Médiévales
Langues, Textes, Histoire

73 | automne 2017
Le texte à l'épreuve du numérique

La reconnaissance des entités nommées dans les bases numériques de chartes médiévales en latin : le cas du *Corpus Burgundiae Medii Aevi* (Xe-XIIIe siècle)

Named Entities Recognition for Digital Bases of Medieval Latin Charters. The Case of the Corpus Burgundiae Medii Aevi *(Tenth-Thirteenth Centuries)*

Sergio Torres Aguilar

Édition électronique
URL : http://journals.openedition.org/medievales/8182
DOI : 10.4000/medievales.8182
ISSN : 1777-5892

Éditeur
Presses universitaires de Vincennes

Édition imprimée
Date de publication : 15 décembre 2017
Pagination : 47-65
ISBN : 978-2-84292-801-8
ISSN : 0751-2708

Référence électronique
Sergio Torres Aguilar, « La reconnaissance des entités nommées dans les bases numériques de chartes médiévales en latin : le cas du *Corpus Burgundiae Medii Aevi* (Xe-XIIIe siècle) », *Médiévales* [En ligne], 73 | automne 2017, mis en ligne le 28 février 2019, consulté le 03 janvier 2020. URL : http://journals.openedition.org/medievales/8182 ; DOI : 10.4000/medievales.8182

Médiévales 73, automne 2017, p. 47-66

Sergio Torres Aguilar

La reconnaissance des entités nommées dans les bases numériques de chartes médiévales en latin : le cas du *Corpus Burgundiae Medii Aevi* (Xe-XIIIe siècle)

À ce jour, plus d'un demi-million de documents médiévaux ont été publiés numériquement, mais ces documents importés « tels quels » sous forme numérique, à partir d'éditions papier le plus souvent, ne sont exploitables que par un œil et un cerveau expérimentés. L'utilisation même de ces documents éveille, et à juste titre, la suspicion parmi les chercheurs en humanités parce qu'ils sont stockés dans des bases de données, soit un concept et un outil qui, pour sa part, relie les collections de manuscrits à un environnement numérique où les résultats obtenus peuvent être reproductibles, transférables et apparaissent univoques[1]. Le paradigme épistémologique, bien établi dans la recherche en humanités, qui favorise à la fois la prise en compte du contexte et la lecture particulière des sources par un spécialiste, semble être remis en cause dans les nouvelles pratiques des humanités numériques[2].

Structurer ces bases de données, autrement dit y incorporer un système qui explicite leurs propriétés, relie les différentes parties à l'ensemble et, le cas échéant, associe cet ensemble à d'autres bases de données, est l'ambition de nombreux projets en cours. Il s'agit d'une condition prioritaire pour transformer une base de données en une base de connaissances[3]. Mais ce travail étant jusqu'à présent fait à la main, il ne peut être adopté à grande échelle, ce qui a rapidement stimulé des initiatives académiques pour l'automatiser en utilisant des outils numériques.

1. J. S. Ward et A. Barker, *Undefined by Data : a Survey of Big Data Definitions*, Ithaca, 2013.

2. J. Drucker, « Humanities Approaches to Graphical Display », *Digital Humanities Quarterly*, 5/1 (2011), p. 1-21.

3. D. Schloen et S. Schloen, « Beyond Gutenberg : Transcending the Document Paradigm in Digital Humanities », *Digital Humanities Quaterly*, 8/4 (2014).

Un des principaux éléments de cette structuration est la reconnaissance des entités nommées qui permet d'attribuer un nom spécifique à tous les éléments physiques et réels présents dans les documents et de les classer. L'identification et la classification correctes de ces unités d'information, qui se rapportent directement au questionnaire élémentaire de la recherche historique (qui, où et quand ?), ouvrent des canaux dynamiques qui permettent d'interroger une base de données.

Les outils informatiques qui permettent d'accomplir cette tâche doivent remplir certaines exigences. Créés pour traiter des données économiques ou biologiques, ils doivent s'adapter aux types de données fournis par la culture textuelle et graphique dans laquelle un manuscrit a été composé. Il s'agit dans ce cas d'objets ou de représentations d'objets dont le contenu est indissociable du contexte social, culturel et sémiotique, catégories difficiles à quantifier, mais qui ne rejettent pas la modélisation, notamment par la découverte de régularités et de stéréotypes.

Dans la réalité aux multiples facettes du document médiéval, certains éléments sont des atouts pour la modélisation informatique, alors que d'autres nécessitent d'importantes adaptations et corrections. Grâce à une modélisation synergique, un processus algorithmique devient un outil puissant, capable de fournir des lignes générales d'information pour les études particulières, ainsi que des détails avérés confirmant des hypothèses générales[4].

Nous allons ici étudier l'exemple de la construction d'un modèle de reconnaissance des entités nommées à partir d'un corpus médiéval régional, le *Corpus Burgundiae Medii Aevi* (*CBMA*)[5]. Nous passerons en revue les opérations informatiques de traitement du langage naturel ; nous décrirons la nature et la composition de notre corpus, exposerons les défis contextuels posés par les textes médiévaux, ainsi que les opérations de correction, d'adaptation et de validation de l'outil que nous avons réalisées.

Les entités nommées

Le concept d'entités nommées recouvre tous les éléments textuels se rapportant à une personne, un lieu ou une organisation. La détection et la classification des entités nommées fournissent des listes de noms qui sont autant d'éléments-clés d'information grâce auxquels il est possible

4. À ce sujet, consulter quelques articles de référence dans A. AMBROSIO, S. BARRET et G. VOGELER éd., *Digital Diplomatics. The Computer as a Tool for the Diplomatist ?*, Wappenkunde, 2015.

5. <http://www.cbma-project.eu/>.

d'indexer la totalité des acteurs sociaux, des éléments spatiaux et des données organisationnelles présents dans un document.

L'identification d'entités nommées prend toute son importance dans l'analyse des grands corpus : elle permet de mieux explorer des ensembles volumineux de données et sert d'analyseur (*parser*) pour y obtenir des portions de données pertinentes. Par conséquent, il s'agit d'une technique largement adoptée dans les domaines traitant d'énormes bases de données. Les données *non structurées* contenues dans des discours formulés en langage courant (fig. 1a) empêchent un moteur de récupération d'obtenir des informations complexes. Au lieu de cela, la même information *classée* dans des champs par catégories d'entités (fig. 1b) peut être interrogée à la fois à un niveau plus granulaire, mais aussi afin d'explorer des domaines plus vastes, en reliant les données à partir d'une ontologie commune.

ITEM	Paris, BNF, coll. De Bourgogne t. 78 n° 123b. Telma N. 1694
DESCRIPTION	Charte originale en parchemin provenant de l'abbaye de Lezat avec sceau en cire. Datée ca. 1073.

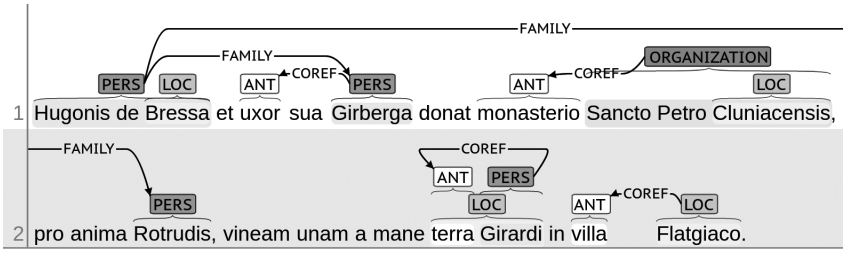
1a. Exemple de données non structurées

N° TELMA	1694	TRADITION	Original
DEPÔT	BNF	DATE	1073
COLLECTION	Coll. Bourgogne	AUTEUR	Abbaye de Lezat
N° COLLECTION	78. n° 123b	MATERIEL	Parchemin
GENRE	Charte	SCEAU	Cire

1b. Exemple de données classées

À partir d'une liste d'entités, un moteur de récupération peut extraire rapidement des informations pertinentes et, à partir de celles-ci, mettre en rapport les différents contextes dans lesquels émergent ces entités pour répondre à des questions complexes. La figure 2 montre en détail la structuration d'un texte à partir d'entités nommées. Une base structurée de cette façon peut fournir de grands volumes d'information sur l'histoire d'un individu ou d'une communauté, mais permet aussi de générer des arbres de relations sociales, de hiérarchies spatiales, ou de suivre l'évolution des noms, des usages, des concepts, etc.

Mais il existe des obstacles importants avant d'arriver à des modèles de reconnaissance d'entités efficients et rapides. Ces modèles sont en effet développés à partir de corpus annotés manuellement, qui sont très peu nombreux et d'un volume limité. Par ailleurs, les questions autour de la représentativité et de la qualité des sources s'appliquent également au



PERS : personne, LOC : location, ANT : antécédent, COREF : co-occurrence.

2. Entités nommées annotées en BRAT (outil d'annotation)

modèle. Enfin, la formation du modèle se compose d'une extraction massive de données langagières qui implique une analyse des différentes couches de texte[6]. La dépendance absolue du modèle aux caractéristiques textuelles et langagières rend chaque modèle très spécifique à chaque état de langue et toute modélisation un exercice *ad hoc*.

Le *Corpus Burgundiae Medii Aevi*

Le corpus utilisé pour la modélisation provient d'une base de données de chartes médiévales de Bourgogne fournies par le projet *CBMA*. Le corpus contient à ce jour près de 29 000 actes provenant de chartriers et de plus de 300 cartulaires produits en Bourgogne entre le XI^e^ et le XVIII^e^ siècle, provenant d'abbayes clunisiennes et de quelques monastères cisterciens[7]. Documents très importants pour l'historiographie, les cartulaires recueillent des transcriptions de documents originaux et des copies d'actes très différents : transactions privées, donations, diplômes royaux, privilèges pontificaux, jugements, censiers, etc. Ces actes ont une structure formalisée qui recueille des informations variables fortement liées au contexte particulier de l'écriture, tels que les noms de personnes, de lieux, d'organisations, conduisant à produire des documents proches dans leurs formules, mais jamais identiques.

Sensibilisé aux propositions des humanités numériques, le *CBMA* a isolé et annoté manuellement les entités de noms de personnes et de lieux dans un ensemble de 5 300 chartes : cet ensemble a constitué notre outil de travail. Au centre de ce dispositif figurent quatre éditions, parmi

6. X. Ren, A. El-Kishky *et al*., « Automatic Entity Recognition and Typing in massive Text Corpora », dans *25th International Conference Companion on World Wide Web*, Montréal, 2016, p. 1025-1028.

7. I. Rosé, « Panorama de l'écrit diplomatique en Bourgogne : autour des cartulaires (XI^e^-XVIII^e^ siècles) », *BUCEMA*, 11 (2007), p. 6-27 ; S. Barret, « Cluny, Note sur le Recueil des chartes de l'abbaye de Cluny d'Auguste Bernard et Alexandre Bruel », *BUCEMA*, Collection CBMA (2009), p. 1-16.

les cinquante-six sur lesquelles s'est fondé le travail du *CBMA* : l'édition monumentale d'Auguste Bernard et Alexandre Bruel des chartes de l'abbaye de Cluny (1884), en particulier les cartulaires A et B qui contiennent des chartes datant du temps de la fondation de l'abbaye en 910 ; l'édition d'Ernest Petit du cartulaire du prieuré de Jully-les-Nonnains (1881) ; celle de Camille Ragut du cartulaire de Saint-Vincent de Mâcon (1864) ; et l'édition critique du cartulaire de l'abbaye cistercienne de Vauluisant par Owen W. Duba (1994). Les actes de ces corpus représentent 85 % des actes sélectionnés. Les 15 % restants proviennent de sept autres éditions de cette époque.

Comme on peut le constater, les éditions diplomatiques et critiques des XIX^e^ et XX^e^ siècles, ainsi que certaines éditions propres au *CBMA*, après un processus d'océrisation[8], forment le matériau textuel de cette base de données. Dans la plupart de ces éditions, la lecture a été facilitée par la modernisation du texte, c'est-à-dire par l'ajout de capitales, de signes de ponctuation, par le développement des abréviations et la restitution des mots manquants, éléments qui n'existent pas dans le manuscrit d'origine.

Tous les documents du corpus qui nous intéresse, qui va de la fin du IX^e^ au milieu du XIII^e^ siècle, sont écrits en latin, mais ils peuvent présenter des traits linguistiques assez différents, notamment la pénétration de mots et/ou groupes de mots provenant des langues vulgaires (ce qui s'inscrit dans un processus de *vernacularisation*) ; de manière plus générale, on constate un affranchissement progressif d'avec les règles et le vocabulaire du latin classique, ainsi qu'une intense variabilité des graphies.

Le contexte du corpus

Une analyse plus approfondie de la constitution du corpus *CBMA* précise ses caractéristiques, et révèle un certain nombre d'avantages et d'inconvénients pour l'élaboration d'un modèle de reconnaissance. Les questions soulevées, que nous détaillons ici, ont motivé un long travail de modifications successives et d'adaptations du corpus.

Représentativité

La première concerne la représentativité. Un corpus de modélisation doit être stable et cohérent. Puisque le modèle généré est discriminant, les règles provenant d'un seul corpus organique peuvent être plus efficaces que celles provenant d'une séquence de petits corpus variés. Mais l'utilisation d'un seul corpus peut affecter la capacité du modèle à reconnaître des

8. Application de l'OCR (*Optical Character Recognition*), technique informatique qui transforme des images numérisées de textes imprimés en fichiers de texte brut.

entités dans des chartes externes au corpus, car celui-ci peut tendre à une uniformisation, conduisant de nombreux phénomènes spécifiques à être mal représentés. Par ailleurs, un corpus régional peut manquer de représentativité par rapport à une réalité beaucoup plus vaste ; en revanche, un corpus provenant d'une institution au centre d'un réseau régional, du fait de son importante circulation, concentrera une pluralité de phénomènes supérieure à la moyenne.

Concentration

La deuxième question concerne la concentration. La plupart des *scriptoria* qui ont produit ces documents sont localisés dans une seule région et appartenaient principalement à deux institutions, Cluny et Saint-Vincent, durant une période bien déterminée, entre le X^e^ et le XIII^e^ siècle. Toutefois, à l'intérieur on peut observer que les chartes proviennent de plus d'une dizaine de cartulaires, qui contiennent des transactions juridiques de plus d'une centaine de petites zones se distribuant selon au moins cinq types d'actes juridiques.

Une vue d'ensemble nous met face à un double problème de perspective : d'une part, nous avons deux institutions centrales de production qui unifient les documents tout en autorisant de grands espaces de variabilité par l'utilisation et l'adaptation de différents formulaires et par l'influence des différents états de la langue et de la tradition ; d'autre part, nous avons une claire surreprésentation de ces deux institutions par rapport à d'autres petits producteurs, ce qui réduit les styles et les phénomènes scripturaux mineurs.

Écriture formulaire

La troisième question concerne l'écriture formulaire. Il y a dans l'écriture des chartes une combinaison particulière d'éléments formels – que l'on désigne par le terme de formulaire – et d'informations spécifiques[9]. La modélisation s'appuie ainsi sur la reconnaissance de certains éléments structurels formant un tronc discursif commun associés à des éléments documentaires variés – les entités nommées – qui permettent une définition statistique. Cependant, ce modèle est pris en défaut lors de son application sur des documents provenant de traditions scripturales qui utilisent des formulaires différents et présentent des états de discours plus atypiques ou des variations personnelles attribuables au scripteur. Dans ces cas, l'itération que le modèle a appris à reconnaître est déformée.

9. Voir à ce sujet M. ZIMMERMANN, *Écrire et lire en Catalogne : IX^e^- XII^e^ siècle*, Madrid, 2003, p. 251-284.

Latin médiéval

La quatrième question concerne le latin médiéval. Dans les études de traitement automatique de la langue, la « boîte à outils » du latin est très pauvre. Certaines études récentes ont porté sur la modélisation du latin, en produisant lemmatiseurs, dictionnaires et index géographiques (*gazzeeters*), mais ceux-ci sont encore expérimentaux et demeurent focalisés sur la littérature en latin classique. Les variantes médiévales, qui présentent des états changeants de la langue imputable au processus de conversion en langues romanes[10], n'ont presque pas été abordées. Cela implique que la modélisation soit effectuée à l'aide d'outils de traitement du langage indépendants de la langue, et par l'adaptation de certains outils développés pour le latin classique.

Numérisation

La cinquième et dernière question concerne la numérisation de sources. L'océrisation des éditions érudites reste le moyen principal d'obtenir un texte brut comme base de travail. Le contenu de ces éditions fournit des éléments de correction syntaxique, graphique et lexicale comme la ponctuation, le développement des abréviations et les majuscules, très utiles pendant la phase de modélisation. Mais en même temps, ils ajoutent des éléments diacritiques et paratextuels : signes conventionnels indiquant abréviations, suppressions, modifications, substitutions, etc. ; caractères spéciaux ; parenthèses ; commentaires et titres, qui sont souvent interprétés comme du « bruit de fond », affectant fortement la reconnaissance.

La construction du modèle

Normalisation du texte

L'une des difficultés du travail sur des textes complexes édités par les érudits est l'excès de ce « bruit de fond ». Un texte comme « Heldevini de Matriolis ([con]cedentis. [Æc]clesia [Vallis lucentis] nunc pos[sid]eat feodum)… »[11] oblige à lister tous les signes diacritiques, à préparer des scripts de correction des signes spéciaux et des diphtongues, ainsi que d'élimination des gloses, titres et commentaires qui font partie de l'appareil textuel, mais pas du texte. Le toilettage du texte réduit les taux d'erreur du

10. Voir T. BRUNNER, « Le passage aux langues vernaculaires dans les actes de la pratique en Occident », *Le Moyen Âge*, 115/1 (2009), p. 29-72.

11. W. O. DUBA, *The Cartulary of Vauluisant : A Critical Edition*, mémoire de maîtrise d'histoire de l'art médiéval, Université Paris IV-Sorbonne, 1996, t. I, p. 235.

système. Le nettoyage automatique n'est pas compliqué, mais il a le défaut de transformer et de simplifier le texte ; il a donc fallu définir des méthodes pour récupérer les données perdues et ainsi restaurer les textes ayant subi un traitement automatisé.

Ventilation chronologique

Un autre aspect complexe touche directement la question de la représentativité. Les méthodes de reconnaissance divisent le corpus principal en deux parties : l'une consacrée à l'entraînement et l'autre à tester le modèle. Cela se fait dans un rapport de 4 : 1, ce qui ici correspond à un corpus d'entraînement de 4 000 unités – plus d'un million de mots – et à un corpus de test de 1 000 documents. Idéalement, les deux parties doivent respecter une distribution interne identique. Cependant, l'homogénéité de la répartition n'est pas assurée dans un corpus avec d'importantes asymétries tel que le *CBMA*, qui accueille des pratiques et des usages textuels fortement associés à des époques, à des copistes et à des institutions diverses. En effet, les statistiques montrent des décennies de grande variété typologique et des décennies de grande pénurie. Afin d'éviter des creux chronologiques, on a proposé une double ventilation : chronologique, à partir d'une clé de 25 ans, et aléatoire. Cependant, la différence entre les deux était quasi nulle et le caractère aléatoire a été retenu pour préserver les valeurs originelles.

Formation de sous-corpus

La ventilation à partir d'une clé constitue néanmoins une solution à considérer pour résoudre les questions posées par la concentration et la représentativité du corpus, et donc pour répondre aux enjeux de l'extension du modèle à des corpus externes. Afin d'étudier l'impact de ces différents facteurs, il a fallu envisager la création de sous-corpus internes au *CBMA* et d'autres corpus provenant d'autres régions, tous construits selon différents critères : la taille, la chronologie et l'origine.

a. Les premiers sous-corpus d'entraînement ont été composés de 4 000, 2 500, 1 000 et 500 chartes afin de trouver le meilleur équilibre entre l'efficacité et la taille, ce qui permet de développer un modèle moins dépendant du corpus local, plus robuste sur des corpus variés et moins exigeant en termes de ressources informatiques.

b. Des sous-corpus ont été formés avec des documents datés du même siècle. On a ainsi créé quatre ensembles servant à la fois de corpus de modélisation et de corpus test. Les documents datés selon une fourchette chronologique à cheval sur deux siècles – par exemple des années 980-1020 – sont inclus dans les sous-corpus des deux siècles. Une validation

croisée entre les quatre ensembles nous a permis de faire des comparaisons plus précises et de vérifier l'effet de la variabilité.

c. Un sous-corpus de 400 documents supplémentaires a été balisé à la main, afin de couvrir les « zones grises » de grande pénurie documentaire au cours des IX^e^, XII^e^ et XIII^e^ siècles. L'objectif est d'éviter de perdre, à cause de ces lacunes chronologiques, certaines variétés scripturales.

d. Enfin, ont été créés des corpus de chartes provenant de quatre régions avec une intense production documentaire : la Castille[12], la Lombardie[13], l'Île-de-France[14] et le Sud de l'Angleterre[15]. Ces corpus de petite taille – 70 à 100 chartes – ont été annotés à la main suivant les mêmes protocoles que le *CBMA*, et ils ont été formés suivant les mêmes critères chronologiques et typologiques que le corpus principal. Tout cela vise à constituer un cadre de validation de la robustesse de notre modèle, avec des documents similaires du point de vue de la typologie et de la diplomatique, mais extérieurs à la Bourgogne.

La modélisation informatique

Sur le plan formel, un texte est une composition relationnelle et séquentielle de signes interdépendants. Au sein de ce cadre les entités nommées correspondent à différentes catégories, elles apparaissent successivement et peuvent être distinguées des autres mots par différents indices syntaxiques et morphologiques. Puisque le défi central de la reconnaissance des entités nommées est de diviser un texte entre les mots constitutifs de ces entités et ceux qui n'en font pas partie, puis de baliser chaque entité, l'exploitation des propriétés de chaque mot et de son contexte est fondamentale.

Conditional Random Fields *(CRF)*

Parmi les différentes méthodes qui peuvent être adoptées pour modéliser cela, la technique des champs aléatoires conditionnels (CRF, *Conditional Random Fields*) semble la plus appropriée. La modélisation de toutes les relations possibles entre les variables en question, de la totalité des états et des attributs possibles, peut conduire à des systèmes très complexes. Mais les modèles CRF font ce travail en conditionnant l'émergence d'une variable à l'émergence d'un certain nombre d'attributs dans un mot et dans les mots voisins. Ainsi, le CRF calcule la probabilité qu'a chaque

12. *Corpus Hispánico y Americano en la Red*, <http://www.corpuscharta.es/consultas.html>.
13. *Codice diplomatico della Lombardia medievale*, <http://cdlm.unipv.it/>.
14. *Cartulaires d'Île-de-France*, <http://elec.enc.sorbonne.fr/cartulaires/>.
15. *Documents of Early England Data Set*, <http://deeds.library.utoronto.ca/>.

séquence de balises d'être correcte selon certaines observations, ce qui est généralement suffisant pour déterminer la classe d'une entité[16].

L'entraînement du modèle

Ainsi, l'explicitation des propriétés qui peuvent être extraites de chaque mot du corpus devient impérative. Le corpus entier devient une base de données dans la mesure où chaque mot qui le compose est entré dans un tableau à sept colonnes (fig. 3).

TOKEN	POS	LEMMA	CASE	SUFIX	ENTITY	ENTITY
Quod	CON	quod	UPPER	uod		
ego %x[2,0]	PRO	ego	LOWER	ego		
Hugo %x[1,0]	NAM	-	UPPER	ugo	B-PERS	
de %x[0,0]	PRE [0.1]	de [0,2]	LOWER [0.3]	de [0,4]	I_PERS[0.5]	[0.6]
Berziaco %x[-1,0]	NAM	-	UPPER	aco	I_PERS	B-LOC
perpendens %x[-2,0]	VBE	perpendeo	LOWER	ens		
.	PON	.	LOWER			

3. Informations séquencées de chaque mot pour l'entraînement du modèle

Les trois premières colonnes donnent au modèle un premier niveau de catégorisation. Ils contiennent la version *tokenizée* – réduite à des unités indivisibles – de chaque mot, la catégorie morphosyntaxique (*Part-of-speech Tagging*), obtenue à partir d'une version de TreeTagger[17] développée par le groupe Omnia en 2013[18], et le *lemma*, version sans flexion de chaque mot. La quatrième colonne indique la présence ou non de majuscules, un indicateur utile de la présence d'entités nommées comme des limites de la phrase, et la cinquième colonne ajoute le morphème final, formé des trois dernières lettres, où figure la déclinaison en latin. Les deux dernières colonnes caractérisent les entités nommées, annotées suivant le format BIO où B-entité, I-entité et O représentent le début (B-entité, *Begin*), la poursuite (I-entité, *Inside*) ou l'absence (O, *Outside*) d'une entité (voir un exemple dans la fig. 3).

On étudie ensuite la combinaison des séquences d'entités nommées dans le texte à l'aide d'un fichier-guide suivant une séquence de *n-grams*. Ce concept, largement utilisé dans le domaine du traitement automatique du langage (TAL), fait référence au nombre de composants intégrant les

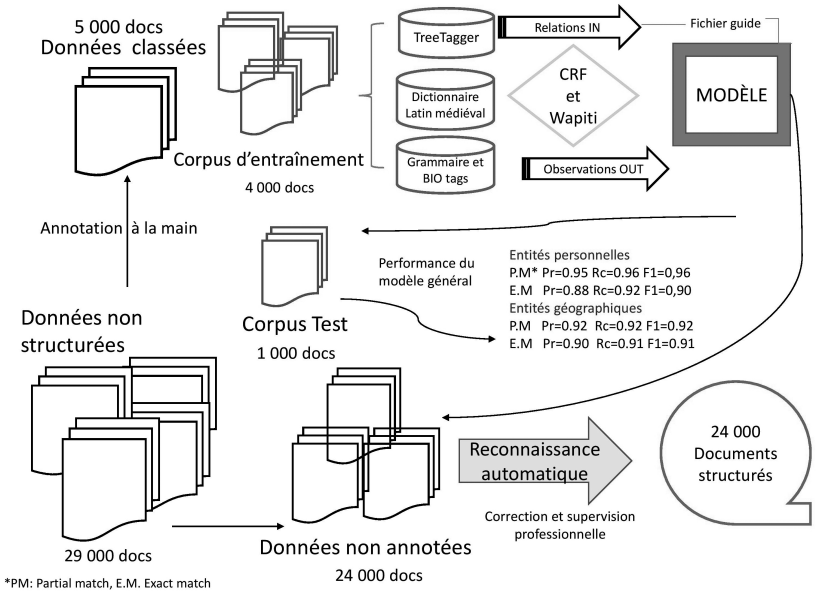
16. É. GAUSSIER et F. YVON, *Modèles statistiques pour l'accès à l'information textuelle, introduction aux CRF*, Paris, 2011, p. 232-247.

17. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>.

18. <http://www.glossaria.eu/treetagger/>.

séquences impliquées dans les observations, dans notre modèle deux lignes avant et deux lignes après la séquence d'intérêt (5 *n-grams*).

Toutes les informations sont finalement traitées par un algorithme d'optimisation L-BFGS incorporé dans Wapiti[19], outil d'étiquetage de séquences mis au point par le laboratoire LIMSI-CNRS.



Pr = Précision Rc = Exactitude

4. Schéma de modélisation sur le *CBMA*

La figure 4 résume le processus : le corpus d'entraînement (5 000 documents annotés) subit les divers processus détaillés ci-dessus, afin d'obtenir notre modèle de reconnaissance d'entités nommées. On applique ensuite ce modèle au corpus de test. On compare les résultats obtenus avec notre modèle à ceux fournis par l'annotation manuelle, en utilisant l'outil Breteval[20] qui analyse les résultats selon les configurations suivantes : *exact match* (E.M.) compte le ratio de parfaites reconnaissances de l'entité (*true positive*) en extension et en catégorie, et *partial match* (P.M.) compte le ratio d'entités détectées mais mal reconnues en extension ou catégorie. Une fois le modèle validé, on peut l'appliquer aux documents non annotés.

19. <https://wapiti.limsi.fr/>.
20. <https://bitbucket.org/nicta_biomed/brateval>.

Analyse des résultats (voir fig. 5)

a. Le modèle général, établi à partir de 4 000 chartes, obtient un taux élevé de reconnaissance sur le corpus test de mille documents. Parmi les analyses visant à déterminer le meilleur rapport performance/taille, on observe que le plus performant est celui de 1 000 chartes, qui, alors qu'il analyse quatre fois moins de documents, présente des résultats inférieurs de seulement 2 points.

Modèles	Match	Entités personnelles	Entités du lieu
4400*	P.M.	0.96	0.92
	E.M.	0.95	0.91
1400*	P.M.	0.92	0.89
	E.M.	0.91	0.88

*À chaque corpus d'entraînement, on a ajouté 400 chartes supplémentaires pour pallier le manque de diversité (voir *supra*).

b. Nous avons obtenu un taux élevé de *match* partiel, indiquant que notre modèle permet de détecter dans le texte des entités nommées, quelle que soit leur taille, avec une efficacité d'au moins 96 % et de les classer correctement avec au moins 90 % de précision. De plus, la différence observée entre le *match* partiel et le *match* exact est de seulement 2 points, ce qui implique que la reconnaissance des entités composées n'a pas été aussi ardue que le laissaient supposer les performances habituelles des modèles. En revanche, on observe que la reconnaissance des entités géographiques complexes reste plus problématique, puisque les *matchs* partiels et exacts sont respectivement de 91 et 92 %.

c. Les résultats croisés des quatre modèles triés par siècles ne montrent pas de grandes différences, et leurs performances sont inférieures, entre 5 et 10 points, au modèle général. Les modèles avec la meilleure performance sont les modèles des siècles centraux (XI^{e} et XII^{e}), et les plus faibles sont ceux des siècles périphériques (X^{e}, XIII^{e}). Cette relative homogénéité et la proximité avec le modèle général suggèrent que la régularité observée au fil des siècles prévaut sur les changements, tout drastiques qu'ils aient pu paraître dans la variété diplomatique comme dans la chronologie. Les différences observées avec le modèle général sont probablement attribuables à des événements spécifiques pouvant correspondre à des dénominations irrégulières. La moindre performance des modèles des X^{e} et XIII^{e} siècles renforce cette hypothèse et indique la présence de plus fortes variations scripturales durant ces deux siècles qu'au cours des deux siècles centraux.

d. Dans la même veine, les résultats obtenus en appliquant le modèle général sur les chartes européennes confirment cette homogénéité : on a

d'excellents résultats en *match* partiel, reproductibles pour l'ensemble des chartes européennes, y compris lorsqu'on applique le modèle développé à partir des sous-corpus triés par siècles : le *match* partiel des entités personnelles et géographiques est quantifié entre 75 et 85 %. Cependant les différences entre *match* partiel et *match* exact sont plus prononcées et les performances en *match* exact sont moindres, entre 65 et 80 %, ce qui indique des problèmes dans la reconnaissance des entités composées, notamment géographiques. Par ailleurs, on remarque à nouveau que la performance des modèles établis à partir des siècles centraux (XI^e^ et XII^e^) est supérieure à celle des siècles périphériques (X^e^ et XIII^e^ siècles).

MODEL/TEST		ANGLO		CASTILE		ILE_FR		LOMB		10th		11th		12th		13th	
		PERS	LOC	PERS	LOC	PERS	LOC	PERS	LOC	PERS	LOC	PERS	LOC	PERS	LOC	PERS	LOC
5000	PM	0,93	0,82	0,93	0,85	0,94	0,85	0,93	0,84								
	EM	0,83	0,76	0,81	0,73	0,87	0,79	0,85	0,81								
1500	PM	0,89	0,82	0,92	0,85	0,94	0,85	0,93	0,83								
	EM	0,83	0,77	0,80	0,73	0,88	0,79	0,88	0,79								
10th	PM	0,93	0,66	0,92	0,75	0,89	0,75	0,93	0,72			0,97	0,89	0,94	0,86	0,84	0,83
	EM	0,86	0,60	0,75	0,63	0,82	0,69	0,87	0,65			0,94	0,87	0,85	0,81	0,75	0,76
11th	PM	0,94	0,80	0,92	0,83	0,91	0,83	0,93	0,83	0,98	0,93			0,96	0,93	0,86	0,88
	EM	0,88	0,74	0,77	0,70	0,85	0,76	0,88	0,78	0,97	0,91			0,90	0,89	0,79	0,82
12th	PM	0,91	0,79	0,88	0,84	0,92	0,85	0,89	0,75	0,96	0,88	0,96	0,90			0,88	0,89
	EM	0,83	0,75	0,76	0,73	0,86	0,79	0,84	0,70	0,95	0,85	0,95	0,88			0,81	0,86
13th	PM	0,84	0,71	0,81	0,77	0,91	0,84	0,85	0,60	0,94	0,87	0,93	0,84	0,92	0,89		
	EM	0,71	0,78	0,69	0,65	0,83	0,78	0,78	0,56	0,93	0,85	0,90	0,82	0,84	0,86		

EM : *Exact match*, PM : *Partial Match*.

5. Résultats de cross-validation entre les modèles et les sous-corpus

Ouvrir la boîte noire

Pourquoi un modèle créé à partir d'un corpus régional est-il en mesure d'obtenir un taux de reconnaissance élevé, même sur des documents provenant d'autres régions ? La création du modèle opère à partir d'une macro-analyse du corpus dans laquelle des millions de caractéristiques sont combinées. Mais l'interprétation fait appel à une analyse plus fine, plus événementielle. Entrouvrir la célèbre boîte noire et tracer des lignes d'explication cohérentes est une tâche qui ne peut être menée qu'en ayant recours aux savoirs des humanités[21].

Les études sur l'anthroponymie médiévale[22] s'accordent à reconnaître un double mouvement dans le panorama des noms médiévaux à la fin du X^e^ siècle : d'une part, une réduction du « stock » de noms et, en conséquence,

21. La relation entre l'informatique et la linguistique pose une question similaire : voir I. TELLIER et M. STEEDMAN, « Préface », *ATALA,* 50/3, (2010), p. 1-20.

22. Voir les études de M. BOURIN et P. CHAREILLE, *Genèse médiévale de l'anthroponymie moderne*, Tours, 1998, vol. 1 et 2, et les actes du colloque dédié à la question : M. BOURIN, J. MARTIN et F. MENANT éd., *L'Anthroponymie : document de l'histoire sociale des mondes méditerranéens médiévaux*, Rome, 1996.

de leur variété, et, d'autre part, une forte extension du nom double à partir de l'association entre le prénom et un second élément, que ce soit le nom du père (*nomen paternum*)[23], un locatif[24] d'origine ou de résidence, ou un surnom (y compris une profession, une qualité ou un titre)[25]. En dépit de différences notables dans la chronologie, les noms doubles commencent à se répandre dans la plupart de l'Europe occidentale dès la fin du Xe siècle.

Les formes simples disparaissent au XIe siècle et le double nom, sous forme prépositionnelle (nom + *de* + entité) ou juxtaposée (nom + nom), devient alors le dénominateur commun. Selon la région, les formes peuvent varier et adopter des déclinaisons, former des syntagmes ou être associées au sein de périphrases (nom + *nexus* + nom). La représentation de ces formes est ordonnée selon une disposition : B-PERS + *de* + B-LOC ou B-PERS + I-PERS. Leur reconnaissance n'est pas très complexe puisque la séquence d'observations nécessaires à cette tâche est limitée.

L'évolution des formes géographiques est plus difficile à déterminer. En principe, la détection des entités géographiques est liée à l'expansion du double nom, dont la forme la plus répandue voit s'ajouter un deuxième élément locatif juxtaposé ou prépositionnel (nom + *(de)* + nom de lieu). Ce locatif n'est pas statique ; il développe différents types selon qu'il s'agisse d'éléments d'origine, de résidence ou de domaine, ce qui, dans de nombreux cas, correspond à des micro-toponymes ou hagio-toponymes composés de deux ou plusieurs éléments.

Cependant, la majorité des entités géographiques apparaissent liées à des descriptions territoriales suivant une relation de co-occurrence, à savoir l'association entre les mots de présentation et les entités ; cela fournit des informations contextuelles particulièrement utiles pour le modèle (voir fig. 6). Ce facteur devient plus important encore, en terme statistique, au sein des structures stéréotypées telles que celles figurant dans les cartulaires. En effet, une partie de ce vocabulaire de présentation, copié, compilé et réutilisé, traduit différents niveaux de représentation de l'espace.

Dans le même ordre d'idées les co-occurrences agissent au niveau des noms. Les différentes catégorisations sociales et professionnelles, les relations familiales, les titres et dignités attachés à l'entité sous la forme d'appositions, d'attributs, de périphrases ou de compléments du nom, développent un vocabulaire qui fonctionne comme un déclencheur, pour le modèle, de la présence d'une entité nommée[26].

23. Par exemple « Rudericus Martinis, Didacus Lupi, Munio Alfonso ».

24. Par exemple « Hugo de Calmonte, Fulco de Brena, Paganus Lumbardus, Gui de la Tour ».

25. Par exemple « Teulfus Pistor, Jocerannus Grossus, Martinus Infantulo ».

26. On peut trouver des informations plus détaillées sur ce sujet dans N. PERREAUX et C. REY, « *CBMA. Chartae Burgundiae Medii Aevi*, VII. Le "vocabulaire courant" en

Entités personnelles	Entités géographiques
Professions : *camerarius*, *magister*, *miles*, *monachus*, *notarius*, *sacerdos*.	Descriptions du paysage : *boscum*, *fluvius*, *locus*, *mons*, *nemus*, *pratus*, *rivus*, *silva*.
Titres séculaires et religieux : *abbas*, *beatus*, *comes*, *dominus*, *domnus*, *dux*, *episcopus*, *papa*, *presbyter*, *princeps*, *rex*.	Division seigneuriale et ecclésiastique : *ager*, *conventus*, *curtilus*, *domus*, *feudus*, *grangia*, *mansus*, *pagus*, *vicus*.
Dignités et surnoms : *benedictus*, *brumus*, *cantor*, *grossus*, *humilis*, *largus*, *normandus*, *paganus*, *servus*, *venerabilis*.	Division légale et juridictionnelle : *areae*, *castrum*, *civitas*, *diœcesis*, *dominus*, *ecclesia*, *provincia*, *sedes*, *terra*, *villa*.
Liens de périphrases : *appellatus*, *cognomen*, *dictus*, *nomen*, *vocatus*.	Micro-espaces : *altar*, *atrium*, *capella*, *capitulum*, *castellum*, *cenobium*, *domus*, *ecclesia*, *hospital*, *monasterius*.
Mots de valeur nominale : *alius*, *ego*, *filius*, *frater*, *idem*, *nepos*, *signum (S.)*, *uxor*.	Termes locatifs, prépositions, adverbes : *ad*, *apud*, *dicitur*, *fines*, *inter*, *manus*, *meridies*, *parte*, *pro*, *supra*, *vocabulum*.

6. Mots de présentation des entités nommés (co-occurrences)

Le modèle détecte et classe avec une grande efficacité les entités simples et les combinaisons limitées, mais sa performance est moindre pour la classification des entités complexes. Ce constat nous amène à analyser certains exemples de cette complexité. Les plus épineux concernent l'imbrication et le chevauchement d'entités, puisque la plupart des classificateurs d'apprentissage automatique ne sont pas conçus pour attribuer plus d'une classe à chaque cas.

Le premier, le génitif possessif (ou nominatif), conduit à la superposition d'une entité géographique et personnelle sous un seul nom. Ceci apparaît souvent dans la description des limites territoriales[27], un phénomène récurrent car les cartulaires recueillent des copies d'actes juridiques territoriaux comme les donations, legs, litiges, etc. Le second, plus problématique, est étroitement associé aux *donationes pro anima*. Le formulaire utilisé apparaît au Xe siècle et crée souvent des entités complexes, de quatre éléments ou plus, en superposant les noms d'un saint, d'une institution et d'une entité géographique[28], puisque, sous un même nom, l'institution détient la fonction de réception, d'intermédiation et de garde des propriétés[29]. Ces deux cas ont besoin d'un long processus de correction manuelle pour bien séparer et distinguer toutes les entités concernées.

diplomatique : techniques et approches comparées », *BUCEMA*, Collection CBMA (2013), p. 4-20.

27. Par exemple « mane terra Bertrannus ; sub domi ipsius Ansaldi ».

28. Par exemple « Ego dono vineam unam Sacrosancte ecclesie Sancti Vincentii Matisconensis ».

29. E. MAGNANI, « Le don au Moyen Âge : pratique sociale et représentations, Perspectives de recherche », *BUCEMA*, 4 (2000), p. 62-79.

Par ailleurs, la diffusion des formes nominales périphrastiques conduit à la création de formes complexes de plus de quatre éléments : B-PERS + (3) I-PERS. Dans les régions lombardes, par exemple, ces entités très complexes à classer sont créées par l'ajout de nombreux liens nominaux avant le deuxième prénom[30]. Un processus similaire est observé dans le centre de la France, où la combinaison de *nexus* et de locatifs peut créer de longues entités à partir du XII^e^ siècle[31]. Au lieu de cela, dans les régions où triomphe sous différentes combinaisons le *nomen paternum* – la Castille, le Sud de l'Angleterre, la France méridionale – sont ajoutés des noms de lieu simples ou doubles et, dans une moindre mesure, des surnoms et des titres, créant ainsi des formes triples ou quadruples, en particulier à la fin du XII^e^ siècle[32]. Afin de détecter de telles compositions, les séquences d'analyse ont été allongées jusqu'à 7 éléments (7 *n-grams*), ce qui élève le nombre d'observations à considérer tout en obligeant à prendre en compte les périphrases comme partie intégrante de l'entité, afin de conserver d'importantes informations nominales.

Cette complexité est encore augmentée par l'état de la langue. D'une part, l'expansion progressive du génitif au XI^e^ siècle produit de longs groupes d'entités sans particule intermédiaire et, d'autre part, l'expansion des prépositions et la disparition du cas latin peuvent conduire le modèle à allonger artificiellement une entité. Le phénomène n'est pas anodin car, en raison de la flexibilité en latin de l'ordre de la phrase, l'entraînement à partir des co-occurrences peut conduire à des faux positifs[33]. Ajouter le cas du mot pourrait pallier ce problème, mais on ne le considère pas pertinent étant donné la variabilité du latin médiéval.

La difficulté de la reconnaissance de ces entités ne réside pas tant dans leur quantité que dans leur extension. En effet, bien qu'il s'agisse de phénomènes minoritaires, ceux-ci ont un impact important sur le taux d'efficacité, car il s'agit d'entités longues de quatre à six éléments. L'impact est encore plus grand si l'on considère que les catégories utilisées par le modèle, en partie dues au balisage originel, ne considèrent que les noms de personnes et de lieux, en rejetant les personnes morales, classées comme des noms géographiques, ce qui augmente la possibilité d'obtenir un classement incomplet.

30. P. CORRARATI, « Nomi, individui, famiglie a Milano nel secolo XI », *Mélanges de l'École française de Rome. Moyen Âge*, 106/2 (1994), p. 459-474.

31. Par exemple « Adamus qui dicitur de Fontanella ; Albertus de loco Praxiate, Teoberto filio Gemardi que vocatur Puncta ».

32. Par exemple « Petrus Nunnez de Salarzar ; Petrus de Sancto Laurentio ; Johannes Paganus de Petra Clausa ».

33. Par exemple « Iohannes Aurelianensis episcopus ; Petrus Sancti Mederici prepositus ; Gauchero milite de Uisione ».

Pour récapituler, les formes simples du nom médiéval s'effacent dès le XI^e siècle alors qu'apparaissent les formes doubles, en particulier les *nomina paterna* et les formes locatives. À partir de celles-ci, des formes triples ou périphrastiques sont produites dès la fin du XII^e siècle, bien qu'avec une plus faible incidence. Formellement, la stabilité relative du nom double, de ses combinaisons proches et du système de co-occurrence permet une modélisation complexe, mais très abordable, conduisant à de hautes performances en matière de détection et de classification des états les plus communs des entités nommées. Les différentes étapes de l'évolution du nom médiéval aident à mieux comprendre la moindre efficacité des modèles établis à partir des sous-corpus des X^e et XIII^e siècles par la forte présence d'entités simples au X^e siècle, ou par la présence croissante de formes complexes au XIII^e siècle.

Dans un autre ordre d'idées, un taux de reconnaissance supérieur à 90 % est favorisé par deux facteurs : tout d'abord, une disposition textuelle modélisée et stéréotypée, qui forme un axe relativement stable à partir duquel la production diplomatique participe à la diffusion de modèles spécifiques relatifs aux changements d'origine et de chronologie[34]. D'autre part, on constate une hétérogénéité supérieure à ce qui était attendu en raison de la richesse du *CBMA*, qui présente un stock et une composition de noms beaucoup plus élevés que la normale parce qu'il provient d'une institution de premier plan, ce qui induit des contributions documentaires variées pour la modélisation.

L'évolution du nom médiéval et l'itération dans le discours, cependant, ne sont qu'une partie de l'explication. Sans doute une exploration plus approfondie, qu'à ce jour nous ne pouvons qu'entrevoir, portant sur la caractérisation typologique des actes, les traditions scripturaires et les mutations dans les pratiques scripturales, comme celles qui affectent la production des chartes privées dans les *scriptoria* européens à partir du X^e siècle[35], offriront des explications beaucoup plus détaillées, puisqu'elles intègreront les changements dans les structures formulaires dans lesquelles le texte s'inscrit.

Notre principale contribution est la création d'un modèle de reconnaissance des entités nommées. Suivant les différentes évaluations que

34. P. BERTRAND, A. MAIREY, O. GUYOTJEANNIN, A. GUERREAU, A. M. EDDÉ et M. BURGHART, « L'historien médiéviste et la pratique des textes : les enjeux du tournant numérique », dans *Actes des congrès de la Société des historiens médiévistes de l'enseignement supérieur public*, 38/1 (2007), p. 273-301.

35. O. GUYOTJEANNIN, « "*Penuria scriptorum*" : le mythe de l'anarchie documentaire dans la France du Nord (X^e-première moitié du XI^e siècle) », *Bibliothèque de l'École des chartes*, 155 (1997), p. 11-44.

nous avons effectuées, notre modèle offre une performance élevée et est assez robuste pour être appliqué à un vaste rayon documentaire. Il devrait permettre de semi-automatiser une étape primordiale de la structuration de grandes quantités de données provenant d'écrits manuscrits, économisant ainsi beaucoup d'efforts humains.

L'introduction de ce type de travail répond à la numérisation massive de documents qui requièrent des moyens efficaces d'exploitation. La modélisation, qui a pour objectif de structurer automatiquement ces données, est effectivement une pratique interdisciplinaire. À l'analyse des différents états de la langue, qui constitue le cœur de la méthode, doit s'ajouter l'observation attentive des spécificités proposées par le document médiéval. L'équilibre méthodologique aide à polir le texte sur lequel l'algorithme est appliqué et produit des observations de nature historique qui déterminent l'application numérique.

Enfin, ajouter au résultat numérique une explication basée sur les études consacrées au nom médiéval n'est pas à négliger, car elle permet d'éclairer des résultats d'abord obscurs et de les lier à la recherche en humanités. Cela nous permet d'observer des pistes de travail, même à partir de travaux de validation, qui visent à tester la robustesse du modèle. À titre d'exemple, les spécificités qui sous-tendent l'hétérogénéité observée, bien que sources d'erreur pour le système, sont l'indicateur d'un terrain fertile pour analyser les divers degrés de singularité et d'identité scripturale. Dans le même temps, l'itération dans le discours permet de mieux distinguer les changements synchroniques et diachroniques dans le modèle proposé pour le formulaire, comme la circulation des formules, l'évolution d'usages et de concepts, ou même les modifications à titre personnel. La confluence des méthodes peut, dans un délai raisonnable, conduire à l'observation de phénomènes encore inconnus et à l'exploitation de nouvelles formes de lecture documentaire.

Sergio Torres Aguilar – DYPAC, Université de Versailles Saint-Quentin (Paris-Saclay)

La reconnaissance des entités nommées dans les bases numériques de chartes médiévales en latin : le cas du *Corpus Burgundiae Medii Aevi* (X^e^-XIII^e^ siècle)

La disponibilité d'une quantité phénoménale de manuscrits médiévaux numérisés nous oblige à chercher des méthodes efficaces pour en réaliser une exploitation à grande échelle. Mais ce travail ne peut être réalisé que dans des bases de données structurées où les propriétés textuelles ont été explicitées et formalisées. Une telle structuration, lorsqu'elle est effectuée à la main, est coûteuse en termes de temps et d'effort, ce qui a conduit à chercher des manières de l'automatiser. Nous en présentons ici un exemple : la création d'un modèle de reconnaissance des entités nommées, qui sont un agent structurant primaire, puisque y sont identifiés tous les sujets et

objets qui adoptent des noms spécifiques. Nous détaillons la création et la mise en œuvre du modèle créé à partir des chartes de Bourgogne (*Corpus Burgundiae Medii Aevi*), produites entre le Xe et le XIIIe siècle ainsi que diverses expériences de validation pour en tester la robustesse sur un large éventail de sources, tout en soumettant les résultats à une discussion qui tente d'exposer les divers avantages et défis qu'offre un corpus de manuscrits médiévaux pour ce type de technique.

Bases de données – chartes – *Corpus Burgundiae Medii Aevi* – humanités numériques – latin – reconnaissance des entités nommées.

Named Entities Recognition for Digital Bases of Medieval Latin Charters. The Case of the *Corpus Burgundiae Medii Aevi* (Tenth-Thirteenth Centuries)

The availability of a vast amount of digitized medieval manuscripts requires to import effective methods for large-scale exploitation. But this work can only be done in structured databases where the textual properties are explicit and formalized. This type of handmade structuring is highly time-consuming, which has led to the search for ways to it. We present an example of this: the creation of a named entities recognition model, which are a primary structuring agent, since it corresponds to all the subjects and objects adopting specific names. We detail the creation and implementation of the model formed from Burgundian charters (*Corpus Burgundiae Medii Aevi*), produced from the tenth to the thirteenth centuries; also, we describe various validation experiments in order to test its robustness on a wide range of sources and at the same time we submit all the results to a discussion that shows the various benefits and challenges of this type of technique on a medieval manuscripts' corpus.

Corpus Burgundiae Medii Aevi – Charters – Digital Bases – Digital Humanities – Latin – Named Entities Recognition.