

ÉCOLE NATIONALE DES CHARTES

Lucas Terriel

Licencié ès histoire

Diplômé de master Histoire, Civilisations, Patrimoine contemporains

Représenter et évaluer les données issues de la structuration et de la transcription automatique d'un corpus.

**L'exemple de la reconnaissance automatique des écritures
manuscrites sur les répertoires de notaires du projet Lectaurep.**

Mémoire pour le diplôme

« Technologies numériques appliquées à l'histoire »

2020





Ce mémoire professionnel/recherche est placé sous les termes de la licence Creative Commons en ces termes : **Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions 4.0 International (CC BY-NC-SA 4.0)**.

Consulter la [licence](https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode.fr)¹ en entier pour plus de détails.

Vous êtes autorisé à :

- **Attribution** : Vous devez créditer l'œuvre, intégrer un lien vers la licence et indiquer si des modifications ont été effectuées à l'œuvre. Vous devez indiquer ces informations par tous les moyens raisonnables, sans toutefois suggérer que l'Offrant vous soutient ou soutient la façon dont vous avez utilisé son œuvre ;
- **Pas d'utilisation commerciale** : Vous n'êtes pas autorisé à faire un usage commercial de cette œuvre, tout ou partie du matériel la composant.
- **Partager copier, distribuer et communiquer** le matériel par tous moyens et sous tous formats ;
- **Adapter remixer, transformer et créer** à partir du matériel.

L'Offrant ne peut retirer les autorisations concédées par la licence tant que vous appliquez les termes de cette licence.

1. *Licence CC 4.0 International*, en ligne : <<https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode.fr>>.

Résumé

Mots-clefs : TAL ; métadonnées ; données ; format pivot ; XML-TEI ; développement applicatif ; similarité syntaxique ; similarité sémantique ; métriques *text-to-text* ; OCR ; HTR ; Machine learning ; Intelligence artificielle ; répertoires de notaires ; valorisation patrimoniale ; Humanités numériques.

Informations bibliographiques : Lucas Terriel, *Représenter et évaluer les données issues de la structuration et de la transcription automatique d'un corpus. L'exemple de la reconnaissance automatique des écritures manuscrites sur les répertoires de notaires du projet Lectaurep.*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Alix Chagué et Thibault Clérice, École nationale des chartes, 2020.

Remerciements

Pour ce stage, qui s'est déroulé durant la période de confinement, je tiens à remercier l'ensemble des personnes qui m'ont aidé et soutenu dans cette situation particulière et qui m'ont encouragé dans mon travail.

En premier lieu, je remercie mon tuteur pédagogique, M. Thibault Clérice, et ma tutrice professionnel, Mme Alix Chagué, ingénieure recherche et développement, pour leurs appuis et leurs conseils avisés.

Je remercie l'ensemble de l'équipe du projet Lectaurep, et le personnel du Minutier central des notaires de Paris aux Archives nationales, Mme Marie-Françoise Limon-Bonnet, responsable du département, Mme Aurélia Rostaing, responsable du pôle instruments de recherche, M. Gaetano Piraino, responsable à la DMOASI, M. Danis Habib, chargé d'études documentaires, Mme Virginie Grégoire, secrétaire de documentation, M. Benjamin Davy, agent technique d'accueil, surveillance et magasinage, et Mme Anna Chéru, stagiaire phase 3.

Je remercie l'équipe ALMAAnaCH d'Inria, qui ont su créer les conditions favorables d'accueil de mon stage et permis un environnement de travail stimulant ; je remercie en particulier M. Laurent Romary, directeur de recherche, pour nos échanges et ses conseils pertinents et Mme Florianne Chiffolleau, ingénieure recherche et développement, pour son aide.

Comme le stage est un travail d'équipe avant tout, je remercie Jean-Damien Généro, collègue de master TNAH à l'École nationale des chartes, qui était stagiaire durant la même période à ALMAAnaCH sur le projet Time us, dont la mise en commun de nos efforts et nos échanges sur les scripts ont permis le bon déroulement technique et pratique de nos stages respectifs.

Enfin je remercie ma famille et mes amis, pour leurs indéfectible soutien tout au long de cette période.

Liste des sigles et abréviations

- A.N. : Archives Nationales
- DMC : Département du Minutier central des notaires de Paris
- DMOASI : département de la maîtrise d'ouvrage du système d'information (direction de l'appui scientifique)

★

- ALMAAnaCH : *Automatic Language Modelling and Analysis & Computational Humanities*
- ANR : Agence Nationale de la Recherche
- EPI : Équipe-Projet Inria
- INRIA : Institut National de Recherche en Informatique et Automatique
- LECTAUREP : Lecture automatique des répertoires
- W3C : World Wide Web Consortium

★

- EN : Entités Nommées
- ML : *Machine Learning*
- REN : Reconnaissance d'Entités Nommées
- RNN : *Recurrent neural network* - Réseau de neurones récurrents
- TAL : Traitement Automatique des Langues

★

- CSS : *Cascading Style Sheets*
- CSV : *Comma-separated values*
- DTD : *Document Type Definition*
- HTML : *HyperText Markup Language*
- HTR : *Handwritten Text Recognition*

- OCR : *Optical Character Recognition*
- ODD : *One Document Does it all*
- PDF : *Portable Document Format*
- RELAXNG : *Regular Language for XML Next Generation*
- TEI : *Text Encoding Initiative*
- XML : *eXtensible Markup Language*
- XSLT : *eXtensible Stylesheet Language Transformations*

Introduction

Première partie

Le projet Lectaurep : un cas d'application de l' « intelligence artificielle » aux documents historiques

Chapitre 1

État de l'art du *machine learning*,
de la reconnaissance automatique
des écritures manuscrites et de
l'implication de ces techniques dans
le cadre de projets de traitement
automatique du langage (TAL)

1.1 Principes élémentaires du *machine learning*

1.1.1 a voir ?

1.2 La reconnaissance automatique des écritures manuscrites : un domaine entre le traitement automatique du langage (TAL) et le *machine learning*

Chapitre 2

Lectaurep, un projet de recherche et développement en reconnaissance automatique des écritures manuscrites

2.1 Des origines à la phase 3

2.2 Une dimension expérimentale

Deuxième partie

Représenter, enrichir et
homogénéiser dans un format pivot
les données et métadonnées au sein
de la chaîne de traitement Lectaurep

Objectifs de la mission ?

Chapitre 3

Enjeux et problématiques liés aux données et aux formats dans le projet

3.1 De l'importance des données et métadonnées

3.2 Les répertoires de notaires ne sont pas que des images numérisées !

- permettre IIIF

Chapitre 4

Le choix du XML TEI (*Text Encoding Initiative*) comme format pivot

- 4.1 Un choix réaliste ? le format XML TEI dans d'autres projets et apports pour LECTAUREP
- 4.2 Esquisse d'un format pivot et premières spécifications TEI grâce l'ODD

Chapitre 5

Simuler la récupération, l'export et la validation d'un format pivot XML TEI

5.1 Objectifs et buts de la simulation

- ne parle forcément aux archivistes ; - besoin de visualiser les données dans un canevas TEI et d'envisager ;

5.2 Un CLI en Python pour générer un format pivot XML TEI et valider grâce à un schéma RELAX NG

Troisième partie

Évaluer et contrôler la transcription
sur des sets d'images comparés :
proposer un vue synthétique des
performances d'un modèle de
transcription

Objectifs de la mission ?

Chapitre 6

État de l'art pour l'évaluation des modèles de transcription entraînés avec le système OCR Kraken

- 6.1 Banc d'essai des outils existants : limites et avantages
- 6.2 Les métriques pour évaluer la transcription en question : définitions et recherche

Chapitre 7

Le développement d'une application : Kraken-Benchmark

7.1 Modélisation

7.2 Conception

7.3 Perspectives d'amélioration pour l'application

Chapitre 8

Tests de Kraken-Benchmark sur les images de répertoires de notaires

- 8.1 Préparation du corpus et mise en place des tests
- 8.2 Déroulement et résultats des tests
- 8.3 Un bilan mitigé ? des propositions pour améliorer les scores

Conclusion

gergr

Table des figures

Table des matières

Résumé	v
Remerciements	vii
Liste des sigles et abréviations	ix
Introduction	3
I Le projet Lectaurep : un cas d’application de l’ « intelligence artificielle » aux documents historiques	3
1 État de l’art du <i>machine learning</i> , de la reconnaissance automatique des écritures manuscrites et de l’implication de ces techniques dans le cadre de projets de traitement automatique du langage (TAL)	5
1.1 Principes élémentaires du <i>machine learning</i>	5
1.1.1 à voir ?	5
1.2 La reconnaissance automatique des écritures manuscrites : un domaine entre le traitement automatique du langage (TAL) et le <i>machine learning</i> .	5
2 Lectaurep, un projet de recherche et développement en reconnaissance automatique des écritures manuscrites	7
2.1 Des origines à la phase 3	7
2.2 Une dimension expérimentale	7
II Représenter, enrichir et homogénéiser dans un format pivot les données et métadonnées au sein de la chaîne de traitement Lectaurep	9
3 Enjeux et problématiques liés aux données et aux formats dans le projet	13
3.1 De l’importance des données et métadonnées	13

3.2	Les répertoires de notaires ne sont pas que des images numérisées!	13
4	Le choix du XML TEI (<i>Text Encoding Initiative</i>) comme format pivot	15
4.1	Un choix réaliste? le format XML TEI dans d'autres projets et apports pour Lectaurep	15
4.2	Esquisse d'un format pivot et premières spécifications TEI grâce l'ODD . .	15
5	Simuler la récupération, l'export et la validation d'un format pivot XML TEI	17
5.1	Objectifs et buts de la simulation	17
5.2	Un CLI en Python pour générer un format pivot XML TEI et valider grâce à un schéma RELAX NG	17
III	Évaluer et contrôler la transcription sur des sets d'images comparés : proposer un vue synthétique des performances d'un modèle de transcription	19
6	État de l'art pour l'évaluation des modèles de transcription entraînés avec le système OCR Kraken	23
6.1	Banc d'essai des outils existants : limites et avantages	23
6.2	Les métriques pour évaluer la transcription en question : définitions et recherche	23
7	Le développement d'une application : Kraken-Benchmark	25
7.1	Modélisation	25
7.2	Conception	25
7.3	Perspectives d'amélioration pour l'application	25
8	Tests de Kraken-Benchmark sur les images de répertoires de notaires	27
8.1	Préparation du corpus et mise en place des tests	27
8.2	Déroulement et résultats des tests	27
8.3	Un bilan mitigé? des propositions pour améliorer les scores	27
	Conclusion	31
	Table des figures	33
	Table des matières	35