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

Deuxième partie

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

Chapitre 1

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, sous leur forme « analogique » ; mais aussi sous leur forme « 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. 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 est reproduite, cela destructurant l'objet livre. 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. Dans le cas de Michel Hart et du Projet Gutenberg, la répétition du texte littéraire reste relativement linéaire et vise une reproduction intégrale, sans problématisation des sources. 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. Cette dimension opératoire peut être qualifiée, 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. Dans le cas des historiens, ce passage implique un véritable travail d'individuation des données : découper des flux documentaires continus en unités discrètes (noms, dates, professions, événements), qui ne préexistent pas à l'opération de transcription mais sont construites par elle. La numérisation est une opération configuratrice où la saisie manuelle se fait l'instrument d'un changement de régime technique du support de l'information. Du côté des historiens, ce travail de transduction informationnelle, plus sophistiqué que la transcription littéraire, 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].

Le travail de saisie manuelle n'est évidemment pas une nouveauté introduite par l'ordinateur. Bien avant l'ère numérique, les historiens s'y adonnaient déjà. Ainsi, à la fin des années 1940, Pierre Chaunu recopiait à la main, sur papier, les données issues des archives et des ouvrages nécessaires à sa thèse, afin de les ordonner et de les exploiter systématiquement [Bertrand Müller]. 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 saisie manuelle constitue en ce sens un exemple singulier, puisqu'il s'oppose radicalement à la logique de la numérisation photographique, dite « en mode image » [Bermès]. Cette opposition 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 la matérialité de l'objet. L'histoire des pratiques documentaires, tant individuelles qu'institutionnelles, témoigne de cette tension durable entre deux paradigmes concurrents [Bermès, 29].

2. En « mode image »

[Développer en mode image et les politiques institutionnelles : Gallica, Google Book autour de 2005. Salmi] À l'opposé du mode texte, la numérisation en mode image repose sur la capture photographique ou le scan des documents, cherchant à restituer leur matérialité visuelle. Le texte, les blancs, les marges, la typographie, les ornements, tout est « figé » dans une matrice de pixels. Cette approche a été privilégiée par les grandes politiques de numérisation institutionnelles à partir des années 1990, avec l'émergence de programmes comme Gallica (BnF, 1997) ou Google Books (lancé officiellement en 2004). Ces projets partagent l'ambition d'une mise à disposition massive du patrimoine imprimé, mais ils diffèrent dans leurs logiques : Gallica, qui s'inscrit dans une mission de service public, avec un accent sur la fidélité documentaire, la conservation et l'interopérabilité avec d'autres bibliothèques numériques ; ou encore Google Books, de son côté, qui met en avant la puissance de l'indexation, cherchant à rendre « trouvable » le contenu des livres plutôt qu'à restituer leur intégrité en tant qu'objets. L'essor du mode image a profondément transformé le rapport aux corpus : il permet d'accéder à la forme originale du document, mais rend l'information brute peu exploitable sans traitements complémentaires (OCR, segmentation, structuration). Contrairement au mode texte de Gutenberg, qui sacrifiait la matérialité au profit de la lisibilité, le mode image fige la matérialité mais laisse l'utilisateur face à des images muettes.

Bien sûr, avec les microfilms voire encore avec les méthodes photographiques anciennes, reproduire un document par le moyen de la photographie n'est pas nouveau. Le fac-similé n'a pas ni été inventé par Google ou Gallica. Ce n'est cependant pas seulement le problème de la reproduction textuelle, photographique ou non dont il est question ; mais bien de la diffusion et à l'accès à l'information capturée. La *numérisation* de documents, en mode texte ou en mode image, comprennent historiquement les dispositifs techniques terminaux qui peuvent recevoir et reconstituer l'information. Ainsi le système technique de la numérisation qui est autant l'affaire de discrétisation de l'information des documents que de leur diffusion – à l'instar du dispositif PLAO ou tout simplement du Web. Il est alors intéressant de relier la notion de numérisation à celle de *datafication* pour souligner l'associativité du travail de mise en données à un milieu technique [Latour, Simondon] et social – c'est-à-dire ici institutionnel.

3. Réseaux et datafication : sphère technique, sphère sociale

La *datafication* est le processus qui vise à quantifier un phénomène de sorte qu'il soit calculable et analysable [Frédéric Clavert]. Elle considère le passage du continu au discret ; la calculabité des phénomènes comme une prérogative du « numérique ». 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].

La datafication ne consiste pas simplement à transformer un document papier en fichier numérique : c'est un processus complexe qui combine numérisation, structuration et mise en base de données. En histoire, cela signifie passer du papier (via OCR ou HTR) à un texte exploitable, puis à des entités ou métadonnées normalisées (XML-TEI, bases relationnelles). Ce processus, qui permet de rendre les corpus calculables et interopérables, ouvre certes des perspectives considérables — recherche plein texte, analyses sérielles, croisement de sources — mais il engage aussi une série de choix méthodologiques et institutionnels qui, étant donné leur lourdeur produisent leurs propres biais.

Ces choix se jouent d'abord dans la sélection des corpus à numériser : les bibliothèques privilégient souvent les imprimés homogènes ou les fonds les plus demandés, laissant de côté des séries manuscrites plus fragiles ou moins visibles. Comme l'a montré le projet TIME-US, cette logique favorise la présence de sources officielles ou institutionnelles (journaux ouvriers imprimés, presse syndicale) et relègue dans l'ombre des archives plus marginales (pétitions manuscrites, traces de travail informel). 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.

S'y ajoutent les biais techniques. L'OCR ou la HTR sont rarement neutres : leur performance varie selon l'état du document, la langue, l'alphabet ou la typographie. Dans TIME-US, il a fallu corriger manuellement des milliers d'occurrences et entraîner des modèles spécifiques pour le français pré-moderne. Par ailleurs, la normalisation des données (dates, métiers, entités) tend à gommer des variations significatives et risque de projeter des catégories anachroniques. De ce point de vue, la datafication ne produit pas des « données brutes », mais bien des données construites, filtrées par des choix techniques.

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 un effet d'« Eldorado numérique » : l'historien travaille sur ce qui est disponible, non sur ce qui est historiquement pertinent. La question devient dès lors : comment documenter ces biais et construire la confiance dans des données issues de systèmes techniques ?

4. Un standard à la croisée de la sphère technique et la sphère sociale : IIIF

C'est dans ce contexte qu'émergent des standards comme le IIIF (International Image Interoperability Framework), né dans les années 2010 sous l'impulsion d'un consortium de grandes bibliothèques (BnF, British Library, Stanford, etc.). IIIF répond à une double exigence, à savoir : uniformiser l'accès aux images numérisées en définissant des API permettant de zoomer, annoter, partager et intégrer les images dans des environnements divers et favoriser l'interopérabilité entre institutions en permettant qu'un même document numérisé à Londres, Paris ou New York puisse être consulté, manipulé et enrichi dans une interface commune.

IIIF illustre bien l'évolution de la numérisation vers la *datafication* : il ne s'agit plus seulement de stocker et montrer des images, mais de les intégrer dans un écosystème où elles deviennent manipulables, annotables, recombinables. Un manuscrit, une affiche ou un numéro du *Journal Officiel* numérisé n'est plus seulement une image : c'est une ressource ouverte, potentiellement enrichie par des métadonnées, des annotations collaboratives ou des algorithmes d'analyse.

5. « Le goût de l'archive à l'ère numérique »

La réflexion sur la datafication des corpus ne peut être dissociée du « goût de l'archive » à l'ère numérique. Le terme, mobilisé par le collectif [retrouver les noms exacts], met en évidence

un paradoxe : alors que la numérisation promet une accessibilité inédite aux archives, elle en modifie profondément l'expérience sensible. L'historien ne se trouve plus confronté à des boîtes, des liasses ou des volumes, mais à des corpus massifs, fragmentés, médiés par des interfaces, des moteurs de recherche ou des API.

Ce déplacement fait écho à un débat ancien dans l'historiographie française. Dès les années 1960, Chaunu ou Le Roy Ladurie, dans le sillage de l'École des Annales, avaient déjà revendiqué l'importance des archives sérielles — registres de baptêmes, testaments, cadastres, minutes notariales — au détriment des sources narratives ou « événementielles » jugées plus séduisantes. Ils affirmaient que l'histoire pouvait (et devait) se nourrir de ces « archives grises », répétitives, sans attrait esthétique ni émotionnel, mais capables, une fois cumulées et quantifiées, de révéler des structures profondes (démographiques, sociales, économiques). Autrement dit, l'absence de « goût » de ces archives constituait paradoxalement leur force heuristique.

L'ère numérique radicalise ce basculement : ce qui, dans les années 1970, nécessitait des dépouillements manuels interminables et le recours à l'informatique balbutiante, peut désormais être automatisé et amplifié à une échelle inédite. La datafication prolonge l'intuition des Annales en rendant ces fonds sériels interrogeables et manipulables à grande vitesse. Mais elle modifie aussi l'expérience sensible : le « goût » ne réside plus dans l'objet matériel de l'archive, mais dans la découverte de motifs et de régularités rendus visibles par des visualisations, des bases de données ou des modèles.

Ainsi, l'historien se trouve à nouveau confronté à un paradoxe : les archives sérielles, longtemps délaissées pour leur monotonie, acquièrent une nouvelle attractivité dans l'espace numérique, mais au prix d'un changement de régime du sensible. Le risque est alors de réduire l'archive à son seul potentiel calculable. Le défi, aujourd'hui comme hier, est de concilier la puissance cumulative de ces données avec la vigilance critique nécessaire à l'interprétation de leurs conditions de production.

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

Paradigme de l'image numérisée ==> OCR. Présentation de la dimension technique de l'image numérique matricielle. Information riche et compliquée. C'est le cas de Gallica et spécifiquement du corpus du JO. Problématique de l'extraction repose sur des stratégies d'obtention du texte, à partir duquel on pourra effectuer des traitements extracttifs

Comme on l'a vu dans le chapitre précédent, les *Tables Annuelles* du Sénat sont dis-

ponibles 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. Une image numérique est un tableau – une *matrice* – d'une largeur et d'une hauteur données, comportant alors $largeur \times hauteur$ 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.

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

Le format .txt, JSON, CSV, XSLX. Envisager

Chapitre 2

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

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

Chapitre 3

Données et FAIRness : de la valuation et de l'évaluation

La dimension « personnelle » des données, la FAIRNESS ==> implique d'expliciter la « sitaution » des données, de leurs valuations et de leur qualité (évaluation. Nécessité d'évaluer.

« Most literary scholars would no more simply use the "results" *of a fellow scholar than they would use her toothbrush* » (Responses to Moretti, p. 4). 5

Troisième partie

Expérimenter et évaluer pour comprendre : une démarche historienne outillée

I. Introduction

The growing use of artificial intelligence by historians **clavert2024histoire** is multiplying the possibilities for producing historical datasets. The advent of large language models (LLMs) is further changing the landscape, especially for processing textual data corpora, with a proliferation of uses and experiments in the humanities and social sciences[1]. Zero-shot LLMs are capable of performing a wide range of tasks without the need for task-specific examples or fine-tuning **kojima2022large** ; **wei2022emergent** ; **zhao2023survey** and have demonstrated their ability to carry out many time-consuming tasks in historical research, such as transcription **humphries2024unlocking**, information extraction **knutsen2024alimenter**, or annotation **yuan2025leveraging**.

The use of large language models (LLMs) opens new perspectives for extracting structured data **liu2024structured** from historical documents. In the context of historical data extraction, a central challenge lies in obtaining structured outputs. One approach is structured generation, which constrains a large language model to directly produce information in a predefined format such as JSON. Alternatively, structure can be imposed through post-processing of free-form text outputs. Regardless of the approach, producing structured data enables traceability back to the original document and facilitates source verification. It also supports downstream uses, such as integration into a database or further computational analysis.

Two fundamental issues still remain : (1) how to move from raw text to an exploitable structured representation, such as a table or CSV file ; and (2) how to assess the quality and reliability of the extracted data. This article addresses both aspects through a concrete case study : the extraction of structured information from the 1931 *Tables nominatives* or *Tables des noms* of the French Senate (index of senatorial activity ordered by name). We explore a lightly constrained generation approach using an LLM and propose a method to represent the target data, guide the extraction process, and evaluate system performance. Beyond this specific case, the study aims to contribute to broader reflections on the feasibility and limitations of generative models for structuring historical data.

The *Tables des noms* of the French Senate was published during the French Third Republic (1870–1940)[2]. Within the broader documentary ecosystem of the *Journal Officiel* — which seeks to reconstruct parliamentary activity and its legal or regulatory outcomes in France —, the Senate's *Tables nominatives* offer a concise and systematic record of senators'

[1]The "DH@LLM : Grands modèles de langage et humanités numériques" conference program, held in Paris in July 2025, is a good illustration of this : https ://www.crihn.org/nouvelles/2025/01/16/colloque-dhllm-grands-modeles-de-langage-et-humanites-numeriques-sorbonne-universite/

[2]These tables are part of the *Tables annuelles* (yearly activity index), which can be consulted on the digital library of the French national library (*BnF*) : https ://gallica.bnf.fr/ark :/12148/cb371291967/date.item.

interventions during public sessions. These indexes were designed to accompany the transcription of debates[3] and to facilitate their consultation. Manually compiled once a year, they recorded each intervention by senators or members of the government who spoke during the sessions, the subject of their speech, and the corresponding page number. While these tables were particularly useful at a time when full-text search in digitized parliamentary debates was not possible, they still hold significant value for historians today. Systematically extracting data from them would make it possible to track parliamentary activity over the long term, quantify the interventions of specific senators affiliated with particular political movements, or support the cross-validation of named entities extracted from the debates themselves. Our objective is to extract structured data from these *Tables*; for our initial experiments, we focus on a single *Table nominative*, namely that of 1931. The early 1930s marked the beginning of the decline of French parliamentarism, culminating in the fall of the Third Republic in 1940 **morel2024parlement**. Analyzing the 1931 *Table* allows us to lay the groundwork for a broader study that will extend across the entire decade, with the aim of capturing the parliamentary activity of the Senate and, subsequently, of the Chamber of Deputies.

After reviewing existing approaches to structured data extraction and evaluation (Section **??**), we present three main contributions.

[3]The complete transcriptions of Senate debates can be consulted via Gallica : https ://gallica.bnf.fr/ark :/12148/cb34363182v/date.

Quatrième partie

Conclusion

Annexe A

Le titre très long de la première annexe

Table des matières