalisée, nous avons appliqué le modèle de base à un ensemble plus large de pages, puis corrigé manuellement les sorties afin de construire un *ground truth* impartial pour l'évaluation.

Cette section détaille le jeu de données obtenu, les variantes générées pour l'analyse, ainsi que le protocole d'évaluation adopté pour mesurer rigoureusement la qualité des prédictions.

1. Jeu de données

Les données de référence (*ground truth*) ont été construites pour évaluer de manière rigoureuse la qualité d'extraction à différents niveaux de fidélité OCR. Pour chaque page sélectionnée, trois variantes distinctes issues du moteur OCR **PERO**

$$8, 10, 11]onttgnres :$$

1. une version corrigée manuellement (or servant de standard de référence),
2. une version basée sur une segmentation manuelle de la mise en page,
3. une version brute, sans correction ni segmentation.

Ce dispositif permet d'évaluer systématiquement la robustesse du pipeline face au bruit et aux artefacts de mise en page.

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.

Chaque page a été traitée indépendamment afin d'éviter tout biais lié au contexte séquentiel. Cela a parfois conduit à des extractions débutant en milieu d'entrée, fournissant au modèle des informations tronquées et testant sa capacité à interpréter des contextes partiels. Bien que les documents sources soient généralement bien numérisés, certaines distorsions (plis, coupures) introduisent des difficultés réalistes, révélant les limites du modèle en conditions imparfaites.

Les sorties structurées ont été générées avec le modèle **Ministral 8B**, en utilisant un prompt fixe et un schéma défini via Pydantic pour garantir la cohérence. Une température nulle a été choisie pour obtenir des résultats déterministes. Pour chaque page et variante OCR, le modèle a produit des sorties JSON, ensuite comparées aux représentations de référence construites manuellement.

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

2. Protocole d'évaluation des sorties structurées

L'objectif de l'évaluation est de comparer rigoureusement les sorties produites par le LLM (notées P) au *ground truth* (noté G). Comme décrit en section 3.3, 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 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 de

$2, 7], utilisant le transport optimal pour tablir une correspondance one-to-one entre prdictions et r frence. Cette$

Distance au niveau des entrées et appariement 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 $d_n(g_i, p_j) \in 0, 1]$ $vaut 0 pour un match exact et 1 pour une dissimilarit totale.$

* **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], $minimisantladistancetotaleetfournissantunebaserigoureusepourl'valuation.$

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.

Integrated Matching Quality (IMQ) : une métrique robuste

Pour dépasser ces limites, nous exploitons directement la distance d_e pour définir 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.

III. 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.

1. 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

* **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.

2. 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 prédites | 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*.

3. Enseignements

* 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.

IV. 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 méta-optimisation**, afin de rendre ces approches reproductibles, transparentes et accessibles à des utilisateurs non spécialistes.

Quatrième partie

Conclusion

Annexe A

Le titre très long de la première annexe

Table des matières