Vers l'indexation automatique du Trésor des chartes : constitution, alignement et utilisation d'un référentiel d'entités nommées au sein du projet Himanis

Virgile Reignier

ÉCOLE NATIONALE DES CHARTES
UNIVERSITÉ PARIS, SCIENCES & LETTRES

Virgile Reignier
licencié ès histoire
diplômé de master mondes médiévaux

Vers l'indexation automatique du Trésor des chartes

Constitution, alignement et utilisation d'un référentiel d'entités nommées au sein du projet Himanis

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

2022

Résumé

Ce mémoire a été réalisé à la suite d'un stage de 4 mois à l'Institut de Recherche et d'Histoire des Textes, au sein de la section de paléographie latine. Il s'inscrit dans la continuité du projet Himanis visant à rendre disponible le contenu des registres du Trésor des chartes en développant des techniques de lecture automatique. L'objectif de notre travail consiste à poursuivre le développement de ces techniques par l'apprentissage du liage d'entités. Ce procédé vise à mettre en lien les entités nommées reconnues préalablement avec le contenu d'un référentiel et à résoudre les ambiguïtés existantes entre ces entités.

Ce mémoire s'attache donc à décrire les différents travaux que nous avons réalisés pour explorer l'utilisation de cette technique dans le cadre du Trésor des chartes. Il s'intéresse principalement à la construction d'un référentiel à partir d'instruments de recherche utilisés comme *legacy medatada*. Il présente les nombreuses problématiques auxquelles nous avons été confrontés par l'utilisation de ces *legacy data*, que ce soit pour le traitement des données, leur alignement ou l'anticipation de leurs usages. Il rend également compte des essais réalisés pour mettre en œuvre le liage d'entités et propose des perspectives pour améliorer les résultats obtenus.

Mots-clés : TNAH, IRHT, Himanis, Trésor des chartes, Archives Nationales, ROC, Numérisation d'instruments de recherche, Alignement de référentiels, REM, REN, Machine learning, Intelligence artificielle, Reconnaissance d'écritures manuscrites, Entity linking.

Informations bibliographiques : Reignier Virgile, *Vers l'indexation automatique du Trésor des chartes. Constitution, alignement et utilisation d'un référentiel 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.

NB : Ce mémoire a fait l'objet d'un premier dépôt au sein d'un répertoire Github rassemblant les rendus techniques et écrits qui le composent. Nous renvoyons donc vers ce dépôt pour tout ce qui concerne la partie technique : https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier.

Remerciements

Je voudrai tout d'abord remercier mon tuteur de stage Dominique Stutzmann pour son accompagnement. Sa confiance, son écoute et ses nombreux conseils ont été essentiels dans la conduite du travail de stage et pour l'apprentissage qui en a été tiré.

Je tiens aussi à remercier toute l'équipe de l'IRHT pour son accueil chaleureux tout au long de ces 4 mois. Je remercie plus particulièrement Nicole Bergk Pinto, Sébastien Barret et Marlène Helias-Baron pour leur soutien et leur aide précieuse.

Je remercie également toute l'équipe pédagogique du master TNAH dont les enseignements et l'engagement ont permis l'apprentissage efficace de compétences fastidieuses. Je remercie notamment Thibault Clérice, mon directeur de mémoire, pour ses nombreux conseils et sa présence indéfectible tout au long de cette année.

Cette année n'aurait pas eu la même saveur sans l'esprit de camaraderie qui a régné au sein de notre promotion de M2 TNAH. Je remercie donc chaleureusement tous mes camarades pour leur soutien moral et pour l'entraide qui a accompagné notre apprentissage et qui n'a pas faibli au cours des stages. Je remercie particulièrement Valentin et Paul pour leurs relectures précieuses.

Pour finir, je dois remercier ma famille et mes amis qui m'accompagnent dans chacune de mes aventures et qui m'épaulent à chaque difficulté. Merci notamment à ma mère, ma grand-mère, mon grand-père, mon père et Thomas pour leur aide dans la rédaction et leurs nombreuses relectures.

Liste des sigles et abréviations

Institutions et organismes

— AN : Archives Nationales
— CNRS : Centre National de la Recherche Scientifique
— CTHS : Comité des Travaux Historiques et Scientifiques
— IRHT : Institut de Recherche et d'Histoire des Textes

Projets développés par l'IRHT

— BVMM : Bibliothèque Virtuelle des Manuscrits Médiévaux
— Himanis : *HIstorical MANuscript Indexing for user-controlled Search*
— HOME : *History of Medieval Europe*

Technologies

— API : *Application Programming Interface*
— CSV : *Comma-separated values*
— EAD : *Encoded Archival Description*
— IIIF : *International Image Interoperability Framework*
— REM : Reconnaissance d'Écriture Manuscrite
— REN : Reconnaissance d'Entités Nommées
— ROC : Reconnaissance Optique de Caractères
— TEI : *Text Encoding initiative*
— XML : *eXtensible Markup Language*
— XSL : *eXtensible Stylesheet Language*

Disciplines

— TAL : Traitement Automatique des Langues

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).

Alrahabi (Motasem), *Tanagra Mapping Tool*, avec la coll. d'Angélique Allaire et OSM, 2022, url : https://obtic.sorbonne-universite.fr/tanagra/map (visité le 01/08/2022).

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.

Bluche (Théodore), Hamel (Sebastien), Kermorvant (Christopher), Puigcerver (Joan), Stutzmann (Dominique), Toselli (Alejandro H.) et Vidal (Enrique), "Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project", dans *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, ISSN : 2379-2140, 2017, t. 01, p. 311-316, doi : 10.1109/ICDAR.2017.59.

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.

— “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.

CCSD, *Principes FAIR*, URL : https://www.ccsd.cnrs.fr/principes-fair/ (visité le 08/11/2022).

Chevalier (Bernard), *Les pays de la Loire moyenne dans le Trésor des chartes : Berry, Blésois, Chartrain, Orléanais, Touraine, 1350-1502 Archives nationale, JJ 80-235*, avec la coll. d’Archives nationales, Paris, 1993 (Collection de documents inédits sur l’histoire de France, 22).

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.

Dossat (Yves), Lemasson (Anne-Marie) et Wolff (Philippe), *Le Languedoc et le Rouergue dans le Trésor des chartes*, Paris, 1983 (Collection de documents inédits sur l’histoire de France, 16).

Dupont (Yoann), *La structuration dans les entités nommées*, Thèse de doctorat, Université Sorbonne Paris Cité, 2017, URL : https://tel.archives-ouvertes.fr/tel-01772268 (visité le 28/03/2022).

Ehrmann (Maud), *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).

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, 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.

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 (1er 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.

KOUDORO-PARFAIT (Caroline), LEJEUNE (Gaël) et BUTH (Richy), "Reconnaissance d'entités nommées sur des sorties OCR bruitées : des pistes pour la désambiguïsation morphologique automatique", dans *Traitement Automatique des Langues Naturelles*, 2022, p. 45-55.

LE CACHEUX (Paul), *Actes de la chancellerie d'Henri VI concernant la Normandie sous la domination anglaise (1422-1435), extraits des registres du Trésor des chartes aux Archives nationales, publiés avec introductions et notes*, Rouen, 1907.

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).

Linhares Pontes (Elvys), Hamdi (Ahmed), Sidère (Nicolas) et Doucet (Antoine), "Impact of OCR Quality on Named Entity Linking", dans *International Conference on Asia-Pacific Digital Libraries 2019*, Kuala Lumpur, Malaysia, 2019, doi : 10.1007/978-3-030-34058-2_11.

— "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.

Longnon (Auguste), *Paris pendant la domination anglaise (1420-1436), documents extraits des registres de la chancellerie de France, par Auguste Longnon*, Paris, 1878.

Maugis (Edouard), *Documents inédits concernant la ville et le siège du bailliage d'Amiens extraits des registres du Parlement de Paris et du Trésor des chartes : XIVe-XVe siècle (1296-1471)*, Amiens Paris, 1908 (Mémoires de la Société des antiquaires de Picardie. Documents inédits concernant la province, t. 17, 19 et 20).

McDonough (Katherine), Moncla (Ludovic) et Camp (Matje), "Named entity recognition goes to old regime France : geographic text analysis for early modern French corpora", *International Journal of Geographical Information Science*, 33 (27 mai 2019), doi : 10.1080/13658816.2019.1620235.

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.

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).

Reignier (Virgile), *De l'index papier à l'indexation automatique*, Himanis, 26 août 2022, url : https://himanis.hypotheses.org/1106 (visité le 26/08/2022).

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.

SAMARAN (Charles) et ROULEAU (Pierre), *La Gascogne dans les registres du Trésor des chartes*, Paris, 1966 (Collection de documents inédits sur l'histoire de France, Vol. 4).

SANTOS (Rui), MURRIETA-FLORES (Patricia), CALADO (Pável) et MARTINS (Bruno), "Toponym matching through deep neural networks", *International Journal of Geographical Information Science*, 32–2 (1er 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 reconnaissanc 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. Sølvberg, 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 (Aicha), 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 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).

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), *Himanis / Home*, avec la coll. de Sebastien Hamel, *et al.*, 2022, URL : https://heurist.huma-num.fr/heurist/?db=stutzmann_himanis&website (visité le 09/11/2022).

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).

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.

VIARD (Jules), *Documents parisiens du règne de Philippe VI de valois : 1328-1350*, avec la coll. de Société de l'histoire de Paris et de l'Ile-de France, Paris, 1899 (Société de l'histoire de Paris et de l'Ile-de-France).

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 la ville 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 se référant à une entité unique de façon autonome [1]. L'analyse du contenu textuel a ainsi largement progressé ces dernières années autour de cette notion avec le développement de deux techniques : la REN (Reconnaissance d'Entités Nommées) qui consiste à repérer ces objets textuels et à leur attribuer une catégorie, puis 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 anciens [2]. 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 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 de généraliser ce travail, en numérisant de nouveaux documents, en associant

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 : https://bvmm.irht.cnrs.fr. 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 http://himanis.huma-num.fr/app, *cf.* Théodore Bluche, Sebastien Hamel, Christopher Kermorvant, Joan Puigcerver, Dominique Stutzmann, Alejandro H. Toselli et Enrique Vidal, "Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project", dans *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, ISSN : 2379-2140, 2017, t. 01, p. 311-316, DOI : 10.1109/ICDAR.2017.59.

chaque texte aux données disponibles les concernant et en déposant les résultats dans une plateforme librement accessible [6].

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 médiévales [7], 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 documents.

C'est dans ce contexte que ce 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 des données issues du projet 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. Nous étudierons 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 pour cela les différentes étapes de construction du référentiel ainsi que les difficultés rencontrées dans ce cadre et dans son utilisation. Un résumé de notre action a déjà été publié au sein du blog Hypothèses du projet Himanis [8] et nous invitons le lecteur à s'y référer pour aborder notre travail de manière plus synthétique et linéaire. Nous avons en effet fait le choix ici de détailler notre action selon un plan thématique afin d'ordonner notre présentation malgré les nombreux aller-retours que nous avons réalisés entre les différentes étapes en fonction de notre compréhension de la complexité des données et de notre capacité à anticiper les étapes suivantes.

Dans cet objectif, nous exposerons dans une première partie le matériel disponible

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

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).

8. Virgile Reignier, *De l'index papier à l'indexation automatique*, Himanis, 26 août 2022, URL : https://himanis.hypotheses.org/1106 (visité le 26/08/2022).

pour mettre en œuvre ce projet. Nous proposerons ainsi un état des lieux des 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.

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 lesquels 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 recherche 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 classifier les références utilisées prend donc dans ce contexte une place centrale qui se perpétue par la suite dans de nombreuses recherches [1]. 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 sélectionné dans ce but [2].

C'est dans ce contexte qu'est né 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 afin de générer 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 » renvoie 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 établi une chaîne de traitement plus complexe en générant et sélectionnant les candidats susceptibles de correspondre à l'entité recherchée [4].

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-ordonnancement *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, soit parce que le référentiel est lacunaire soit 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 clef 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 caractérisent [5]. Dans les deux cas, cette base de connaissances peut être amenée 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 freiné 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 connaissances [6]. Pourtant, ces techniques

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/

sont particulièrement pertinentes 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 corpus [7] 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 large [8]. 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 source [9].

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.

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.

au moment de la création d'un algorithme de liage d'entités depuis la base Europeana [10]. D'autres travaux se sont également portés sur la comparaison entre les principales ontologies disponibles en fonction du résultat obtenu pour des corpus précis [11].

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 rare [12] 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-langue [14]. 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 évolué dans 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 DBpedia [15]. 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 pu [16]. Pour 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 Munnelly 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/jdmdh.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 d'en comparer l'efficacité dans la tâche de liage d'entités [17].

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 main [18]. Dans une logique similaire, l'outil REDEN permet d'accroître les possibilités de liage des entités nommées par la multiplication des ontologies tout en permettant à l'utilisateur d'en ajouter manuellement [19]. 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 main [20]. 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 existant [21].

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 pour lesquelles 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 utilisée [22]. 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 services [23].

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 absents de cette base, on observe que c'est pour la reconnaissance de ces dernières que le programme est particulièrement pertinent. 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 19ème 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 (1er juin 2015), p. 262-279, DOI : 10.1093/llc/fqt067.

Kea - qui ne savent pas reconnaître les entités absentes de la base de connaissances [24]. 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'entrainement 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 choisi [25]. 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 améliorer dans certains cas le résultat obtenu [26].

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 vise à approfondir l'analyse de la répartition genrée 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 étrangers [27]. 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 annotés [28].

Dans le même temps, un certain nombre d'études se sont portées sur la localisation automatique des toponymes historiques mentionnés dans les textes. C'est le cas par

24. Gary Munnelly 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…".

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 conséquent [29]. Ces localisations automatiques peuvent également permettre de générer un certain nombre de productions cartographiques 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 période [30]. 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 répertoire géographique en ligne [31]. 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'utilisateur [32].

Ces différents travaux ont également participé à l'évolution de certaines technologies couramment utilisés pour décrire des documents historiques. C'est le cas de la TEI (*Text Encoding Initiative*) 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 entités [33]. 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 connaissances [34]. 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 format EAD (*Encoded Archival Description*) afin d'enrichir leur

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. Sølvberg, 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 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).

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 (1er févr. 2018), Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/13658816.2017.1390119, p. 324-348, DOI : 10.1080/13658816.2017.1390119.

32. Motasem Alrahabi, *Tanagra Mapping Tool*, avec la coll. d'Angélique Allaire et OSM, 2022, URL : https://obtic.sorbonne-universite.fr/tanagra/map (visité le 01/08/2022).

33. Francesca Frontini, Carmen Brando, Marine Riguet, 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.

contenu de liens vers des référentiels externes décrivant ces mêmes entités [35].

Conclusion

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 corpus Himanis.

35. Florence Clavaud, Laurent Romary, Pauline Charbonnier, Lucas Terriel, Gaetano Piraino et Vincent Verdese, "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 une des évolutions récentes du TAL appliqué aux sources historiques. Il s'insère dans une série de travaux portant sur la lecture automatique des textes et dont les évolutions récentes se concentrent sur la REM 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 langage [1]. La mise en pratique de la REM et REN dans le cadre de la lecture des textes médiévaux représente un enjeu pour lequel les projets Himanis et HOME ont tenté d'apporter leur contribution. Ces recherches fournissent ainsi la base sur laquelle s'est appuyé le travail réalisé pendant le stage.

Ce chapitre sera donc consacré à la présentation des différents résultats disponibles grâce aux études menées sur les registres du Trésor des chartes depuis le projet Himanis. Nous présenterons dans un premier temps les modèles de REM et REN qui ont été développés au cours de ces travaux. Nous exposerons ensuite le travail réalisé à propos de la structure des documents. Enfin, nous décrirons les étapes de formation du fichier permettant d'associer chaque élément de cette structure aux métadonnées disponibles le concernant.

2.1 Des modèles de REM et REN appliqués aux registres du Trésor des chartes

2.1.1 Processus de travail

Mis en œuvre entre 2015 et 2017, le projet Himanis s'est dans un premier temps concentré sur l'indexation des manuscrits numérisés du Trésor des chartes. Ce travail a

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).

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 texte [2]. L'édition de Paul Guérin des actes royaux du Poitou, une fois numérisée et structurée pour son édition électronique [3], 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 assurer son application à l'ensemble du corpus. En utilisant les outils développés par le projet Oriflamms, les membres du projet Himanis ont pu proposer une transcription complète des registres du Trésor des chartes. Plutôt que de restituer une transcription linéaire des textes, le choix a été fait de conserver chaque mot comme unité isolée 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 facilitent la recherche textuelle au sein du corpus [4].

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 Flair et spaCy couramment usités [5].

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 à des corpus de documents manuscrits médiévaux et plus récents [6].

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"…

6. A ce propos, v. aussi Hugo Scheithauer, *La reconnaissanc 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

Ces recherches ont évalué 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'entrainement en quantité limitée [7].

Une autre étude s'est proposé 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 montré que la qualité de la REM peut avoir une influence conséquente sur la qualité de la REN lorsque le taux d'erreurs dans la reconnaissance des lettres et des mots est élevé, mais aussi que l'approche combinée REM et REN génère dans tous les cas des résultats plus intéressants que l'approche séparée [8]. Nous disposons donc actuellement de modèles de lecture automatique des textes manuscrits médiévaux applicables pour le Trésor des chartes. Ils permettent de transcrire automatiquement le texte et d'y reconnaître les entités nommées présentes à l'intérieur. Dans le cadre de la mise en œuvre de ces travaux et afin de faciliter la navigation dans les documents et l'intégration de nouvelles fonctionnalités dans les modèles utilisés, les textes ont été chargés dans une interface dédiée au traitement d'images et à l'application des modèles : Arkindex.

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 caractères et d'écritures manuscrites, extraction d'entités nom-

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.

mées) [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.teklia.com/api-docs.

Pour faciliter la navigation dans les documents, 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, des techniques de lecture automatique ont été utilisées 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 segmenté en fonction de sa structure physique par des éléments qui découlent directement des éléments pages. Or cette méthode ne permet pas d'associer directement le résultat au contenu des registres. En effet les descriptions disponibles concernent essentiellement la structure logique des textes. Pour faire le lien entre un acte et les éléments qui le concernent dans l'inventaire, il faut donc transformer la segmentation des zones de texte en fonction de la structure logique. En effet, cette dernière ne correspond pas toujours à la structure physique : il est fréquent qu'une page contienne plusieurs actes ou qu'un acte soit présent dans plusieurs pages. Il arrive même que les pages portant les parties d'un même acte ne soient pas à la suite les unes des autres et que d'autres actes entrecoupent ces parties.

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. Celles-ci correspondent aux différents composants d'un acte au sein des pages. Elles sont donc à la fois une composante des actes et des pages et forment une interface entre la structure physique et la structure logique des registres.

9. Claire Bizon Monroc, Blanche Miret, Marie-Laurence Bonhomme, *et al.*, "A Comprehensive Study of Open-Source Libraries for Named Entity Recognition on Handwritten Historical Documents"…

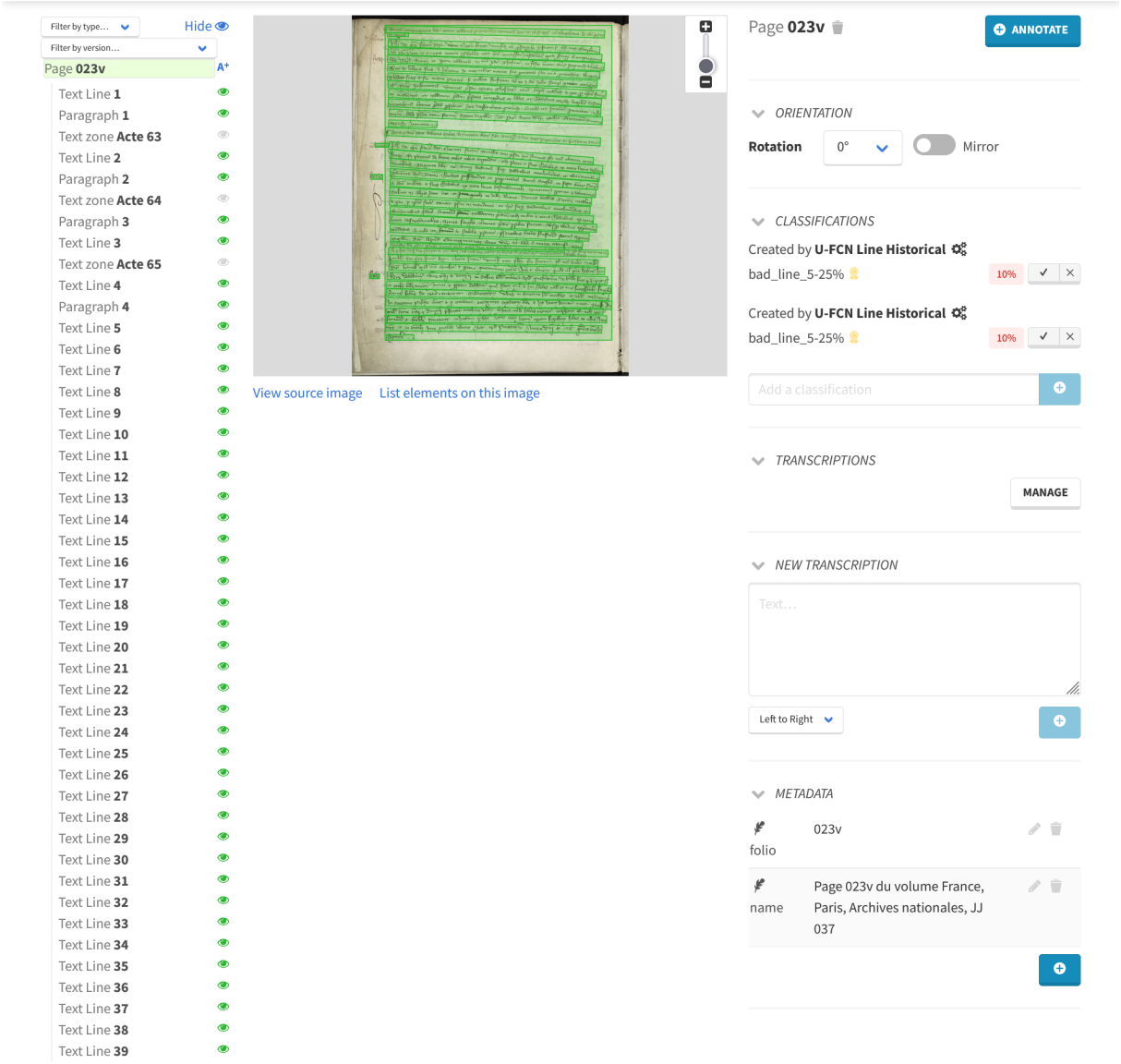

Figure 1 – *Segmentation initiale des pages dans la plateforme Arkindex. L'élément page comprend à gauche les éléments enfants, au centre l'image de la page avec les éléments « Paragraph » et « Text Line » en surbrillance et à droite les métadonnées associées à la page.*

Figure 2 – *Exemple d'un acte contenu par le fichier JSON sous sa forme initiale.*

2.3 Des métadonnées prêtes à l'import

2.3.1 Alignement des données issues d'Arkindex et des inventaires

Les inventaires des registres du Trésor des chartes [10] ont donc été utilisés pour extraire les différentes données permettant de décrire ces actes et les aligner avec celles issues de la plateforme Arkindex. Ensuite, Sergio Torrès Aguilar a utilisé la segmentation réalisée par Paul Chaffenet et José Ramon Prieto pour l'aligner avec la segmentation des inventaires réalisée par Sébastien Hamel et former un fichier JSON au sein duquel les actes ont été ordonnés par volume puis par leur rang au sein du volume.

Les actes se présentent sous la forme d'un élément dictionnaire (*cf.* figure 2) dont les clefs sont le numéro du volume, le numéro de l'acte dans le volume, le folio de début de l'acte, le folio de fin de l'acte, la langue, la référence de l'inventaire, le numéro de l'acte dans cet inventaire, le nom de l'image dans la BVMM, la date, la description de l'acte par l'inventaire et plusieurs numérotations provisoires utilisées au cours de ce travail. Une dernière clef contient sous forme d'une liste de dictionnaires l'ensemble des données sur les zones de texte qui forment ces actes. Ces données sont issues de la segmentation en acte de la transcription automatique des images des manuscrits [11]. Chaque zone de texte est composée de la transcription ligne à ligne, de la référence de l'image, des coordonnées de la zone dans l'image, de l'url correspondant à ces coordonnées et de la partie de l'acte que

10. Jean Glénisson, Jean Guerout, Jules Viard, Aline Vallée-Karcher et Henri-Frédéric Jassemin, *Registres du Trésor des chartes : inventaire analytique*, dir. Robert Fawtier, 6 t., Paris, France, 1958. Pour une description plus précise du contenu des instruments de recherche, *cf.* chapitre 3.

11. Il s'agit ici d'une transcription inédite réalisée à l'aide du logiciel Transkribus.

représente cette zone. Cette dernière métadonnée peut prendre plusieurs formes : « AC » pour les actes composés d'une seule zone de texte, « AI », « AM » et « AF » pour les actes composés de plusieurs zones de textes successives ou « AS » pour les zones de textes qui ne sont pas à la suite des zones précédentes.

2.3.2 Normalisation des éléments

Afin d'envisager l'import de toutes ces données au sein d'Arkindex [12], nous avons été confrontés au problème de la normalisation des données. Par exemple, les dates peuvent se présenter sous différentes formes :

'Date': '1317, 16 février'

'Date': '10 janvier 1359'

'Date': '1304-1305'

Nous avons donc écrit un script python contenant un arbre de décision pour transformer ces dates en fonction de la manière dont elles se présentent [13]. Une fois normalisée, la date prend la forme d'une liste contenant soit un élément dictionnaire unique dans le cas d'une date précise :

```
'Date-normalisee': [
    {
    'type': 'when',
    'year': 1359,
    'month': 1,
    'day': 10
    }
]
```

12. Nous décrirons plus précisément ce travail au chapitre 8.
13. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Normalize_json/normalize_date.py.

soit deux éléments dans le cas d'un intervalle :

```
'Date-normalisee': [
    {
    'type': 'notBefore'
    'year': 1304,
    'month': null,
    'day': null
    },
    {
    'type': 'notAfter',
    'year': 1305,
    'month': null,
    'day': null
    }
]
```

Un certain nombre de formes uniques ont également été repérées au cours de cette étape puis normalisées manuellement, comme par exemple :

```
'Date': '1323, 3 septembre (corriger 13 septembre ?)'
```

La normalisation des langues a suivi un processus similaire car elles se présentaient initialement sous plusieurs formes :

```
'Language': 'lat. et fr.'
```

```
'Language': 'francais'
```

Nous avons ici aussi rédigé un script python pour normaliser ces données sous une forme unique[14]. Le résultat se présente donc sous la forme d'un élément dictionnaire contenant la norme utilisée[15] et la liste des langues reconnues encodées en fonction de cette norme :

```
'normalized_language': {
    'norme': 'iso-639-3',
    'language': [
        'lat',
        'frm'
    ]
}
```

La dernière étape de ce travail consistait à revoir les numérotations provisoires établies précédemment. Le contenu du fichier a en effet été modifié au cours du temps, ce qui a affecté leur forme. L'attribut « provisory_index_2 » s'est ainsi trouvé avec une valeur vide à plusieurs reprises. Au contraire, des valeurs « bis » ont été ajoutées à l'attribut « provisory_index_1 » qui n'est donc pas complètement linéaire :

'Provisory_index_1': '8774'

'Provisory_index_1': '8774bis'

'Provisory_index_1': '8775'

Nous avons donc créé un nouvel attribut « Provisory_index_3 » permettant d'associer chaque acte à un entier unique correspondant à son rang dans le fichier :

'Provisory_index_3': 8774

'Provisory_index_3': 8775

'Provisory_index_3': 8776

Conclusion

Ce chapitre nous a permis d'aborder les différents éléments disponibles grâce aux recherches réalisées au sein du projet Himanis et dans les opérations qui lui ont succédé. Ces recherches ont abouti au développement de modèles de REM et de REN applicables

14. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Normalize_json/normalize_language.py.

15. Pour la liste des codes décrits par la norme iso-639-3, *cf.* https://fr.wikipedia.org/wiki/Liste_des_codes_ISO_639-2.

```
{
  "Acte_70": {
    "Provisory_index_1": "2311",
    "Volume": "JJ047",
    "Act_N": "70",
    "Folio_start": "48r",
    "VOL_FOL_START": "JJ047_48r",
    "Folio_end": "",
    "Language": "lat.",
    "Inventory_Name": "Paris_AN_JJ_inventaire_JJ37-50.xml",
    "Inventory_Nr": "1598",
    "ImageStart_UPVLC_BVMM": "FRCHANJJ_JJ047_0053R_A.TIF",
    "ImageStart_READ": "FRCHANJJ_JJ047_0053R_A.TIF",
    "F": "",
    "Provisory_index_2": 2308.0,
    "Text_Region": [
      {
        "Reference": "FRCHANJJ_JJ047_0053R_A",
        "type_act": "AC",
        "Region": "TextRegion_1501749765009_41",
        "Graphical_coord": "2507,2012 0,2012 0,658 2507,658",
        "Address_bvmm": "https://iiif.irht.cnrs.fr/iiif/France/Paris/Archives_nationales/FRCHANJJ_JJ047/DEPOT/FRCHANJJ_JJ047_0053R_A/0,658,2508,1355/600,/0/default.jpg",
        "Texte": [
          "littera regis super amortizacione viginti librates terre facta J de castello, pro quadam capellania per ipsum fundum apud Vienam",
          "Dei gracia Francorum rex. Notum facimus universis, presentibus et futurs, quod cum.",
          "de Castello, dominus de Vienna, miles, quamdam capellaniam apud Viemna fundare et",
          "eamdem de Nignili libratis terre seu redditus annui capiendis in et super terra seu redditibus",
          "quos idem miles habere noscitur in vulla de Challoma in parrochia de Bailliaco, ac super censibus",
          "uis de Vienna, dotare intendat, nos ipsius militis in hac parte laudabile, propositum com",
          "mendantes, ob nostre progenitorumque nostrorum incitereque memorie Johane Francu et Navarre",
          "regine, quondam consortis nostre carissime annuar remedium et salutem concedimus, ut ea",
          "pellam predicte capellanie, qui pro tempore fuerint predictas viginti libratas Parisiensius, terre seu red",
          "dicus annui tenere possint percipere et habe pacifice et quiete, sine coactione vendendi",
          "vel extra manum suam ponendi aut prestandi nobis vel successorbus nostris pro premissis financiam",
          "qualemcunque. Nolumus tamen quod prefatu capellami aliquam justiciam habeant quomodolibet in",
          "premissis. quod ut finum et stabile permaneat in futurum, presentibus litteris nostrum fecimus",
          "apponi sigillum. Salvo in aliis jure nostro et quolibet alieno. actum apud Vivarium in",
          "Gniaaonno domini m. ccc. decimo mense mprcto. P. . de Virns",
          "J. de Templo"
        ]
      }
    ],
    "Date": "1311, mars",
    "Regeste": "Amortissement — pour le salut des âmes du roi, de ses aïeux et de feue la reine Jeanne — en faveur du desservant de la chapellenie que Jean du Châtel, seigneur de Vienne, chevalier, se propose de fonder à Vienne, des 20 livrées par. de terre ou de revenu, dont ledit Jean veut doter ladite chapelle",
    "Date-normalisee": [
      {
        "type": "when",
        "year": 1311,
        "month": 3,
        "day": null
      }
    ],
    "normalized_language": {
      "norme": "iso-639-3",
      "language": [
        "lat"
      ]
    },
    "Provisory_index_3": 2315
  }
},
```


Figure 3 – *Exemple d'un acte contenu par le fichier JSON après normalisation des éléments*

aux textes médiévaux. Par la suite, ces travaux ont permis de proposer une transcription automatique des registres du Trésor des chartes et un travail de transformation a été réalisé pour associer la structure physique à la structure logique des textes et permettre un alignement de cette transcription avec les données présentes dans les inventaires. Après quelques modifications réalisées pendant le stage, nous disposons maintenant d'un fichier complet décrivant l'ensemble du contenu des registres JJ 35 à JJ 91. La figure 3 présente un exemple du format final des actes décris dans ce fichier.

Chapitre 3

Legacy Metadata : numérisation et exploitation des instruments de recherche

Comme nous l'avons vu, les travaux réalisés à partir du projet Himanis se sont appuyés à plusieurs reprises sur les instruments de recherche qui concernent le Trésor des chartes, que ce soit pour la construction du modèle de REM à partir des éditions de Paul Guérin ou pour la description du contenu des registres à partir des inventaires systématiques. Ces ouvrages contiennent en effet tout un ensemble de données disponibles pour appréhender ces textes. Conçus selon une structure logique facilement identifiable, ces outils de travail forment une masse de *legacy data* permettant d'associer chaque élément de la structure logique à toute une série de métadonnées. Leur traitement numérique prend donc toute sa place dans le cadre de la recherche sur la lecture automatique des registres et compose à ce titre le principal matériel sur lequel s'est appuyé notre travail de stage.

Ce chapitre sera consacré à l'utilisation des instruments de recherche du Trésor des chartes dans le cadre des travaux initiés par le projet Himanis. Nous présenterons dans un premier temps les différents outils disponibles et leurs lacunes. Puis nous exposerons les transformations réalisées à partir de ces ouvrages afin de structurer leur contenu sous format numérique. Pour finir, nous décrirons le projet de structuration de l'index dans lequel notre stage s'est inséré.

3.1 Description du matériel disponible

3.1.1 Inventaires systématiques et géographiques

Corpus incontournable pour aborder le pouvoir des rois de France à la fin du Moyen Âge, le Trésor des chartes a été l'objet de nombreuses études qui ont permis d'en diffuser largement le contenu. Ces travaux ont été complétés par plusieurs projets archivistique de description systématique des registres pour faciliter l'accès au contenu de ce fonds [1]. L'outil le plus complet et le plus précis est l'inventaire analytique des registres du Trésor des chartes publié par les Archives Nationales entre 1958 et 1999 [2]. Il propose une analyse systématique des actes contenus dans les registres JJ 37 à JJ 79B avec le rang de l'acte dans l'inventaire, les dates de temps et de lieu, un résumé de l'acte, le numéro de l'acte dans le volume, le folio de début (et parfois de fin) et de potentiels renvois (*cf.* figure 4).

Ces ouvrages disposent également d'une suite sous forme manuscrite ou dactylographiée pour les registres JJ 80 à JJ 98 réalisée par Y. Lanhers, A. Vallée, S. Clémencet et P. Luc entre 1945 et 1985 [3]. Le travail d'inventaire systématique ne recouvre cependant pas l'ensemble du corpus et les seuls instruments disponibles pour les registres de la fin du Moyen Âge sont les inventaires thématiques qui recensent systématiquement les actes selon un axe précis. Les plus complets sont trois inventaires géographiques rassemblant les actes qui concernent la Gascogne, le Languedoc, le Rouergue et la Loire Moyenne [4].

3.1.2 Documentation complémentaire : inventaires papier, éditions et indexations

Un autre moyen de compléter ces lacunes est d'utiliser les instruments de recherche anciens dressés au XVIII^{e} siècle et complétés au XIX^{e}. Leur niveau de détail est cependant assez faible puisqu'ils se contentent de compiler les tables des actes contenues dans les registres [5]. On retrouve également des informations précieuses dans les éditions partielles des actes du Trésor des chartes. Outre celle de Guérin citée plus haut, il en existe 4

1. Dominique Stutzmann, Jean-François Moufflet et Sébastien Hamel, "La recherche en plein texte dans les sources manuscrites médiévales…".

2. Jean Glénisson, Jean Guerout, Jules Viard, *et al.*, *Registres du trésor des chartes…*

3. Archives Nationales, IR 49 à IR 59, IR 5127 et IR 50000.

4. Charles Samaran et Pierre Rouleau, *La Gascogne dans les registres du Trésor des chartes*, Paris, 1966 (Collection de documents inédits sur l'histoire de France, Vol. 4) ; Yves Dossat, Anne-Marie Lemasson et Philippe Wolff, *Le Languedoc et le Rouergue dans le Trésor des chartes*, Paris, 1983 (Collection de documents inédits sur l'histoire de France, 16) ; Bernard Chevalier, *Les pays de la Loire moyenne dans le Trésor des chartes : Berry, Blésois, Chartrain, Orléanais, Touraine, 1350-1502 Archives nationale, JJ 80-235*, avec la coll. d'Archives nationales, Paris, 1993 (Collection de documents inédits sur l'histoire de France, 22).

5. Inv. ms. de JJ 1 à JJ 264, dressé au début du XVIIIe s., révisé et complété par A. Longnon et A. Coulon, 1880-1900, copie en quatre vol., IR421-IR424 et IR 429 ; Acta omissa. Inv. somm. ms. de la série JJ, actes non mentionnés dans l'inv. de A. Longnon et A. Coulon, 1898-1900, IR33-IR34.

1598. 1311, mars. Le Vivier-en-Brie.

Amortissement — pour le salut des âmes du roi, de ses aïeux et de feue la reine Jeanne — en faveur du desservant de la chapellenie que Jean du Châtel, seigneur de Vienne, chevalier, se propose de fonder à Vienne, des 20 livrées par. de terre ou de revenu, dont ledit Jean veut doter ladite chapellenie, à percevoir dans le village de Chaillot, la paroisse de Bailly et sur les cens de Vienne. Mais le roi entend que le chapelain n'ait aucune justice. *Per dominum de Virmis. J. de Templo.* (Fol. 48, n° 70.)

1599. 1310, mai. Compiègne.

Vidimus et confirmation — avec amortissement pour le salut des âmes du roi, de ses aïeux et de feue la reine Jeanne — de la donation entre vifs [1310, 30 janvier (1)] par Michel de Morienval, chanoine de Laon, aux prieur et frères de Royallieu-lès-Compiègne, au diocèse de Soissons, ordre du Val-des-Écoliers, de sa vicomté de Pierrefonds avec toutes ses dépendances, excepté ce qu'il possède à Retheuil et dans son terroir, à charge de célébrer divers services.

Le roi fait en outre remise auxdits religieux, qui sont ses chapelains, des 6 d. de cens que ledit Michel lui devait chaque année pour sa vicomté, mais se réserve la haute justice. *Per dominum Regem. Maillardus.* (Fol. 48-48 v°, n° 71.)

1600. 1311, mars. Les Carrières, près Paris.

Amortissement des acquisitions réalisées, il y a environ 30 ans, par les maître et frères de l'Hôtel-Dieu de Bourges : un manoir dit La Motte-de-Nohant-en-Goût, dans leur censive; un pré et une terre à Nohant-en-Goût, mouvant des censives du roi et de la dame de Contremoret, et la grange de Troussin; le tout valant environ 20 l. par. de revenu. *Per dominum Regem. Maillardus.* (Fol. 48 v°-49, n° 72.)

(1) Ms. : *feria tertia ante festum [Purificationis] beate Marie Virginis.* Le mot *Purificationis* est fourni par le cartulaire de Royallieu (Bibl. nat., lat. 5434, fol. 28; éd. Paul Guynemer, Compiègne, 1911, in-4°, n° XX, p. 47).

1601. 1311, mars. Le Vivier-en-Brie.

Sur la prière de Charles de Valois, seigneur de Tournan-en-Brie, rattachement à sa châtellenie de Tournan de toute la partie de la chaussée d'un vivier ou étang — récemment créé dans sa terre près de sa maison du Vivier-en-Brie, aux limites de ladite châtellenie — qui empiète sur la châtellenie royale de Melun, avec un arpent de terre tout autour. *Per dominum Regem. Guido.* (Fol. 49, n° 73.)

Original : J 377, n° 6.

1602. 1311, février. Gaye-en-Champagne.

Don — pour le salut des âmes du roi, de ses aïeux et de feue la reine Jeanne — aux chapelains royaux les prieur et religieux de Royallieu-lès-Compiègne, ordre du Val-des-Écoliers, de deux arpents de bois contigus à leur terrain de Vieux-Moulin pour élargir et rectifier la clôture dudit terrain, de manière à ce qu'elle ne fasse ni angles, ni brisures, et du ruisseau qui vient de Pierrefonds (auj. le rû de Berne) avec la pêcherie et l'usage de cette pêcherie, le roi se réservant la haute justice; avec injonction aux forestiers de Cuise de faire borner les deux arpents précités. *Per dominum Regem. Maillardus.* (Fol. 49-49 v°, n° 74.)

V. n° 1534.

1603. 1311, mars. Paris.

Assiette sur les biens et revenus de Challeau, près Moret-sur-Loing, qui furent à Guillaume de Villebéon, chevalier, des 80 l. par. de rente que Gilles Grange, écuyer du roi, percevait jusqu'alors, par don du roi, sur les revenus de la prévôté de Paris. La terre de Challeau avait été saisie, avec d'autres biens, sur ledit Guillaume pour non-payement d'une amende de 3.000 l. par. dont il avait été frappé pour délits dans les forêts du roi, et mise aux enchères sans résultat; par la suite, sur la prière de Guillaume et de sa femme, le roi avait décidé de réunir à son domaine cette terre, dont les revenus seraient déduits de ladite amende de 3.000 l. Une prisée (s.d.n.l.) — dont le texte,

Figure 4 – *Exemple d'analyses contenues dans l'inventaire systématique des registres du Trésor des chartes publié par les Archives Nationales.*

autres consacrées à Amiens, Paris pendant le règne de Philippe VI et la Normandie et Paris pendant l'occupation anglaise au début du XV^{e} siècle [6]. Un autre moyen d'accéder au contenu des registres est d'utiliser un index des sujets, noms de personnes et noms de lieux. La plupart des ouvrages que nous venons de citer en contiennent, à l'exception notable du tome II de l'inventaire analytique des registres du Trésor des chartes pour lequel un index des noms de lieux est en préparation. Enfin, deux instruments permettent d'aborder les actes des registres à partir des noms de lieux, personnes et sujets pour les registres de Charles VI et Henri VI [7]. Ils sont disponibles respectivement sous forme de fiches papiers et d'un fichier texte.

Ainsi donc, le matériel disponible pour étudier ce corpus est à la fois conséquent et très lacunaire. Alors que des pans entiers sont connus de manière précise grâce aux inventaires systématiques et géographiques, un grand nombre de registres ne disposent que d'instruments très primaires pour repérer le contenu qui intéresse le chercheur et éviter un dépouillement systématique de ces archives. L'automatisation de l'analyse archivistique est donc un enjeu essentiel car elle permet de rendre accessible la lecture des registres et d'envisager leur dépouillement ponctuel et précis. L'objectif du processus est ainsi de faciliter la recherche sur certains sujets comme par exemple l'étude du pouvoir du roi de France.

3.2 Formats de l'information

3.2.1 Segmentation en XML

Un certain nombre des outils que nous avons décrits ont déjà été numérisés et convertis par les Archives Nationales sous format EAD, une norme utilisant le langage XML pour structurer des descriptions de manuscrits ou de documents d'archives. Ils présentent cependant quelques variations dans leur format, c'est pourquoi les membres du projet Himanis ont fait ici le choix de tout transformer sous un format homogène utilisant la norme TEI, un autre format basé sur le langage XML. L'objectif est de pouvoir combiner

6. Paul Guérin, *Actes Royaux du Poitou (1302-1464)…* ; Jules Viard, *Documents parisiens du règne de Philippe VI de valois : 1328-1350*, avec la coll. de Société de l'histoire de Paris et de l'Ile-de France, Paris, 1899 (Société de l'histoire de Paris et de l'Ile-de-France) ; Auguste Longnon, *Paris pendant la domination anglaise (1420-1436), documents extraits des registres de la chancellerie de France, par Auguste Longnon*, Paris, 1878 ; Paul Le Cacheux, *Actes de la chancellerie d'Henri VI concernant la Normandie sous la domination anglaise (1422-1435), extraits des registres du Trésor des chartes aux Archives nationales, publiés avec introductions et notes*, Rouen, 1907 ; Edouard Maugis, *Documents inédits concernant la ville et le siège du bailliage d'Amiens extraits des registres du Parlement de Paris et du Trésor des chartes : XIVe-XVe siècle (1296-1471)*, Amiens Paris, 1908 (Mémoires de la Société des antiquaires de Picardie. Documents inédits concernant la province, t. 17, 19 et 20).

7. Archives Nationales, IR 1810 et Projet d'inventaire par C. Gut, 2010.

```xml
<text xml:id="JJ27199">
   <front>
      <head>1552</head>
      <docDate>
         <date when="1310-12">1310, décembre</date>
         <origPlace>Abbaye Notre-Dame du Lys près de Melun</origPlace>
      </docDate>
      <argument>
         <p>Licence à M<hi rend="sup">e</hi> Raoul de Meulan, clerc du roi, chanoine de
            Paris, de transporter à quiconque sa maison, avec dépendances, dans la Cité
            de Paris, près du port Saint-Landry, avec amortissement pour le ou les
            bénéficiaires de ce transport. Le roi se réserve toutefois la haute et basse
            justice et tout cens ou revenu pouvant lui appartenir. <quote type="extrasigillum">
               <hi rend="italic">Per dominum Regem. Chalop</hi>.</quote>
         </p>
      </argument>
      <div type="tradition" org="uniform" sample="complete">
         <listWit sortKey="copie">
            <witness>
               <msDesc status="draft">
                  <msIdentifier>
                     <repository>Archives nationales</repository>
                     <idno>JJ47</idno>
                  </msIdentifier>
               </msDesc>
               <locus scheme="folio">17r</locus>
               <idno>24</idno>
               <lang opt="false">lat.</lang>
            </witness>
         </listWit>
      </div>
      <index>
```

Figure 5 – *Exemple d'analyse d'un acte du Trésor des chartes encodée sous format XML-TEI.*

les transcriptions des textes avec les métadonnées les concernant [8].

Comme présenté par la figure 5, chaque acte y est encodé au sein d'un élément <text> disposant d'un attribut @xml:id permettant de lui associer un identifiant unique. Les balises enfants contiennent ensuite le rang de l'acte dans l'inventaire, les dates de lieu et de temps, l'analyse du texte et sa manifestation physique dans les registres du Trésor des chartes. Les fichiers ont par la suite été mis en ligne dans le répertoire Github du projet [9] et leur contenu a servi à la constitution d'un fichier sous format JSON décrivant les actes des registres inventoriés [10].

3.2.2 Numérisation de l'index

Ce travail de segmentation du contenu des instruments de recherche pour l'organiser sous le format XML-TEI a également été réalisé pour les différents index qui accompagnent ces ouvrages. Chaque entrée est insérée dans une balise <p> qui peut contenir une ou

8. Dominique Stutzmann, Jean-François Moufflet et Sébastien Hamel, "La recherche en plein texte dans les sources manuscrites médiévales…".
9. https://github.com/oriflamms/himanis/tree/master/Inventories.
10. *Cf.* chapitre 2.

```
<p>
   <seg>
      <term>Empire (mère et mixte)</term>
      <num>
         <idno>1022</idno>, <idno>1032</idno>, <idno>1504</idno>, <idno>1508</idno>,
            <idno>1738</idno>, <idno>1761</idno>, <idno>1807</idno>, <idno>1901</idno>,
            <idno>1956</idno>, <idno>1965</idno>, <idno>2041</idno>, <idno>2082</idno>,
            <idno>2091</idno>, <idno>2138</idno>, <idno>2147</idno>, <idno>2188</idno>,
            <idno>2238</idno>.</num>
   </seg>
</p>
```

Figure 6 – *Exemple d'une entrée d'index encodée sous format XML-TEI.*

plusieurs balises <seg> en fonction du nombre de sous-entrées présentes. L'entité et sa description sont ensuite insérées dans une balise <term> et les renvois vers les actes décrits par l'index sont rassemblés dans une balise <num>. Chaque renvoi est ensuite inséré dans une balise <idno> (*cf.* figure 6).

Une des difficultés de ce travail a été de reconstituer les sous-entrées de l'index. Le corps de certaines d'entre elles est en effet directement dépendant de celui de l'entrée principale, ce qui rend délicat sa manipulation comme une entité isolée. Par exemple l'entrée :

Récupérations : des arrérages de bois

est suivie de la sous-entrée :

des droits royaux

Cette dernière n'a pas de sens prise isolément, il faut donc ici reconstituer la notion dans son ensemble : « Récupérations des droits royaux ». Ce travail de reconstitution a été réalisé en s'appuyant autant que possible sur la structure de l'index : les « : » servent ici de séparateur entre l'entrée principale et les sous-entrées, tandis que ces dernières sont systématiquement séparées par un « ; — ». Le travail réalisé a donc consisté à éliminer le premier séparateur et à transformer le second en « , — » afin de bien marquer les entrées dont le corps dépend de celui d'une entrée générale. Pour les cas exposés, la segmentation en TEI donne le résultat suivant :

<term>Récupérations des arrérages de bois</term>

<term>Récupérations , — des droits royaux</term>

3.3 Les index comme compléments aux métadonnées décrivant les actes ?

3.3.1 Le projet : mettre à plat les entrées d'index

L'objectif de ce travail était ainsi de faciliter l'association de chaque entrée d'index aux entrées d'inventaire qui lui correspondent. En effet, le but du processus était d'utiliser cet index comme vérité terrain pour développer l'apprentissage du liage d'entités. L'index a donc été mis à plat en fonction de cette logique : toutes les balises <idno> dans les entrées d'index ont été utilisées pour identifier l'entrée d'inventaire contenant la notion décrite par cette entrée. Afin que le lien soit bien manifeste, les entrées d'inventaires décrites en TEI ont été enrichies d'une balise <index> contenant l'ensemble des entrées d'index renvoyant vers l'acte en question. La figure 7 donne un exemple de ces entrées d'index encapsulées dans des éléments <term> et disposant d'un attribut @type dont la valeur est « place », « person » ou « subject » en fonction du type d'entrée dont il s'agit [11].

Le projet était d'inscrire directement les entrées d'index au sein des métadonnées de l'acte afin qu'elles puissent être comprises comme autant d'entités nommées contenues dans le texte. Cela devait faciliter l'association entre texte et entités nommées et façonner la vérité terrain permettant l'apprentissage du liage d'entités. Cette vision de l'index mis à plat comme une extension de l'inventaire jointe aux métadonnées des actes a largement guidé la segmentation des entrées car il fallait que chacune se suffise à elle-même pour qu'on puisse la comprendre comme une entité nommée à reconnaître.

3.3.2 Perte de qualité et de complexité dans les données

Cette vision des choses n'est cependant pas sans poser quelques problèmes dans la mise en œuvre du liage d'entités. L'index ainsi mis à plat correspond en effet moins à une base de connaissances proprement dite qu'à un ensemble d'entités nommées associées au texte. Elle ne permet pas l'apprentissage du liage puisqu'elle ne se fonde pas sur un référentiel en tant qu'entité autonome du texte. Les entités nommées sont ici simplement répétées aux lieux où elles ont été reconnues et ne disposent pas de métadonnées pour les décrire (définition du concept, synonymes, coordonnées géographiques, ...). L'absence de référentiel unique et autonome ne permet donc pas le traitement *a posteriori* des entrées pour les segmenter en différentes parties ou les enrichir par alignement avec d'autres référentiels. C'est ainsi que l'absence d'identifiant unique pour les entrées d'index condamne le travail à rester dans l'état où il est au moment de la mise à plat alors que le liage d'entités peut aussi conduire à augmenter le nombre d'entités nommées identifiées dans

11. A propos de la typologie des entrées, *cf.* chapitre 4.

```
    <term type="person">Robert II, comte d'Artois</term>
    <term type="place">Vendôme [Loir-et-Cher],
       archidiacre</term>
    <term type="person">Alain de Lamballe, chanoine de Laon</term>
    <term type="place">Antoing [Belgique, prov. Hainaut], chanoine</term>
    <term type="place">Artois [province], comte</term>
    <term type="place">Artois [province], v. Mahaut d'Artois</term>
    <term type="place">Artois [province], Mahaut d'</term>
    <term type="place">Cambrai [Nord]</term>
    <term type="place">Cambrai [Nord], évêque</term>
    <term type="place">Chartres [Eure-et-Loir], archidiacres</term>
    <term type="person">Festu (Simon), clerc et trésorier du roi, archidiacre de
       Vendôme, évèque de Meaux</term>
    <term type="place">Lamballe [Côtes-du-Nord, ar. Saint-Brieuc], Alain de</term>
    <term type="place">Lamballe [Côtes-du-Nord, ar. Saint-Brieuc], Mathias
       de</term>
    <term type="place">Laon [Aisne], chanoines</term>
    <term type="person">Mahaut d'Artois, comtesse palatine d'Artois et de
       Bourgogne, dame de Salins, femme d'Othon IV</term>
    <term type="place">Marigny [Seine-Maritime, con Gournay, cne Cuy-Saint-Fiacre
       ou Dampierre], Philippe de</term>
    <term type="person">Mathias de Lamballe, chanoine d'Antoing</term>
    <term type="person">Paris, actes passés à</term>
    <term type="person">Philippe de Marigny, clerc du Roi, commissaire aux
       nouveaux-acquêts dans la prévôté de Paris, évêque de Cambrai</term>
    <term type="person">Pierre de Grez (Me), chancelier du roi Louis de Navarre,
       chantre de Paris</term>
    <term type="person">Robert II, comte d'Artois</term>
    <term type="place">Vendôme [Loir-et-Cher], archidiacre</term>
    <term type="subject">Concessions de pleins pouvoirs</term>
    <term type="subject">Hommes [sujets, tenanciers des seigneurs]</term>
    <term type="subject">Parisis [monnaie réelle] (bons)</term>
    <term type="subject">Procurations (lettres de)</term>
    <term type="subject">Procureurs des communautés ou des particuliers</term>
    <term type="subject">Promesses des communautés ou des particuliers</term>
    <term type="subject">Ratifications par les parties contractantes</term>
    <term type="subject">Sentences du roi</term>
    <term type="subject">Sentences arbitrales</term>
    <term type="subject">Sujets des seigneurs</term>
    <term type="subject">Tournois [monnaie réelle] , anciens (petits)</term>
    <term type="subject">Trésoriers du roi de France à Paris, au Louvre ou au
       Temple</term>
</index>
```

Figure 7 – *Exemple d'une liste d'entrées d'index contenue dans une balise <index> insérée dans l'entrée d'inventaire sous format XML-TEI qui lui correspond.*

les textes.

De plus, l'index se caractérise par un ensemble de relations entre les entrées : associations entre entrées générales et sous-entrées, renvois contenus dans les données, ou encore entrées rejetées vers une autre entrée. Or ces relations, qui ne peuvent être reportées par la mise à plat de l'index, sont essentielles à plusieurs titres pour construire le travail de liage d'entités. Par exemple, les entrées rejetées ne contenant pas de renvois vers l'inventaire mais uniquement un renvoi vers une autre entrée d'index sont complètement invisibilisées lors de ce travail de mise à plat alors qu'elles fournissent autant d'alias d'une même notion mobilisables au moment de la construction de la base de connaissances. De même les renvois contenus dans les entrées et les relations entre entrées générales et sous-entrées permettent d'appréhender l'association entre les entités et les possibles synonymies ou co-occurrences de celles-ci, elles sont donc nécessaires à la bonne réalisation du liage. Il apparaît donc que l'utilisation de l'index comme un ensemble de métadonnées associées aux actes ne fournit pas un cadre optimal à la mise en œuvre du liage d'entités. La bonne compréhension du projet initial en vue duquel les données ont été manipulées est cependant essentielle pour saisir l'enjeu des étapes déjà réalisées et aborder celles que nous avons menées pour tenter d'avancer dans une direction plus propice à la constitution d'un référentiel.

Conclusion

Nous avons vu que le Trésor des chartes dispose d'un certain nombre d'instruments de recherche. Ces ouvrages sont une mine d'information pour développer la lecture automatique des textes puisqu'ils fournissent un grand nombre de métadonnées décrivant le contenu des registres. Les membres du projet Himanis ont donc travaillé à leur numérisation et leur segmentation pour les inclure dans un modèle sous format XML-TEI. Cette transformation a placé l'acte comme élément central de description du corpus, chacun étant associé à l'ensemble des métadonnées disponibles pour le décrire. Ce travail a cependant trouvé sa limite au moment de la mise à plat de l'index comme devant accompagner l'ensemble de métadonnées. Les entrées d'index ne peuvent en effet pas se réduire à un ensemble d'entités nommées reconnues dans un texte et doivent constituer un référentiel propre afin de modéliser la complexité de leur définition et de leurs relations entre elles et avec les textes, puis assurer l'apprentissage du liage d'entités à partir de données d'entraînement. Nous avons donc consacré un temps conséquent du stage à travailler sur ces entrées d'index pour comprendre leur complexité et la meilleure manière de la prendre en compte pour leur utilisation future. Dans sa finalité, le projet serait de permettre de compléter les lacunes de ces instruments par l'indexation automatique des registres dont le contenu n'a pas été décrit de manière systématique.

Conclusion partielle

Cette première partie nous a permis d'exposer dans son ensemble le contexte de travail dans lequel nous avons réalisé notre stage. Notre développement a commencé par un premier chapitre consacré à l'état de l'art sur le liage d'entités et son utilisation dans l'étude des documents patrimoniaux. Ensuite, nous avons dédié un second chapitre à la description des différents travaux réalisés au sein du projet Himanis ou à sa suite et des résultats disponibles pour poursuivre les recherches sur le liage d'entités. Nous avons enfin terminé cet exposé par un troisième chapitre portant sur l'utilisation des instruments de recherche disponibles comme *legacy metadata* permettant de décrire le contenu des registres ainsi que les limites de cette logique.

Tout cela nous a permis de voir à quel point le corpus traité par le projet Himanis fournit un terreau favorable au développement du liage d'entités. Cette technique est en effet difficile à mettre ne œuvre pour des documents anciens car elle nécessite une base de connaissances dédiée. Or les instruments de recherches déjà numérisés par les acteurs du projet Himanis facilitent largement la constitution de cette base de connaissances. Les index disponibles se prêtent à la constitution d'un référentiel vers lequel pointer au moment de la réalisation du liage. En plus de fournir un matériel idéal pour l'entraînement d'un modèle, le Trésor des chartes dispose également de tout un ensemble de textes non-indexés qui se prêtent à la mise en œuvre et à l'évaluation du modèle.

Deuxième partie

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

L'objectif final du stage que nous avons réalisé étant de préparer l'apprentissage du liage d'entités à partir des données issues du projet Himanis, notre action s'est principalement concentrée sur la création d'un référentiel pouvant servir de base de connaissances. Nous consacrerons donc cette deuxième partie au travail de modélisation et de formalisation de ce référentiel à partir d'un index papier, activité qui a occupé la majeure partie de notre stage. Nous nous sommes concentrés ici sur l'index du premier volume de l'inventaire analytique des registres du Trésor des chartes portant sur les volumes JJ 37 à JJ 50 et paru en 1958. Pour rendre compte de ce travail, nous décrirons dans un premier chapitre les différentes difficultés rencontrées pour comprendre et manipuler cet instrument de recherche. Puis nous dédierons un second chapitre à l'analyse des relations entre les entités. Nous consacrerons enfin un troisième chapitre à la transformation de l'index en une base de données relationnelle.

Chapitre 4

Appréhender les *legacy data*

La mobilisation d'un instrument de travail papier pour construire un référentiel numérique implique de prendre le temps de comprendre les données qu'on utilise. Leur structuration peut en effet suivre une logique complexe, passant de l'utilisation de caractères séparateurs à un grand nombre de renvois implicites. Or cette utilisation n'est pas toujours systématique et la quête de la perfection peut parfois demander un temps de relecture très conséquent pour repérer toutes les erreurs et incohérences qui peuvent se glisser dans les données de manière plus ou moins répétitive. Dans le cadre de ce travail, l'index que nous avons manipulé est un ouvrage imprimé qui a été numérisé, transcrit par ROC et préalablement structuré sous la forme d'un fichier au format XML-TEI. Il nous a donc fallu prendre le temps de saisir le projet qui a guidé les différentes étapes de ce travail [1] et son influence sur l'état des données. Ce temps est nécessaire pour faire le lien entre le fichier que nous avons manipulé et le contenu intellectuel de l'index que nous voulons reporter dans le référentiel.

Ce chapitre sera donc consacré aux différentes difficultés auxquels nous avons été confrontés dans l'analyse de ces *legacy data* et aux solutions mises en œuvre pour y faire face. Pour cela, nous décrirons dans un premier temps les différents types de contenus présents dans les entrées. Puis nous reporterons les différentes exceptions qu'il nous a fallu traiter pendant notre travail. Enfin, nous nous intéresserons aux différentes erreurs qui ont pu être mises au jour tout au long de la chaîne de traitement.

1. *Cf.* chapitre 3.

4.1 Des entrées composées de différents éléments

4.1.1 Reconnaître la typologie

Au début de notre stage, le fichier contenait 20 136 balises <term> correspondant à tous les éléments contenus dans l'index. Ces éléments se répartissent en plusieurs catégories. On retrouve tout d'abord quelques éléments titres qui sont nécessaires à l'organisation de l'index mais qui n'appartiennent pas au contenu :

<term>INDEX DES MATIÈRES</term>

<term>T</term>

Nous avons ici profité de l'étape de création d'identifiants uniques pour éliminer ces éléments de la suite du processus [2]. Nous avons cependant utilisé l'indication offerte par ces titres sur la séparation entre les deux parties de l'index. Celui-ci est en effet constitué d'abord d'un index des matières qui recense tous les sujets mentionnés dans les actes, puis d'un index des noms de personnes et de lieux.

Cette séparation a servi de base à l'établissement d'une typologie des entrées. Toutes les entrées présentes dans l'index des matières ont été associées au type « subject » tandis que les entrées de l'index des noms de personnes et de lieux ont dû être analysées pour discriminer celles qui relèvent du type « person » et celles qui relèvent du type « place ». Cette étape avait déjà été anticipée avant le début du stage à partir d'un principe simple : les entrées disposant de crochets ont été marquées comme des « place » et celles n'en disposant pas comme des « person ». Cela donne ainsi le résultat suivant :

<term type='place'>Montreuil [Pas-de-Calais]</term>

<term type='person'>Montreuil (Pierre de)</term>

Si cette séparation est pertinente dans la plupart des cas, il arrive cependant que cela conduise à un certain nombre d'erreurs :

<term type='place'>Dreux (Jean) [à Orléans]</term>

<term type='person'>Agulhou (pièce de terre dite), dans le territoire de Béziers</term>

Nous avons donc tenté de repérer les erreurs systématiques qui peuvent exister dans ces données. Pour cela, nous avons notamment cherché toutes les occurrences de « terre », « bien », « bois », « près », ... pour rétablir la valeur de l'attribut @type vers « place ».

2. Nous traiterons plus précisément toutes les étapes associées à la gestion des liens entre les entrées au chapitre 5.

Un autre problème systématique concerne les personnes dont le nom renvoie vers un toponyme :

> <term type='place'>Aigues-Vives [Lot-et-Garonne, con Monclar, cne Saint-Pastour?] (Bérenger d')</term>

Nous avons ici fait le choix de considérer ces noms comme relevant du type « person » [3]. Nous avons utilisé l'expression régulière « [A-Z].* de\) » (et ses déclinaisons avec « d' », « des », ...) pour les repérer et avons procédé par une méthode semi-automatisée pour les transformer une à une tout en gardant un regard sur la réalisation afin d'éviter les effets de bords pouvant toucher les quelques exceptions de noms de lieu correspondant à ce modèle, comme par exemple :

> <term type='place'>Bannes (mas de) [Aveyron, ar. et con Villefranche-de-Rouergue, cne Morlhon-le-Haut]</term>

Les autres erreurs repérées ont été corrigées ponctuellement au fur et à mesure de leur identification au cours de notre travail. Il est possible qu'il en existe encore.

4.1.2 Segmentation des entrées

Une fois ce travail réalisé et l'index mis sous forme d'une table [4], nous avons pu repérer le contenu des entrées pour en segmenter les différentes parties en fonction du type de données. Nous avons donc rédigé un script python permettant de réaliser cette étape de manière automatisée [5]. Les éléments mis entre crochets ont ainsi été ajoutés dans une colonne « Détails » et ceux qui sont précédés d'une virgule dans la colonne « Supplément ». Cette étape a également permis de construire une colonne « Entrée_simplifiée » construite à partir de l'entrée originelle après extraction des éléments utilisés pour les autres colonnes. L'entrée :

> Chambre aux Deniers, Camera Denariorum [terme encore employé à cette époque pour désigner la Chambre des Comptes]

a été transformée sous la forme suivante :

Entrée_simplifiée	Détails	Supplément
Chambre aux Deniers	terme encore employé à cette époque pour désigner la Chambre des Comptes	Camera Denariorum

3. Sur les différentes formes de noms de personnes, *cf.* chapitre 5.
4. Nous décrirons plus précisément cette étape au chapitre 6.
5. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Index/simplification_entrees.py.

Cette étape a également été accompagnée d'une rationalisation des sous-entrées afin d'éviter la répétition inutile des éléments contenus dans les entrées générales. L'entrée

> Cens cotage [acquitté pour un tènement en roture] , — morts [ne comportant pas de lods et ventes]

a ainsi été simplifiée afin que la colonne « Entrée_simplifiée » devienne

> Cens cotage , — morts

Les données ainsi segmentées sont disponibles dans le tableau contenant les données de l'index et mis en ligne sur le Github du projet Himanis [6].

4.2 Comment utiliser des *semi-structured data* ?

4.2.1 Des cas spécifiques tout au long de la chaîne de traitement

Le travail d'analyse et de traitement des entrées d'index nous a confronté à quelques difficultés dans l'appréhension des données. Les entrées bénéficient en effet d'une structure sensible et très utile pour comprendre les différents éléments qu'elles contiennent, mais celle-ci n'est jamais pleinement systématique. Nous avons donc fait face à plusieurs surprises tout au long de ce travail. Par exemple, les renvois explicites sont généralement séparés par des virgules lorsqu'ils sont plusieurs :

> Abandons, v. donations, renonciations, remises

Il arrive toutefois que la virgule située après un « v. » ne sépare pas deux renvois mais soit au contraire une partie de la définition de l'entrée à laquelle la chaîne de caractères fait référence :

> France , — v. Jeanne, reine de France et de Navarre

Cette situation a donc créé quelques erreurs dans le travail de segmentation automatique des renvois internes aux entrées. Il arrive même que ces derniers soient sous une forme factorisée :

> Lettres , — de grâce, v. lettres d'abolition, de rappel de ban et de rémission

> Ambianensis (ballivia, ballivus), v. Amiens (bailli, bailliage)

6. https://github.com/oriflamms/himanis/blob/master/Inventories/Systematic/Paris_AN_JJ_inventaire_JJ37-50_index.xlsx.

Il faut alors développer manuellement ces éléments pour que les liens se réalisent selon le schéma voulu.

Nous avons aussi repéré quelques incohérences orthographiques qui ont perturbé la bonne compréhension des renvois. Par exemple « v. main-morte » fait référence à une entrée contenant le même mot, mais orthographié « mainmorte ». On peut également remarquer quelques variantes dans les accords comme par exemple « v. demeure épiscopale » qui correspond à l'entrée « demeures épiscopales ». Ou encore des raccourcis comme pour l'entrée « Passy (Marie de) » qui constitue un renvoi implicite vers l'entrée « Marie, femme de Jean de Passy ».

On peut noter pour finir un certain nombre d'exceptions dans la formalisation des données comme par exemple l'entrée « Coutumes [droits et redevances] » dont les sous-entrées sont les seules à disposer de numéros :

> Coutumes [droits et redevances coutumiers , — 1. droits de péage levés sur les bateaux]
>
> Coutumes [droits et redevances coutumiers , — 2. redevances ou exactions]
>
> Coutumes [droits et redevances coutumiers , — 3. revenus des taxes sur les denrées mises en vente dans les foires et marchés]

L'une de ces sous-entrées numérotées dispose même à son tour de sous-sous-entrées. On retrouve le même type d'exceptions pour l'entrée « Paris » qui contient un nombre conséquent de sous-entrées. C'est pourquoi certains renvois souhaitant faire référence directement à celles-ci peuvent prendre la forme « v. Paris (évêché) ». Or l'entrée correspondante ne prend pas cette forme ni dans l'inventaire ni dans le fichier TEI, mais plutôt « Paris , — évêché ». Pour finir, quelques renvois sont agrémentés d'une précision sur le positionnement de l'entité correspondante : « v. dans l'index des matières » ou « v. plus bas ».

4.2.2 Analyse fine des caractères

La compréhension du contenu des données nécessite donc d'étudier finement les différents séparateurs utilisés entre les éléments et de comprendre les différentes significations qu'ils peuvent prendre. Par exemple, les éléments entre crochets peuvent à la fois servir à définir une notion :

> Acaptes [droits de mutation acquittés au seigneur par le successeur d'un tenancier défunt (dans le midi)]

à détailler sa situation :

> Montferrier [Aveyron, con Laissac, cne Bertholène]

ou à désambiguïser des entités homonymes :

Droit [ensemble des lois et des règlements]

Droit [impositions ou taxes]

Droit [privilèges, pouvoirs]

De la même manière, un renvoi vers une entrée peut être signifié par plusieurs séparateurs sans que cela n'ait une réelle signification sur le lien entre les entités :

Élargissements de prisonniers, v. liberté, relâchements

Envoyés, v. aussi ambassadeurs

Recours , — au roi, cf. recours à l'arbitrage

On trouve également quelques entrées tellement conséquentes qu'elles peuvent contenir plusieurs niveaux de détail. Il faut alors dans ce cas savoir faire la différence entre les blocs qui sont uniquement des sous-entrées successives et ceux qui constituent un niveau de détail supplémentaire dépendant d'une même sous-entrée (*cf.* Annexe A). De même, certains renvois vers l'inventaire sont présents sous forme factorisée, par exemple « 1852-1854 ». Il faut alors développer ce renvoi pour permettre la réalisation des liens entre l'entrée d'index et les entrées d'inventaire « 1852 », « 1853 » et « 1854 ».

Pour finir, on compte quelques éléments de simplification systématique des données au sein des entrées qu'il nous a fallu rétablir pour faciliter au mieux l'automatisation de la création des liens entre les entités. Par exemple l'entrée :

Acy, v. J. d', Isabelle d'

comporte deux renvois vers « J. d'Acy » et « Isabelle d'Acy ». Si le lecteur peut facilement comprendre qu'il y a une lacune à compléter lorsqu'il réalise sa recherche, l'automatisation de la procédure nécessite que le renvoi corresponde exactement à l'entrée vers laquelle il pointe. Nous avons donc pour cela identifié toutes les entrées de cette forme avec l'expression régulière « >([^,\s]*)(.*[^o]v. [A-Z].*)\s(d[^\)]1,3)(\)?)</term> » et nous les avons transformées pour qu'elles prennent la forme :

Acy, v. J. d'Acy, Isabelle d'Acy

Nous avons ici aussi réalisé cette étape de manière semi-automatisée pour éviter les éventuels effets de bord.

4.3 Multiplication des erreurs avec l'allongement de la chaîne de traitement

4.3.1 Erreurs de ROC et erreurs originelles

Les éléments que nous venons d'aborder sont autant de freins à la réutilisation des données contenues dans l'index, mais ils sont tous issus de la mise en forme de l'ouvrage par l'auteur. Cette complexité n'est cependant pas la seule que nous avons dû affronter au cours de notre stage. Nous avons aussi été confrontés à un certain nombre d'erreurs au sein des entrées qui ont également freiné notre action. Les plus fréquentes sont celles issues de la transcription du texte par ROC : confusion « l » et « i », «] » et « l » ou «] » et «) ». La première erreur peut nuire notamment dans la bonne réalisation des renvois, les deux suivantes sont plus gênantes encore car elles nuisent à la segmentation des données. Nous avons donc repéré systématiquement toutes les entrées qui possèdent un crochet ouvrant sans crochet fermant ou le contraire afin de rétablir l'état originel des données et permettre la réalisation automatique de l'étape de segmentation que nous avons décrite plus haut. Ces erreurs s'accompagnent également de quelques erreurs natives de l'index comme l'utilisation de parenthèses à la place de crochets, l'oubli de crochets fermants ou encore le mauvais rangement alphabétique des mots (*cf.* Annexe B)

Quelques erreurs se sont également glissées dans la séparation entre les entrées :

```
<seg>
    <term>Fondations pieuses de Philippe le Bel :</term>
    <num>
        (...)
        <idno>1003</idno>, <idno>1098</idno>, <idno>1117</idno>,
        <idno>1118</idno>, <idno>1170</idno>, <idno>1215</idno>,
        <idno>1279</idno>, <idno>1349</idno>, <idno>1393</idno>,
        <idno>1436</idno>, <idno>1547</idno>, <idno>1659</idno>,
        <idno>1744</idno>, <idno>1790</idno>, <idno>1814</idno>,
        <idno>1839</idno>, <idno>2113</idno>, <idno>2148</idno>,
        <idno>2153</idno>, <idno>2257</idno>, <idno>2285</idno>,
        — de ses prédécesseurs : <idno>1151</idno>, <idno>2210</idno>,
        — de la reine-mère : <idno>1321</idno>, <idno>1565</idno>,
        <idno>1939</idno>
    </num>
</seg>
```

Leur identification nous a permis de rétablir un certain nombre d'entrées qui étaient soit

cachées au sein de la balise « term » (elles sont dans ce cas difficile à repérer car il n'y a pas de séparation systématique entre les deux entrées) soit à la fin de la balise « num » (nous les avons dans ce cas trouvées avec l'expression régulière « [A-Za-z] » appliquée à la forumule xpath « num/text() »).

4.3.2 Erreurs automatiques et erreurs manuelles

Ces erreurs dans le découpage des données s'accompagnent également d'erreurs dans leur traitement automatique. Le travail de reconstitution des sous-entrées que nous avons décrit au chapitre 3 s'est accompagné d'une génération importante d'incohérences que nous avons dû appréhender et corriger. Celles-ci se présentent notamment dans le choix de la chaîne de caractères qui fonde le morceau d'entrée générale à répéter pour former la sous-entrée :

> Autorisation par le roi de France , — par le roi d'Angleterre

La répétition de « par le roi de France » est ici inutile, et la chaîne doit être simplifiée en :

> Autorisation , — par le roi d'Angleterre

Il est également arrivé que l'erreur concerne le choix du mot à répéter, ce qui a par exemple formé l'entrée :

> Aumônes du roi de France ou à son nom , — v. confirmations par Philippe le Bel

à la place de :

> Autorisations du roi de France ou à son nom , — v. confirmations par Philippe le Bel

Des erreurs se sont également glissées dans la création des liens au moment du choix de l'entité vers laquelle pointer. La méthode utilisée consiste en effet à trouver dans l'index la première chaîne de caractères correspondant à l'élément recherché afin d'éviter les ambiguïtés provoquées par la répétition des entrées générales dans le corps des sous-entrées. Or ce processus a généré quelques incohérences : par exemple le renvoi « v. Agen » pointait dans un premier temps vers l'entrée « Agent » de la table des matières car c'était la première dans l'ordre du fichier à commencer par la chaîne de caractères « Agen ».

Ces erreurs automatiques sont accompagnées d'erreurs manuelles issues d'une mauvaise compréhension de certaines situations. Par exemple, dans l'entrée :

> <term type='place'>Fontaine-le-Port [Seine-et-Marne, con Le Châtelet-en-Brie] , — v. Port (Le) [2037]</term>

la référence « 2037 » a dans un premier temps été comprise comme un renvoi vers l'inventaire qui aurait été indiqué d'une manière inhabituelle. Nous l'avons donc rétabli sous la même forme que les autres :

```
<num>
    <idno>2037</idno>
</num>
```

Or ce renvoi a ici une toute autre signification. La chaîne « Port (Le) » peut renvoyer vers pas moins de quatre entrées différentes et l'indication « [2037] » permet de désambiguïser la situation en précisant l'entrée vers laquelle pointe ce renvoi, ici :

```
<p>
    <seg>
        <term type='place'>Port (Le) [Port de l'abbaye de Barbeau à
Fontaine-le-Port, Seine-et-Marne, con Le Châtelet-en-Brie]</term>
        <num>
            <idno>2037</idno>.
        </num>
    </seg>
</p>
```

Conclusion

Nous avons vu que la compréhension d'un instrument de travail papier n'est pas un processus évident. Le traitement numérique de l'index nous a ici permis d'aborder les différents problèmes liés à l'analyse et à la segmentation des entrées. Nous avons tout d'abord travaillé à la reconnaissance de la typologie, puis à celle des différentes parties qui les composent. Nous avons ensuite décrit les différents cas spécifiques qui se sont présentés à nous ainsi que les problèmes liés à la compréhension des caractères utilisés. Nous avons enfin traité les cas d'erreurs repérées au cours du processus, qu'elles soient issues du document originel, de la transcription par ROC, des traitements automatiques antérieurs ou même de nos propres manipulations.

Ces différents éléments nous ont permis de voir que les *legacy data* peuvent être complexes à appréhender lorsque leur forme n'est pas pleinement systématisée. L'édition repose ici sur une forte économie de caractères qui tend à accroître les cas de polysémie. Si les recherches récentes ont permis de montrer que les modèles d'apprentissage de liage d'entités disposent d'une forme de plasticité qui leur permet de rester efficaces malgré

la présence de bruit dans les données[7], notre travail de constitution d'un référentiel a largement été impacté par ces difficultés. Nous avons exposé ici celles qui sont le plus marquantes pour comprendre l'action que nous avons menée, mais nous aurons l'occasion d'en présenter d'autres tout au long de nos explications sur le processus de travail.

7. Elvys Linhares Pontes, Ahmed Hamdi, Nicolas Sidère et Antoine Doucet, "Impact of OCR Quality on Named Entity Linking", dans *International Conference on Asia-Pacific Digital Libraries 2019*, Kuala Lumpur, Malaysia, 2019, DOI : 10.1007/978-3-030-34058-2_11 ; Caroline Koudoro-Parfait, Gaël Lejeune et Richy Buth, "Reconnaissance d'entités nommées sur des sorties OCR bruitées : des pistes pour la désambiguïsation morphologique automatique", dans *Traitement Automatique des Langues Naturelles*, 2022, p. 45-55.

Chapitre 5

Analyser le lien entre les entités

Au cours du chapitre précédent consacré aux difficultés associées à la segmentation et au traitement des données de l'index, nous avons eu l'occasion d'aborder à plusieurs reprises des problématiques liées à la restitution automatique des liens entre les entités. Ces liens sont en effet une composante essentielle de cet instrument de travail et permettent de transformer une simple liste d'entités nommées repérées dans les textes en une réelle ontologie sous forme papier. Leur prise en compte est donc un enjeu essentiel pour développer la tâche du liage d'entités et représente la principale valeur ajoutée de notre travail au projet Himanis. Nous avons déjà décrit au chapitre 3 le projet originel qui a guidé les premiers traitements numériques de l'index, ce chapitre sera donc consacré à la prise en compte des relations entre les entrées de l'index dans le cadre de leur transformation en un référentiel mobilisable pour le liage d'entités.

Nous retracerons les différentes étapes de compréhension et d'utilisation de ces liens. Pour cela, nous décrirons dans un premier temps le processus de transformation utilisé pour prendre en compte les relations entre les entités dans le traitement numérique de l'index. Ensuite, nous proposerons une description des différentes formes possibles que peuvent prendre ces relations. Enfin, nous nous intéresserons à leur caractérisation en vue de la création d'une base de données relationnelle.

5.1 Numériser les relations entre les entrées d'index

5.1.1 Travail préparatoire

La première étape de notre travail de stage s'est concentrée sur la rédaction d'une feuille de style XSL permettant de transformer les liens décrits par des chaînes de carac-

tères en des liens manifestés par des attributs insérés dans les données [1]. Nous avons dans un premier temps utilisé la fonction « generate-id() » afin d'insérer dans chaque balise <seg> un attribut @xml:id contenant un identifiant unique et normalisé :

```
<p>
    <seg xml:id=' d1e368617'>
        <term type='person'>Sanche de Labatut</term>
        <num>
            <idno>107</idno>.
        </num>
    </seg>
</p>
```

Cette étape a également été l'occasion d'effectuer un tri entre les entrées pour éviter la récupération d'éléments inutiles. Nous avons donc fait le choix d'appliquer la génération d'identifiants uniquement aux balises <seg> disposant soit d'une balise <num> pleine soit de la chaîne « v. » dans la balise <term> soit d'une balise sœur <seg>. Outre les titres de l'index [2], ce tri a également permis de ne pas accorder d'identifiants aux noms de personnes qui ne sont pas suivis de renvois vers l'inventaire car l'utilisation de ces entrées nous a semblé trop complexe [3].

Nous nous sommes ensuite efforcés de reconnaître les différents types de relations qui peuvent exister entre les données. Nous en avons retenu trois : relation depuis une entrée rejetée vers une entrée retenue, relation depuis une sous-entrée vers une entrée générale et relation depuis une entrée vers une autre entrée associée. Les relations depuis une entrée rejetée vers une entrée retenue ont été identifiées par deux critères : une balise <num> vide et la présence de la chaîne de caractères « v. » ou « v. aussi ». Les mots suivants cette chaîne ont été insérés comme valeur de l'attribut @sameAs. Le caractère « , » est ici utilisé comme séparateur entre des renvois successifs, nous avons donc séparé ces derniers par un « | » :

```
<seg xml:id='d1e71' sameAs='donations|renonciations|remises'>
    <term>Abandons, v. donations, renonciations, remises</term>
    <num/>
</seg>
```

Nous avons décrit au chapitre 3 le travail préalable de reconstruction des sous-entrées pour qu'elles disposent dans leurs corps des parties de l'entrée générale nécessaires à leur

1. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Index/ajout_id.xsl.
2. *Cf.* chapitre 4.
3. *Cf.* plus bas.

bonne compréhension. Le séparateur « , — » a donc été utilisé ici pour récupérer les mots correspondants à l'entrée générale dans un attribut @ana :

```
<seg xml:id='d1e533' ana='Accensements'>
    <term>Accensements , — par les particuliers</term>
    <num>
        <idno>515</idno>, <idno>1168</idno>, <idno>1589</idno>,
        <idno>1828</idno>, <idno>1843</idno>, <idno>1941</idno>,
        <idno>2247</idno>.
    </num>
</seg>
```

Les renvois vers les entrées associées sont eux aussi manifestés par la présence des séparateurs « v. » ou « v. aussi » mais se différencient des premières relations par la présence de balises <idno> à l'intérieur de la balise <num>. Nous avons ici inséré ces renvois dans un attribut @rend :

```
<seg xml:id='d1e101495' rend='Quints'>
    <term>Quintaines [semble un simple synonyme de quints, v. Quints]</term>
    <num>
        <idno>2287</idno>.
    </num>
</seg>
```

Ce travail nous a permis de repérer un certain nombre d'entrées correspondant à la forme suivante :

```
<seg xml:id='d1e133394' ana='Agen' prev='d1e133291' rend='Agenais'>
    <term type='place'>Agen [Lot-et-Garonne] , baile , — v. Agenais</term>
    <num>
        <idno>107</idno><idno>1526</idno><idno>1818</idno>
        <idno>2030</idno><idno>2100</idno><idno>107</idno>
    </num>
</seg>
```

Or l'observation des données nous a permis de voir que les renvois vers une entrée associée situés juste après une entrée disposant de renvois vers l'inventaire prennent généralement la forme suivante :

```
<p>
    <seg xml:id='d1e273'>
        <term>Abus</term>
        <num>
            <idno>784</idno>, <idno>1158</idno>, <idno>1159</idno>,
            <idno>1757</idno>, <idno>2220</idno>
        </num>
    </seg>
    <seg xml:id='d1e297' sameAs='excès|violences' ana='Abus'>
        <term>Abus , — v. excès, violences</term>
        <num/>
    </seg>
</p>
```

Pour maintenir la cohérence des données, nous avons donc rétabli tous les cas de « , — v. » comme une entrée supplémentaire contenant uniquement la relation entre l'entrée précédente et l'entrée associée [4].

Cette étape nous a permis d'observer que la séparation entre les relations depuis une entrée rejetée vers une entrée retenue et les relations depuis une entrée vers une entrée associée n'est pas vraiment pertinente dans le cadre de la construction du référentiel [5]. Si nous avons laissé les deux types de relations dans des attributs séparés, nous les avons traités indifféremment dans la suite de notre travail. Nous avons également été confrontés à un certain nombre d'imbrications, que ce soient des sous-entrées disposant elles-mêmes de sous-sous-entrées, des renvois contenus dans les entrées générales et répétées dans le corps de la sous-entrée [6] ou encore des renvois dans les sous-entrées. Nous avons donc adapté les scénarios de notre feuille de style en fonction de ces situations [7] et corrigé certains éléments *a posteriori* lorsque cela s'est avéré nécessaire.

5.1.2 Réalisation de la relation

Une fois les balises contenant les entrées de l'index pourvues d'un identifiant et d'attributs décrivant ses relations avec les autres entrées, il nous reste encore à récupérer les identifiants de chacune des entrées correspondantes à ces relations. Cette étape a

4. Nous avons ici utilisé le script suivant pour isoler les entrées contenant un terme associé et disposant de renvois propres vers l'inventaire : https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Index/verification_termes_associes.py.

5. De plus ces dernières sont très peu nombreuses : après réalisation de la tâche précédente, nous n'en comptons que 12.

6. Nous avons travaillé sur la rationalisation de ces situations à l'aide de la feuille de style : https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Index/correction_inventaire.xsl.

7. A ce propos, v. aussi Virgile Reignier, *De l'index papier à l'indexation automatique…*

été réalisée à l'aide de la fonction XSL « $doc//div[@n=$type_index]//p[child::seg[starts-with(lower-case(replace(term, ' , — ', ' ')), lower-case($token))]][1]/seg[starts-with(replace(lower-case(term), ' , — ', ' '), lower-case($token))][1]/@xml:id » dans lequel $doc correspond à l'ensemble du fichier TEI, $type_index à la portion de ce fichier dans laquelle l'entrée à l'origine du lien se situe (index des sujets ou des noms de personnes et de lieux) et $token à la chaîne de caractères recherchée. Afin d'éviter au maximum les ambiguïtés, la fonction contient deux filtres : un premier qui sélectionne le premier ensemble d'entrées contenant la chaîne recherchée et un second qui sélectionne la première entrée correspondante au sein de cet ensemble. L'objectif de ces filtres est d'éviter au maximum la sélection accidentelle d'une sous-entrée au lieu d'une entrée générale parce qu'elle commencerait par la même chaîne de caractères.

Nous avons également tiré parti d'une incohérence décrite au chapitre 4 (sélection de l'entrée « Agent » à la place de « Agen ») pour ajouter un autre filtre dans la sélection de l'entrée vers laquelle pointe la relation : il faut qu'il soit dans la même partie de l'index que l'entrée depuis laquelle part la relation. Les renvois ne permettant pas de trouver un @xml:id correspondant à la chaîne de caractères ont été marqués par « Not Found », puis résolus manuellement [8]. Les identifiants trouvés sont ensuite ajoutés dans un attribut @corresp pour les relations depuis une entrée rejetée vers une entrée retenue et dans un attribut @rendition pour les relations depuis une entrée vers une entrée associée. Les relations depuis une sous-entrée vers une entrée générale sont ici un peu différentes puisque chaque entrée n'en dispose toujours que d'une et qu'elle ne nécessite pas l'élimination des séparateurs ' , — ' dans la fonction de recherche. Les identifiants trouvés sont cette fois ajoutés dans un attribut @prev. Cette opération aboutit ainsi au résultat suivant :

```
<seg xml:id='d1e799' sameAs='compositions|finances' ana='Accords avec
le roi ou ses gens' prev='d1e591' corresp='d1e37468 d1e63204'>
    <term>Accords avec le roi ou ses gens , — v. aussi compositions,
finances</term>
    <num/>
</seg>
```

8. Il y en avait 740 dont 275 dans des entrées lieux, 275 dans les personnes, 159 dans les sujets. A la fin de notre travail, il reste encore trois renvois pour lesquels nous n'avons pas pu retrouver l'entrée correspondante.

5.2 Des liens implicites à prendre en compte

5.2.1 Différentes formes de noms de personnes

Les relations explicites que nous venons de présenter ne sont cependant pas les seules au sein de cet index : on y compte aussi un grand nombre de relations implicites qu'il nous faut appréhender. Une grande partie d'entre elles concernent les noms de personnes, c'est pourquoi nous allons nous attarder dans un premier temps sur les différentes formes sous lesquelles ces noms peuvent apparaître. On en compte 4 principales :

1. Colard Wilequart

2. Wilequart (Colard), chevalier

3. Walincourt [Nord, con Clary] , — v. Jean de —.

4. Villars-Montroyer [Haute-Marne, con Auberive] (Guion de)

Les deux premières manifestent d'un problème récurrent dans l'anthroponymie médiévale : est-ce le nom ou le prénom qui définit l'individu ? Pour éviter tout problème à ce sujet, les auteurs de l'index ont ici fait le choix de prendre en compte les deux et de ranger « Colard Wilequart » à la fois à « Colard » et à « Wilequart ». Le choix a cependant été fait de n'indiquer les entrées d'inventaire contenant ce nom uniquement pour l'entrée « Wilequart », l'autre étant restée vide. L'entrée « Colard Wilequart » est donc ici à comprendre comme un renvoi implicite vers « Wilequart (Colard) » si l'on veut retrouver les actes mentionnant cette personne. Comme évoqué plus haut, nous avons jugé plus simple d'éliminer ces entrées ne disposant ni de sous-entrées ni de renvois explicites vers une autre entrée ni de renvois vers l'inventaire au moment de la création des identifiants.

La troisième forme est plus facile à appréhender : il s'agit d'une entrée toponymique accompagnée d'une sous-entrée renvoyant vers une personne associée à ce même lieu (et qui est donc ici rangée à son prénom). Le caractère « — » est à comprendre comme signifiant la répétition de l'entrée générale, nous l'avons donc remplacée par le mot correspondant pour permettre le traitement automatique du lien entre les entrées. La quatrième forme est une variante de cette dernière car elle est aussi formée à partir d'une entrée toponymique, mais celle-ci ne dispose pas de renvois vers des entrées d'inventaires. L'entrée n'existe que pour former un renvoi implicite vers l'entrée formée à partir du prénom (ici « Guion de Villars-Montroyer »). Des exceptions peuvent exister, notamment lorsque plusieurs personnes sont associées à un même lieu : dans ce cas le renvoi est explicite comme pour la forme 3.

5.2.2 Une chaîne de traitement complexifiée

Ces liens implicites forment autant de relations à numériser dans notre travail. Cependant, les traitements antérieurs sur les données avaient déjà commencé à les prendre en compte en vue de la mise à plat de l'index. L'objectif étant d'insérer toutes les entités nommées repérées par l'index dans les métadonnées des actes, le choix avait été fait de recopier les renvois vers les entrées d'inventaire dans les entrées contenant ces renvois implicites :

```
<seg xml:id='d1e392309'>
    <term type='person'>Villar [-en-Val] (Raimond de)</term>
    <!–NUM ajouté automatiquement–>
    <num>
        <idno>1</idno>.</num>
</seg>
```

Nous avons donc dû éliminer ces renvois rajoutés (il y en avait 718) et les avons remplacés par une relation depuis une entrée rejetée vers une entrée retenue entre les deux formes du même nom (ici vers « Raimond de Villar ») [9].

Une dernière forme de renvoi a été repérée à cette occasion :

> Aragon [ancien royaume d'Espagne] , — roi : Jaime II, Pierre III

Le caractère « : » fait ici l'office du « v. » employé habituellement et permet de faire le lien vers les entrées « Jaime II, roi d'Aragon » et « Pierre III, roi d'Aragon ». Comme nous avons ignoré ces liens lors des étapes précédentes, nous les avons ajoutés en même temps que les autres liens implicites.

5.3 Caractériser les relations

5.3.1 Différentes formes de « sous-entrées »

Nous avons vu les différentes méthodes utilisées par l'index pour formuler les relations entre les entités. Mais ces relations n'ont pas toute la même valeur. Nous devons donc prendre soin d'ajouter des qualités aux relations dans le référentiel pour prendre en compte ces caractéristiques. Le caractère le plus visible est celui qui lie les sous-entrées aux entrées générales dans l'index des matières. Les sous-entrées de cette table sont en effet construites systématiquement comme des termes spécifiques dépendant d'un terme

9. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Index/comparaison_entrees.py.

générique. Ces termes génériques n'ont parfois pas de renvois vers l'inventaire qui leur sont propres et ne sont indiqués que pour former un bloc de plusieurs termes proches :

```
<seg xml:id='d1e1062'>
    <term>Accroissements, augmentations</term>
    <num/>
</seg>
<seg xml:id='d1e1070' ana='Accroissements, augmentations' prev='d1e1062'>
    <term>Accroissements, augmentations , — d'assignation</term>
    <num>
        <idno>470</idno>
    </num>
</seg>
<seg xml:id='d1e1082' ana='Accroissements, augmentations' prev='d1e1062'>
    <term>Accroissements, augmentations , — de don</term>
    <num>
        <idno>256</idno>
    </num>
</seg>
```

Cette relation entre terme générique et terme spécifique a été spécifiée en ajoutant la caractéristique « isNarrowerConceptOf/isBroaderConceptOf » dans le modèle de données du référentiel.

Pour ce qui est des sous-entrées de la table des noms de personnes et de lieux, la relation est plus délicate à caractériser. Outre les renvois vers les personnes dont le nom est associé au lieu, les sous-entrées dépendant d'entrées toponymiques peuvent ainsi être soit des localités plus précises au sein de ces lieux :

Clermont [Puy-de-Dôme, cne Clermont-Ferrand] , — église cathédrale

soit des personnes exerçant une fonction attachée au lieu :

Clermont [Puy-de-Dôme, cne Clermont-Ferrand] , — évêque

soit un groupe de personnes associées au lieu :

Clermont [Puy-de-Dôme, cne Clermont-Ferrand] , — chanoines

D'autres encore rassemblent plusieurs éléments en même temps :

Pujols (Les) [Ariège, con Pamiers], dame, terre et châtellenie

Faute de pouvoir les distinguer de manière automatique, nous avons continué de considérer toutes ces sous-entrées comme relevant du type « lieu » et n'avons ajouté aucune caractéristique les reliant à l'entrée générale. Cela fait partie, avec la discrimination entre les entrées toponymiques et anthroponymiques, des principales voies d'amélioration de notre travail.

5.3.2 Retracer le lien entre les entrées d'index et les entrées d'inventaire

Les sous-entrées dépendant d'entrées anthroponymiques offrent encore des perspectives différentes car elles expriment la plupart du temps une fonction exercée par la personne :

> Benoît de Saint-Gervais , — auditeur au Châtelet

Cette fonction représente ici à son tour une relation : celle qui lie la personne à sa mention dans l'acte. Nous avons donc fait le choix de ne considérer que l'entrée générale au sein du référentiel et de lui attribuer toutes les relations entre les sous-entrées et les entrées d'inventaire. La fonction des personnes représente ici un élément descriptif de ces relations [10]. Cette solution ne permet cependant pas la prise en compte des entrées associant plusieurs personnes ni des personnes qui ne sont connues que par un lien familial par rapport à une autre :

> Furon et sa femme

> Fessac (Jean), bourgeois de La Rochelle , — sa femme

Ce dernier cas est d'ailleurs particulièrement difficile à prendre en compte puisqu'il est possible que l'entrée générale ne dispose pas de renvois vers l'inventaire et donc que la personne concernée ne soit mentionnée dans les textes que par son lien avec une autre personne.

Pour finir, la numérisation du lien entre les entrées d'index et les entrées d'inventaire s'est concentrée sur les éléments insérés dans les balises <idno> et a de ce fait éliminé quelques indications présentes dans l'index, comme par exemple des renvois vers la partie « Additions et corrections » :

10. Nous exposerons plus précisément la formalisation du référentiel au chapitre suivant.

```
<seg xml:id='d1e341326'>
    <term type='place'>Prouvais [Aisne, con Neufchâtel-sur-Aisne]</term>
    <num>
        <idno>401</idno>, <idno>1594</idno> (v. add. et corr.).
    </num>
</seg>
```

Conclusion

Nous avons vu que la prise en compte des relations entre les entités représente un enjeu crucial pour comprendre et réutiliser les données contenues dans l'index. Leur numérisation dans le cadre de la création du référentiel a été rendue possible par la mise en place d'une chaîne de traitement permettant de préparer les données puis de formaliser les liens au sein des balises <seg> du fichier TEI. Les liens explicites ont ensuite été complétés par la prise en compte de toute une série de liens implicites. Enfin, certaines relations ont été enrichies par l'ajout d'une description ou d'un type.

La prise en compte des liens entre les entités fournit ici un niveau supplémentaire d'analyse des textes et permet à la fois d'augmenter le potentiel d'apprentissage du liage d'entités et d'envisager d'autres possibilités. Lorsqu'il sera possible d'associer les entités du référentiel à leur manifestation dans les textes, les relations entre entités pourront en effet être utilisées comme des données d'entraînement pour développer des modèles associant la REN et l'extraction des relations [11]. Cette perspective permettrait d'améliorer encore notre capacité à lire les textes de manière automatique et d'envisager l'augmentation des données présentes dans le référentiel par l'ajout de nouvelles entités nommées et de nouvelles relations.

11. Yoann Dupont, *La structuration dans les entités nommées*, Thèse de doctorat, Université Sorbonne Paris Cité, 2017, URL : https://tel.archives-ouvertes.fr/tel-01772268 (visité le 28/03/2022), p. 169–180.

Chapitre 6

Transformer l'index en une base de données relationnelle

En l'état actuel de notre travail, nous avons vu comment manipuler automatiquement les entrées de l'index et comment prendre en compte les relations entre ces entrées. Cependant la mise en forme utilisée au début de notre travail ne permet pas de tirer pleinement parti de ces relations. Leur manifestation dans le fichier TEI se limite en effet à un ensemble d'attributs dans les balises entourant les entrées. Afin d'envisager la mise en place d'une base de connaissances pour l'apprentissage du liage d'entités, nous avons donc transformé ces données pour obtenir une base de données relationnelle contenant les éléments à notre disposition. L'objectif est de pouvoir réutiliser facilement ces données et les mettre à disposition pour les chercheurs qui en ont l'utilité.

Ce chapitre sera donc consacré à la modélisation et à la mise en œuvre de cette base de données. Pour cela, nous aborderons dans un premier temps le travail préparatoire à la modélisation des données. Puis nous exposerons la mise en place du modèle et l'import des données dans la base de données. Enfin, nous analyserons le cas spécifique des entités toponymiques.

6.1 Préparation du modèle

6.1.1 Transformer l'index sous forme de table

La première étape de la mise en place de la base de données est de transformer le fichier TEI pour construire une table à partir des entrées de l'index. L'objectif de ce travail est à la fois de préparer l'import dans la base de données et de fournir un cadre qui se prête plus facilement à la manipulation des entrées pour en corriger certains éléments. Nous avons donc rédigé une feuille de style XSL afin de construire un fichier au

format CSV qui rassemble toutes les entrées d'index [1]. Outre la colonne « entrée », la table dispose de 3 colonnes par type de relation (ces types ont été nommés « terme retenu » pour les relations depuis une entrée rejetée vers une entrée retenue, « terme générique » pour une relation depuis une sous-entrée vers une entrée générale et « terme associé » pour une relation depuis une entrée vers une entrée associée) : une pour la chaîne de caractères qui forme le renvoi, une pour l'identifiant de l'entrée correspondant à cette chaîne de caractères et une pour le contenu de cette entrée. Ce dernier élément permet ainsi de comparer visuellement le contenu avec celui de la chaîne de caractères qui forme le renvoi pour vérifier la bonne correspondance entre les deux et l'absence d'erreur dans le processus. Une dernière colonne contient l'ensemble du contenu des balises <idno> de l'entrée, séparées par un « | »

Nous avons dans un premier temps créé deux tables pour les deux parties de l'index (sujets et noms de personnes et de lieux) accompagnées d'une table supplémentaire contenant toutes les entrées avec leur id et leur type. Cette table supplémentaire a été utilisée pour réaliser un certain nombre de corrections que nous avons évoquées dans les chapitres 4 et 5 à propos de la forme des données ou des renvois. Par la suite, nous nous sommes rendus compte que le fichier CSV dispose d'une limite dans le nombre de caractères compris dans un champ. Les champs de la colonne « idno » pouvant être très longs lorsqu'une entrée dispose d'un grand nombre de renvois vers l'inventaire, certains n'ont pas été reportés dans les tables. Nous avons donc changé de format de fichier et utilisé Excel pour la suite des opérations [2]. Les éléments de la colonne « idno » qui ont été perdus ont pu être restitués depuis le fichier TEI en opérant une transformation via le site https://regex101.com et les expressions régulières « <idno>([^<]*)<\/idno> ( )*(,)?( )*(\n)?( )* » et « $1| ». Après le travail de segmentation des entrées que nous avons décrit au chapitre 4, le résultat de ce travail forme le fichier Excel déposé sur le Github du projet Himanis [3].

6.1.2 Un lieu unique pour rassembler les données

L'objectif final de notre travail est de pouvoir associer les entités nommées reconnues dans les images aux entités du référentiel construit à partir de l'index. Le modèle de la base de données doit donc prendre en compte à la fois les éléments décrits par l'index, les actes et les images. Nous avons construit 5 tables dont 3 pour les sujets, personnes et lieux décrits par les entrées d'index. Ces tables sont formées à partir des entrées segmentées

1. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Index/index_to_csv.xsl.

2. Ce changement a aussi entraîné un changement dans la librairie python utilisée pour manipuler les données. Nous avons utilisé la librairie Pandas pour la suite de notre travail.

3. https://github.com/oriflamms/himanis/blob/master/Inventories/Systematic/Paris_AN_JJ_inventaire_JJ37-50_index.xlsx.

selon la correspondance suivante :

Colonne du fichier Excel	Attribut de la table
Entrée_simplifiée	Name
Détails	Note
Supplément	Short summary

Les tables des actes et des zones d'image sont construites à partir du fichier JSON que nous avons décrit au chapitre 2. Voici la correspondance entre les attributs des actes dans le fichier JSON et dans la table :

Attribut du fichier JSON	Attribut de la table
Volume	Register
Act_N	Num
normalized_language/language	Language
Regeste	Regeste
Date	Date (not standard)
Date-normalisee	Date
Folio_start	Folio start
Inventory_Name + '#' + Inventory_Nr	Inventory Reference

La table contenant les zones d'images est construite à partir du contenu des attributs « Text_Region » de ces mêmes actes décrits par le fichier JSON. La relation entre les actes et les zones d'image est formalisée par un attribut « Id Act » qui contient l'identifiant de l'enregistrement correspondant dans la table des actes. La correspondance est ensuite la suivante :

Attribut du fichier JSON	Attribut de la table
Reference	Base image
type_act	Part
Graphical_coord	Coordinates
Address_bvmm	Address bvmm
Texte	Transcription

Un dernier attribut a été ajouté aux zones d'image pour indiquer le folio sur lequel se situe la zone. Cette donnée n'était pas présente en l'état dans les données décrites par le fichier JSON, sa récupération a nécessité la construction d'une table de concordance entre les folios et les images. Pour cela, nous avons rédigé un script python afin de chercher tous les éléments « page » dans le corpus « Himanis | TEKLIA processing » d'Arkindex et en

extraire une table contenant le nom de l'élément (qui correspond au numéro de folio) et l'url de l'image dans la BVMM [4]. Cette étape nous a également permis de mettre au jour un certain nombre d'erreurs dans les url des images indiquées dans le fichier JSON. Nous avons corrigé celles qui ont pu l'être et fait remonter celles qui concernent directement la BVMM.

6.2 Mise en place de la base

6.2.1 Un modèle relationnel

Une fois les tables définies, il faut aussi prendre en compte l'ensemble des relations qui peuvent exister entre ces tables. Les tables sujets, personnes et lieux disposent ainsi chacune de relations les liant à la table des actes, mais aussi de relations entre elles. Ces relations correspondent au lien entre les entrées de l'index et peuvent se construire soit d'une table vers une autre soit d'une table vers elle-même. Les différents types de relations que nous avons décrits précédemment sont ici indifférenciées, à l'exception des relations depuis une sous-entrée vers une entrée générale au sein de la table des sujets qui disposent d'une qualification spécifique [5]. Comme certaines entrées contiennent plusieurs entités, nous avons longuement hésité sur la possibilité de créer une table dédiée aux entrées de l'index et qui permettrait de faire le lien entre les entités contenues dans ces entrées et les actes décrits par l'inventaire, mais il a finalement été convenu que cette table était superflue et que nous pouvons lier directement les entités aux actes. La figure 8 représente le modèle utilisé pour générer la base de données.

6.2.2 L'import des données

Nous avons donc créé une base de données au sein de la plateforme Heurist [6] permettant la gestion de données issues de corpus en sciences humaine, leur diffusion et leur enrichissement collaboratif. La plateforme dispose d'un certain nombre de modèles de données déjà intégrés et personnalisables en fonction des besoins (nous avons ici utilisé les entités « Person » et « Place ») et propose également d'en créer de nouveaux (nous avons créé les entités « Subject », « Act » et « Image zone »). Elle propose aussi un modèle de relation N-N à partir de l'entité « Record relationship ». Ces relations peuvent être qualifiées à l'aide d'un « Relationship type » personnalisable et peuvent disposer d'attributs descriptifs.

4. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Index/table_concordance_image_folio.py.

5. *Cf.* chapitre 5.

6. https://heuristnetwork.org.

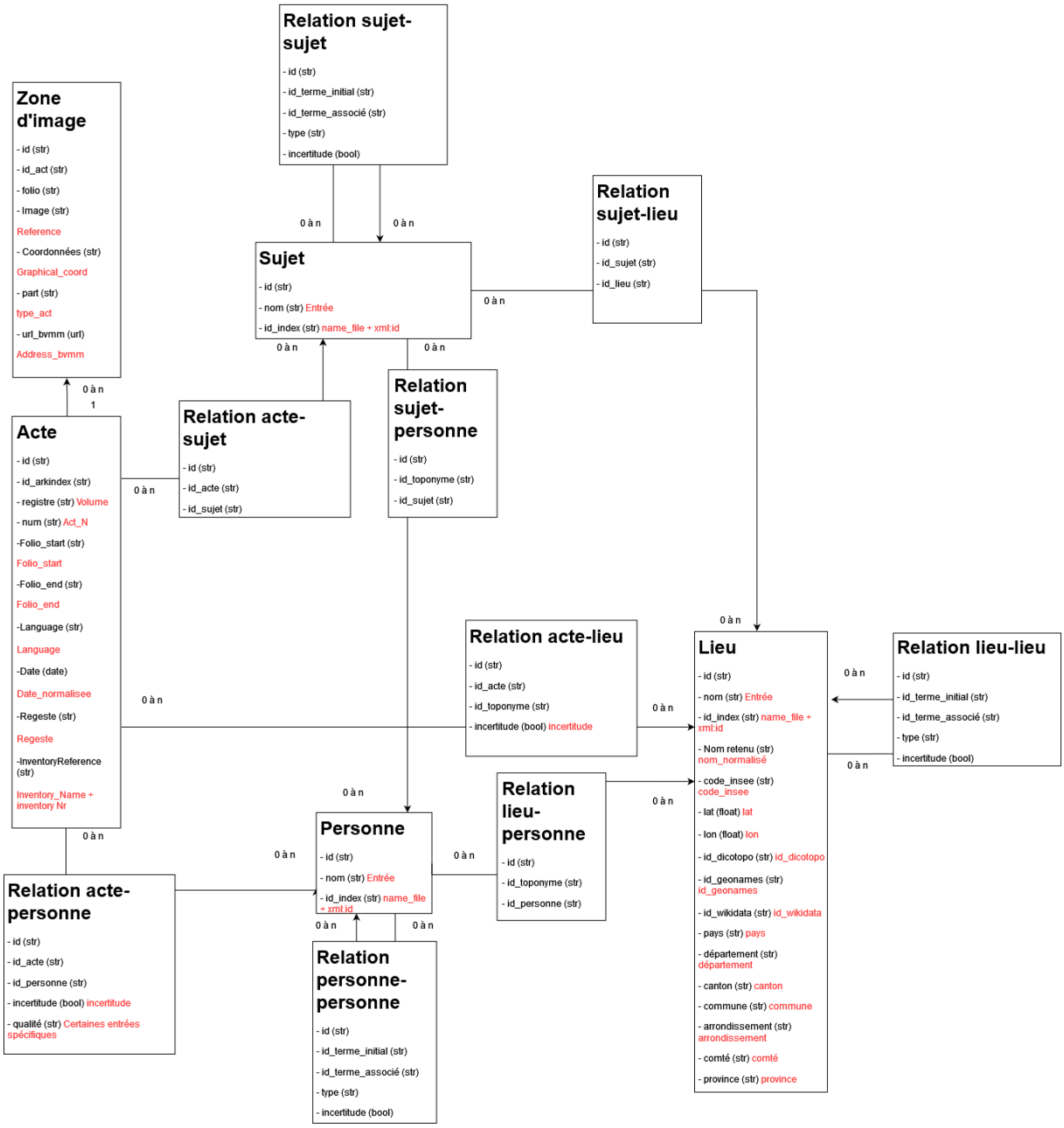

Figure 8 – *Modèle de la base de données relationnelle permettant de décrire les registres du Trésor des chartes avec les zones d'images, les actes et les entités contenus dans les actes.*

La préparation des données pour l'import a été réalisée en plusieurs étapes à l'aide d'un script python [7]. La première étape consiste à créer à partir du fichier Excel les tables Person, Place et Subject et les importer dans Heurist. Il est alors possible de les exporter avec leurs identifiants local H-ID nécessaires à la construction des tables de relations. Ces dernières sont au nombre de 6 : Person-Person, Person-Subject, Person-Place, Place-Subject, Place-Place, Subject-Subject. L'étape suivante consiste à utiliser le fichier JSON pour créer la table Act et l'importer dans Heurist. Les enregistrements de cette table sont ensuite exportés avec leurs H-ID nécessaires pour construire la table Image zone et les tables de relation Act-Person, Act-Subject et Act-Place.

6.3 Le cas particulier des noms de lieu

6.3.1 Segmentation des différentes parties de l'entrée

Importées au sein de la table Place de la base de données, les entrées toponymiques de l'index ont la particularité d'être composées de plusieurs éléments permettant de les localiser :

> Lincques [Pas-de-Calais, con Guînes, cne Licques]

La segmentation de ces éléments permet dans un premier temps de les intégrer parmi les attributs de la table. Il devient alors possible de filtrer les recherches en fonction de ces attributs. Cette segmentation est également un travail nécessaire pour aligner ces entités avec d'autres référentiels afin de les enrichir de coordonnées [8].

Cette étape a été réalisée à partir du fichier CSV contenant les entrées d'index. Nous avons rédigé un script python afin de sélectionner automatiquement les entrées toponymiques et constituer un autre fichier au même format pour les traiter à part [9]. Ce script permet ensuite de segmenter automatiquement les éléments qui composent l'entrée. La première étape de ce travail a été de former un élément « entrée » permettant de simplifier son contenu et de faciliter l'alignement avec les autres référentiels. Cette forme simplifiée a été constituée par l'élimination des éléments de l'entrée situés entre crochets, puis par l'élimination des éléments restants placés après une virgule. Par exemple l'entrée :

> Sainte-Menehould, Sancta Manalhidis [Marne]

a pour forme simplifiée « Sainte-Menehould ». Nous avons également créé un élément

7. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Index/upload_heurist.py.

8. Nous décrirons ce travail plus précisément au chapitre 7.

9. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Index/toponymes.py.

« nom_normalise » afin de proposer une forme du nom potentiellement plus facile à retrouver. Nous avons par exemple réinséré les particules devant le nom : L'entrée « Sauvement (Le) » est ainsi transformée en « Le Sauvement ». Nous avons également traité les noms composés dont certains morceaux peuvent parfois être glissés au sein des crochets. Ils sont alors repérable par la présence d'un « - » :

Vailly [-sur-Aisne, Aisne, ar. Soissons]

nous avons dans ce cas reconstitué le nom complet : « Vailly-sur-Aisne ». Il existe cependant ici une irrégularité dans la construction de ces noms composés : certains maintiennent l'espace présent habituellement avant le crochet ouvrant (comme l'exemple ci-dessus). D'autres sont au contraire construits au plus proche du nom composé et ne disposent pas de cet espace :

Bar[-le-Duc, Meuse] , châtelain

La deuxième étape consiste ensuite à catégoriser automatiquement les différents éléments qui sont situés entre crochets. Ceux-ci peuvent être segmentés à partir des virgules qui sont utilisées comme séparateurs. Différents scénarios peuvent alors exister, mais le cas le plus courant est celui de « Lincques » présenté plus haut. Le département est alors repéré par sa présence au sein d'une liste pré-établie, le canton par la chaîne de caractères « con » et la commune par la chaîne « cne ». Le résultat forme la table suivante :

nom_normalise	département	canton	commune
Lincques	Pas-de-Calais	Guînes	Licques

Le repérage des départements s'est ici confronté à deux difficultés. Tout d'abord un certain nombre d'entre eux sont mentionnés sous une forme abrégée comme « Tarn-et-G. » ou « Seine-Mme ». Nous les avons donc développées au fur et à mesure de la découverte de ces formes abrégées. Une autre difficulté repose sur les quelques évolutions administratives qui ont lieu depuis la parution de l'index. Ainsi le département « Côtes-du-Nord » a par exemple changé de nom pour devenir « Côtes-d'Armor ». Nous avons donc opéré la transformation automatiquement. Une situation plus complexe est celle de la réorganisation en 1968 des départements de la « Seine » et « Seine-et-Oise » pour former les 7 départements qui composent, avec la ville de Paris, l'actuelle région Île-de-France. Faute de pouvoir attribuer automatiquement un de ces nouveaux départements, nous avons conservé les anciens au sein du référentiel.

Certaines entrées disposent également d'une indication à propos de l'arrondissement (marquée par « ar. »). Cette indication, celle du canton et celle de la commune peuvent parfois être factorisées lorsque la ville concernée est la même :

Osmeaux [Eure-et-Loir, ar. et con Dreux, cne Cherisy]

Les entrées peuvent aussi disposer d'une forme actuelle du toponyme, elle est alors marquée à l'aide de l'indication « auj. » :

> Genlis [auj. Viliequier-Aumont, Aisne, con Chauny] , monastère de Sainte-Élisabeth

Ou simplement par un nom différent de celui de l'entrée comme premier élément entre crochets :

> Agarne (L') [Notre-Dame-de-l'Agarne, Gard, con et cne Marguerittes]

Ces éléments ont la particularité de reprendre parfois certains mots de l'entrée, ils peuvent donc prendre une forme factorisée ou abrégée :

> Labastide-de-Montfort [auj. L.-de-Lévis, Tarn, con Gaillac]

Nous avons donc développé ces formes à la main (ici « Labastide-de-Lévis ») pour disposer de la donnée entière. Lorsqu'elle est présente, cette forme actuelle a été utilisée comme valeur de l'attribut « nom_normalise ». Si la commune n'est pas précisée, nous avons aussi fait le choix d'entrer d'office la valeur de cet attribut dans l'attribut « commune ».

Nous avons également dû prendre en compte un nombre conséquent de toponymes étrangers. Pour cette raison, nous avons ajouté un élément « country » dont la valeur est systématiquement « France » sauf indication contraire. Nous avons créé un référentiel de pays, puis avons opéré une conversion au moment de l'import dans Heurist pour que le nom corresponde aux entités prédéfinies dans la plateforme. Certaines localités sont situées dans des pays qui ont cessé d'exister depuis (par exemple la Yougoslavie), mais leur nombre était suffisamment faible pour que nous puissions chercher leur situation actuelle manuellement. Ces toponymes étrangers sont parfois agrémentés d'une précision sur la province où ils se situent (marquée par « prov. ») ou sur le comté (marqué « co. ») s'il s'agit d'un lieu en Grande-Bretagne [10].

Un dernier problème se pose à partir des toponymes qui ne sont pas situés en un seul lieu. C'est le cas par exemple de certaines localités à cheval sur deux communes ou englobant deux communes, comme par exemple :

> Abbans [-Dessus et -Dessous, Doubs, con Boussières]

Nous avons dans ce cas conservé les deux entités dans la colonne correspondante (ici « Abban-Dessus » et « Abban-Dessous » sont des communes). Dans un second cas, la localité dispose d'indications permettant à l'index de la situer relativement :

> Le Maréchal (ferme qui fut à Richard) [près de Rouen]

10. Ces toponymes étant au nombre de 12, leur existence nous avait échappé au moment de la rédaction de notre script. La segmentation de cet élément a donc été réalisée manuellement *a posteriori*.

Ces indications ne sont cependant pas suffisantes pour que l'on puisse retrouver précisément le lieu où aligner ces données avec d'autres référentiels. Pour cette raison, nous avons conservé les attributs « Note » et « Short summary » décrits précédemment afin de conserver l'ensemble des indications fournies par l'index. Dans un dernier cas, le toponyme n'a pas pu être identifié avec certitude et plusieurs propositions sont formulées :

> Angles [Angle, Charente-Maritime, con Aulnay, cne Blanzay-sur-Boutonne,
> ou Anglas, Charente-Maritime, con Aulnay, cne Nuaillé-sur-Boutonne]

Dans ce cas, nous avons dédoublé l'entrée pour disposer des informations de localisation sur chacune des propositions et réaliser l'alignement.

6.3.2 Modéliser les lieux non-identifiés

Cette dernière situation a largement participé au questionnement que nous avons évoqué plus haut sur la modélisation d'une table pour les entrées d'index qui servirait d'interface entre les entités et les actes. Les propositions de localisation sont en effet des éléments qui ne sont pas superposables aux toponymes qui constituent l'entrée d'index, pourtant elles ne sont formulées que pour accompagner cette entrée toponymique. Nous avions imaginé un moment créer une table pour les lieux et une pour les localisations, mais il a finalement été convenu que chacune de ces propositions devienne une entité à part entière et distincte de celle formée à partir de l'entrée. Ces entités ont la particularité de ne pas être directement une entrée d'index et de n'être liées à aucun acte, elles sont simplement liées à leur entité de référence via une relation Lieu-Lieu. Contrairement aux cas de coordonnées multiples générées par l'alignement avec les référentiels externes, les propositions multiples de localisation présentes dans l'index sont encore à importer dans la base de données. Pour être complet dans la numérisation des indications de l'index, il faudrait également prendre en compte les marques d'incertitude dans les entrées d'index (généralement un « ? », nous les avons marqués dans le fichier CSV par une colonne « incertitude ») et les reporter comme attribut de la relation entre l'entité correspondante de la base de données et l'acte auquel l'entrée d'index renvoi.

Conclusion

Par ce chapitre, nous avons abordé le processus de transformation du fichier TEI contenant les entrées d'index en une base de données relationnelle. Nous avons pour cela décrit dans un premier temps les étapes de préparation de ce travail par la transformation préalable du fichier TEI en un fichier au format CSV puis Excel et par la description des attributs des tables. Nous avons ensuite décrit la mise en place du modèle relationnel à

travers la modélisation de la table et le travail d'import. Pour finir, nous avons exposé le cas spécifique des entrées toponymiques au travers des problèmes de segmentation et de modélisation des lieux non-identifiés.

La création de cette base de données relationnelle constitue donc une forme d'aboutissement de notre travail de stage puisqu'elle prend en compte tous les éléments que nous avons décrit auparavant. Nous verrons cependant aux chapitres suivants comment ce travail a pu être enrichi par l'ajout de lien avec les autres données du projet et avec des ontologies web. Avec cette base de données, nous disposons donc d'un référentiel complet comprenant 7 503 entités Place, 5 814 entités Person, 3 692 entités Subject, 22 217 entités Act, 38 839 entités Image zone et 60 042 Record relationship mis en ligne sur la plateforme Heurist et librement utilisables pour la suite de ce travail [11]. Le choix de la plateforme permet également de favoriser la publication de ce travail, mais aussi son accroissement futur par l'ajout de nouvelles entrées au sein du référentiel.

11. Dominique Stutzmann, *Himanis / Home*, avec la coll. de Sebastien Hamel, *et al.*, 2022, URL : https://heurist.huma-num.fr/heurist/?db=stutzmann_himanis&website (visité le 09/11/2022).

Conclusion partielle

Après une première partie consacrée au contexte de travail dans lequel nous avons réalisé notre stage, cette seconde partie nous a permis d'exposer toutes les étapes de la construction du référentiel d'entités nommées destiné à produire les données d'entraînement pour l'apprentissage du liage d'entités. Notre développement a commencé par un premier chapitre consacré au travail de compréhension des *legacy data* constituées par l'index et aux difficultés liées à leur traitement numérique. Ensuite, nous avons dédié un second chapitre aux différentes relations qui caractérisent les entrées d'index et à leur prise en compte dans la construction du référentiel. Nous avons enfin terminé cet exposé par une troisième partie portant sur la modélisation et la concrétisation du référentiel sous la forme d'une base de données relationnelle. Nous disposons donc aujourd'hui d'un ensemble d'entités prêtes à être utilisées pour le liage avec les entités nommées présentes dans les registres concernés.

Ce long travail nous a permis de voir que la réutilisation de *legacy data* dans un cadre numérique est associée à certain nombre de problématiques qui freinent largement ce processus. Chaque étape doit en effet prendre en compte un grand nombre d'exceptions, qu'elles viennent de l'ouvrage en lui-même ou d'erreurs réalisées au cours de la chaîne de traitement. Ces problématiques sont à l'origine d'une réduction conséquente de l'ambition du stage. Nous avions en effet envisagé dans un premier temps poursuivre ce travail de transformation des index avec ceux des inventaires des registres JJ 65 à JJ 79B puis avec ceux des inventaires géographiques. Au vu du temps consacré à la transformation du premier ouvrage, il a été décidé de se concentrer sur celui-ci pour aller le plus loin possible et fournir une base de travail aux prochains travaux sur ce corpus.

Troisième partie

Alignement, diffusion et utilisation du référentiel

Si la constitution d'un référentiel à partir d'entrées d'index constitue l'objectif premier du stage, ce travail n'est pas suffisant pour fournir un outil pertinent afin d'étudier le contenu du Trésor des chartes. Les données constituées, pour être utiles, doivent en effet pouvoir rejoindre et compléter les différentes ontologies web déjà disponibles. Nous consacrerons donc cette troisième et dernière partie à la création de liens à partir du référentiel et à ses utilisations. L'objectif étant de l'insérer dans les connaissances générales disponibles sur les documents historiques. Pour cela, nous décrirons dans un premier chapitre le processus d'enrichissement du référentiel par son alignement avec des ontologies web. Puis nous dédierons un second chapitre à la mise à disposition du référentiel et aux liens réalisés avec les autres données du corpus. Enfin, nous consacrerons un troisième chapitre aux différents essais réalisés pour la mise en œuvre du liage d'entités.

Chapitre 7

Enrichir les données à l'aide de référentiels externes

Nous disposons à ce stade d'une table au format CSV contenant les toponymes identifiés dans les registres JJ 37 à JJ 50 du Trésor des chartes. Nous avons pour cela utilisé l'index de l'inventaire analytique de ces registres. De plus, nous avons segmenté les différents éléments présents dans ces entrées toponymiques de l'index et en avons notamment extrait les circonscriptions administratives qu'elles contiennent. Il manque cependant un élément pour localiser précisément ces entités : des coordonnées. Celles-ci ne sont en effet pas présentes dans l'index, nous devons donc les extraire depuis un référentiel externe pour permettre l'enrichissement des toponymes contenus dans notre base de données.

Ce chapitre sera donc consacré au travail d'alignement entre les données de l'index et plusieurs référentiels en ligne pour permettre la sélection de coordonnées et enrichir ces données. Pour cela, nous présenterons dans un premier temps les interfaces que nous avons utilisées. Puis nous exposerons le processus de construction des requêtes utilisé pour obtenir les coordonnées. Enfin, nous décrirons différents cas spécifiques qui se sont présentés à nous au cours de ce travail.

7.1 Présentation des référentiels utilisés

7.1.1 GeoNames

Parmi les référentiels géographiques mobilisables pour enrichir des toponymes de coordonnées, GeoNames (https://www.geonames.org) dispose de plusieurs qualités essentielles pour notre travail : les données sont libres de droit, elles couvrent le monde entier et chaque entité dispose d'un identifiant unique permettant de le retrouver en cas de besoin.

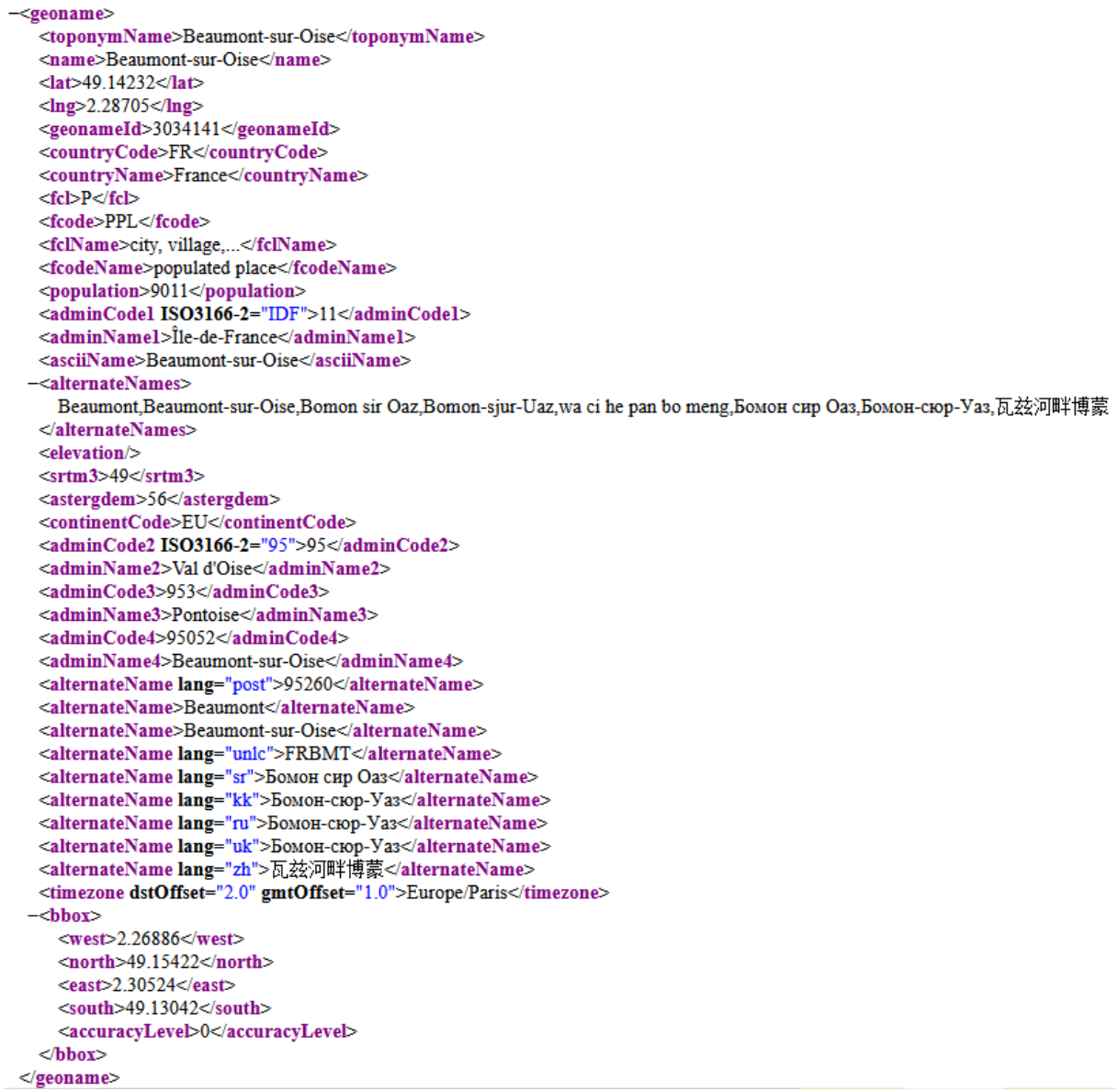

Figure 9 – *Exemple de retour de l'API GeoNames.*

Outre l'interface web permettant de naviguer parmi les localités, le site dispose également d'une API pour faciliter la récupération automatique d'information [1]. Cette API est un service client qui nécessite de s'inscrire sur le site pour disposer d'un compte utilisateur. Ce compte permet ensuite d'adresser des requêtes qui retournent la donnée correspondante sous format XML (*cf.* figure 9). Quelques limites sont à prendre en considération pour les comptes gratuits : la taille des réponses est limitée à 500 lignes et le nombre de requêtes à 1 000 par heure et 20 000 par jour. Les besoins de notre travail permettent de s'accommoder assez facilement de ces contraintes, notamment par l'ajout d'un minuteur lorsque le retour d'une requête indique que la limite est dépassée. La fonction python « time.sleep » permet de mettre l'exécution du programme en pause le temps nécessaire.

1. https://www.geonames.org/export/web-services.html.

GeoNames est donc l'outil le plus complet pour travailler sur des localités françaises et étrangères. Les subdivisions administratives et leurs codes sont respectivement encodés dans des balises <adminName[N]> et <adminCode[N]> où [N] est le rang de la subdivision. Cette méthode permet de disposer d'un encodage unique pour toutes les localités et facilite la navigation d'un pays à un autre. Néanmoins, ces balises ne sont pas accompagnées d'un attribut permettant de savoir à quelle entité administrative correspond chaque donnée (département, région, conté, province, ...), ce qui limite la précision des recherches pour certains pays.

7.1.2 DicoTopo

Le *Dictionnaire topographique de la France* est un projet initié en 1859 et dirigé par le CTHS. Son objectif est de collecter l'ensemble des noms de lieu français dans leur graphie ancienne et moderne. Publié sous la forme d'un volume par département, ce travail encore inachevé a permis de rendre disponible un ensemble conséquent de données sur les localités situées dans 35 départements français. Depuis 2010, le CTHS a entrepris de rendre disponible ces données sous format numérique à l'aide d'une application dédiée (https://dicotopo.cths.fr). Elle fournit donc à ce jour l'outil disponible le plus précis pour chercher des localités à l'échelle infra-communale et des formes anciennes de toponyme. L'application DicoTopo dispose elle aussi d'une API librement utilisable à l'aide de requêtes qui retournent un résultat au format JSON (*cf.* figure 10).

Contrairement à GeoNames, l'API de DicoTopo ne dispose d'aucune limite dans son utilisation [2]. De plus, la présence de localités infra-communales et de formes anciennes rendent cet outil plus adapté que GeoNames pour retrouver les lieux mentionnés dans le Trésor des chartes. La principale limite de l'application réside donc dans le caractère inachevé du travail de collecte. Pour cette raison, nous avons fait le choix d'utiliser en priorité DicoTopo pour les lieux situés dans les départements couverts par l'application et d'utiliser GeoNames pour ceux situés hors de France ou dans les départements non-couverts.

2. https://dicotopo.cths.fr/documentation.

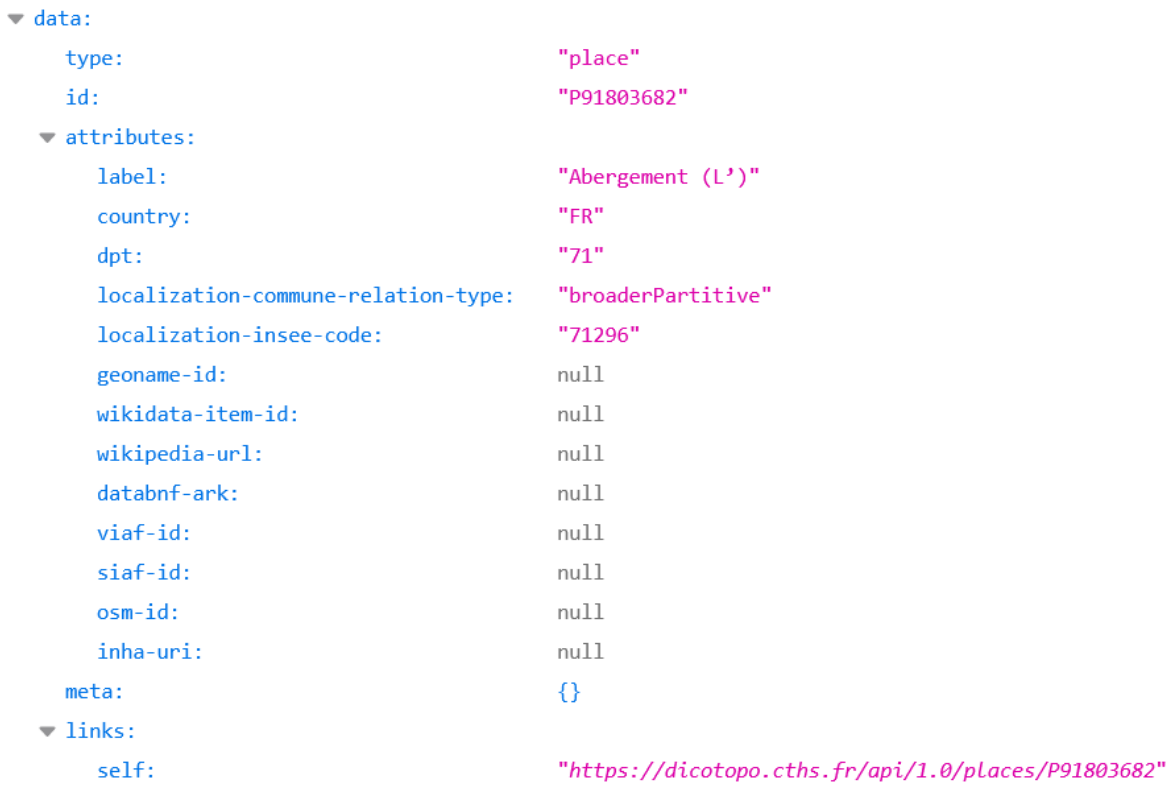

Figure 10 – *Exemple de retour de l'API DicoTopo.*

7.2 Construction de la requête

7.2.1 Choisir le nom à chercher

La recherche des toponymes a donc été réalisée à l'aide d'un algorithme python permettant d'écrire les requêtes en fonction des situations[3]. Une étape centrale consiste à vérifier si le toponyme est situé dans un département couvert par DicoTopo car elle détermine la suite de la procédure. Le cas échéant, la requête est construite à partir de l'url « https://dicotopo.cths.fr/api/1.0/search?query=label: » suivie du nom du lieu recherché. La réponse comprend toutes les entités du référentiel contenant ce nom. Des filtres sont donc opérés pour trier ces entités en fonction du département ou de la commune s'ils sont indiqués par l'index. Si aucune commune n'est indiquée, toutes les entités portant ce nom dans le département sont conservées comme des localisation potentielles. Par la suite, toutes les entités retenues sont transformées sous la forme d'un élément liste contenant leurs codes insee, leurs coordonnées et leurs identifiants dans DicoTopo.

Si le toponyme n'est pas situé dans un département couvert par DicoTopo, la recherche est effectuée sur l'API GeoNames. La requête est construite différemment : au lieu de filtrer les entités contenues dans la réponse en fonction du département ou de la com-

3. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Index/geonames.py. Lire aussi Virgile Reignier, *De l'index papier à l'indexation automatique…*

mune, nous avons mis en place les filtres dès la requête afin de prendre en compte les limites posées par l'interface. La requête est construite sous la forme : « http://api.geonames.org/search?style=FULL&name=[NAME]&username=[USERNAME] » où [NAME] correspond au nom du toponyme et [USERNAME] à l'identifiant du compte utilisateur. Les arguments supplémentaires sont ensuite « country=[COUNTRY] » où [COUNTRY] correspond au pays et « adminCode2=[NUM] ». Dans le contexte d'un toponyme français, [NUM] correspond au numéro du département. Des filtres sont néanmoins opérés pour vérifier que le nom correspond bien à celui recherché ou au moins à un élément du nom [4] (la comparaison est également étendue au contenu des balises <alternateName>) et pour ne garder qu'une seule entité par commune. Comme avec DicoTopo, les entités retenues sont ensuite transformées sous la forme d'un élément liste contenant leurs codes insee, leurs coordonnées et leurs identifiants GeoNames. Pour les localités dont le département renseigné correspond à « Seine » ou « Seine-et-Oise », la recherche doit être adaptée au contexte actuel. Les communes des deux anciens départements ayant été réparties entre les 6 nouveaux (Yvelines, Val-d'Oise, Essonne, Hauts-de-Seine, Seine-Saint-Denis, Val-de-Marne) [5], les requêtes doivent être démultipliées pour chercher le toponyme dans chacun des départements correspondants.

Il faut également prendre en compte des ambiguïtés parfois présentes entre l'entrée de l'index, la forme actualisée du nom et la commune. Malgré les catégorisations que nous avons présentées au chapitre 6, il est parfois délicat de différencier automatiquement des formulations similaires :

> Mauves-sur-Huine [Orne, con Mortagne]
>
> Mauvers [Tarn-et-Garonne, con Verdun, cne Aucamville]
>
> Mazis-Bocquet (Les) [Les Mazis, Seine-Maritime, con Neufchâtel, cne Saint-Saire]
>
> Molerie [Molières, Tarn-et-Garonne, ar. Montauban] , homines et universitas

Dans le premier cas, Mauves-sur-Huine correspond à une commune dans l'Orne. Dans le second cas, Mauvers est une localité située dans la commune d'Aucamville dans le Tarn-et-Garonne. Dans le troisième cas, Les Mazis Bocquet est une forme ancienne de la localité Les Mazis située dans la commune de Saint-Serre en Seine-Maritime. Dans le troisième cas, Molerie est une forme ancienne de la commune de Molières dans le Tarn-et-Garonne. On peut donc voir que la place du mot comme entrée index ou comme premier élément

4. Il arrive parfois que des parties du nom de la commune soient modifiées au cours du temps. Par exemple la ville de Mantes-la-Jolie actuellement dans les Yvelines s'appelait simplement Mantes au moment de la parution de l'index.

5. A l'exception de la commune de Paris qui forme une collectivité autonome.

entre crochet ne suffit pas à définir sa sémantique. Il faut donc réaliser plusieurs tests successifs pour s'assurer que la recherche sur les API trouve la bonne localité ou aligner le toponyme avec la commune dans laquelle il est situé en cas d'absence dans le référentiel.

Pour cela, nous avons établi une recherche préliminaire basée sur la commune lorsqu'elle est renseignée afin de disposer de ses coordonnées. Cette recherche est effectuée sur l'API GeoNames car celle-ci dispose d'un argument de recherche dédié aux entités disposant d'un code postal : « postalCodeSearch? ». La recherche s'effectue ensuite de la même manière que celle décrite précédemment, à ceci près que la réponse ne contient pas l'identifiant GeoNames de l'entité. Il faut donc construire une nouvelle requête à partir du résultat de la dernière sur le modèle « http://api.geonames.org/search?style=FULL &q=[POSTALCODE]&name=[NAME]&username=[USERNAME] » pour obtenir toutes les indications voulues sur la commune. Ici, l'utilisation de [POSTALCODE] permet de s'assurer que la commune contenue dans le retour est bien la même que celle trouvée avec la requête précédente. Si la commune n'est pas trouvée dans GeoNames et qu'elle se situe dans un département couvert par DicoTopo, alors une tentative est également effectuée via son API.

Cette étape permet ainsi de prendre en compte les différents éléments contenus dans les entrées. La recherche par toponyme est ainsi réalisée uniquement si son nom est différent de celui de la commune. Si la réponse ne contient pas de résultat satisfaisant, une nouvelle recherche est effectuée à partir de l'élément « nom_normalise ». Si aucune des deux ne permettent d'obtenir un résultat, alors les éléments extraits de la recherche par commune sont conservés pour enrichir le toponyme.

7.2.2 Extraction des coordonnées

Le résultat de ce processus est ensuite enregistré dans un fichier CSV qui reprend les éléments contenus dans le fichier formé à partir de la segmentation des entrées toponymiques (*cf.* chapitre 6) accompagné de colonnes supplémentaires pour y insérer le code INSEE, la latitude, la longitude et l'identifiant de l'entité dans le référentiel utilisé. De plus, le résultat des requêtes a été complété par une série de messages d'erreur permettant de suivre certains résultats associés aux différentes étapes du processus. Ces indications nous ont permis d'améliorer la rédaction de notre algorithme par le suivi de certaines variables et de leur adéquation avec les données sources.

Dans la plupart des cas, le résultat de ce processus permet d'associer chaque entrée d'index à une série complète de données supplémentaires. Mais quelques cas spécifiques peuvent se présenter : les localités étrangères ne disposent pas de code INSEE et quelques entités renseignées par DicoTopo ne disposent pas de coordonnées. Il arrive également que certaines localités ne disposent pas de coordonnées propres et s'étendent sur plusieurs

communes :

> Canteloup (fief de) [Calvados, ar. et con Lisieux, cnes Marolles et Fumichon]

Dans ce cas, les coordonnées associées au toponyme sont celles des deux communes Marolles et Fumichon. Cette situation est ici à distinguer des situations d'incertitude : dans ce dernier cas la multiplication des coordonnées correspond à la présence d'hypothèses différentes pour situer le lieu. Au contraire, la présence de coordonnées multiples signifie que l'entité s'étend sur la surface de plusieurs entités dans le référentiel utilisé. Nous avons donc ajouté une colonne « étendue » afin de pouvoir opérer cette distinction. Au moment de l'import des données dans Heurist, les coordonnées retenues des toponymes correspondant à cette situation sont calculées à partir de la moyenne des valeurs présentes dans les colonnes « latitude » et « longitude ». Dans le cas où ce sont des communes, tous les codes INSEE sont importés dans Heurist.

Comme exposé au chapitre 5, les sous-entrées dépendant d'entrées toponymiques n'ont pas pu être différenciées les unes des autres et ont toutes été catégorisées comme relevant du type « lieu ». En suivant cette logique, nous avons systématiquement enrichi chacune de ces sous-entrées avec les coordonnées de l'entité donc elles dépendent. Il arrive cependant que quelques unes de ces sous-entrées correspondent à des entités qui disposent de coordonnées propres :

> Bacqueville [Eure, con Fieury-sur-Andelle] , — forêt [Eure, con Fleury-sur-Andelle, cnes Bacqueville et Radepont]

Comme nous ne les avions pas identifiées au moment de la création du fichier CSV préliminaire au travail d'alignement, nous les avons repérées *a posteriori* par la présence de crochets après le séparateur « , — » et avons procédé manuellement à leur alignement avec les référentiels GeoNames et DicoTopo. Dans cette situation, ces coordonnées ont été préférées à celles de l'entrée générale au moment de l'import dans Heurist.

7.3 Gestion des cas spécifiques

7.3.1 Résoudre les ambiguïtés

Comme nous l'avons vu, la prise en compte des différentes possibilités de composition des entrées nous a poussés à tester plusieurs scénarios au moment de la construction des requêtes. Nous avons ainsi dû arbitrer entre une récupération large de résultats pour s'assurer de trouver la bonne entité et des filtres efficaces afin de limiter la présence d'entités superflues. Malgré nos efforts, un certain nombre de reprises ont été effectuées.

Figure 11 – *La recherche pour la commune « Devesset » dans GeoNames renvoie deux entités : une pour la ville et une pour la circonscription administrative*

C'est le cas notamment des requêtes opérées sur DicoTopo dont les retours contiennent régulièrement des doublons que nous avons dû éliminer. Quant à GeoNames, le retour des requêtes contient parfois plusieurs résultats pour une même entité. C'est le cas par exemple des communes françaises qui sont référencées à la fois comme des lieux d'habitation et comme des circonscriptions administratives (*cf.* figure 11). Nous avons fait ici le choix d'éliminer systématiquement les circonscriptions administratives lorsqu'il existe une autre entité portant le même nom. Cette situation se retrouve également pour un certain nombre de grandes villes étrangères dont le nom est le même que celui de l'entité administrative qui la contient (par exemple la ville de Milan est le chef-lieu de la ville métropolitaine de Milan qui contient 134 communes).

D'autres ambiguïtés sont présentes à partir de certaines rivières dont le nom est présent dans celui de plusieurs communes riveraines ou dans le nom d'un département (par exemple la rivière Allier donne son nom au département du même nom ainsi qu'à la commune Varennes-sur-Allier). De la même manière le nom de certaines régions historiques est encore porté dans celui de quelques communes (par exemple Laurac-en-Vivarais, située dans le Vivarais historique). Pour éviter les ambiguïtés, nous avons repris manuellement l'ensemble de ces entrées (repérées par la présence de mots-clefs comme « pays », « rivière », « région », ...) pour vérifier que les entités avec lesquelles elles ont été alignées

correspondent bien à celles voulues. Une fois ce travail accompli, il reste encore quelques ambiguïtés issus du processus d'alignement qui n'ont pas pu être résolues de manière automatique ou semi-automatique [6]. Nous avons donc conservé chacune des entités issues du processus d'alignement et les avons considérées de la même manière que les propositions de localisation formulées par l'index au moment de l'import dans Heurist [7].

7.3.2 Les régions historiques

Parmi les cas particuliers présents dans notre corpus, les régions historiques sont particulièrement difficiles à aligner avec des coordonnées [8]. Si certaines d'entre elles sont présentes en tant que telles dans GeoNames ou DicoTopo (par exemple l'Armagnac, GeoNames ID 3036926), d'autres ont été alignées avec la commune qui constituait le chef-lieu :

> Termenès [pays de l'Aude dont Termes était le chef-lieu] , juge

L'entité « Termenès » a été dotée d'un attribut « Name chosen » avec pour valeur « Termes » et les coordonnées choisies pour l'entité sont celles du chef-lieu. Nous avons également pu trouver quelques coordonnées supplémentaires dans Wikidata comme par exemple pour le Léon (Wikidata ID Q12178). Un dernier cas spécifique est celui des régions historiques situées à cheval entre plusieurs pays actuels :

> Quatre-Métiers (les) [partie de la Flandre relevant de l'Empire et composée des paroisses d'Hulst (Pays-Bas, prov. Zélande), Axel (idem), Bouchaute (Belgique, prov. Flandre orientale) et Assenede (idem)]

Cette région est présente dans Wikidata (ID Q2789324), mais ne dispose pas de coordonnées. Nous avons donc aligné l'entité avec cet identifiant, ainsi qu'avec celui des communes d'Hulst, Axel, Bouchate et Assenede dans GeoNames et nous avons calculé des coordonnées à partir de la moyenne des coordonnées de ces communes.

Nous avons également été confronté à quelques ambiguïtés dans les formulations proposées par l'index :

> Dunois (le) [région de Dun, Ariège, con Mirepoix] , châtelain

> Rabonitum [Ariège, région Pamiers]

6. A ce propos, v. aussi *Ibid.*
7. *Cf.* chapitre 6.
8. A ce propos, v. aussi Katherine McDonough, Ludovic Moncla et Matje Camp, "Named entity recognition goes to old regime France : geographic text analysis for early modern French corpora", *International Journal of Geographical Information Science*, 33 (27 mai 2019), DOI : 10.1080/13658816.2019.1620235.

Si la première entrée correspond sans conteste à une région dont le chef-lieu est Dun, la seconde correspond *a priori* à une localité située près de Pamiers sans que la distinction entre les deux possibilités ne puisse être définie de manière certaine.

Conclusion

Ce chapitre nous a permis d'aborder les différentes étapes du travail d'alignement entre les entités créées à partir de l'index et plusieurs référentiels en ligne disposant de coordonnées. Nous avons pour cela présenté les caractéristiques des deux principaux outils que nous avons utilisés, puis nous avons décrit le processus de construction de la requête à partir des différents noms à chercher et le choix des coordonnées récupérées, enfin nous avons exposé les différents cas spécifiques qui se sont présentés au cours de notre travail.

Cette étape constitue donc la dernière de la mise en place du référentiel d'entités nommées destiné à nourrir l'entrainement d'un modèle de liage d'entités sur les actes du Trésor des chartes. Sur les 4 808 toponymes isolés au sein du fichier CSV, 4 428 (92 %) ont été enrichis et sont venus compléter les éléments importés sur Heurist. L'ajout de coordonnées permet également d'envisager de dépasser le simple cadre du travail sur le liage d'entités. Les données ainsi localisées permettent en effet de dresser un certain nombre de visualisations cartographiques à partir des données du Trésor des chartes et utilisables à des fins multiples.

Chapitre 8

Mise à disposition d'un nouveau référentiel

Une fois constituée, la base de données Heurist contenant les entités décrites par l'index accompagnant l'inventaire analytique des registres JJ 37 à JJ 50 du Trésor des chartes peut être utilisée pour plusieurs usages. Si certains sont liés directement aux techniques de lecture automatique des textes, d'autres concernent plus directement le contenu des documents. Cet outil permet par exemple de faire des études croisées sur l'utilisation d'une notion ou de visualiser l'évolution du lien entre le pouvoir royal et une région au cours de la période. Il paraît donc nécessaire de rendre ces données disponibles aux usagers qui souhaitent parcourir le contenu de ces archives afin de les réutiliser dans leurs propres recherches.

Ce chapitre sera donc consacré au travail de mise à disposition des données manipulées pendant le stage afin de faciliter leur usage pour la suite du projet et pour la communauté scientifique. Pour cela, nous détaillerons dans un premier temps les éléments mis en ligne pendant le stage et les liens visibles entre les différents fichiers. Puis nous présenterons les différentes visualisations cartographiques permises par la présence de coordonnées. Enfin, nous rendrons compte du travail préparatoire à la suite du projet par l'import des données dans Arkindex.

8.1 Mise en ligne des données collectées

8.1.1 Des compléments au dépôt Github

Le travail réalisé au cours du stage à partir de l'index nous a permis d'enrichir les données déjà en ligne sur le dépôt Github du projet Himanis (https://github.com/oriflamms/himanis). Nous y avons notamment ajouté le fichier XML final contenant les

```
<index>
   <term type="subject">Actes en français</term>
   <term type="subject">B. [abréviation de <hi rend="italic">Bona</hi>, note
      signalant les actes remarquables par leur formulaire]</term>
   <term type="subject">Chevaliers de Philippe le Bel, <hi rend="italic">milites
        regis</hi>
   </term>
   <term type="subject">Chirurgiens</term>
   <term type="subject">Commissions du roi de France</term>
   <term type="subject">Droit [public et privé] d'aubaine</term>
   <term type="subject">Enquêteurs réformateurs</term>
   <term type="subject">Maisons [au sens actuel, ou au sens de manoirs]</term>
   <term type="subject">Portes de villes</term>
   <term type="subject">Receveurs du roi</term>
   <term type="subject">Tournois [monnaie réelle] , petits --</term>
   <term type="subject">Ventes par le roi ou à son nom</term>
   <term type="subject">Vidimus et confirmations de Philippe le Bel</term>
</index>
```

Figure 12 – *Exemple d'entrées d'index insérées dans une entrée d'inventaire après suppression des éléments issus du fichier XML pour ne laisser que ceux issus du tableau des Archives Nationales.*

entrées d'index [1]. Ce fichier permet de montrer un état intermédiaire de la transformation de l'index car il est construit de manière linéaire et contient l'ensemble des éléments présents dans la publication papier. De plus, chaque entrée utilisée par la suite dispose d'un identifiant unique et de liens visibles avec les autres entrées avec laquelle elle est associée. Nous avons également ajouté dans le même dépôt le fichier Excel utilisé pour l'import des données dans Heurist [2].

Ce travail a également été l'occasion de mettre à jour le fichier XML contenant les entrées de l'inventaire [3]. Nous avons en effet décrit au chapitre 3 le travail de mise à plat des entrées d'index au sein de l'inventaire qui avait été réalisé dans les précédentes étapes du projet (*cf.* figure 7). Une fois les données de l'index importées dans Heurist, nous avons donc procédé à l'élimination systématique des entrées ajoutées dans les entrées d'inventaire. Nous avons cependant conservé un certain nombre d'éléments de type « subject » car ils ne sont pas directement issus de ce travail de mise à plat mais de l'exploitation d'un tableau produit par les Archives Nationales. Ces entrées ont été transformées de manière différentes et des comparaisons pourront être réalisées avec les éléments contenus dans la base de données. Le résultat prend la forme d'une balise <index> au contenu allégé comme présenté par la figure 12.

1. https://github.com/oriflamms/himanis/blob/master/Inventories/Systematic/Paris_AN_JJ_inventaire_JJ37-50_index.xml.
2. https://github.com/oriflamms/himanis/blob/master/Inventories/Systematic/Paris_AN_JJ_inventaire_JJ37-50_index.xlsx.
3. https://github.com/oriflamms/himanis/blob/master/Inventories/Systematic/Paris_AN_JJ_inventaire_JJ37-50.xml.

8.1.2 Des liens visibles entre les données

La mise en ligne de la base de données sur la plateforme Heurist permet d'envisager la diffusion du référentiel à large échelle et la démultiplication de ses usages. Cependant, les données ne forment pas actuellement l'état final visé [4]. Leur publication est donc encore à l'état de projet. Nous avons toutefois pensé les données importées dans Heurist en fonction de cette future mise à disposition. Pour cette raison, un attribut « Index reference » a été ajouté aux entités Place, Person et Subject pour les associer aux entrées d'index au sein des fichiers de travail. Cet attribut est construit à partir du nom du fichier contenant cette entrée d'index et de l'identifiant de l'entrée en question, sous la forme :

Paris_AN_JJ_inventaire_JJ37-50_index.xlsx#d1e132079

De la même manière, les entrées de la table Act disposent d'un attribut « Inventory Reference » qui contient le nom du fichier contenant cette entrée d'inventaire et le numéro de l'acte dans le même inventaire sous la forme :

Paris_AN_JJ_inventaire_JJ37-50.xml#433

Ces éléments permettent de faciliter le suivi des données depuis les documents sources vers la base de données afin de retracer les différentes étapes de sa transformation et de retrouver les incohérences qui pourraient s'y être glissées. Ce suivi paraît particulièrement pertinent en cas d'ouverture de la base de données au *crowdsourcing* car il permettra à chacun de vérifier la cohérence des données avec les documents sources et de proposer les ajouts ou corrections qu'il jugera nécessaire.

8.2 Un résultat direct : visualiser les lieux mentionnés dans les registres JJ 37-60

8.2.1 Génération d'une carte de chaleur

L'alignement des entrées toponymiques avec des référentiels externes pour les enrichir de coordonnées permet d'apporter une dimension spatiale au contenu de l'index. Chacune des entités enrichies peut en effet être projetée sur une carte et offrir autant de visualisation du contenu des registres. Nous avons donc travaillé à la production d'une carte de chaleur permettant une visualisation complète des lieux mentionnés par l'index. Cette réalisation permet à la fois de montrer un état de notre travail en cours et de proposer une cartographie des lieux mentionnés dans les registres JJ 37 à 50 du Trésor des chartes et donc concernés d'une manière directe ou indirecte par l'action du roi de France

4. Nous avons formulé quelques propositions d'amélioration au sein de l'annexe C.

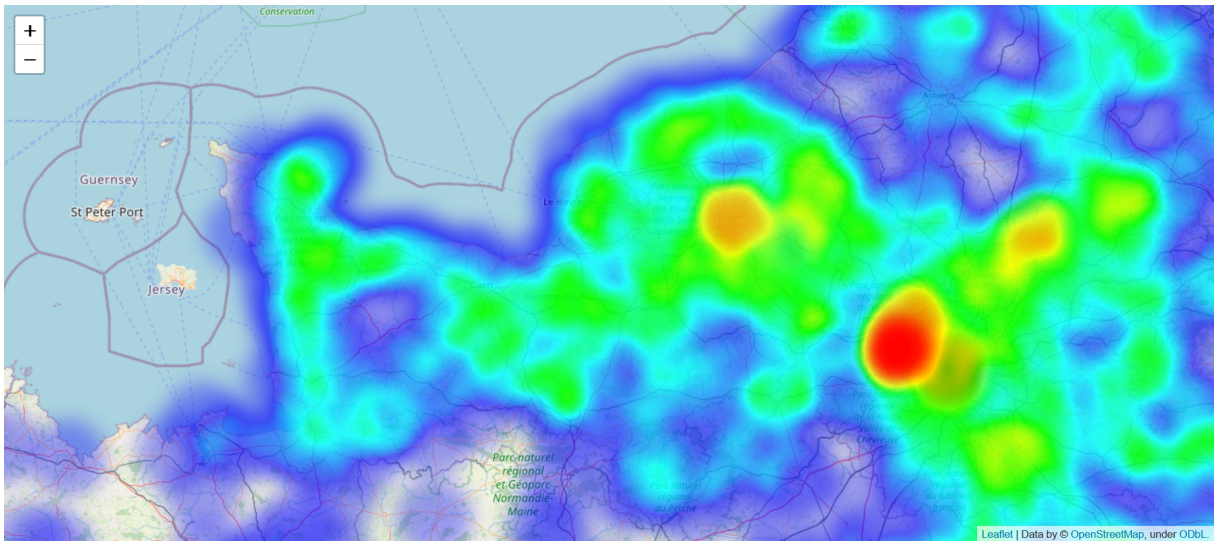

Figure 13 – *Carte des lieux mentionnés dans les registres JJ 37 à JJ 50 du Trésor des Chartes, zoom sur la Normandie.*

entre 1303 et 1314. Cette carte a été réalisée à l'aide d'un script python [5] et est disponible à l'adresse suivante : https://virgile-reignier.github.io/Carte-JJ37-50.

Les données ont été extraites au cours de notre travail à partir du fichier Excel contenant les entrées de l'index et du fichier CSV contenant les toponymes enrichis de coordonnées. Chaque entrée de type « place » a ainsi été alignée avec les coordonnées qui lui correspondent et pondérée en fonction du nombre de renvois vers l'inventaire qu'elle contient. Dans le cas d'une sous-entrée ne disposant pas de coordonnées propres, nous avons utilisé celles de l'entrée générale. Trois essais ont ainsi été réalisés à l'aide de la librairie python folium qui permet l'utilisation de la librairie javascript leaflet et du service OpenStreetMap [6]. Le premier contient l'ensemble des résultats, le deuxième exclut les résultats qui contiennent « Paris » dans leur nom et le troisième ajoute au second une étiquette pour chaque point. Seul le deuxième a été retenu pour la publication finale car le troisième nous a semblé trop peu clair pour être utilisable et le premier rendu illisible par la place prépondérante prise par Paris. Cette production permet ainsi de visualiser globalement les lieux mentionnés dans les registres JJ 37 à JJ 50 et de réaliser des zooms ciblés sur des régions ou des localités pour lesquelles la présence du roi est significative (comme par exemple la Normandie pour la figure 13).

8.2.2 Cartographie intégrée à Heurist

La plateforme Heurist permet également de produire des visualisations cartographiques à partir des données géolocalisées en utilisant la même librairie javascript leaflet.

5. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Index/carte_toponymes.py.

6. https://python-visualization.github.io/folium.

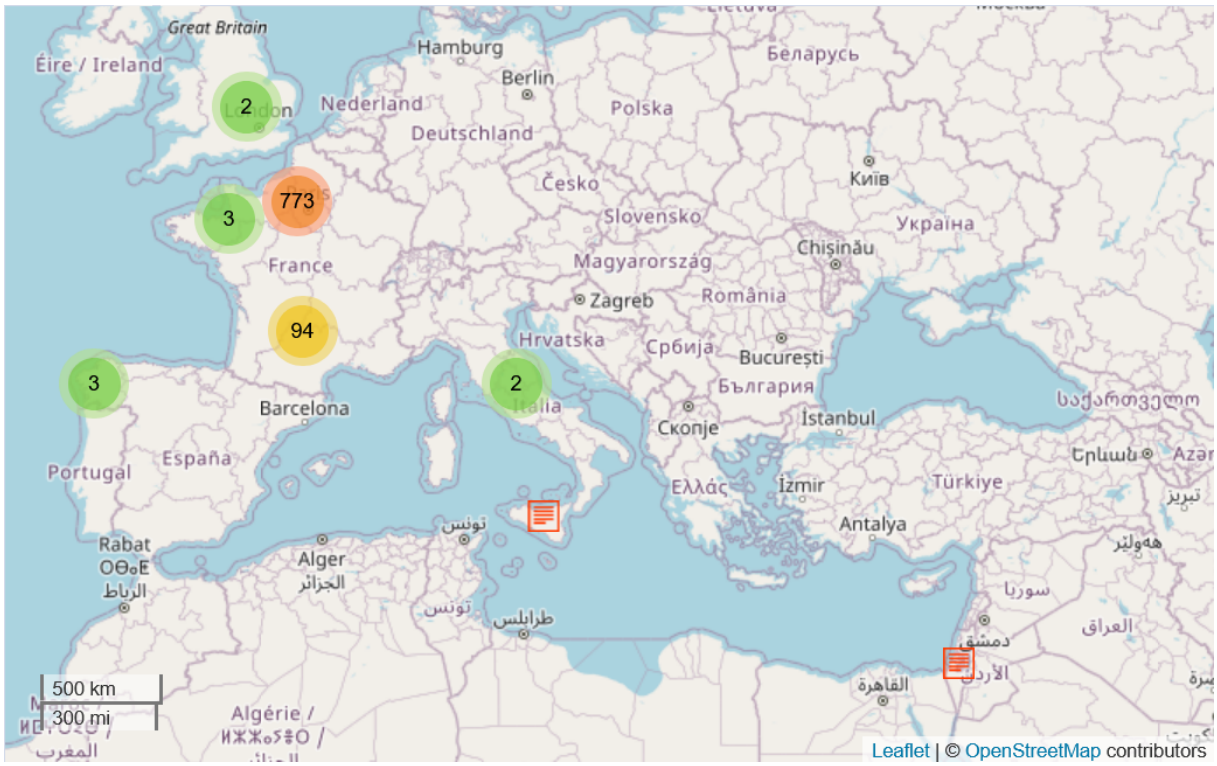

Figure 14 – *Carte des lieux mentionnés dans les actes contenant le nom d'Enguerrand de Marigny.*

L'interface offre de nombreuses possibilités car il est possible de trier les données en fonction des éléments que l'on veut observer en utilisant les jointures entre plusieurs tables. On peut par exemple visualiser les lieux mentionnés dans les actes où un nom précis apparaît (par exemple Enguerrand de Marigny pour la figure 14) ou visualiser les lieux contenus dans les actes réalisés pendant une période précise (par exemple au cours du pontificat de Benoît XI pour la figure 15). Ces requêtes jointes démultiplient ainsi les possibles études sur les liens entre des entités ou sur des collections d'actes du Trésor des chartes établis en fonction de critères précis.

8.3 Import des actes dans Arkindex

8.3.1 Les fonctionnalités de la plateforme

La dernière étape de la mise en ligne des données concerne la plateforme Arkindex utilisée pour l'entraînement des modèles de lecture automatique des textes. Les données issues de la segmentation en acte des registres transcrits ont en effet été intégrées dans un fichier JSON (*cf.* chapitre 2) et doivent être chargées dans la plateforme pour faire le lien entre le texte et les entités contenues dans la base de données. Cette étape a été réalisée à

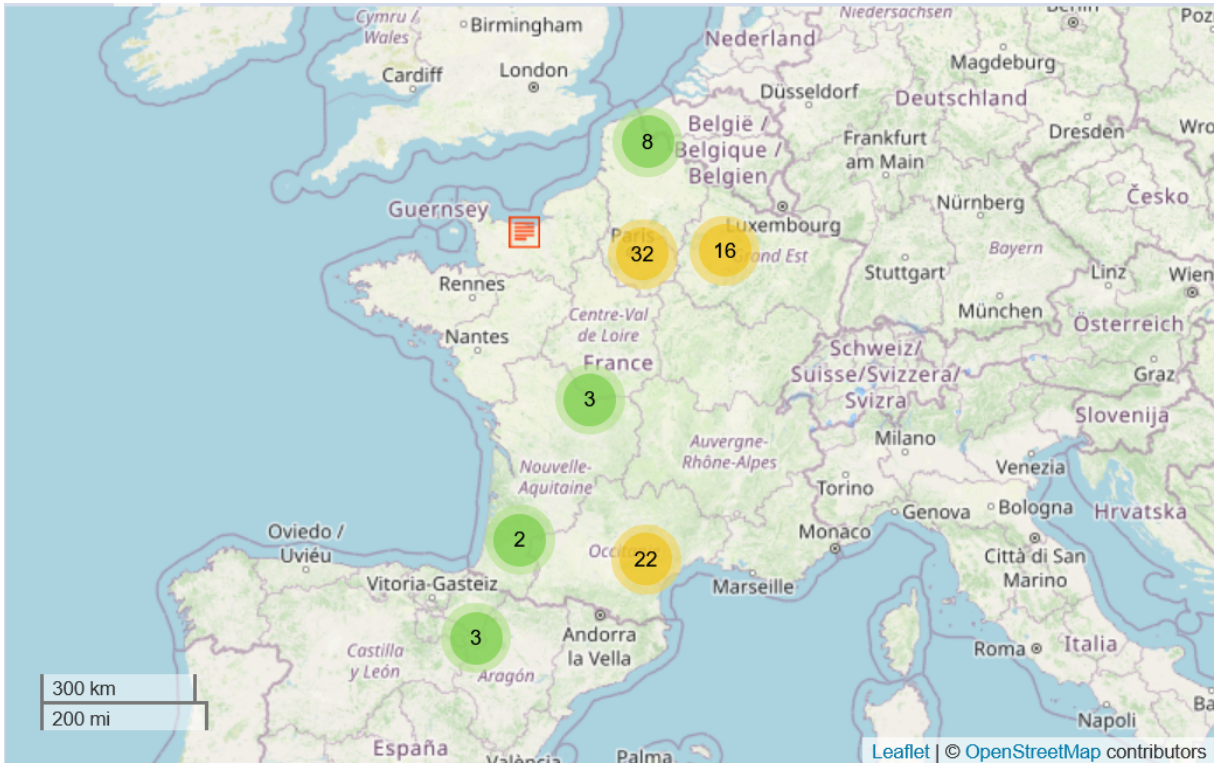

Figure 15 – *Carte des lieux mentionnés dans les actes réalisés pendant le pontificat de Benoît XI (27 octobre 1303 - 7 juillet 1304).*

partir de la librairie python arkindex-client qui fournit un client API pour la plateforme [7].

L'utilisation de ces fonctions nécessite de déclarer au préalable des identifiants comme variables d'environnement afin de se connecter à un compte Arkindex. Les différentes fonctions utilisables par l'API sont décrites par la documentation suivante : https://arkindex.teklia.com/api-docs. Dans le cadre de notre travail, nous avons utilisé la méthode « paginate » dans le cadre des requêtes GET pour chercher des éléments dans un corpus Arkindex. Cette méthode de requête est particulièrement utile pour manipuler un grand nombre de données car le contenu est organisé dans des éléments « Page » qui ne sont chargés qu'au moment où ils sont utilisés par le programme. Le retour est ensuite rassemblé dans un élément python ResponsePaginator et peut être manipulé et utilisé pour d'autres opérations. Les arguments utilisés pour les recherches sont « ListElements » si on cherche des éléments en fonction d'un attribut ou « ListElementChildren » si on cherche leurs éléments enfants.

Les requêtes POST utilisent quant à elles la méthode « request » et permettent d'envoyer du contenu dans la plateforme. Les arguments utilisés dépendent de l'action souhaitée : « CreateElement » permet de créer un nouvel élément en renseignant son titre, son nom, l'identifiant du corpus et si besoin l'identifiant de l'élément parent ; « CreateMetaData » permet de créer une métadonnée en renseignant son type, son nom, son

7. https://pypi.org/project/arkindex-client.

contenu et l'identifiant de l'élément concerné ; « CreateElementParent » permet de créer une relation enfant-parent en renseignant les identifiants des éléments enfants et parents concernés ; « CreateTranscription » permet de créer une transcription en renseignant le texte et l'identifiant de l'élément concerné. Tous les contenus sont décrits à l'aide d'un élément dictionnaire dont les clés correspondent au type de données nécessaires à l'établissement de l'action souhaitée.

8.3.2 Mettre en lien les éléments connus sur les images et les actes qu'elles contiennent

Dans ce cadre, nous avons donc élaboré en collaboration avec Teklia un script python permettant d'organiser l'import des données contenues dans le fichier JSON. L'algorithme se calque sur la structure du fichier que nous avons décrite au chapitre 2 (*cf.* figure 3) et itère des opérations successives avec l'API pour chaque élément correspondant à un acte. La première opération consiste à rechercher l'identifiant du volume dans lequel se trouve l'acte à l'aide de ListElements, puis à créer un élément Act comme enfant de ce volume avec CreateElement. CreateMetaData permet ensuite d'intégrer un à un les attributs de l'acte dans le fichier JSON comme métadonnée de l'élément Act selon la concordance suivante :

Attribut dans le fichier JSON	Metadata dans Arkindex
Provisory_index_3	himanisId
Inventory_Name + '_' + Inventory_Nr	inventoryReference
Regeste	abstract
Date	date_orig
normalized_language/language	language

Une métadonnée supplémentaire « date » est créée à partir de l'attribut « Date-normalisee » transformé sous la forme YYYY-MM-DD s'il s'agit d'une date exacte ou YYYY-YYYY s'il s'agit d'un intervalle.

Les zones de textes sont ensuite décrites sous forme d'un élément liste dans l'attribut « Text_Region » et permettent de décrire toutes les parties des actes au sein des pages. Leur modélisation au sein d'Arkindex doit donc prendre en compte cette situation d'interface entre la structure physique et la structure logique des registres. Pour cette raison, nous avons d'abord cherché le nom du folio dans lequel se situe cette zone de texte à l'aide de l'adresse de l'image dans la BVMM et de la table de concordance entre les folios et les images que nous avons décrite au chapitre 6, puis nous avons cherché

l'identifiant de ce folio dans Arkindex avec ListElements. Nous avons ensuite créé un élément Text zone comme enfant de ce folio à l'aide de CreateElement. L'élément est doté d'un nom qui reprend la valeur de l'attribut « Act_N » précédé de la chaîne « Acte » et d'une image qui est la même que celle de l'élément parent. Cette image est découpée à partir du contenu de l'attribut « Graphical_coord » pour n'en retenir que la partie qui concerne l'acte. Ces zones de texte sont aussi liées aux actes que nous avons importés précédemment, nous avons pour cette raison rajouté un lien enfant-parent depuis l'élément Text zone vers l'élément Act à l'aide de CreateElementParent. Nous avons pour finir utilisé CreateTranscription pour importer la transcription du texte contenu dans l'image et CreateMetaData pour intégrer la valeur de l'attribut « type_act comme Metadata « part » selon la correspondance suivante :

Valeur de l'attribut « type_act »	Valeur de la Metadata « part »
AI	initial
AF	final
AM	middle
AS	supplemental
AC	complete

Cette situation permet aux éléments « Text zone » d'être à la fois des enfants des éléments « Page » et des éléments « Act ». Ces derniers représentent donc le contenu intellectuel du texte et ne sont associés à aucune image de manière directe. En revanche, il est possible de visualiser le contenu physique d'un élément « Act » dans Arkindex à travers l'affichage successif de chacun des éléments « Text zone » qu'il contient comme présenté par la figure 16.

De la même manière, nous avons mis en lien les éléments présents dans Heurist avec ceux chargés dans Arkindex. L'attribut « Address bvmm » a notamment été modifié pour devenir du type « File ». Ce type permet notamment de considérer les url en tant que tels et, dans le cas où il s'agit d'une image, d'en visualiser le contenu. Grâce à cela, il est possible d'explorer les images et leurs transcriptions dans l'interface Heurist de la même manière que dans l'interface Arkindex. Pour finir, nous avons rédigé un script python pour extraire automatiquement les identifiants des entités Act dans Arkindex et les aligner avec les entités correspondantes dans Heurist en utilisant l'attribut « provisory_index_3 » comme clé d'alignement [8]. Ces identifiants Arkindex ont ensuite été ajoutés aux entités Act dans Heurist comme valeur de l'attribut « Id Arkindex ».

8. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Normalize_json/export_id_arkindex.py.

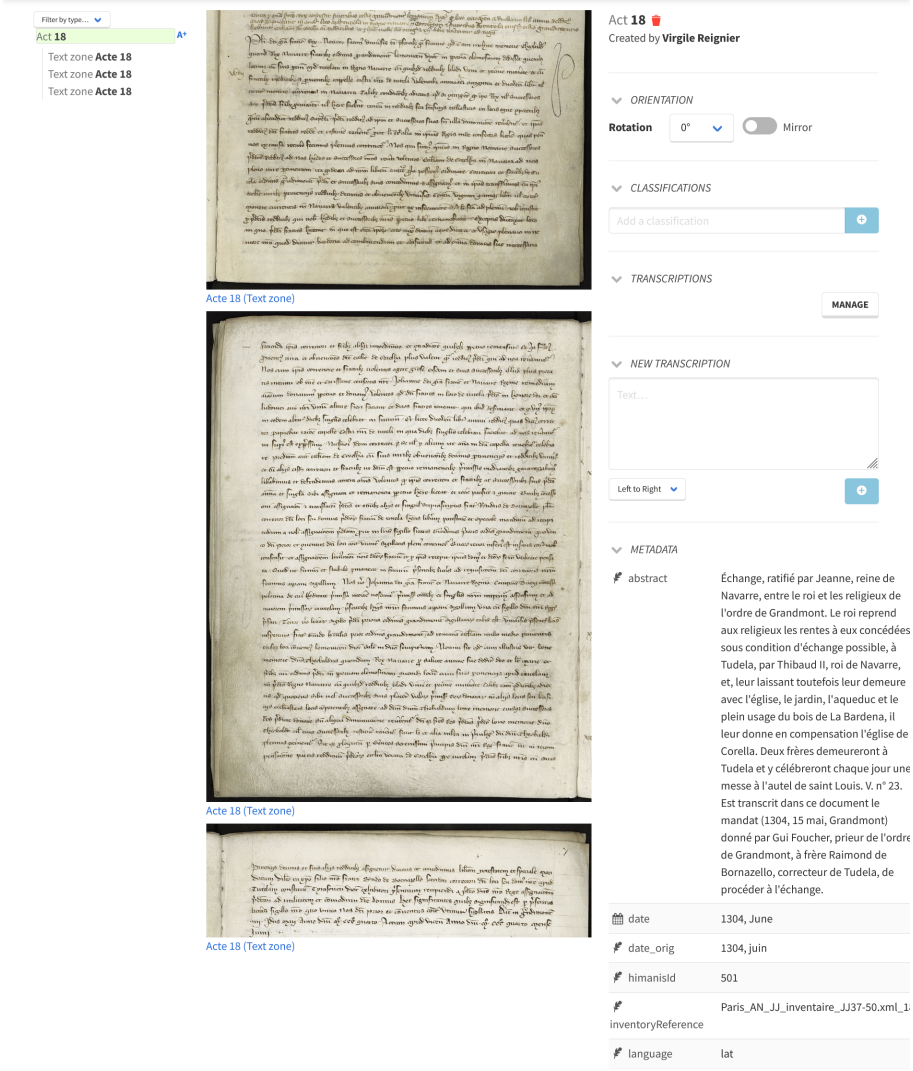

Figure 16 – *Visualisation d'un élément « Act » dans Arkindex. Cet élément est le parent de plusieurs éléments « Text zone » qui représentent chacun une portion du texte décrit par l'élément « Act ».*

Conclusion

Ce chapitre nous a permis d'aborder les différents usages possibles à partir des données que nous avons manipulées au cours du stage et les travaux préparatoires que nous avons réalisés pour anticiper ces usages. Nous avons pour cela décrit dans un premier temps la mise en ligne des données à travers la mise à jour des fichiers sur le dépôt Github du projet Himanis et l'ajout de renvois vers ces fichiers dans les entités présentes dans la base de données. Puis nous avons détaillé les différentes visualisations cartographiques disponibles à partir de la carte que nous avons publiée et de l'outil dédié dans la plateforme Heurist. Enfin, nous avons exposé le travail d'import des données issues des inventaires dans la plateforme Arkindex et la création de liens entre les données hébergées dans les deux plateformes.

La mise à disposition des données - que ce soit pour leur publication ou leur intégration dans Arkindex pour de prochains travaux - crée donc un réseau global d'informations sur le Trésor des chartes qui s'inscrit dans la logique du web de données. Si la partie chargée dans Arkindex est réservée aux usagers du projet, les données dans Heurist sont quant à elles destinées à devenir publiques. La création de liens entre les entités issues de l'index, des référentiels libres de droits et des documents librement accessibles sur Github participe donc à la démarche d'accessibilité et d'interopérabilité des connaissances décrite par les principes FAIR[9]. Les données collectées pourront ainsi rejoindre les différentes ressources disponibles et serviront de base pour de multiples usages.

9. CCSD, *Principes FAIR*, URL : https://www.ccsd.cnrs.fr/principes-fair/ (visité le 08/11/2022).

Chapitre 9

Mise en œuvre du liage d'entités

Une fois notre référentiel modélisé, formalisé, mis en ligne et lié aux autres données du corpus, la dernière étape de notre stage repose sur la mise en application du liage d'entités dans le cadre des registres du Trésor des chartes. Nous avons à cet effet réalisé quelques essais à partir de la librairie python spaCy pour évaluer différentes manières de procéder. Cette étape étant intervenue à la toute fin de notre stage, nous n'avons pas eu l'occasion de poursuivre jusqu'à obtenir des résultats concluants.

Ce chapitre nous permettra donc de rendre compte de l'état de notre réflexion sur la mise en œuvre du liage d'entités. Pour cela, nous exposerons dans un premier temps la méthode utilisée à partir de l'outil spaCy et l'objectif visé. Puis nous présenterons les différents essais que nous avons réalisés et les conclusions que nous pouvons en tirer. Enfin, nous proposerons un bilan de notre travail et nous émettrons quelques suggestions pour le poursuivre.

9.1 Appliquer le liage d'entités au contexte des données du projet Himanis

9.1.1 Présentation de l'outil spaCy

Nous avons décrit au chapitre 1 le principe du liage d'entités et les perspectives pour sa mise en œuvre dans le cadre de l'étude des documents historiques. Le défi de la mise en place d'une base de connaissances apte à décrire les entités nommées reconnues dans le corpus ayant été résolu par la construction de la base de données à l'aide de l'index, il nous reste encore à trouver une méthode de travail adéquate à la réalisation complète du processus. Nous avons pour cela utilisé la librairie python spaCy spécialisée dans le

TAL [1]. Pour nous familiariser avec ses fonctionnalités, nous avons également pu profiter d'une séance de tutoriel organisée le 19 mai 2022 par le réseau Mate-SHS [2] et consacrée à cette librairie python [3].

Cet outil propose notamment de charger un modèle de TAL et de l'appliquer sur un texte afin de mettre en œuvre toute un chaîne de traitements. Les principales fonctions que nous avons utilisées sont la tokenisation (c'est-à-dire le découpage du texte en unités lexicales), la lemmatisation (c'est-à-dire la réduction des mots à leur forme canonique), l'analyse morphologique et l'étiquetage de la partie du discours. Ces annotations sont visibles respectivement à l'aide des attributs « lemma_ », « morph » et « pos_ » conservées dans les variables tokens. Par exemple l'initiation du modèle et du texte suivant :

```
nlp = spacy.load("fr_core_news_sm")
doc = nlp("Vienne est une ville")
```

permet de réaliser les affichages d'attributs suivant :

```
for token in doc:
    print(token, token.lemma_, token.pos_, token.morph)
```

et d'obtenir ce résultat :

```
Vienne Vienne PROPN
est être AUX Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin
une un DET Definite=Ind|Gender=Fem|Number=Sing|PronType=Art
ville ville NOUN Gender=Fem|Number=Sing
```

Les modèles peuvent également intégrer d'autres traitements comme la REN. Les entités reconnues sont conservées dans l'attribut « ents » du texte analysé. A leur tour, les variables entités contiennent un texte et une étiquette conservés dans les attributs « text » et « label_ ». Pour l'exemple initié ci-dessus, la commande d'affichage :

```
for entity in doc.ents:
    print(entity.text, entity.label_)
```

permet d'obtenir le résultat suivant :

```
Vienne LOC
```

1. https://spacy.io.
2. https://mate-shs.cnrs.fr.
3. La retransmission de la séance est disponible ici : https://www.youtube.com/watch?v=Z58g33GglR0 et le support d'accompagnement ici : https://github.com/clement-plancq/tuto-mate.

9.1.2 L'objectif visé : construire une vérité terrain

Cet apprentissage a ensuite été complété par un autre tutoriel plus spécifique au liage d'entités [4]. Le nœud central de cette technique repose sur l'établissement de la base de connaissances. Celle-ci est construite avec la classe « KnowledgeBase » qui est initialisée à l'aide de la classe « Vocab » extraite depuis le modèle utilisé et le vecteur de mot qui lui correspond (ici [N] correspond à la taille du vecteur de mots) :

```
kb = KnowledgeBase(vocab=nlp.vocab, entity_vector_length=[N])
```

Les entités de la base de connaissances sont ensuite insérées en deux étapes. Dans un premier temps, on renseigne des entités à l'aide d'un identifiant « qid » et d'une description « desc » pré-traitée par le modèle utilisé (« 342 » est ici une valeur arbitraire) :

```
desc_doc = nlp(desc)
kb.add_entity(entity=qid, entity_vector=desc_doc.vector, freq=342)
```

Dans un second temps, on renseigne les alias permettant de mentionner les entités. Les deux étapes sont séparées car plusieurs mentions peuvent correspondre à une même entité et inversement. Chaque mention « alias » est donc ajoutée à partir d'une chaîne de caractères « name », de la liste des entités correspondantes « qids » et de la liste des probabilités que cet alias corresponde à chaque entités « probs » :

```
kb.add_alias(alias=name, entities=qids, probabilities=probs)
```

Les listes « qids » et « probs » doivent donc être de taille égale et la somme des probabilités doit être inférieure à 1.

Il est alors possible de constituer des données d'entraînement à partir d'un ensemble de textes au sein desquels les entités nommées ont été reconnues et associées aux entités de la base de connaissances. Ces données permettent d'entraîner la reconnaissance de ces entités et d'ajouter cette nouvelle fonctionnalité au modèle utilisé. La dernière étape du processus consiste à tester l'efficacité du modèle en chargeant des textes contenant des entités nommées déjà connues et identifiées et en vérifiant qu'il les associe à la bonne entité de la base de connaissances. L'objectif poursuivi est donc d'utiliser à terme la base de données Heurist et le lien entre textes et entités reconnues dans ces textes comme vérité terrain pour entraîner un modèle de liage d'entités et l'appliquer sur les registres non-indexés du Trésor des chartes.

4. https://spacy.io/universe/project/video-entity-linking. Le support utilisé est disponible ici : https://github.com/explosion/projects/tree/master/nel-emerson.

9.2 Des réalisations aux résultats mitigés

9.2.1 Le processus de travail

Nous avons donc rédigé un script python calqué sur celui proposé par le tutoriel afin de mettre en application les fonctionnalités décrites pour notre contexte de travail. Nos essais ont été réalisés à partir de deux modèles différents : le modèle « fr_core_news_md » fourni par spaCy et dédié au traitement des textes en français et le modèle « multi-home-c3po4-LOC-model-best » utilisé par Teklia pour réaliser de la REN sur les textes en latin et en langue vernaculaire contenues dans les registres du Trésor des chartes. La première difficulté que nous avons rencontrée concerne l'interopérabilité des modèles avec notre environnement de travail : le premier a été entraîné avec la version 3.4.0 de spaCy et le second avec une version 2.3 qui elle-même nécessite d'installer python en version 3.7.

Nous avons ensuite réalisé un script python à partir de celui proposé par le tutoriel pour mettre en application ces deux modèles. Cet algorithme met en œuvre plusieurs fichiers que nous avons créés pour l'occasion [5]. Nous avons centré notre test sur la reconnaissance de deux groupes d'entités porteuses d'ambiguïtés. Le premier groupe concerne la chaîne « Vienne » qui est présente dans 4 toponymes de la table Place dans la base de données (Mesnil-sous-Vienne dans l'Eure, Aixe-sur-Vienne dans la Haute-Vienne, Vienne dans l'Isère et Vienne dans la Seine-et-Marne). Le second groupe concerne la chaîne « France » présent dans 2 toponymes (France, le territoire correspondant au pays actuel et la ville de Saint-Denis-en-France, ancien nom de Sain-Denis dans l'actuelle Seine-Saint-Denis). Ces entités sont rassemblées dans un fichier « kb_Vienne.csv » qui est ensuite chargé pour nourrir les entités de la base de connaissances. Ce fichier est complété par un autre fichier « alias.csv » qui contient les toponymes tels qu'ils sont renseignés et sous une forme simplifiée au seul mot ambigu dans le cas des mots composés (« France » pour Saint-Denis-en-France et « Vienne » pour Mesnil-sous-Vienne et Aixe-sur-Vienne). Les toponymes y sont à chaque fois associés à leurs identifiants. Trois alias ont été ensuite rajoutés pour l'entité France afin de tenter l'utilisation du modèle pour la lecture de morceaux de textes issus du Trésor des chartes : « francorum », « france » et « fancie ». Par la suite, deux lots de textes contenant des données avec des entités nommées reconnues et liées aux entités du référentiel sont constitués. Le fichier « entites_vienne.csv » contient des textes inventés contenant une mention de la ville de Vienne tandis que le fichier « extrait_français_france.csv » contient des textes extraits du Trésor des chartes.

5. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/tree/main/Scripts/Entity_linking/input.

9.2.2 Analyse des résultats

L'objectif de ce travail est de pouvoir tester le rendu des modèles utilisés sur des données simples afin de prendre en main les fonctionnalités de spaCy et préparer de futurs travaux plus ambitieux. « multi-home-c3po4-LOC-model-best » a ici posé quelques difficultés au moment de l'initiation de la base de connaissances et au moment de l'ajout du liage d'entités au modèle. En effet, il ne dispose pas d'un vecteur de mots pouvant être utilisé au moment de l'initiation de la classe « KnowledgeBase », ce qui réduit les performances du modèle. De plus, la documentation spaCy spécifie que l'ajout d'un vecteur de mots *a posteriori* nuit grandement à son efficacité [6]. De la même manière, l'ajout du liage d'entités comme fonctionnalité du modèle a nécessité l'ajout préalable d'une fonction « sentencizer » permettant la segmentation des phrases. Ces éléments n'ont pas empêché la réalisation du processus en entier mais ils sont probablement à l'origine de la faiblesse des résultats obtenus.

Nous avons donc commencé par un premier essai à partir de données d'entraînement vides [7]. Cet essai nous a permis d'observer les limites de l'utilisation du modèle « fr_core_news_md » dans le traitement des textes anciens : le liage d'entités reposant sur la reconnaissance préalable des entités nommées, l'utilisation du modèle sur les extraits proposés du Trésor des chartes s'est avérée inopérante faute de trouver les bons mots à lier. Pour ce qui est du modèle « multi-home-c3po4-LOC-model-best », cet essai a permis de montrer qu'il pouvait lier une entité nommée reconnue à l'entité correspondante à la condition que cette entité soit composée exactement de la même chaîne de caractères qu'un des alias renseignés dans la base de connaissance. Par exemple pour l'extrait :

> philippus dei gratia francorum rex notum facimus universis tam presentibus quam futuris nos infrascriptas litteras vidisse formam

la chaîne « francorum » a bien été reconnue, étiquetée comme un lieu et identifiée comme correspondant à l'entité France.

Un second essai a ensuite été réalisé pour intégrer des données d'entraînement dans l'ajout du liage d'entités comme fonctionnalité du modèle [8]. Le modèle « fr_core_news_md » s'est ici trouvé confronté aux mêmes limites d'analyse du langage que pour l'essai précédent. Quant au modèle « multi-home-c3po4-LOC-model-best », son utilisation pour le traitement des données d'entrainement génère une erreur (*cf.* figure 17). Notre stage s'étant achevé, nous n'avons pas eu l'occasion d'approfondir plus avant les raisons de cette erreur ni de proposer un moyen de la résoudre.

6. https://spacy.io/usage/embeddings-transformers.
7. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Entity_linking/scripts/el_Himanis_without_gold_entity.py.
8. https://github.com/virgile-reignier/Memoire-TNAH-2022-Reignier/blob/main/Scripts/Entity_linking/scripts/el_Himanis.py.

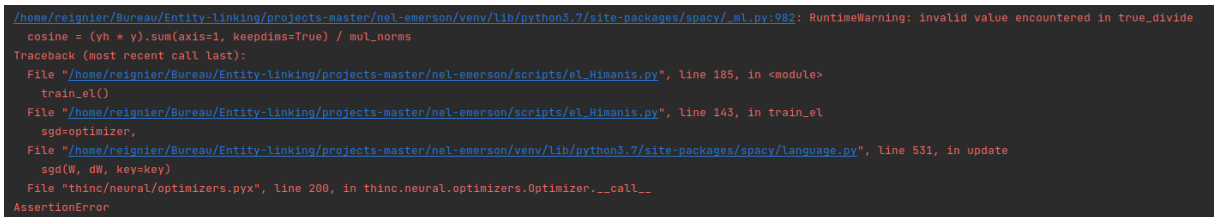

Figure 17 – *« RuntimeWarning: Invalid value encountered in true_divide cosine = (yh * y).sum(axis=1, keepdims=True) / mul_norms » : Erreur affichée par PyCharm lors de l'ajout de données d'entraînement pour développer le liage d'entités avec le modèle « multi-home-c3po4-LOC-model-best ».*

Nous pouvons résumer l'état de nos travaux par le tableau suivant :

Modèle utilisé	fr_core_news_md	multi-home-c3po4-LOC-model-best
Vecteur de mots	Oui	Non
REN sur le Trésor des chartes	Non	Oui
Liage d'entité sans entraînement	Oui	Oui
Intégration de données d'entrainement	Oui	Non

9.3 Quelles perspectives d'amélioration ?

9.3.1 Entraîner un nouveau modèle

Nos difficultés dans la mise en œuvre du liage d'entités provient ici en bonne partie d'un problème d'interopérabilité entre le modèle que nous avons utilisé et les fonctions proposées par spaCy. Une des principales voies d'amélioration serait donc d'entraîner un nouveau modèle qui dispose dès l'origine des caractéristiques nécessaires à la mise en place du liage d'entités. De plus, les résultats du modèle sur la REN pour le Trésor des chartes sont assez mitigés. Par exemple le texte dont nous avons proposé un extrait ci-dessus contient d'autres noms de lieux, mais seul « francorum » a été reconnu comme tel [9].

Une piste pour l'entraînement d'un nouveau modèle serait d'utiliser le corpus HOME-Alcar produit au sein du projet HOME pour entraîner des modèles de NER à partir d'archives datées entre le XII^{e} et le XIV^{e} siècles [10]. Une fois l'entraînement réalisé, le modèle pourrait être testé sur le corpus Arkindex « Himanis | TEKLIA processing » et les résultats comparés à ceux déjà existants. Nous avons à ce titre réalisé un export du projet sous la forme d'une base de données SQLite [11] afin de former une table à partir des entités

9. Cette chaîne de caractères constitue à elle seule 24 759 des 48 954 entités nommées reconnues dans le corpus Arkindex « Himanis | TEKLIA processing », soit 51 %.
10. https://zenodo.org/record/5600884#.Yt5NT3VBzUQ.
11. https://doc.arkindex.org/howto/export.

nommées déjà reconnues dans le corpus. Cet export nous a également permis de repérer quelques transcriptions réalisées à partir du modèle « Kaldi IAM » spécialisé dans la lecture de textes anglais. Les résultats de ce modèle étant incohérents, nous avons listé les textes concernés et transmis ces informations à Teklia.

9.3.2 Vers d'autres horizons

Une fois le modèle entraîné pour la REN et apte à intégrer le liage d'entités parmi ses fonctionnalités, les entités contenues dans la base de données Heurist pourront être exportées au sein d'une table et former la base de connaissances dans spaCy. Plusieurs voies d'améliorations se dessinent ensuite : outre un travail sur la formulation et la description des entités déjà importées dans Heurist [12], d'autres données peuvent encore être ajoutées au référentiel. L'ouvrage que nous avons utilisé ne constitue en effet qu'une partie des instruments de recherche disponibles pour décrire le contenu du Trésor des chartes que nous avons présentés au chapitre 2. Il est donc possible d'augmenter encore la quantité de données utilisées pour l'entraînement en intégrant l'ensemble du contenu des inventaires analytiques : les volumes décrivant les registres JJ 37 à JJ 91 disposent tous d'un index à l'exception des deux volumes imprimés décrivant les registres JJ 50 à JJ 64 pour lesquels un index des noms de lieu est en préparation. D'autres données encore peuvent être extraites à partir des inventaires thématiques et géographiques ainsi que de quelques travaux préparatoires.

Une des difficultés de ce travail réside dans l'alignement systématique entre les entités de l'index et leurs manifestations dans les textes. Les seules données que nous avons à notre disposition concernent en effet les actes au sein desquels chaque entité est présente, mais sans que nous ne connaissions avec exactitude le segment de texte concerné. Il est donc possible que la transformation des données exportées depuis Heurist en données d'entraînement nécessite plusieurs étapes. Il est en effet difficilement envisageable de faire correspondre le nom des entités avec des chaînes de caractères présentes dans les textes car ils sont mentionnés dans des formes latines ou vernaculaires et selon des graphies qui peuvent varier d'un texte à l'autre. Un premier enjeu consiste donc à faire cohabiter ces différentes graphies comme autant d'alias potentiels repérés grâce à la REN. La préparation manuelle d'une première série de données devrait cependant permettre d'intégrer cette partie de l'apprentissage et de faciliter l'alignement entre le contenu des registres indexés et les entités de la base de connaissances.

Il est à ce titre possible que les fonctionnalités offertes par spaCy ne suffisent pas à cet usage. On pourra donc se tourner vers les outils et modèles développés par les chercheurs

12. *Cf.* Annexe C.

travaillant sur le liage d'entités multi-langue [13]. Les entités sont en effet décrites en français alors que les textes sont tous en latin ou en vernaculaire. Mais il existe cependant quelques entrées pour lesquelles la langue d'origine est utilisée en complément ou à la place de la version française et qui pourront être utilisées pour l'entraînement du modèle. De plus, certaines entités sont décrites par un ensemble de mots qui peut contenir d'autres entités (par exemple « rex Francorum et Navarre » contient les entités France et Navarre mais désigne également la personne qui est en règne à ce moment) et peuvent poser problèmes pour la reconnaissance comme pour le liage [14]. Il faut également envisager que la NER permette de dévoiler de nouvelles entités qu'il faudra alors soit associer à un autre référentiel soit intégrer à Heurist. Pour cette raison, les essais que nous avons réalisés et que nous envisagions de poursuivre concernent uniquement les noms de lieux car leur identification manuelle est facilitée par la présence de référentiels complets pour la période actuelle.

Une autre voie à explorer concerne la prise en compte des relations entre les entités. Cette fonctionnalité n'est en effet pas permise par spaCy alors qu'elle constitue le principal atout du référentiel que nous avons constitué. Les relations connues entre les entités permettent en effet de prédire de probables co-occurrences. Par exemple Agnès du Bois est déjà connue comme étant la femme de Guillaume de Flavacourt. La mention de ce dernier dans un texte augmente donc la probabilité que sa femme le soit aussi, notamment si le texte contient le mot « Agnès ». Ces co-occurrences peuvent également être déterminantes pour former les alias des entités. Dans le cas précédent, la même Agnès du Bois peut être mentionnée par l'expression « femme de Guillaume de Flavacourt ». Un autre élément parfois présent dans les textes politiques est le phénomène de métonymie entre une entité dotée d'un pouvoir et un lieu ou une idée qui lui est associée. Cette figure de style est par exemple employée dans le langage courant lorsqu'on dit « Londres a déclaré la guerre à Buenos Aires le 2 avril 1982 ». Les deux villes ne sont pas mentionnées comme des acteurs directs de la guerre mais comme les lieux de pouvoir des États britanniques et argentins. La connaissance du lien entre les entités peut ainsi aider à déceler ces expressions métonymiques.

Pour finir, la mise en œuvre du liage d'entités doit également prendre en compte la possible intégration d'éléments nouveaux repérés par le modèle. Ces éléments concernent en premier lieu les entités NIL qui désignent des entités nommées identifiées comme étant absentes de la base de connaissances. Il est alors possible d'étudier ces entités et de sélectionner celles qui pourront intégrer notre référentiel. D'autres recherches permettent

13. A ce sujet, v. notamment Elvys Linhares Pontes, Jose G. Moreno et Antoine Doucet, "Linking Named Entities across Languages using Multilingual Word Embeddings"… et Shruti Rijhwani, Jiateng Xie, Graham Neubig, *et al.*, "Zero-Shot Neural Transfer for Cross-Lingual Entity Linking"…

14. Sur le sujet, v. Katherine McDonough, Ludovic Moncla et Matje Camp, "Named entity recognition goes to old regime France…".

également d'envisager l'extraction des relations entre les entités [15]. Ces relations peuvent être utilisées pour enrichir la base de données d'éléments nouveaux. Cette opération est assez importante dans le cadre de l'utilisation du modèle sur les textes non-indexés du Trésor des chartes. Il est en effet probable que plus les textes à partir desquels on tente d'opérer le liage d'entités sont éloignés géographiquement et chronologiquement des textes utilisés pour établir la base de connaissances, plus le nombre d'entités NIL sera important (notamment pour les noms de personnes). On peut donc envisager de procéder par étapes afin d'intégrer progressivement ces entités NIL à la base de données et permettre leur reconnaissance dans les textes suivants.

Conclusion

Ce chapitre nous a permis de présenter les différents éléments relatifs à la mise en œuvre du liage d'entités à partir du corpus du projet Himanis et de la base de données que nous avons constituée. Pour cela, nous avons dans un premier présenté les fonctionnalités offertes par spaCy et son usage pour le liage d'entités. Puis nous avons décrit les différents essais que nous avons réalisés et les résultats obtenus. Enfin, nous avons proposé des perspectives d'amélioration par l'entraînement d'un nouveau modèle et par la mise en œuvre de plusieurs techniques.

Cette étape de notre travail présente donc un bilan mitigé puisqu'elle n'a pas permis de mettre en œuvre le liage d'entités à proprement parler. Cependant, nos balbutiements ont été l'occasion d'aborder différents aspects de cette technique et de prendre la mesure des besoins du modèle souhaité. Pour être pleinement opérationnel, celui-ci doit en effet être interopérable avec les outils utilisés et proposer de bons résultats en REN. La reconnaissance des entités est primordiale pour envisager de les lier à une base de connaissances. Nous espérons ainsi que ces essais permettrons de baliser les travaux futurs et d'offrir quelques perspectives à la mise en œuvre du liage d'entités pour envisager l'indexation automatique du Trésor des chartes.

15. A ce propos, v. Yoann Dupont, *La structuration dans les entités nommées…*, p. 169–180.

Conclusion partielle

Cette dernière partie a été l'occasion d'aborder les différents aspects liés à l'utilisation et à la mise en relation du référentiel que nous avons constitué. Notre développement s'est d'abord concentré sur l'enrichissement des entités toponymiques à l'aide de référentiels externes afin d'y ajouter des coordonnées. Nous avons ensuite exposé l'intérêt de la mise en ligne des données que nous avons manipulées, que ce soit pour compléter les dépôts déjà constitués ou pour offrir des éléments de visualisation. Pour finir, nous avons présenté nos essais de mise en œuvre du liage d'entités et les différentes perspectives disponibles malgré nos résultats mitigés.

Ces éléments nous ont permis d'aborder différentes problématiques liées à l'interopérabilité et à l'accessibilité des données. L'utilisation des entrées de l'index pour un alignement avec un autre référentiel nécessite en effet de caractériser précisément le contenu de chaque élément. Le fait de réutiliser des données déjà traitées au préalable ajoute à ce titre une difficulté puisque cela nécessite de comprendre le sens de chaque élément. Cette difficulté a également été présente au moment de la mise en œuvre du liage d'entités et a participé à notre proposition d'entrainer un nouveau modèle qui intègre dès le début toutes les fonctionnalités nécessaires à cette étape. C'est dans ce même objectif que nous avons pris le temps d'insérer dans chaque entrée de la base de données des indications permettant de la retrouver dans les fichiers de travail mis en ligne. Le suivi des différentes étapes de traitement par ce biais facilitera ainsi l'appropriation des données pour d'autres usages et la compréhension des actions antérieures.

Conclusion générale

Ce stage s'est inscrit dans le cadre du développement de techniques de lecture automatique des textes initiés par le projet Himanis à partir du Trésor des chartes. Après des travaux sur la REM et la REN, l'objectif était de poursuivre sur le liage d'entités afin d'associer chaque entité nommée reconnue dans les textes à une référence et de résoudre les ambiguïtés présentes. Pour mener à bien ce projet, nous avons commencé par préparer des données afin de servir de référence au moment de la réalisation du liage d'entités. Ces données sont issues des entrées d'index décrites par l'index du premier volume de l'inventaire analytique des registres du Trésor des chartes relatif aux cotes JJ 37 à JJ 50. Cette étape avait été initialement conçue comme la première d'un long processus visant à l'intégration d'un grand nombre de données issues des instruments de travail décrivant le contenu du Trésor des chartes. Mais ce premier travail s'est avéré beaucoup plus complexe que nous l'avions initialement pensé, à tel point qu'il a absorbé la quasi-totalité de notre temps de stage. Le traitement des autres instruments de recherche a donc été abandonné afin de pouvoir consacrer un peu de temps à la mise en œuvre du liage d'entités. Malgré ce choix, les tests effectués à cet effet ont été très réduits et n'ont pas abouti à des résultats probants. De plus, le référentiel mis en ligne pourrait encore être amélioré par d'autres traitements sur les entrées issues de l'index.

Ces difficultés ont toutefois été l'occasion d'étudier la structure des données contenues dans l'ouvrage utilisé et nous ont fourni de nombreux apprentissages dans le traitement automatique des *legacy data*. Nous avons donc commencé notre développement par un exposé du contexte de travail dans lequel nous nous sommes insérés. Nous avons pour cela développé l'état de la recherche sur le liage d'entités, présenté l'état des données traitées sous l'impulsion du projet Himanis et détaillé l'utilisation possible d'instruments de recherche comme *legacy metadata* pour décrire le contenu des registres du Trésor des chartes. Nous avons ensuite consacré notre seconde partie au développement du référentiel destiné à être utilisé pour le liage d'entités à partir de l'index retenu. Pour cela, nous avons commencé par rendre compte des difficultés que nous avons rencontrées dans le traitement automatique de ces données, puis nous avons présenté les problématiques propres à la prise en compte des relations entre les entités, enfin nous avons décrit la transformation de ces données en une base de données relationnelle. La troisième partie de notre développe-

ment a été consacrée à l'utilisation du référentiel constitué pour l'enrichir de coordonnées, mettre son contenu à disposition et engager l'entraînement du liage d'entités. Pour cela, nous avons présenté le processus d'alignement mis en œuvre entre les entrées d'index et des référentiels en ligne, puis nous avons décrit différents aspects de la mise en ligne des données par l'augmentation des dépôts disponibles, l'utilisation d'outils cartographiques et leur insertion dans la plateforme Arkindex, enfin nous avons présenté les différents essais effectués à partir de la librairie python spaCy et les perspectives envisageables pour approfondir ce travail.

Notre travail a donc entraîné la mise en ligne d'un grand nombre de données qui sont autant de moyens d'approfondir la connaissance sur le contenu des registres du Trésor des chartes. Une fois la publication de la base de données effectuée, son contenu pourra également être réutilisé pour d'autres recherches. De surcroît, nous avons eu l'occasion d'initier quelques analyses sur l'utilisation du liage d'entités dans le cadre du projet en proposant notamment un tour d'horizon des travaux réalisés sur le sujet. Cette synthèse a ainsi servi de base aux différentes propositions que nous avons suggérées pour tirer parti des essais effectués et finaliser la mise en œuvre du liage d'entités sur les registres du Trésor des chartes. Pour finir, nous avons mis en place un processus de travail en plusieurs étapes que nous avons décrites dans notre développement et illustrées par nos scripts de travail mis en ligne sur Github. Ce processus de travail est ainsi largement réutilisable pour intégrer les données des autres instruments de recherche au référentiel ou de manière générale pour traiter automatiquement des données issus d'ouvrages papier décrivant des documents d'archives.

Annexe A

Les sous-entrées de « Paris »

Comme on peut le voir dans les figures 18 et 19, les sous-entrées de Paris sont complexes à interpréter de manière automatique. En effet, les séparateurs utilisés sont les mêmes au sein d'un même bloc, que celui-ci soit formé comme un ensemble de sous-sous-entrées dépendant d'une même sous-entrée ou tout simplement comme un ensemble de sous-entrées successives.

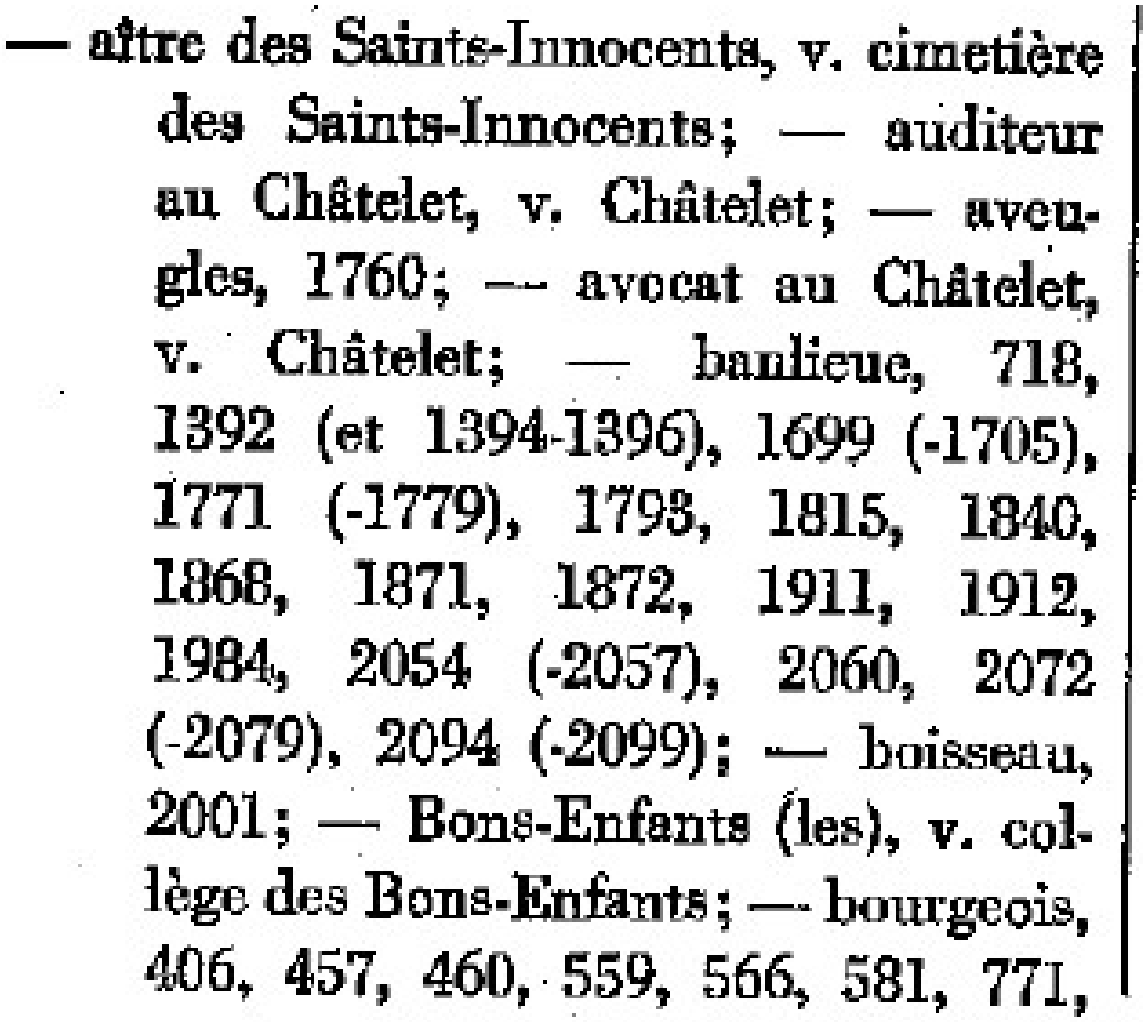
— aître des Saints-Innocents, v. cimetière des Saints-Innocents; — auditeur au Châtelet, v. Châtelet; — aveugles, 1760; — avocat au Châtelet, v. Châtelet; — banlieue, 718, 1392 (et 1394-1396), 1699 (-1705), 1771 (-1779), 1793, 1815, 1840, 1868, 1871, 1872, 1911, 1912, 1984, 2054 (-2057), 2060, 2072 (-2079), 2094 (-2099); — boisseau, 2001; — Bons-Enfants (les), v. collège des Bons-Enfants; — bourgeois, 406, 457, 460, 559, 566, 581, 771,

Figure 18 – *Ensemble de sous-entrée de « Paris » dont les éléments sont rangés par ordre alphabétique.*

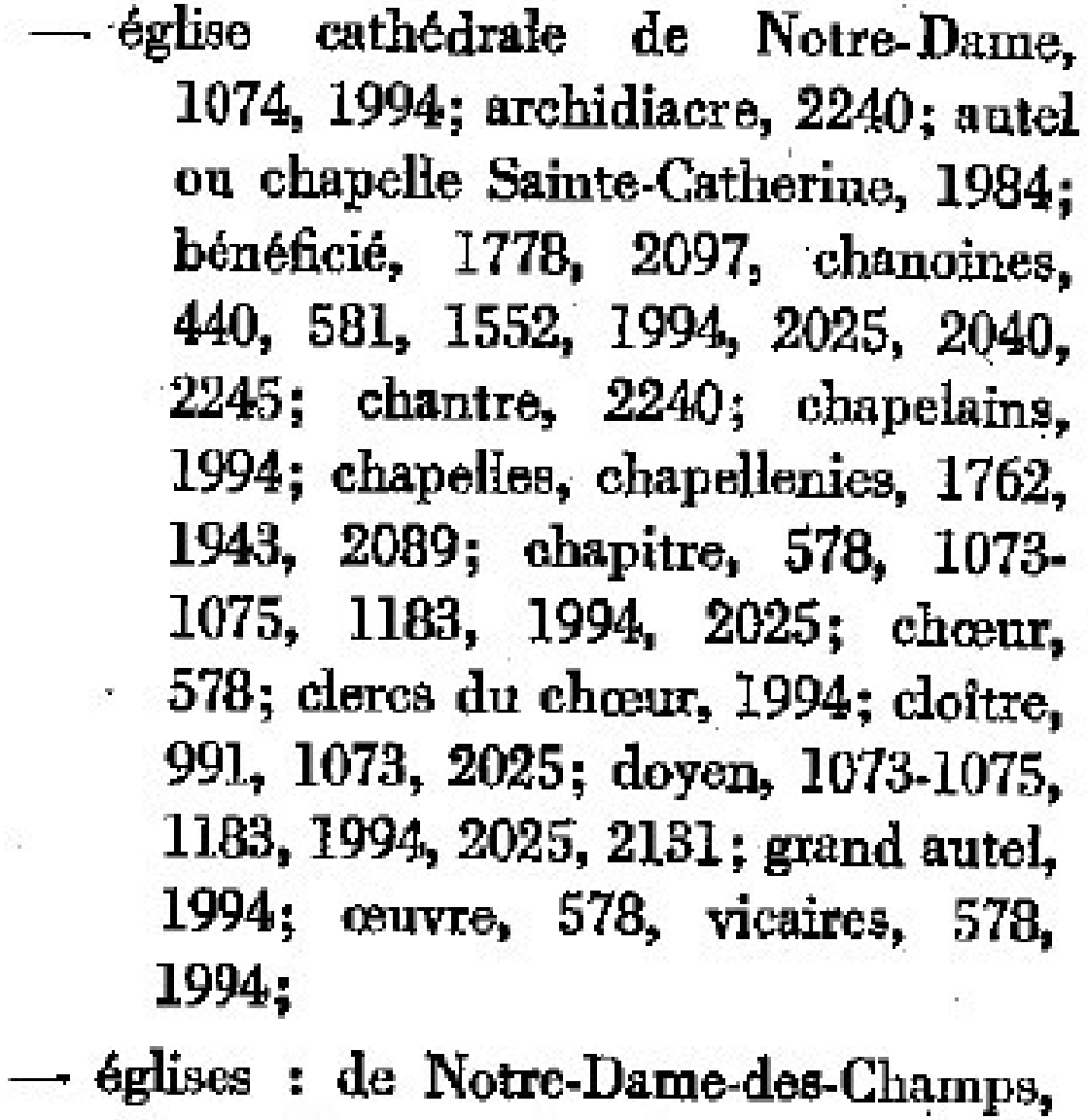

Figure 19 – *Ensemble de sous-entrée de « Paris » dont les éléments sont des sous-sous-entrées dépendant de la sous-entrée « église cathédrale de Notre-Dame ».*

Annexe B

Les erreurs natives dans l'index

Nous reportons ici différentes erreurs rencontrées au cours de notre travail. Les figures 20 et 21 sont des erreurs de typographie liées à l'absence d'un caractère ou à l'utilisation d'un caractère pour un autre. Les figures 22 et 23 sont des erreurs dans la systématicité des données liées à l'usage indifférent de plusieurs caractères pour signifier une même chose ou à l'ordonnancement des entrées.

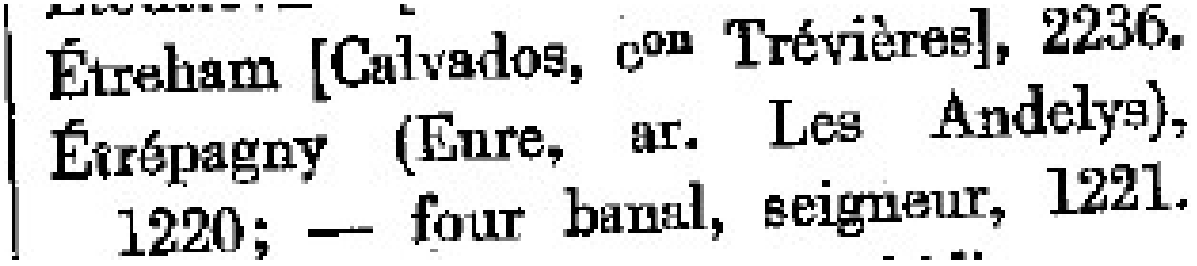
Étreham [Calvados, c^on Trévières], 2236.
Étrépagny (Eure, ar. Les Andelys),
1220; — four banal, seigneur, 1221.

Figure 20 – *Contrairement à la forme utilisée tout au long de l'index, l'entrée « Etrepagny » contient des indications géographiques entourées de parenthèses.*

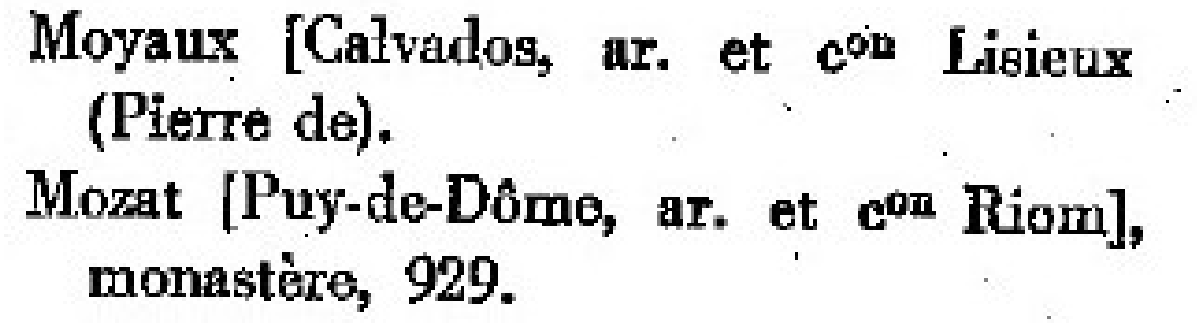
Moyaux [Calvados, ar. et c^on Lisieux
(Pierre de).
Mozat [Puy-de-Dôme, ar. et c^on Riom],
monastère, 929.

Figure 21 – *Déficit d'un crochet fermant.*

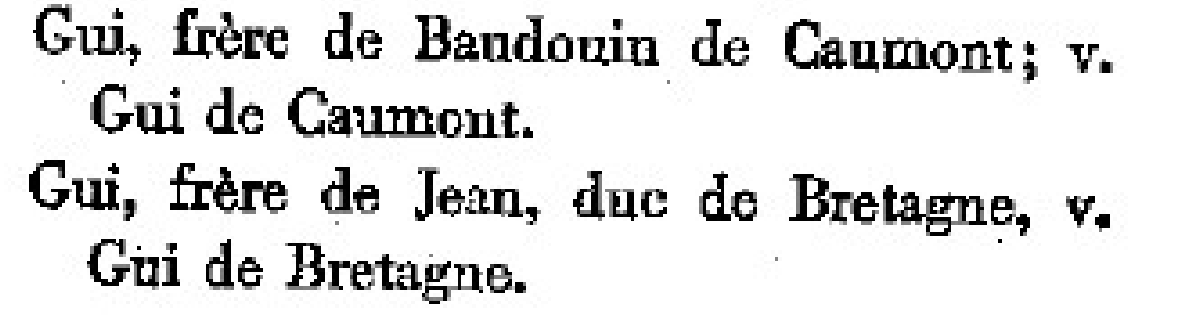
Gui, frère de Baudouin de Caumont; v.
Gui de Caumont.
Gui, frère de Jean, duc de Bretagne, v.
Gui de Bretagne.

Figure 22 – *Utilisation de deux caractères différents pour un même usage.*

Figure 23 – *« Saint-Savin » est ici placé par erreur entre « Saint-Omer » et « Saint-Ouen »*

Annexe C

Suggestions pour l'amélioration du contenu de la base de données mise en ligne sur Heurist

Malgré notre travail, la durée du stage ne nous a pas permis d'aller au bout de la formalisation du référentiel constitué à partir des entrées d'index. Pour que celui-ci soit pleinement utilisable, il faudrait encore améliorer certains éléments. Voici la liste de nos propositions :

- Les nom des entités issues de sous-entrées de l'index sont actuellement sous la forme « Abandons , — spontanés au roi ». Si le séparateur « , — » nous a permis de repérer facilement ces entrées et de travailler sur leurs spécificités, ces noms pourraient être simplifiés par son élimination systématique.
- Un tableau a été produit par les Archives Nationales à partir de la mise à plat de l'index des sujets. Dans le cadre de l'étape précédente, on peut envisager son usage comme référence pour les noms de sujet et proposer une fonction de comparaison automatique.
- Certaines marques de renvoi n'ont pas été prises en compte par les traitements réalisés, comme par exemple « Agnès du Bois, femme de Guillaume de Flavacourt ». Guillaume de Flavacourt dispose bien d'une entrée à son nom dans l'index, mais la relation entre les deux entités reste à réaliser.
- La discrimination entre les éléments de la table Place et ceux de la table Person n'a pas pu être définie de manière parfaite. Il est donc probable qu'il demeure quelques incohérences parmi ces entrées.
- Les relations entre sous-entrées et entrées générales n'ont été qualifiées que pour la table des sujets. Pour la table des lieux et personnes, les sous-entrées dépendant d'entrées catégorisées comme des lieux ont toutes été transformées de la même

façon indépendamment de leur contenu. Elles forment donc toutes des entrées de la table Place. De la même manière, toutes les sous-entrées dépendant d'entrées catégorisées comme des personnes ont été transformées comme des descriptions de la relation entre l'entrée principale et les entrées d'inventaire associées à ces sous-entrées. La catégorisation de ces sous-entrées permettrait donc d'améliorer largement l'utilisation de ces données.

— Lorsque les circonscriptions administratives auxquelles sont rattachés les éléments de la table Place disposent elles aussi d'une entrée dans cette même table, il serait envisageable de créer une relation entre ces deux entrées.

— Certaines entrées de la table Place disposant de plusieurs propositions de localisation sont en réalité des erreurs. Par exemple l'entrée « Bruetel (maison de) [région de Marigny, Seine-Maritime, con Gournay, cne Dampierre] » a été alignée avec les deux communes nommées Dampierre dans le département de la Seine-Maritime alors que l'une d'elle n'est pas située dans le canton de Gournay et aurait pu être éliminée des propositions de localisations. Ces données, comme le script utilisé pour l'alignement, peuvent donc être améliorées à cet effet.

— Les marques d'incertitude qui sont décrites dans la colonne « incertitude » du fichier geonames_final.csv n'ont pas été reportées dans la base de données. La prise en compte de ces incertitudes, ainsi que des marques présentes dans les autres types d'entrées (notamment les « ? ») est donc encore à réaliser.

— Les entrées pour lesquelles l'index propose plusieurs localisations ont été dédoublées dans le fichier geonames_final.csv. Mais seule l'une des propositions a été importée dans la base de données. L'import des autres propositions reste donc à réaliser.

— Les éléments insérés parmi les renvois vers les entrées d'inventaire, comme par exemple les références à la section « Additions et corrections », n'ont pour le moment pas été pris en compte.

— Actualiser le nom de départements des entités situées dans les anciens départements de la Seine et Seine-et-Oise.

— Charger les attributs « Provisory_index_3 » dans les entités Act sur Heurist.

Table des figures

Table des matières