

Table des matières

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

II.	Forme et organisation des tables	7
III.	Informations sémantiques et usages	8
1.	Index des orateurs	9
2.	Table des matières thématiques	9
3.	Références législatives et réglementaires	9

II L'enjeu des données structurées : des sources à la base de données 13

3 Une histoire par les données 15

I.	Numériser les sources	17
1.	En « mode texte »	17
2.	En « mode image »	20
II.	Ce que la numérisation fait aux sources	22
1.	La « datafication »	22
2.	Pratiques historiennes : sphère technique, sphère sociale	24
3.	Documents sériels et approches quantitatives	27

4 La reconnaissance optique de caractères 29

I.	Une vue d'ensemble	29
II.	Les grandes technologies d'OCR	35
III.	CER, WER : métriques pour évaluer la qualité de l'OCR	36
IV.	Numérisation du <i>Journal Officiel</i> : une politique documentaire partagée entre la BnF et les institutions parlementaires	38

5 Données brutes, données structurées 41

I.	Modéliser des données structurées	42
II.	Produire des données structurées à partir de texte	45
1.	Approche à motifs explicites : les ReGex	46
2.	Approches extractives : BERT	49
3.	Approches génératives : les LLMs	51
III.	La sortie structurée via LLM pour le <i>Journal Officiel</i>	52
1.	Simplicité	53
2.	Sortie structurée, génération structurée	53
3.	L'outil « Corpusense » du projet Mezanno : une pipeline de l'image numérisée à la donnée structurée	55

III Expérimenter et évaluer pour comprendre : une démarche historienne outillée	59
6 L’outil Corpusense : une chaîne de traitement pour les sources historiques	61
I. Une instance de pipeline « classique »	62
II. Le travail sur Corpusense	65
III. Ateliers à l’EHESS et à la BnF	66
7 La sortie structurée via LLM appliquée à la Table des Noms du Sénat : une approche empirique	67
I. Expérimentations	68
1. Prise en main intuitive du problème de la génération de données . . .	68
2. Design, prompt et vérité terrain : trouver le bon modèle de données .	70
3. Préparer l’évaluation : comparer, apparier	72
II. Recouper des données pour l’analyse historienne	75
1. Des pages aux dates : utilisation de l’API Gallica et des métadonnées des manifestes	75
2. Datavisualisations	75
8 Livrables	77
I. Jeux de données pour le protocole d’évaluation de la sortie structurée	77
1. Vérité terrain, prompt et schéma : un modèle simple pour l’évaluation	77
2. Prédiction obtenues avec le modèle Mistral 8b avec l’extraction guidée par schéma	79
3. Protocole d’évaluation de la sortie structurée	79
II. Une métrique pour l’évaluation des données générées par la sortie structurée via LLM	80
1. Echantillon	80
2. Mise en correspondance des prédictions et de la vérité terrain avec le transport optimal	81
3. Limites des métriques classiques : précision, rappel et F1	82
4. Integrated Matching Quality (IMQ)	82
5. Résultats et analyse	83
III. Conclusion	84

IV Conclusion	91
A Titre court	93