ÉCOLE NATIONALE DES CHARTES
UNIVERSITÉ PARIS, SCIENCES & LETTRES

Virgile Reignier
licencié.e ès histoire
diplômé.e de master mondes médiévaux

Vers l'indexation automatique du Trésor des chartes

Constitution, alignement et utilisation de référentiels d'entités nommées au sein du projet Himanis

Mémoire pour le diplôme de master
« Technologies numériques appliquées à l'histoire »

2022

Résumé

Blablabla résumé du mémoire.

Mots-clés : TNAH, IRHT, Himanis, Trèsor des Chartes, Archives Nationales, OCR, Numérisation d'instruments de recherche, Alignement de référentiels, HTR, REN, Machine learning, Intelligence artificielle, Reconnaissance d'écriture manuscrite, identity linking.

Informations bibliographiques : Reignier Virgile, *Vers l'indexation automatique du Trésor des chartes. Constitution, alignement et utilisation de référentiels d'entités nommées au sein du projet Himanis.*, mémoire de master « Technologies numériques appliquées à l'histoire », dir. Dominique Stutzmann et Thibault Clérice, École nationale des chartes, 2022.

Remerciements

Blablabla remerciements...
Faire une liste des abréviations avec : REN, TAL, HTR, IRHT

Bibliographie

BVMM, URL : https://bvmm.irht.cnrs.fr/ (visité le 16/07/2022).

Ehrmann (Maud), *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*, These de doctorat, Paris 7, 2008, URL : https://hal.archives-ouvertes.fr/tel-01639190/document (visité le 14/04/2022).

Himanis - Chancery Indexing and Search, URL : http://himanis.huma-num.fr/app// (visité le 16/07/2022).

Holtz (Louis), "Les premières années de l'Institut de recherche et d'histoire des textes", *La revue pour l'histoire du CNRS*–2 (5 mai 2000), ISBN : 9782271057082 Number : 2 Publisher : CNRS Éditions, DOI : 10.4000/histoire-cnrs.2742.

Potin (Yann), *La mise en archives du trésor des chartes (XIIIe-XIXe siècle)*, Positions de thèse pour le diplôme d'archiviste-paléographe, Paris, Ecole nationale des chartes, 2007, URL : http://theses.enc.sorbonne.fr/2007/potin (visité le 16/07/2022).

Stutzmann (Dominique), Moufflet (Jean-François) et Hamel (Sébastien), "La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique", *Médiévales. Langues, Textes, Histoire*, 73–73 (15 déc. 2017), p. 67-96, DOI : 10.4000/medievales.8198.

Torres Aguilar (Sergio) et Stutzmann (Dominique), "Named Entity Recognition for French medieval charters", dans *Workshop on Natural Language Processing for Digital Humanities*, Helsinki, Finland, 2021 (Workshop on Natural Language Processing for Digital Humanities Proceedings of the Workshop), URL : https://hal.archives-ouvertes.fr/hal-03503055 (visité le 17/07/2022).

Introduction

> A ce sujet papa avait une plaisanterie. (...) Il disait, quand il présentait maman, « je l'ai connue et épousée à Paris » et (...) il attendait avant de dire « Texas » que tout le monde ait cru, que tout le monde ait pensé qu'il parlait de Paris, France. Ça faisait tordre de rire toutes les fois.

Si le mot "Paris" évoque en premier lieu la capitale française, il désigne également d'autres villes à travers le monde. C'est en exploitant l'homonymie entre cette première et une ville du Texas que la citation ci-dessus, extraite du film "Paris, Texas" (1984) de Wim Wenders, construit la plaisanterie. L'information "Paris" ne suffit en effet pas à identifier le lieu où lesdits parents se sont rencontrés. Utilisé seul, le mot renvoie naturellement vers la ville située en France. C'est seulement en précisant l'État dans lequel elle se situe que l'on peut identifier le lieu exact où les protagonistes se sont rencontrés et mariés. Ce jeu d'ambiguïté manifeste ainsi d'une difficulté rencontrée dans le langage naturel : l'identification des références utilisées. La connaissance lexicale ne suffit en effet pas à elle seule pour comprendre un discours, il faut également que les références renvoient vers une réalité qui soit clairement identifiée.

Cet enjeu est également présent au sein du Traitement Automatique des Langues à travers la notion d'Entité Nommée qui désigne une expression linguistique qui se réfère à une entité unique de façon autonome [1]. L'analyse du contenu textuel a ainsi largement progressé ces dernières années autour de cette notion par le développement de deux techniques : la REN (Reconnaissance d'Entités Nommées) qui consiste à repérer ces objets textuels et à leur attribuer une catégorie et le liage d'entités qui permet d'associer ces objets textuels à un élément décrit par une ressource référentielle. Si un grand nombre de ces travaux concernent des corpus contemporains, quelques chercheurs s'intéressent également à leur application pour la lecture des archives anciennes et rencontrent ainsi les recherches menées par les spécialistes de ces corpus.

1. Sur la définition des entités nommées, *cf.* Maud Ehrmann, *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*, These de doctorat, Paris 7, 2008, URL : `https://hal.archives-ouvertes.fr/tel-01639190/document` (visité le 14/04/2022), p. 167-170.

Contexte scientifique de travail

L'Institut de Recherche et d'Histoire des Textes (IRHT) est un laboratoire de recherche fondé en 1937 par Félix Grat et rattaché au CNRS dans le but de faciliter l'accès des chercheurs aux manuscrits et imprimés anciens [2]. Les recherches qui y sont menées portent également sur la transmission des textes et l'étude des écriture et connaissent à ce titre des développements récents à propos de la lecture automatique des documents anciens. Initiés par sa collaboration au sein du projet GRAPHEM, les travaux en "paléographie artificielle" développés par la section de paléographie latine sont menés conjointement avec des chercheurs en informatique spécialisés dans l'analyse de l'image. Les projets développés prennent deux directions principales : la caractérisations des écritures médiévales (Oriflamms, ECMEN, CrEMe) d'une part et la lecture automatique des archives (Himanis, HOME, HORAE) d'autre part. Pilotés par Dominique Stutzmann, ces recherches ont permis le développement d'outils informatiques et de modèles d'intelligence artificielle qui ont largement renouvelé l'accès aux textes anciens.

Parmi les corpus étudiés par ces travaux, le Trésor des Chartes occupe une place centrale puisqu'il constitue le matériel source du projet Himanis et participe à celui du projet HOME. Conservé au sein de la série JJ des Archives Nationales, ce fonds se compose d'une immense collection de titres rassemblée par les rois de France. Il se présente sous la forme de registres contenant des actes organisés de manière plus ou moins systématique et linéaire [3]. Le projet Himanis (HIstorical MANuscript Indexing for user-controlled Search) a ainsi permis de numériser les registres et de convertir les inventaires et éditions disponibles afin de les structurer en un format homogène et unique [4]. Ces éléments ont ensuite servi de base au développement d'un modèle d'indexation automatique des mots présents dans le corpus [5]. Par la suite, le projet HOME (History of Medieval Europe) s'est proposé d'amplifier et généraliser ce travail en numérisant de nouveaux documents, en associant chaque texte aux données disponibles les concernant et en déposant les résultats dans une plateforme librement accessible [6].

2. Sur la fondation de l'IRHT, *cf.* Louis Holtz, "Les premières années de l'Institut de recherche et d'histoire des textes", *La revue pour l'histoire du CNRS*–2 (5 mai 2000), ISBN : 9782271057082 Number : 2 Publisher : CNRS Éditions, DOI : 10.4000/histoire-cnrs.2742.

3. Sur la constitution du trésor des chartes, *cf.* Yann Potin, *La mise en archives du trésor des chartes (XIIIe-XIXe siècle)*, Positions de thèse pour le diplôme d'archiviste-paléographe, Paris, Ecole nationale des chartes, 2007, URL : http://theses.enc.sorbonne.fr/2007/potin (visité le 16/07/2022)

4. Les registres numérisés ont été intégrés à la Bibliothèque Virtuelle des Manuscrits Médiévaux *BVMM*, URL : https://bvmm.irht.cnrs.fr/ (visité le 16/07/2022). Tous les fichiers issus de ces travaux sont disponibles ici : https://github.com/oriflamms/himanis.

5. Dominique Stutzmann, Jean-François Moufflet et Sébastien Hamel, "La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspectives du projet HIMANIS pour l'édition électronique", *Médiévales. Langues, Textes, Histoire*, 73–73 (15 déc. 2017), p. 67-96, DOI : 10.4000/medievales.8198. Les résultats sont disponibles dans l'interface *Himanis - Chancery Indexing and Search*, URL : http://himanis.huma-num.fr/app// (visité le 16/07/2022).

6. https://github.com/oriflamms/Home

Problématique du stage

Ces différents travaux ont ainsi permis de diffuser largement les textes qui composent le Trésor des chartes et de progresser dans l'analyse automatique des écritures qu'ils contiennent. Il reste néanmoins une problématique à approfondir : l'identification des références utilisées au sein des documents. Si les travaux réalisés permettent de faciliter la lecture des textes, celle-ci se trouve encore freiné par la difficile compréhension des références utilisées. Après des travaux récents portant sur la REN dans les chartes médiévales [7], l'objectif poursuivi est de parvenir à développer un modèle de liage d'entités afin d'enrichir et désambiguïser les entités nommées reconnues dans les textes.

C'est dans ce contexte que mon stage, effectué dans le cadre du Master 2 Archives - Technologies Numériques Appliquées à l'Histoire de l'Ecole Nationale des Chartes, s'est donné pour objectif de rassembler les données disponibles au sein du corpus Himanis pour avancer sur la problématique de l'identification des entités nommées. A partir des inventaires déjà convertis, des registres numérisés et des travaux préliminaires en HTR et REN, nous avons ainsi travaillé sur la construction d'un référentiel et d'une méthode de travail pour lier les entités nommées reconnues en limitant au maximum les ambiguïtés possibles. Le présent mémoire se propose donc de décrire les travaux effectués et la manière dont ils s'insèrent dans un contexte de travail. Quels sont les apports des données fournies par le projets Himanis et HOME pour apprendre à désambiguïser automatiquement les entités nommées reconnues dans un texte médiéval ? Nous aborderons les différentes étapes de construction du référentiel ainsi que les difficultés rencontrées dans ce cadre et dans son utilisation.

Dans cet objectif, nous exposerons dans une première partie le matériel disponible pour mettre en œuvre ce projet. Nous proposerons ainsi un état des lieux sur les recherches en cours à propos du liage d'entités, puis nous décrirons plus précisément les avancées permises par le projet Himanis dans l'accès au corpus du Trésor des Chartes, enfin nous analyserons l'apport des instruments de recherches convertis sous format numérique. Notre deuxième partie sera consacrée à la formalisation du référentiel. Nous développerons pour cela les différents enjeux liés à l'utilisation d'un instrument papier, puis nous proposerons une analyse du lien entre les entités décrites, enfin nous décrirons l'insertion des éléments dans une base de données relationnelle. Notre troisième et dernière partie se portera sur les différents traitements mis en œuvre afin de compléter et diffuser ce référentiel. Nous décrirons ainsi l'enrichissement des données à partir de référentiels externes, puis la mise à disposition du référentiel et enfin les premiers pas de son utilisation.

7. Sergio Torres Aguilar et Dominique Stutzmann, "Named Entity Recognition for French medieval charters", dans *Workshop on Natural Language Processing for Digital Humanities*, Helsinki, Finland, 2021 (Workshop on Natural Language Processing for Digital Humanities Proceedings of the Workshop), URL : `https://hal.archives-ouvertes.fr/hal-03503055` (visité le 17/07/2022).

Première partie

De la legacy data à l'identity linking : quel matériel disponible pour entraîner un modèle ?

Chapitre 1

Etat des lieux de la recherche sur le liage d'entités

1.1 Un enjeu actuel

Pour le premier chapitre : thèse Stern + https ://hal.archives-ouvertes.fr/hal-03034492

Faire le point sur la distinction liage - désambiguisation (je trouve que c'est pas si clair dans mon intro)

Bilan journée Sorbonne

1.2 Les outils disponibles

1.3 Des résultats ?

Chapitre 2

Un autre chapitre

2.1 Structuration du mémoire

Le mémoire se structure en plusieurs parties :

1. tout d'abord, les pièces liminaires : page de titre, résumé, remerciements, bibliographie, introduction
2. ensuite, le corps du texte, suivi d'une conclusion
3. après, les annexes (documentation, extraits de code, etc.)
4. enfin, les pièces finales : index (si besoin), glossaire (si besoin) ; tables (des figures et des tables, si nécessaire) ; table des matières

Ce mémoire s'accompagne d'une autre partie très importante : les **données**.

2.2 Les données

Elles constituent une partie primordiale du travail à rendre. Ce sont :

— les données traitées (textes, images, vidéos, BDD, etc.)

— les scripts de traitement

— la documentation associée aux données et au script

— tout autre document qui semble nécessaire au traitement du sujet.

Ces données doivent être ordonnées et accompagnées d'un fichier `lisezMoi` (format `.txt` ou `.md`), présent à la racine du dossier contenant les données. Ce fichier doit décrire l'arborescence des fichiers et dossiers et la fonction de chacun des fichiers.

Le principe important à retenir est celui de la **reproductibilité** du travail.

Deuxième partie

Une autre partie

Conclusion

Annexe A

Première annexe

Table des matières