

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, ROC, Numérisation d'instruments de recherche, Alignement de référentiels, REM, REN, Machine learning, Intelligence artificielle, Reconnaissance d'écriture manuscrite, identity linking.

Informations bibliographiques : Reignier Virgile, *Vers l'indexation automatique du Trèsor 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

B Lablabla remerciements...

Faire une liste des abréviations avec : REN, TAL, REM, IRHT

Bibliographie

- ABADIE (Nathalie), ESCOBAR (Carmen Brando) et FRONTINI (Francesca), “Evaluation de la qualité des sources du Web de Données pour la résolution d’entités nommées”, *Revue des Sciences et Technologies de l’Information - Série ISI : Ingénierie des Systèmes d’Information*, 21–5 (8 févr. 2017), URL : <https://iieta.org/download/file/fid/27476> (visité le 27/07/2022).
- AGIRRE (Eneko), BARRENA (Ander), LACALLE (Oier Lopez de), SOROA (Aitor), FERNANDO (Samuel) et STEVENSON (Mark), “Matching Cultural Heritage items to Wikipedia”, dans *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey, 2012, p. 1729-1735, URL : http://www.lrec-conf.org/proceedings/lrec2012/pdf/1021_Paper.pdf (visité le 21/07/2022).
- BLUCHE (Théodore), STUTZMANN (Dominique) et KERMORVANT (Christopher), “Automatic Handwritten Character Segmentation for Paleographical Character Shape Analysis”, dans *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, Santorini, France, 2016 (2016 12th IAPR Workshop on Document Analysis Systems (DAS)), p. 42-47, DOI : 10.1109/DAS.2016.74.
- BOEGLIN (Noémie), DEPEYRE (Michel), JOLIVEAU (Thierry) et LE LAY (Yves-François), “Pour une cartographie romanesque de Paris au XIXe siècle. Proposition méthodologique”, dans *Conférence Spatial Analysis and GEOMatics*, Nice, France, 2016 (Actes de la conférence SAGEO’2016 - Spatial Analysis and GEOMatics), URL : <https://hal.archives-ouvertes.fr/hal-01619600> (visité le 23/07/2022).
- BOROŞ (Emanuela), ROMERO (Verónica), MAARAND (Martin), ZENKLOVÁ (Kateřina), KŘEČKOVÁ (Jitka), VIDAL (Enrique), STUTZMANN (Dominique) et KERMORVANT (Christopher), “A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters”, dans *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2020, p. 79-84, DOI : 10.1109/ICFHR2020.2020.00025.
- BRANDO (Carmen), FRONTINI (Francesca) et GANASCIA (Jean-Gabriel), “Disambiguation of Named Entities in Cultural Heritage Texts Using Linked Data Sets”, dans *New Trends in Databases and Information Systems*, dir. Tadeusz Morzy, Patrick Valduriez et Ladjel Bellatreche, Series Title : Communications in Computer and

- Information Science, Cham, 2015, t. 539, p. 505-514, DOI : 10.1007/978-3-319-23201-0_51.
- BRANDO (Carmen), FRONTINI (Francesca) et GANASCIA (Jean-Gabriel), “REDEN : Named Entity Linking in Digital Literary Editions Using Linked Data Sets”, *Complex Systems Informatics and Modeling Quarterly*-7 (29 juill. 2016), p. 60, DOI : 10.7250/csimq.2016-7.04.
- BVMM, URL : <https://bvmm.irht.cnrs.fr/> (visité le 16/07/2022).
- CLAVAUD (Florence), ROMARY (Laurent), CHARBONNIER (Pauline), TERRIEL (Lucas), PIRAINO (Gaetano) et VERDESE (Vincent), “NER4Archives (named entity recognition for archives) : Conception et réalisation d’un outil de détection, de classification et de résolution des entités nommées dans les instruments de recherche archivistiques encodés en XML/EAD” (), p. 23, URL : <https://hal.archives-ouvertes.fr/hal-03625734/document>.
- 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).
- FRONTINI (Francesca), BRANDO (Carmen) et GANASCIA (Jean-Gabriel), “Domain-adapted named-entity linker using Linked Data”, dans 2015, URL : <https://hal.archives-ouvertes.fr/hal-01203356> (visité le 27/07/2022).
- FRONTINI (Francesca), BRANDO (Carmen), RIGUET (Marine), JACQUOT (Clémence) et JOLIVET (Vincent), “Annotation of Toponyms in TEI Digital Literary Editions and Linking to the Web of Data”, *MALTIT : Materialities of literature*-2 (juill. 2016), DOI : 10.14195/2182-8830_4-2_3.
- GLÉNISSON (Jean), GUEROUT (Jean), VIARD (Jules), VALLÉE-KARCHER (Aline) et JASSEMIN (Henri-Frédéric), *Registres du trésor des chartes : inventaire analytique*, dir. Robert Fawtier, avec la coll. d’Archives nationales, 6 t., Paris, France, 1958.
- GUÉRIN (Paul), *Actes Royaux du Poitou (1302-1464)*, avec la coll. de Léonce Celier, Frédéric Glorieux et Vincent Jolivet, 1881, URL : <http://corpus.enc.sorbonne.fr/actesroyauxdupoitou/> (visité le 04/08/2022).
- HEINO (Erkki), TAMPER (Minna), MÄKELÄ (Eetu), LESKINEN (Petri), IKKALA (Esko), TUOMINEN (Jouni), KOHO (Mikko) et HYVÖNEN (Eero), “Named Entity Linking in a Complex Domain : Case Second World War History”, dans *Language, Data, and Knowledge*, dir. Jorge Gracia, *et al.*, Cham, 2017 (Lecture Notes in Computer Science), p. 120-133, DOI : 10.1007/978-3-319-59888-8_10.
- 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.

- HOOLAND (Seth van), DE WILDE (Max), VERBORGH (Ruben), STEINER (Thomas) et WALLE (Rik Van de), “Exploring entity recognition and disambiguation for cultural heritage collections”, *Digital Scholarship in the Humanities*, 30–2 (1^{er} juin 2015), p. 262-279, DOI : 10.1093/llc/fqt067.
- HOSSEINI (Kasra), NANNI (Federico) et COLL ARDANUY (Mariona), “DeezyMatch : A Flexible Deep Learning Approach to Fuzzy String Matching”, dans 2020, DOI : 10.18653/v1/2020.emnlp-demos.9.
- HUET (Thomas), BIEGA (Joanna) et SUCHANEK (Fabian M.), “Mining history with Le Monde”, dans *Proceedings of the 2013 workshop on Automated knowledge base construction*, New York, NY, USA, 2013 (AKBC ’13), p. 49-54, DOI : 10.1145/2509558.2509567.
- LINHARES PONTES (Elvys), MORENO (Jose G.) et DOUCET (Antoine), “Linking Named Entities across Languages using Multilingual Word Embeddings”, dans *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, New York, NY, USA, 2020, p. 329-332, URL : <https://doi.org/10.1145/3383583.3398597> (visité le 23/07/2022).
- Location*, URL : <https://obtic.sorbonne-universite.fr/tanagra/map> (visité le 01/08/2022).
- MENDES (Pablo N.), JAKOB (Max), GARCÍA-SILVA (Andrés) et BIZER (Christian), “DBpedia spotlight : shedding light on the web of documents”, dans *Proceedings of the 7th International Conference on Semantic Systems*, New York, NY, USA, 2011 (I-Semantics ’11), p. 1-8, DOI : 10.1145/2063518.2063519.
- MONROC (Claire Bizon), MIRET (Blanche), BONHOMME (Marie-Laurence) et KERMORVANT (Christopher), “A Comprehensive Study of Open-Source Libraries for Named Entity Recognition on Handwritten Historical Documents”, dans *Document Analysis Systems*, dir. Seiichi Uchida, Elisa Barney et Véronique Eglin, Cham, 2022 (Lecture Notes in Computer Science), p. 429-444, DOI : 10.1007/978-3-031-06555-2_29.
- MUNNELLY (Gary) et LAWLESS (Seamus), “Investigating Entity Linking in Early English Legal Documents”, dans *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, New York, NY, USA, 2018 (JCDL ’18), p. 59-68, DOI : 10.1145/3197026.3197055.
- PONTES (Elvys Linhares), CABRERA-DIEGO (Luis Adrián), MORENO (José G.), BOROS (Emanuela), HAMDY (Ahmed), SIDÈRE (Nicolas), COUSTATY (Mickaël) et DOUCET (Antoine), “Entity Linking for Historical Documents : Challenges and Solutions”, dans *22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020*, 2020 (Lecture Notes in Computer Science), t. 12504, p. 215-231, DOI : 10.1007/978-3-030-64452-9_19.

- 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).
- RIJHWANI (Shruti), XIE (Jiateng), NEUBIG (Graham) et CARBONELL (Jaime), “Zero-Shot Neural Transfer for Cross-Lingual Entity Linking”, *Proceedings of the AAAI Conference on Artificial Intelligence*, 33–1 (17 juill. 2019), Number : 01, p. 6924-6931, DOI : 10.1609/aaai.v33i01.33016924.
- RUIZ (Pablo) et POIBEAU (Thierry), “Mapping the Bentham Corpus : Concept-based Navigation”, *Journal of Data Mining and Digital Humanities*, Atelier Digit_Hum (mars 2019), Publisher : Episciences.org, DOI : 10.46298/jdmdh.5044.
- SANTOS (Rui), MURRIETA-FLORES (Patricia), CALADO (Pável) et MARTINS (Bruno), “Toponym matching through deep neural networks”, *International Journal of Geographical Information Science*, 32–2 (1^{er} févr. 2018), Publisher : Taylor & Francis _eprint : <https://doi.org/10.1080/13658816.2017.1390119>, p. 324-348, DOI : 10.1080/13658816.2017.1390119.
- SCHEITHAUER (Hugo), *La reconnaissance d’entités nommées appliquées à des données issues de la transcription automatique de documents manuscrits patrimoniaux. Expérimentations et préconisations à partir du projet LECTAUREP*, Mémoire de master “Technologies numériques appliquées à l’histoire”, Ecole nationale des chartes, 2021, URL : https://raw.githubusercontent.com/HugoSchtr/memoire_TNAH_M2_HugoScheithauer/main/memoire_Hugo_Scheithauer_TNAH.pdf (visité le 29/05/2022).
- SMITH (David A.) et CRANE (Gregory), “Disambiguating Geographic Names in a Historical Digital Library”, dans *Research and Advanced Technology for Digital Libraries*, dir. Panos Constantopoulos et Ingeborg T. Sølvsberg, réd. par Gerhard Goos, Juris Hartmanis et Jan van Leeuwen, Series Title : Lecture Notes in Computer Science, Berlin, Heidelberg, 2001, t. 2163, p. 127-136, DOI : 10.1007/3-540-44796-2_12.
- SOUDANI (Aïcha), MEHERZI (Yosra), BOUHAFS (Asma), FRONTINI (Francesca), BRANDO (Carmen), DUPONT (Yoann) et MÉLANIE-BECQUET (Frédérique), “Adaptation et évaluation de systèmes de reconnaissance et de résolution des entités nommées pour le cas de textes littéraires français du 19^{ème} siècle”, dans *Atelier Humanités Numériques Spatialisées (HumaNS’2018)*, Montpellier, France, 2018, URL : <https://hal.archives-ouvertes.fr/hal-01925816> (visité le 21/07/2022).
- STERN (Rosa), *Identification automatique d’entités pour l’enrichissement de contenus textuels*, Thèse de doctorat, Université Paris-Diderot - Paris VII, 2013, URL : <https://tel.archives-ouvertes.fr/tel-00939420> (visité le 28/03/2022).
- STUTZMANN (Dominique), MOUFFLET (Jean-François) et HAMEL (Sébastien), “La recherche en plein texte dans les sources manuscrites médiévales : enjeux et perspec-

- tives du projet HIMANIS pour l'édition électronique", *Médiévales. Langues, Textes, Histoire*, 73–73 (15 déc. 2017), p. 67-96, DOI : 10.4000/medievales.8198.
- SUÁREZ (Pedro Javier Ortiz), DUPONT (Yoann), MULLER (Benjamin), ROMARY (Laurent) et SAGOT (Benoît), "Establishing a New State-of-the-Art for French Named Entity Recognition", dans 2020, URL : <https://hal.inria.fr/hal-02617950> (visité le 21/07/2022).
- Teklia - Arkindex*, URL : <https://teklia.com/solutions/arkindex/> (visité le 05/08/2022).
- 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).
- USBECK (Ricardo), NGONGA NGOMO (Axel-Cyrille), RÖDER (Michael), GERBER (Daniel), COELHO (Sandro Athaide), AUER (Sören) et BOTH (Andreas), "AGDISTIS - Graph-Based Disambiguation of Named Entities Using Linked Data", dans *The Semantic Web – ISWC 2014*, dir. Peter Mika, *et al.*, Series Title : Lecture Notes in Computer Science, Cham, 2014, t. 8796, p. 457-471, DOI : 10.1007/978-3-319-11964-9_29.
- ZHOU (Shuyan), RIJHWANI (Shruti) et NEUBIG (Graham), "Towards Zero-resource Cross-lingual Entity Linking", dans *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, Hong Kong, China, 2019, p. 243-252, DOI : 10.18653/v1/D19-6127.

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 est naturellement associé à la 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 soient comprises et associées à une réalité clairement identifiée.

Cet enjeu est également présent au sein du TAL (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 autonome¹. 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*, Thèse 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 anciens². Les recherches qui y sont menées portent également sur la transmission des textes et l’étude des écritures 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 caractérisation 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 linéaire³. 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 unique⁴. Ces éléments ont ensuite servi de base au développement d’un modèle d’indexation automatique des mots présents dans le corpus⁵. Par la suite, le projet HOME (History of Medieval Europe) s’est proposé d’amplifier et de 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 accessible⁶.

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, cette dernière se trouve encore freinée par la difficile compréhension des références utilisées. Après des travaux récents portant sur la REN dans les chartes médiévales⁷, l’objectif poursuivi est de parvenir à développer un modèle de liage d’entités afin d’enrichir et de 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 mission de rassembler les éléments 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 REM (Reconnaissance d’Écritures Manuscrites) 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 les 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* au liage d'entités : quel matériel disponible pour entraîner un modèle ?

Avant d'aborder plus précisément les actions menées au cours de ce stage, il convient d'exposer dans cette première partie les différents éléments contextuels dans lequel il s'inscrit. Nous consacrerons donc un premier chapitre à la description des enjeux scientifiques actuels autour de la problématique du liage d'entités afin de mieux appréhender les perspectives d'évolution. Un second chapitre permettra de résumer les différents résultats offerts par le projet Himanis et leur utilisation possible dans le cadre du stage. Enfin, le troisième chapitre sera consacré à l'utilisation des instruments de recherches papier pour construire un référentiel numérique.

Chapitre 1

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

Initiée par les *Message Understanding Conferences* qui se réunissent entre 1987 et 1998, la REN est directement associée aux techniques d'extractions d'informations. L'objectif est en effet d'automatiser la lecture des textes afin d'en comprendre au mieux la substance. Reconnaître et classer les références utilisées prend donc dans ce contexte une place centrale qui se perpétue par la suite dans de nombreuses recherches¹. Dans un objectif similaire, d'autres travaux portant sur l'annotation sémantique des textes, c'est à dire l'enrichissement des contenus textuels à partir de métadonnées, ont mis en valeur la nécessité de construire un lien entre les entités nommées reconnues dans le texte et un référentiel à disposition dans ce but².

C'est dans ce contexte qu'est née le principe du liage d'entités. Il se définit comme une technique permettant d'associer chaque élément reconnu comme devant être expliqué à un nœud d'une base de connaissances permettant la génération de ladite explication. La conception de cette technique procède donc de deux éléments : la construction d'une base de connaissances utilisée comme référence et la reconnaissance des entités à mettre en lien avec cette base. Son enjeu principal est de permettre la résolution des ambiguïtés qui peuvent exister entre les entités, soit parce qu'un même mot peut renvoyer vers plusieurs entrées (polysémie), soit au contraire parce qu'une même entité peut s'exprimer de plusieurs façons différentes (synonymie)³.

Nous tenterons donc dans ce chapitre d'exposer succinctement l'état de l'art autour des problématiques associées au liage d'entités. Pour cela, nous décrirons dans un

1. Maud Ehrmann, *Les entités nommées, de la linguistique au TAL...*, p. 17–19.

2. Sur les enjeux de l'Annotation Sémantique, cf. Rosa Stern, *Identification automatique d'entités pour l'enrichissement de contenus textuels*, Thèse de doctorat, Université Paris-Diderot - Paris VII, 2013, URL : <https://tel.archives-ouvertes.fr/tel-00939420> (visité le 28/03/2022), p. 15–16. Sur sa mise en œuvre, *Ibid.*, p. 96–99.

3. *Ibid.*, p. 110–114.

premier temps son fonctionnement général puis les problématiques de son application aux sources historiques. Nous proposerons ensuite une analyse des propositions abordées dans différents travaux. Enfin, nous décrirons les résultats obtenus par ces travaux et les perspectives d'application pour l'étude des textes historiques.

1.1 Mise en œuvre du liage d'entités

1.1.1 Méthodologie

Une méthode utilisée naturellement pour résoudre les ambiguïtés est de considérer que ces entités se rapportent *a priori* à leur sens par défaut, qui se définit généralement en fonction de sa fréquence d'apparition. Si on en revient à l'exemple utilisé en introduction, le fait de savoir qu'il existe plusieurs "Paris" à travers le monde ne dispense pas de penser que la phrase "je l'ai connue et épousée à Paris" renvoi par défaut vers la capitale française puisque c'est le sens le plus couramment utilisé pour ce mot. Pourtant cette méthode paraît ici très insatisfaisante puisqu'elle échoue à lier correctement la mention "Paris" vers l'entité qui lui correspond, à savoir "Paris, Texas". Les chercheurs ont donc établis une chaîne de traitement plus complexe en générant et sélectionnant les candidats susceptibles de correspondre à l'entité recherchée⁴.

La première étape consiste à construire un sous-ensemble de la base de connaissances composé des entités susceptibles de correspondre à la mention. Elle est nécessaire car elle permet d'éviter de travailler avec l'ensemble d'une base de connaissances qui peut compter plusieurs milliers ou millions d'entrées. Mais la sélection doit aussi être suffisamment large pour s'assurer que l'entité recherchée est bien dans cette sous-base. Il faut donc établir des critères de sélection basés sur la relation supposée entre la mention et sa correspondance dans la base de connaissances. La méthode d'usage consiste à se baser sur les variantes lexicales des entités : est considéré comme candidat toute entité qui dispose d'une variante lexicale correspondante à la mention recherchée. Cette étape peut également s'accompagner d'un pré-ordonnement *a priori* des candidats en fonction de critères comme la popularité par exemple. On peut ainsi considérer par défaut que la mention "Paris" a plus de chance d'être un renvoi vers l'entité "Paris, France" que vers "Paris, Texas".

Cet ordonnancement *a priori* ne peut cependant être considéré comme suffisant pour réaliser le liage. Pour être juste, il faut également comparer le contexte d'apparition de la mention avec les métadonnées associées à chaque entité candidate. L'objectif est d'ordonner les entités en fonction de leur proximité avec le contexte de la mention afin de sélectionner celle qui a le plus de chance de lui correspondre. Cette proximité peut s'établir

4. *Ibid.*, p. 117–125.

en fonction de plusieurs critères comme la co-occurrence de certaines entités par exemple. Il faut également envisager la possibilité que cette mention ne soit pas disponible au sein de la base de connaissances, ou parce que le référentiel est lacunaire ou parce qu’il s’agit d’une variante lexicale qui n’a pas encore été référencée. Ces cas doivent être clairement identifiés car ils représentent autant de potentiels ajouts à la base de connaissances.

Cette base de connaissances constitue donc ici la clé du processus. Elle se présente comme un ensemble d’entrées associées à des informations dont la structure est systématisée. Similaire à une ontologie, elle peut comme cette dernière se construire de deux façons. On peut l’envisager tout d’abord selon une logique de mise en place d’un ensemble général de connaissances sur un domaine, que ce soit dans un contexte industriel ou participatif. Elle peut au contraire être contextuelle au corpus et se nourrir d’un repérage préalable - manuel ou automatique - des concepts pertinents et des relations qui les caractérisent⁵. Dans les deux cas, cette base de connaissances peut être emmenée à évoluer au cours du travail de liage par l’intégration de nouvelles entités qui ne correspondent à aucune entité de la base de connaissances.

1.1.2 Un enjeu pour les sources historiques

Le développement des techniques de liage d’entités est apparu dans un contexte d’étude de textes contemporains, mais il peut aussi s’appliquer dans le cadre de documents historiques. L’appropriation des outils numériques par les acteurs de la recherche en histoire et du patrimoine a permis d’accroître largement la disponibilité des textes et de faciliter l’extraction d’information via des techniques de ROC (Reconnaissance Optique de Caractères) ou REM et d’études statistiques. L’accès au contenu des textes est cependant freinée par des problématiques propres à ces documents. Tout d’abord, le passage par un processus de ROC peut altérer pour partie le texte. De plus, les conventions orthographiques peuvent varier largement en fonction des lieux et époques, ce qui rend la reconnaissance de certains mots encore plus délicate.

Le cas de confusion le plus courant se place entre le *f* et le *s* long présent dans de nombreux textes manuscrits et imprimés. D’autres cas de confusion portent sur le mélange des langues (par exemple un nom de lieu en français dans un texte en latin) ou sur des variations orthographiques d’un même mot qui peuvent exister au sein d’un même document. Tous ces éléments rendent d’autant plus complexe la tâche de reconnaissance d’entités nommées et de liage avec une base de connaissances⁶. Pourtant, cette tâche

5. *Ibid.*, p. 33.

6. Sur les enjeux du liage d’entités pour les documents historiques et les différentes propositions pour y répondre, cf. Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, José G. Moreno, Emanuela Boros, Ahmed Hamdi, Nicolas Sidère, Mickaël Coustaty et Antoine Doucet, “Entity Linking for Historical Documents : Challenges and Solutions”, dans *22nd International Conference on Asia-Pacific Digital Libraries, ICADL 2020*, 2020 (Lecture Notes in Computer Science), t. 12504, p. 215-231, DOI : 10.1007/

est particulièrement pertinente dans ce contexte où de nombreuses ambiguïtés existent, notamment pour identifier les personnes et lieux qui sont mentionnés par les documents.

1.2 Les pistes pour l'application sur des corpus patrimoniaux

1.2.1 Un défi : bien établir la base de connaissances

Plusieurs travaux de recherches ont donc été menés ces dernières années afin de pallier ces difficultés et améliorer les techniques de liage d'entités pour les adapter au contexte des documents historiques. Ces travaux se sont souvent nourris d'autres recherches parallèles portant sur des problématiques proches. C'est le cas par exemple des recherches sur le liage d'entités multi-langue, c'est à dire un modèle dans lequel la langue des données sources n'est pas la même que celle de la base de connaissances. Des chercheurs ont proposé des modèles spécifiques développés à partir de l'incorporation de mots étrangers dans le corpus⁷ ou, s'il existe quelques éléments pour produire une base de connaissances dans la langue source, à partir du mélange entre ces derniers et un modèle de liage issu d'une langue disposant d'une base de connaissances plus large⁸. Une dernière méthode consiste à construire un modèle se passant de toute ressource bilingue par l'utilisation d'une langue pivot suffisamment proche pour qu'il soit pertinent de construire un modèle à partir de celle-ci puis de l'utiliser sur la source⁹.

Une des problématiques rencontrées par les chercheurs est le choix de la base de connaissances à utiliser au moment du processus. Un certain nombre de travaux ont ainsi procédé au liage des entités nommées présents dans leur corpus avec des ontologies web pré-existantes (Wikidata, DBpedia, ...). Celles-ci ont l'avantage d'être très fournies, ce qui est particulièrement utile dans le cadre de données qui n'ont pas de contexte chronologique ou géographique précis. Mais cette situation comporte aussi des inconvénients : ces ontologies sont porteuses de nombreuses ambiguïtés, notamment liées à un grand nombre d'homonymies. Ces caractéristiques ont par exemple été décrites pour Wikipedia

978-3-030-64452-9_19.

7. Elvys Linhares Pontes, Jose G. Moreno et Antoine Doucet, "Linking Named Entities across Languages using Multilingual Word Embeddings", dans *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, New York, NY, USA, 2020, p. 329-332, URL : <https://doi.org/10.1145/3383583.3398597> (visité le 23/07/2022).

8. Shuyan Zhou, Shruti Rijhwani et Graham Neubig, "Towards Zero-resource Cross-lingual Entity Linking", dans *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, Hong Kong, China, 2019, p. 243-252, DOI : [10.18653/v1/D19-6127](https://doi.org/10.18653/v1/D19-6127).

9. Shruti Rijhwani, Jiateng Xie, Graham Neubig et Jaime Carbonell, "Zero-Shot Neural Transfer for Cross-Lingual Entity Linking", *Proceedings of the AAAI Conference on Artificial Intelligence*, 33-1 (17 juill. 2019), Number : 01, p. 6924-6931, DOI : [10.1609/aaai.v33i01.33016924](https://doi.org/10.1609/aaai.v33i01.33016924).

au moment de la création d'un algorithme de liage d'entités depuis la base Europeana¹⁰. D'autres travaux se sont également portés sur la comparaison entre les principales ontologies disponibles en fonction du résultat obtenu pour des corpus précis¹¹.

Il existe cependant un certain nombre de ressources documentaires dont le contenu ne dispose pas de base de connaissances préétablies, que ce soit parce qu'il s'agit d'une langue rare¹² ou parce que les entités nommées reconnues sont propres au contexte. C'est le cas par exemple d'une étude basée sur un corpus de témoignages de citoyens irlandais concernant la rébellion de 1641 et pour laquelle le nombre d'entités absentes de la base de connaissances utilisée s'élève à 77%¹³. Pour compenser ce manque, les chercheurs ont utilisé un outil permettant d'étendre la recherche à partir d'un principe similaire à ceux mis en œuvre pour le liage d'entités multi-langue¹⁴. Un autre cas problématique est celui du changement de sens de certains mots au cours du temps. C'est ainsi qu'un projet portant sur les manuscrits du philosophe Bentham a dû modifier sa méthode de travail après l'observation de nombreuses incohérences entre les mentions du texte et les entités DBpedia utilisées pour l'annotation du corpus. Ils ont donc amélioré leur modèle par l'utilisation de techniques d'extraction de phrases-clés associant les principales notions à des séquences de mots. Puis ils ont construit des annotations se basant en priorité sur le repérage de mention de ces concepts plutôt que leur alignement avec DBpedia¹⁵. Un autre problème rencontré est celui des données incomplètes. Il a notamment été abordé lors de l'identification de lieux, personnes et unités militaires mentionnés dans des archives de la seconde guerre mondiale. Les chercheurs ont donc adopté une démarche heuristique afin de résoudre les ambiguïtés présentes du mieux qu'ils ont pu¹⁶. Afin de faciliter le choix

10. Eneko Agirre, Ander Barrena, Oier Lopez de Lacalle, Aitor Soroa, Samuel Fernando et Mark Stevenson, "Matching Cultural Heritage items to Wikipedia", dans *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012, p. 1729-1735, URL : http://www.lrec-conf.org/proceedings/lrec2012/pdf/1021_Paper.pdf (visité le 21/07/2022).

11. Aicha Soudani, Yosra Meherzi, Asma Bouhafs, Francesca Frontini, Carmen Brando, Yoann Dupont et Frédérique Mélanie-Becquet, "Adaptation et évaluation de systèmes de reconnaissance et de résolution des entités nommées pour le cas de textes littéraires français du 19ème siècle", dans *Atelier Humanités Numériques Spatialisées (HumaNS'2018)*, Montpellier, France, 2018, URL : <https://hal.archives-ouvertes.fr/hal-01925816> (visité le 21/07/2022).

12. cf. plus haut.

13. Gary Munnely et Seamus Lawless, "Investigating Entity Linking in Early English Legal Documents", dans *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, New York, NY, USA, 2018 (JCDL '18), p. 59-68, DOI : 10.1145/3197026.3197055.

14. A ce propos, v. aussi Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Michael Röder, Daniel Gerber, Sandro Athaide Coelho, Sören Auer et Andreas Both, "AGDISTIS - Graph-Based Disambiguation of Named Entities Using Linked Data", dans *The Semantic Web - ISWC 2014*, dir. Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandečić, Paul Groth, Natasha Noy, Krzysztof Janowicz et Carole Goble, Series Title : Lecture Notes in Computer Science, Cham, 2014, t. 8796, p. 457-471, DOI : 10.1007/978-3-319-11964-9_29.

15. Pablo Ruiz et Thierry Poibeau, "Mapping the Bentham Corpus : Concept-based Navigation", *Journal of Data Mining and Digital Humanities*, Atelier Digit_Hum (mars 2019), Publisher : Episciences.org, DOI : 10.46298/jdmhdh.5044.

16. Erkki Heino, Minna Tamper, Eetu Mäkelä, Petri Leskinen, Esko Ikkala, Jouni Tuominen, Mikko Koho et Eero Hyvönen, "Named Entity Linking in a Complex Domain : Case Second World War History",

parmi les ontologies web disponibles, des chercheurs ont proposé un certain nombre de mesures permettant d'évaluer la qualité de la ressource et de comparer leur efficacité dans la tâche de liage d'entités¹⁷.

1.2.2 Les outils disponibles

Malgré les difficultés que nous venons d'évoquer, l'utilisation du liage d'entités s'est largement diffusée au sein des travaux portant sur des archives historiques grâce au développement d'outils spécifiques. C'est le cas notamment de DBpedia Spotlight, une application qui permet d'annoter automatiquement les entités nommées reconnues dans un texte à partir de l'ontologie web DBpedia. Elle permet notamment de spécifier le type d'entités qui nous intéresse afin de faciliter le processus de résolution des ambiguïtés et dispose d'une interface utilisateur afin d'accompagner sa prise en main¹⁸. Dans une logique similaire, l'outil REDEN permet d'accroître les possibilités de liage des entités nommées par la multiplication des ontologies et en permettant à l'utilisateur d'en ajouter manuellement¹⁹. Plus récemment, d'autres outils se sont développés afin de permettre l'apprentissage d'un modèle de liage d'entités à partir des sources étudiées. Pour notre travail, nous avons choisi d'utiliser la librairie python Spacy car elle est aujourd'hui l'outil le plus répandu et réputé le plus facile à prendre en main²⁰. Il existe cependant d'autres outils similaires permettant de développer son propre modèle comme par exemple la librairie python DeezyMatch qui peut s'utiliser autant pour entraîner un nouveau modèle sur un contexte précis que pour s'intégrer dans un workflow déjà existant²¹.

dans *Language, Data, and Knowledge*, dir. Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos et Sebastian Hellmann, Cham, 2017 (Lecture Notes in Computer Science), p. 120-133, DOI : 10.1007/978-3-319-59888-8_10.

17. Nathalie Abadie, Carmen Brando Escobar et Francesca Frontini, "Evaluation de la qualité des sources du Web de Données pour la résolution d'entités nommées", *Revue des Sciences et Technologies de l'Information - Série ISI : Ingénierie des Systèmes d'Information*, 21-5 (8 févr. 2017), URL : <https://iieta.org/download/file/fid/27476> (visité le 27/07/2022).

18. Pablo N. Mendes, Max Jakob, Andrés García-Silva et Christian Bizer, "DBpedia spotlight : shedding light on the web of documents", dans *Proceedings of the 7th International Conference on Semantic Systems*, New York, NY, USA, 2011 (I-Semantics '11), p. 1-8, DOI : 10.1145/2063518.2063519.

19. Francesca Frontini, Carmen Brando et Jean-Gabriel Ganascia, "Domain-adapted named-entity linker using Linked Data", dans 2015, URL : <https://hal.archives-ouvertes.fr/hal-01203356> (visité le 27/07/2022).

20. Nous développerons plus avant les fonctionnalités de spacy au chapitre 9.

21. Kasra Hosseini, Federico Nanni et Mariona Coll Ardanuy, "DeezyMatch : A Flexible Deep Learning Approach to Fuzzy String Matching", dans 2020, DOI : 10.18653/v1/2020.emnlp-demos.9.

1.3 Les avancées actuelles de la recherche

1.3.1 Quels résultats pour les modèles proposés ?

A partir des éléments que nous avons présentés, plusieurs travaux ont ainsi mis en œuvre des techniques de liage d'entités sur des sources historiques et proposent une évaluation des résultats obtenus. C'est le cas par exemple d'une étude basée sur un corpus de textes littéraires français du XIX^e siècle, qui obtient un taux de rappel des candidats - c'est-à-dire la proportion des ensembles de candidats contenant la bonne référence par rapport au nombre de mentions auxquelles il existe une référence dans la base de connaissances - entre 0,63 et 0,83 en fonction de l'ontologie utilisée. Quant à la précision des candidats - c'est-à-dire la proportion des ensembles de candidats contenant la bonne référence par rapport au nombre d'ensembles de candidats -, elle atteint même 1 en utilisant DBpedia. La mesure de l'exactitude globale - c'est-à-dire la proportion de références correctement assignée pour chaque mention d'entité nommée disposant d'une référence pertinente dans la base de connaissances - est située entre 0,7 et 0,85 en fonction de l'ontologie utilisée²². Une autre étude basée sur les champs descriptifs du Smithsonian Cooper-Hewitt National Design Museum à New York parvient à un taux de rappel entre 0,08 et 0,44 et une précision entre 0,24 et 0,80 en fonction de l'application utilisée. Cette étude a également permis de mesurer la complémentarité entre ces ressources : si DBpedia Spotlight produit des scores très bas, seules 4% des entités trouvées l'ont été communément par les 3 applications utilisées. De plus, 54% des entités trouvées l'ont été uniquement par l'un des autres outils (34% par Zemanta et 20% par Alchemy API), ce qui manifeste d'une bonne complémentarité entre ces services²³.

Ces résultats peuvent également varier en fonction des cas particuliers que nous avons évoqués plus haut. Par exemple l'étude sur les témoignages irlandais permet de mettre en valeur un outil (AGDISTIS) qui se caractérise par d'excellents résultats globaux pour le liage d'entités. Cependant, si on sépare les entités liées à des éléments de la base de connaissances des entités reconnues à juste titre comme absentes de cette base, on observe que c'est pour la reconnaissance de ces dernières que le programme est particulièrement efficace. Or ils forment ici 77% du corpus. Pour ce qui est de la première tâche, son efficacité est largement supplantée par celle de deux autres programmes - Dexter et

22. Aicha Soudani, Yosra Meherzi, Asma Bouhafs, *et al.*, "Adaptation et évaluation de systèmes de reconnaissance et de résolution des entités nommées pour le cas de textes littéraires français du 19^{ème} siècle"... A propos des critères d'évaluation des modèles, *cf.* Nathalie Abadie, Carmen Brando Escobar et Francesca Frontini, "Evaluation de la qualité des sources du Web de Données pour la résolution d'entités nommées"...

23. Seth van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner et Rik Van de Walle, "Exploring entity recognition and disambiguation for cultural heritage collections", *Digital Scholarship in the Humanities*, 30-2 (1^{er} juin 2015), p. 262-279, DOI : 10.1093/llc/fqt067.

Kea - qui ne savent pas reconnaître les entités absentes de la base de connaissances²⁴. Une autre étude basée sur un corpus composé de cinq langues a permis de mettre en œuvre plusieurs approches pour compléter le travail de liage d’entités au moment de l’entraînement du modèle : exploration des résultats pour différentes variations orthographique et linguistique d’un même mot puis filtrage des candidats obtenus en fonction de critères comme le type d’entité ou des métadonnées qui lui sont associées (par exemple la date de naissance pour les personnes). Ces différents tests ont permis de largement augmenter la précision des candidats et le taux de rappel lorsque ces approches sont ajoutées au modèle entraîné avec des variations en fonction de la langue et du scénario choisi²⁵. Pour finir, une dernière étude utilisant REDEN a permis de montrer la variation de la correction des résultats en fonction de l’ajout d’un poids aux relations entre les entités. Cette opération permet de modifier les caractéristiques du graphe calculé pour opérer le liage d’entités et améliore dans certains cas le résultat obtenu²⁶.

1.3.2 De nouvelles perspectives pour la recherche historique

Ces résultats ont ainsi permis d’accroître la portée de certaines analyses historiques en automatisant l’identification des entités nommées reconnues dans les textes. C’est le cas par exemple d’une étude sur les archives du journal *Le Monde* (1944-1986) qui a permis d’approfondir l’analyse de la répartition en genre des personnalités mentionnées dans les articles. Plutôt que d’analyser uniquement les occurrences des mots ”homme” et ”femme”, l’utilisation du liage d’entités a permis de relier chaque nom de personne à une entrée de la base de connaissances YAGO et d’évaluer plus précisément la répartition genrée des personnalités mentionnées par le journal. Cette étude a également pu calculer les variations de l’âge en fonction des différentes catégories de personnes ainsi que les occurrences des pays étrangers²⁷. Dans une optique similaire, l’étude sur Bentham que nous avons citée plus haut a permis de produire un certain nombre de graphes pour visualiser sous forme de réseau les concepts utilisés dans les manuscrits annotés²⁸.

Dans le même temps, un certain nombre d’études se sont portées sur la localisa-

24. Gary Munnely et Seamus Lawless, “Investigating Entity Linking in Early English Legal Documents”...

25. Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, José G. Moreno, *et al.*, “Entity Linking for Historical Documents...”.

26. Carmen Brando, Francesca Frontini et Jean-Gabriel Ganascia, “Disambiguation of Named Entities in Cultural Heritage Texts Using Linked Data Sets”, dans *New Trends in Databases and Information Systems*, dir. Tadeusz Morzy, Patrick Valduriez et Ladjel Bellatreche, Series Title : Communications in Computer and Information Science, Cham, 2015, t. 539, p. 505-514, DOI : 10.1007/978-3-319-23201-0_51.

27. Thomas Huet, Joanna Biega et Fabian M. Suchanek, “Mining history with *Le Monde*”, dans *Proceedings of the 2013 workshop on Automated knowledge base construction*, New York, NY, USA, 2013 (AKBC ’13), p. 49-54, DOI : 10.1145/2509558.2509567.

28. Pablo Ruiz et Thierry Poibeau, “Mapping the Bentham Corpus...”.

tion automatique des toponymes historiques mentionnés dans les textes. C’est le cas par exemple du projet Perseus qui rassemble des données historiques concernant plusieurs périodes. Les essais de localisation automatique ont permis d’observer de fortes disparités entre les corpus : le processus de liage d’entités est plus efficace pour les textes anciens (Grèce et Rome) que pour les textes modernes (Angleterre et États-Unis) parce que le nombre d’ambiguïtés est bien moins conséquent²⁹. Ces localisations automatiques peuvent également permettre de générer un certain nombre de production cartographique afin de mieux visualiser la répartition de ces toponymes. C’est le cas par exemple d’une étude portant sur les rues de Paris mentionnées dans 31 romans écrits au XIX^e siècle. Les essais de cartographie de ces rues permettent de comparer efficacement les quartiers qui sont mentionnés et sélectionner les romans qui sont susceptibles de contenir des données sur l’état d’un lieu précis à cette période³⁰. Pour finir, des travaux ont permis de développer des modèles de liage d’entités par apprentissage machine afin d’associer les toponymes mentionnés dans un texte avec un gazetier en ligne³¹. A la suite de ces travaux, le projet Tanagra Mapping Tool propose une interface pour visualiser les entités présentes dans n’importe quel texte importé par l’utilisateur³².

Ces différents travaux ont également permis de participer à l’évolution de certaines technologies couramment utilisés pour décrire des documents historiques. C’est le cas de la TEI utilisée notamment pour l’édition de textes et qui peut également contenir des éléments pour décrire les liens entre les entités repérées dans un texte et un référentiel en ligne contenant une description plus complète de ces entités³³. Il est alors possible de compléter cette tâche par l’utilisation d’un modèle de liage d’entités permettant d’enrichir automatiquement les balises de la TEI à partir d’une base de connaissances³⁴. Selon le même principe, le projet NER4Archives a permis de développer des outils pour repérer automatiquement les entités nommées présentes dans des inventaires d’archives sous for-

29. David A. Smith et Gregory Crane, “Disambiguating Geographic Names in a Historical Digital Library”, dans *Research and Advanced Technology for Digital Libraries*, dir. Panos Constantopoulos et Ingeborg T. Sølvsberg, réd. par Gerhard Goos, Juris Hartmanis et Jan van Leeuwen, Series Title : Lecture Notes in Computer Science, Berlin, Heidelberg, 2001, t. 2163, p. 127-136, DOI : 10.1007/3-540-44796-2_12.

30. Noémie Boeglin, Michel Depeyre, Thierry Joliveau et Yves-François Le Lay, “Pour une cartographie romanesque de Paris au XIX^e siècle. Proposition méthodologique”, dans *Conférence Spatial Analysis and GEOMatics*, Nice, France, 2016 (Actes de la conférence SAGEO’2016 - Spatial Analysis and GEOMatics), URL : <https://hal.archives-ouvertes.fr/hal-01619600> (visité le 23/07/2022).

31. Rui Santos, Patricia Murrieta-Flores, Pável Calado et Bruno Martins, “Toponym matching through deep neural networks”, *International Journal of Geographical Information Science*, 32-2 (1^{er} févr. 2018), Publisher : Taylor & Francis __eprint : <https://doi.org/10.1080/13658816.2017.1390119>, p. 324-348, DOI : 10.1080/13658816.2017.1390119.

32. *Location*, URL : <https://obtic.sorbonne-universite.fr/tanagra/map> (visité le 01/08/2022).

33. Francesca Frontini, Carmen Brando, Marine Riguët, Clémence Jacquot et Vincent Jolivet, “Annotation of Toponyms in TEI Digital Literary Editions and Linking to the Web of Data”, *MALTIT : Materialities of literature-2* (juill. 2016), DOI : 10.14195/2182-8830_4-2_3.

34. Carmen Brando, Francesca Frontini et Jean-Gabriel Ganascia, “REDEN : Named Entity Linking in Digital Literary Editions Using Linked Data Sets”, *Complex Systems Informatics and Modeling Quarterly-7* (29 juill. 2016), p. 60, DOI : 10.7250/csimq.2016-7.04.

mat EAD afin d'enrichir leur contenu de liens vers des référentiels externes décrivant ces mêmes entités³⁵.

Conclusion

En conséquence, nous avons vu dans ce chapitre les différents éléments fondateurs de la technique de liage d'entités, son application dans l'étude des archives anciennes et les différents résultats qui ont pu être obtenus. Cet état des lieux nous permet ainsi de situer notre travail par rapport à la recherche actuelle et de prendre en compte les enjeux mis au jour par les autres travaux. Ces éléments sont cruciaux pour envisager l'application du liage d'entités dans le cadre du projet Himanis.

35. Florence Clavaud, Laurent Romary, Pauline Charbonnier, Lucas Terriel, Gaetano Piraino et Vincent Verdesse, “NER4Archives (named entity recognition for archives) : Conception et réalisation d'un outil de détection, de classification et de résolution des entités nommées dans les instruments de recherche archivistiques encodés en XML/EAD” (), p. 23, URL : <https://hal.archives-ouvertes.fr/hal-03625734/document>.

Chapitre 2

Les avancées du projet Himanis

Le liage d'entités constitue donc une des évolutions récentes de l'application du TAL appliqué aux sources historiques. Il s'insère ainsi dans une série de travaux portant sur la lecture automatique des textes et dont les évolutions récentes se concentrent sur la REM (Reconnaissance d'Écriture Manuscrite) et la REN. Cette dernière est aujourd'hui une technique bien maîtrisée et plusieurs travaux ont permis de l'intégrer aux algorithmes d'apprentissage d'analyse du langage¹. La mise en pratique de la REM et REN dans le cadre de la lecture des textes médiévaux représente un enjeu pour lequel le projet Himanis a tenté d'apporter sa contribution. Ces travaux fournissent ainsi une base nécessaire au travail réalisé pendant le stage.

Ce chapitre sera donc consacré à la présentation des différents résultats disponibles grâce aux travaux menés sur les registres du Trésor des Chartes par le projet Himanis ou à la suite de ce dernier. Nous présenterons dans un premier temps les modèles de REM et REN qui ont été développés au sein de ces travaux. Nous exposerons ensuite la structure des documents étudiés et sa mise en forme dans l'interface Arkindex. Enfin, nous rendront compte du travail d'import des différentes métadonnées concernant ces textes au sein de la même interface.

1. Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary et Benoît Sagot, “Establishing a New State-of-the-Art for French Named Entity Recognition”, dans 2020, URL : <https://hal.inria.fr/hal-02617950> (visité le 21/07/2022).

2.1 Des modèles de REM et REN appliqués au Trésor des Chartes

2.1.1 Processus de travail

Initié en 2015, le projet Himanis s’est dans un premier temps concentré sur l’indexation des manuscrits numérisés du Trésor des Chartes. Ce travail a notamment été permis par le développement préalable au sein du projet Oriflamms de techniques d’alignement automatique entre un texte et des images porteuses de ce textes². L’édition de Paul Guérin des actes royaux du Poitou, une fois numérisée et structuré pour édition électronique³, a ici servi de vérité terrain à la mise en œuvre de cet alignement pour le trésor des chartes. Le résultat produit par le logiciel de REM à partir des images a ainsi été optimisé pour être le plus proche possible de cette vérité terrain et optimiser son application à l’ensemble du corpus. En utilisant le logiciel Transkribus, les membres du projet ont pu proposer une transcription complète des registres du trésor des chartes. Plutôt que de proposer une transcription linéaire des textes, le choix a été fait d’isoler chaque mot et de rendre disponible toutes les interprétations possibles pour chacun accompagné d’un indice de confiance pour chaque hypothèse. Ces atomes d’informations permettent ainsi la constitution d’un index général des occurrences de mots parmi ces hypothèses et facilite ainsi la recherche textuelle au sein du corpus⁴.

Ces modèles de REM ont par la suite été complétés par d’autres travaux concernant la REN sur des textes médiévaux. Une partie du corpus utilisé pour le projet HOME et deux autres ensembles de textes ont été préparés et annotés pour apprendre à reconnaître automatiquement les entités nommées présentes dans ces textes. Les travaux se sont concentrés sur la reconnaissance des personnes et des lieux et ont permis d’atteindre des résultats très satisfaisants : tous les tests réalisés sur des corpus d’évaluation ont obtenu une précision supérieure à 0,85 et un taux de rappel supérieur à 0,88. Plusieurs modèles ont été utilisés et leur évaluation comparative a permis de mettre en valeur un modèle personnalisé à partir d’un modèle Bi-LSTM comme obtenant de meilleurs résultats que les modèles Fair et Spacy couramment usités⁵.

2. Théodore Bluche, Dominique Stutzmann et Christopher Kermorvant, “Automatic Handwritten Character Segmentation for Paleographical Character Shape Analysis”, dans *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, Santorini, France, 2016 (2016 12th IAPR Workshop on Document Analysis Systems (DAS)), p. 42-47, DOI : 10.1109/DAS.2016.74.

3. Paul Guérin, *Actes Royaux du Poitou (1302-1464)*, avec la coll. de Léonce Celier, Frédéric Glorieux et Vincent Jolivet, 1881, URL : <http://corpus.enc.sorbonne.fr/actesroyauxdupoitou/> (visité le 04/08/2022).

4. Dominique Stutzmann, Jean-François Moufflet et Sébastien Hamel, “La recherche en plein texte dans les sources manuscrites médiévales...”

5. Sergio Torres Aguilar et Dominique Stutzmann, “Named Entity Recognition for French medieval charters”...

2.1.2 Une chaîne de traitement presque complète

Ces résultats ont ainsi permis l'émergence de travaux associant les modèles de REM et de REN appliqués aux corpus médiévaux comme cela a pu déjà être réalisé sur d'autres documents contemporains⁶. Ces travaux ont ainsi permis d'évaluer l'influence du traitement préalable de l'image sur la tâche de REN. Il a été notamment observé que la qualité de cette dernière ne varie que faiblement lors du passage de la transcription manuelle à la transcription par REM. Au contraire, la qualité de la détection des lignes a un impact conséquent sur la qualité de la REN. L'évaluation des modèles a également permis de mettre en valeur la performance des modèles multi-langues et leur utilisation possible dans les cas où on dispose de données d'entraînement en quantité limitée⁷.

Une autre étude a permis de comparer l'efficacité des modèles de REM et REN en fonction de leur utilisation successive ou combinée au sein d'un même modèle. Elle a permis de montrer que la qualité de la REM peut avoir une influence conséquente sur la qualité de la REN lorsque le taux d'erreur dans la reconnaissance des lettres et des mots est élevé, mais aussi que l'approche combinée REM et REN a dans tous les cas des résultats plus intéressants que l'approche séparée⁸. Nous disposons donc actuellement de modèles de lecture automatique des textes manuscrits médiévaux applicables pour le Trésor des Chartes qui permettent de transcrire automatiquement le texte et d'y reconnaître les entités nommées présentes en son sein. Afin d'envisager l'amélioration de ces travaux par l'intégration d'un modèle de liage d'entités, les textes ont été chargés dans une interface dédiée au traitement d'image et à l'application des modèles : Arkindex.

6. A ce propos, cf. Hugo Scheithauer, *La reconnaissance d'entités nommées appliquées à des données issues de la transcription automatique de documents manuscrits patrimoniaux. Expérimentations et préconisations à partir du projet LECTAUREP*, Mémoire de master "Technologies numériques appliquées à l'histoire", Ecole nationale des chartes, 2021, URL : https://raw.githubusercontent.com/HugoSchtr/memoire_TNAH_M2_HugoScheithauer/main/memoire_Hugo_Scheithauer_TNAH.pdf (visité le 29/05/2022)

7. Claire Bizon Monroc, Blanche Miret, Marie-Laurence Bonhomme et Christopher Kermorvant, "A Comprehensive Study of Open-Source Libraries for Named Entity Recognition on Handwritten Historical Documents", dans *Document Analysis Systems*, dir. Seiichi Uchida, Elisa Barney et Véronique Eglin, Cham, 2022 (Lecture Notes in Computer Science), p. 429-444, DOI : 10.1007/978-3-031-06555-2_29.

8. Emanuela Boroş, Verónica Romero, Martin Maarand, Kateřina Zenklová, Jitka Křečková, Enrique Vidal, Dominique Stutzmann et Christopher Kermorvant, "A comparison of sequential and combined approaches for named entity recognition in a corpus of handwritten medieval charters", dans *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2020, p. 79-84, DOI : 10.1109/ICFHR2020.2020.00025.

2.2 Structure physique et logique du texte

2.2.1 Les éléments déjà présents dans Arkindex

Arkindex est une interface créée par l'entreprise Teklia dans le but de gérer le traitement automatique d'un grand nombre de documents numérisés. Elle permet l'import d'images via des manifestes sous format IIIF (International Image Interoperability Framework), leur annotation manuelle et leur analyse automatique (structure et composition de l'image, reconnaissance de caractère et d'écriture manuscrites, extraction d'entités nommées)⁹. Dans le cadre de notre travail, nous avons principalement utilisé l'API d'Arkindex pour importer les données nécessaires à la mise en place du liage d'entités. La documentation associée à cette API est disponible ici : <https://arkindex.tekليا.com/api-docs/>.

Au cours des phases précédentes du projet Himanis, le corpus du Trésor des Chartes a été chargé au sein de l'interface pour former la collection "Himanis | TEKLIA processing" contenant 200 registres numérisés. Les images ont ensuite été segmentées en fonction des zones de texte sous forme d'éléments "Paragraph" et "Text Line" comme présenté dans la Figure 1. Par la suite, un logiciel de lecture automatique a été utilisé pour lire le texte contenu dans ces éléments via un modèle de REM.

2.2.2 Des zones de texte comme interface entre actes et pages

Le texte est ici contenu par des éléments qui découlent directement des éléments pages. Or ce n'est pas ici la manière la plus pertinente d'aborder le contenu des registres. En effet la description disponible de ces derniers concerne essentiellement les actes qu'ils contiennent et non les pages. Pour faire le lien entre chaque acte et les éléments qui le concernent contenus dans l'inventaire des registres, il faut donc transformer la segmentation des zones de texte en fonction de la séparation des actes. Or cette séparation est ne suit que rarement celle des pages : il est fréquent qu'une page contienne plusieurs actes ou qu'un acte soit présents dans plusieurs pages. Il arrive même que les pages portant les morceaux d'un même acte ne soient pas à la suite les unes des autres et que d'autres actes entrecoupent ces morceaux.

La segmentation du contenu des registres a donc été revue pour être organisée en fonction des actes qu'ils contiennent et non en fonction des pages. Il a donc été imaginé un niveau supplémentaire de segmentation au travers des zones de textes. Ceux-ci correspondent aux différents composants d'un acte au sein des pages. Ils sont donc à la fois une composante des actes et des pages et forment une interface entre le contenu physique et le contenu intellectuel des registres.

9. *Teklia - Arkindex*, URL : <https://tekليا.com/solutions/arkindex/> (visité le 05/08/2022).

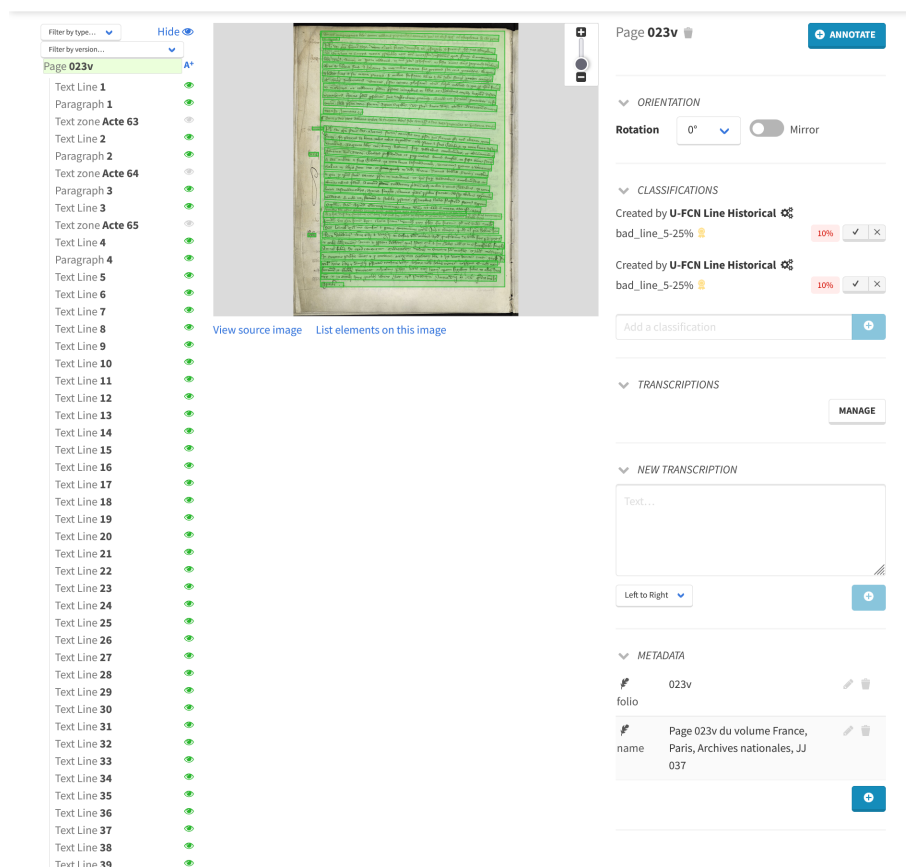


Figure 1 — *Segmentation initiale des pages dans la plateforme Arkindex. L'élément page comprend à gauche les éléments enfants présents dans l'image, au centre l'image avec les éléments initiaux en surbrillance et à droite les métadonnées associées à l'image.*

2.3.2 Normalisation des éléments (dates et langues)

Afin d'envisager l'import de toutes ces données au sein d'Arkindex¹¹, nous avons été confrontés au problème de la normalisation des données.

Conclusion

11. Nous décrirons plus précisément ce travail au chapitre 8

Chapitre 3

***Legacy Metadata* : Des instruments de recherche précis mais incomplets**

Il faudrait que les mots des titres ne soient pas séparés (2.3.1)

Il y a deux articles de Caroline Parfait sur l'influence de l'OCR dans la NER

Article sur l'influence de la qualité d'OCR dans l'entity linking : <https://hal.archives-ouvertes.fr/hal-02557116/document>

Thèse Yoann Dupont parle de la structure des entités nommées et de du modèle en relation des entités (intéressant par rapport à ce qu'on a fait avec Heurist)

Je sais plus si j'en parle ailleurs dans mon plan, mais pour la notion de legacy data, il faut aussi voir Scheitauer (aussi pour la REN d'ailleurs)

Pour la suite, sur le chapitre sur l'alignement d'entités géographiques avec Geonames et DicoTopo, ça peut être intéressant et sur le parsing des noms de lieux : <https://hal.archives-ouvertes.fr/hal-02141257> -> + ça parle aussi du lien noms personnes / lieux et des difficultés avec les lieux non-anglais et non-modernes

Deuxième partie

Modéliser et formaliser un référentiel à partir d'un instrument de recherche papier

Troisième partie

Alignement, diffusion et utilisation du référentiel

Chapitre 4

Enrichir les données à l'aide d'un référentiel externe

Chapitre 5

Mise à disposition d'un nouveau référentiel ?

Nous avons donc créé des éléments "Act" comme des enfants direct des registres et contenant tous les morceaux de textes d'un même acte. Ces éléments sont eux-mêmes composés d'éléments "Text zone" qui contiennent les morceaux de ces actes dans les pages, ils sont les enfants à la fois des éléments "Act" et des éléments "Page". Ces "Text zone" sont à l'interface entre le contenu physique et le contenu intellectuel des registres puisqu'ils représentent la manifestation des textes dans une page.

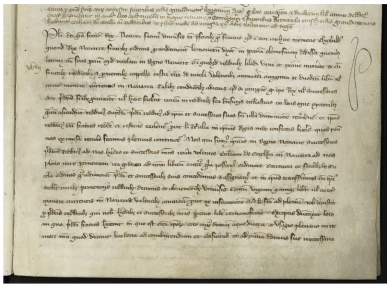
Filter by type...

Act 18

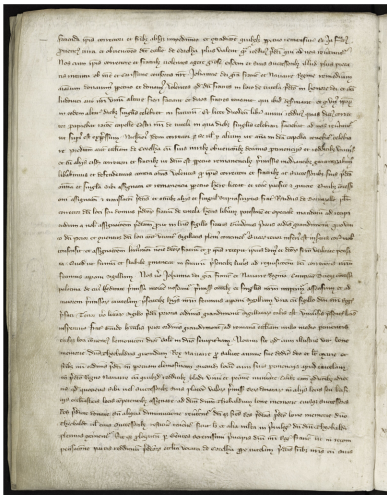
Text zone Acte 18

Text zone Acte 18

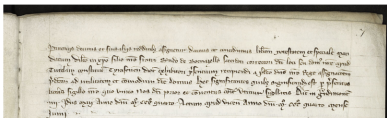
Text zone Acte 18



Acte 18 (Text zone)



Acte 18 (Text zone)



Acte 18 (Text zone)

Act 18

Created by Virgile Reigner

ORIENTATION

Rotation

0°

Mirror

CLASSIFICATIONS

Add a classification

TRANSCRIPTIONS

MANAGE

NEW TRANSCRIPTION

Text...

Left to Right

METADATA

abstract

Échange, ratifié par Jeanne, reine de Navarre, entre le roi et les religieux de l'ordre de Grandmont. Le roi reprend aux religieux les rentes à eux concédées sous condition d'échange possible, à Tudela, par Thibaud II, roi de Navarre, et, leur laissant toutefois leur demeure avec l'église, le jardin, l'aqueduc et le plein usage du bois de La Bardena, il leur donne en compensation l'église de Corella. Deux frères demeureront à Tudela et y célébreront chaque jour une messe à l'autel de saint Louis. V. n° 23. Est transcrit dans ce document le mandat (1304, 15 mai, Grandmont) donné par Gui Foucher, prieur de l'ordre de Grandmont, à frère Raimond de Bornazello, correcteur de Tudela, de procéder à l'échange.

date	1304, June
date_orig	1304, juin
himanisId	501
inventoryReference	Paris_AN_JJ_inventaire_JJ37-50.xml_1
language	lat

Figure 3 – Visualisation d’un élément ”Act” dans Arkindex. Cet élément est le parent de plusieurs éléments ”Text zone” qui représentent chacun une portion du texte décrit par l’élément ”Act”.

Conclusion générale

Tout ce qui suit ce sont des conseils de Lucence Ing, utile à garder pour le moment mais il faudra les supprimer.

Chapitre 6

Un autre chapitre

6.1 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**.

6.2 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.

Quatrième partie

Une autre partie

Conclusion

Annexe A

Première annexe

Table des matières

Résumé	i
Remerciements	iii
Introduction	xi
I De la <i>legacy data</i> au liage d’entités : quel matériel disponible pour entraîner un modèle ?	1
1 État des lieux de la recherche sur le liage d’entités	3
1.1 Mise en œuvre du liage d’entités	4
1.1.1 Méthodologie	4
1.1.2 Un enjeu pour les sources historiques	5
1.2 Les pistes pour l’application sur des corpus patrimoniaux	6
1.2.1 Un défi : bien établir la base de connaissances	6
1.2.2 Les outils disponibles	8
1.3 Les avancées actuelles de la recherche	9
1.3.1 Quels résultats pour les modèles proposés ?	9
1.3.2 De nouvelles perspectives pour la recherche historique	10
2 Les avancées du projet Himanis	13
2.1 Des modèles de REM et REN appliqués au Trésor des Chartes	14
2.1.1 Processus de travail	14
2.1.2 Une chaîne de traitement presque complète	15
2.2 Structure physique et logique du texte	16

2.2.1	Les éléments déjà présents dans Arkindex	16
2.2.2	Des zones de texte comme interface entre actes et pages	16
2.3	Des métadonnées prêtes à l'import	18
2.3.1	Transformation en métadonnées pour les éléments Arkindex	18
2.3.2	Normalisation des éléments (dates et langues)	19
3	<i>Legacy Metadata</i> : Des instruments de recherche précis mais incomplets	21
II	Modéliser et formaliser un référentiel à partir d'un instrument de recherche papier	23
III	Alignement, diffusion et utilisation du référentiel	25
4	Enrichir les données à l'aide d'un référentiel externe	27
5	Mise à disposition d'un nouveau référentiel ?	29
6	Un autre chapitre	33
6.1	Structuration du mémoire	33
6.2	Les données	33
IV	Une autre partie	35
	Conclusion	37
A	Première annexe	39