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 ou le sociologue, les archives font figure centrale. 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'extraire et de fournir, par exemple, à son public de généalogistes. 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) semble pouvoir faire endosser aux ordinateurs ce labeur de traduction et de structuration de l'information et sans

dépendre en amont de données déjà bien formées.

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 dormants" – ou du moins ces documents ingrats qu'il est difficile d'exposer étant donné leur austérité – 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.

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 : de l'image à la base de données

Troisième partie

Expérimenter et évaluer pour comprendre : une démarche historienne outillée

Quatrième partie

Conclusion

Annexe A

Le titre très long de la première annexe

Table des matières