

THÈSE DE DOCTORAT DE

NANTES UNIVERSITÉ

ÉCOLE DOCTORALE N° 641
*Mathématiques et Sciences et Technologies
de l'Information et de la Communication*
Spécialité : *Informatique*

Par

Thomas SEBBAG

**Analyse des dynamiques d'économie territoriale via l'extraction
automatique d'informations issues des documents administratifs
français**

Thèse présentée et soutenue à Nantes, le 23 septembre 2026

Unité de recherche : UMR6004 – Laboratoire des Sciences et du Numérique de Nantes (LS2N)

Rapporteurs avant soutenance :

Nathalie AUSSENAC-GILLES	Directrice de Recherche, CNRS - IRIT
Vincent LABATUT	Maître de Conférences, Avignon Université - LIA

Composition du Jury :

Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse

Président :	Prénom NOM	Fonction et établissement d'exercice (<i>à préciser après la soutenance</i>)
Examineurs :	Pascale Sebillot	Professeure des Universités, INSA Rennes - IRISA
	Antoine Doucet	Professeur des Universités, La Rochelle Université - L3I
Dir. de thèse :	Emmanuel MORIN	Professeur des Universités, Nantes Université - LS2N
Encadrant :	Solen QUINIOU	Maître de conférences, Nantes Université - LS2N

Invité(s) :

Jeffery PETIT Docteur, Explore

Titre : Analyse des dynamiques d'économie territoriale via l'extraction automatique d'informations issues des documents administratifs français

Mot clés : de 3 à 6 mots clefs

Résumé : La production documentaire des administrations publiques françaises constitue un cas singulier à l'échelle mondiale. La multiplicité des acteurs administratifs et l'autonomie dont chacun dispose dans la rédaction de ses contenus ont engendré des conventions propres à ce genre, parfois marquées par des singularités stylistiques ou structurelles, qui posent un certain nombre de défis pour le traitement automatique du langage (TAL).

Cette thèse porte sur l'extraction d'information dans les documents administratifs français, dans la perspective d'étudier les dynamiques territoriales. Elle met l'accent sur la création de ressources, le développement de méthodes d'extraction et la conception de modèles de langue dédiés à ce domaine. Nos contributions sont les suivantes.

Afin d'enrichir le TAL en ressources nouvelles, nous présentons AdminSET, un large corpus de documents segmentés rassemblant une variété de documents administratifs en français, conçu pour représenter au mieux la diversité de ce domaine. Nous introduisons

également la famille de modèles AdminBERT, entraînés sur ce corpus et fondée sur l'architecture CamemBERT, qui constitue la première famille de modèles de langue adaptés au domaine administratif. En parallèle, nous proposons AdminSet-NER, premier jeu de données annoté pour la reconnaissance d'entités nommées dans des documents administratifs français. Ces travaux démontrent l'intérêt de disposer de modèles dédiés à un domaine encore largement sous-exploité par le TAL. Ils mettent également en évidence l'existence d'un plafond de performances que la seule modélisation sémantique ne suffit pas à dépasser.

Enfin, nous présentons une méthode d'identification de relations inconnues entre entités, accompagnée des catégories de relations identifiées et d'un corpus de textes issus de dix villes françaises. L'ensemble de ces travaux constitue le socle d'un projet de construction de ressources pour l'extraction de relations dans ce type de documents.

Title: Analysis of regional economic dynamics through the automatic extraction of information from French administrative documents

Keywords: de 3 à 6 mots clefs

Abstract: The documentary output of French public administrations represents a singular case on a global scale. The multiplicity of administrative actors and the autonomy each enjoys in producing its own content have given rise to genre-specific conventions, sometimes

marked by stylistic or structural idiosyncrasies, that pose a number of challenges for natural language processing (NLP).

This thesis addresses information extraction in French administrative documents, with the aim of studying territorial dynamics. It

focuses on resource creation, the development of extraction methods, and the design of domain-specific language models. Our contributions are as follows.

To enrich NLP with new resources, we present AdminSET, a large corpus of segmented documents bringing together a wide variety of French administrative texts, designed to best capture the diversity of this domain. We also introduce the AdminBERT model family, trained on this corpus and built upon the CamemBERT architecture, which constitutes the first family of language models adapted to the administrative domain. Alongside these, we propose AdminSET-NER, the

first annotated dataset for named entity recognition in French administrative documents. This body of work demonstrates the value of having models dedicated to a domain that remains largely underexplored in NLP, while also revealing the existence of a performance ceiling that semantic modelling alone is insufficient to overcome.

Finally, we present a method for identifying unknown relations between entities, together with the relation categories identified and a corpus of texts drawn from ten French cities. Taken together, these contributions form the foundation of a resource-building project for relation extraction in this type of document.