Un modèle de reconnaissance automatique des entités nommées et des structures textuelles pour les corpus diplomatiques médiolatins.

Sergio Torres Aguilar

université PARIS-SACLAY

UNIVERSITÉ DE VERSAILLES ST-QUENTIN-EN-YVELINES

université PARIS-SACLAY

NNT : 2019SACLV081

Thèse de doctorat

Un modèle de reconnaissance automatique des entités nommées et des structures textuelles pour les corpus diplomatiques médiolatins.

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université de Versailles-Saint-Quentin-en-Yvelines

École doctorale n°578 Sciences de l'Homme et de la Société (SHS)
Spécialité de doctorat : histoire, histoire de l'art et archéologie

Thèse présentée et soutenue à Paris, le 05 décembre 2019, par

SERGIO TORRES AGUILAR

Composition du Jury :

Chloé Clavel Professeur HDR, Télécom Paris	Président
Aude Mairey Directrice de recherche, CNRS, LAMOP	Rapporteur
Julien Velcin Professeur des Universités, Université de Lyon 2	Rapporteur
Eliana Magnani Chargée de Recherche, CNRS, LAMOP	Examinateur
Miguel Calleja Puerta Professeur des Universités, Universidad de Oviedo	Examinateur
Pierre Chastang Professeur des Universités, Université de Versailles-Saint-Quentin-en-Yvelines	Directeur de thèse
Xavier Tannier Professeur des Universités, Sorbonne Université	Co-directeur de thèse

Résumé

Nous présentons dans cette thèse deux modèles informatiques développés pour délivrer de l'information structurée et applicables à de grandes bases de données de textes médiévaux. Les deux modèles, l'un appliqué à la reconnaissance des entités nommées, l'autre à la détection des parties du discours diplomatique, ont suivi un apprentissage supervisé utilisant la méthode des Champs aléatoires conditionnelles (CRF) sur un corpus manuellement annoté de actes médiévaux (*Corpus Burgundiae Medii Aevi* ou CBMA).

Notre modèle principal de reconnaissance d'entités nommées a prouvé sa robustesse lorsqu'il a été appliqué sur des échantillons de corpus de taille, chronologie et origine très variés. Le modèle secondaire détectant les parties du discours diplomatique, bien que moins performant, s'est montré valide comme outil de structuration. Ils peuvent à présent être utilisés pour l'indexation et l'étude d'une grande variété de sources diplomatiques, économisant, ainsi des considérables efforts humains.

Nous avons développé différentes solutions destinées à trouver un juste équilibre entre la dépendance du modèle à son corpus d'origine et sa capacité à être appliqué à d'autres corpus. De même, différents ajouts et corrections ont été opérés sur le corpus de référence à partir de plusieurs observations de type historique et linguistique concernant les documents utilisés, ce qui a permis d'améliorer la performance initiale.

Nous avons ensuite appliqué les outils ainsi générés à la reconnaissance de noms de personnes, de lieux et de parties du discours diplomatique sur des milliers d'actes du CBMA afin d'étudier différentes questions intéressant la science historique et la diplomatique. Ces études concernent la datation semi-automatique d'un cartulaire qui en était dépourvu ; l'évolution du vocabulaire spatial dans les actes du Moyen Âge Central ; et l'indexation des documents à partir des modules les intégrant, notamment les formules du protocole des actes. Par ces études nous poursuivons un double objectif : illustrer différentes stratégies permettant d'abstraire et d'adapter au traitement automatique des données des méthodes de recherche classiques en Histoire ; démontrer que nos outils de traitement massif permettent la génération de connaissances pertinentes pour la science historique.

Abstract

In this thesis, we present two computer models to structure textual information for large databases of medieval charters. The two models, one applied to the recognition of named entities, the other to the detection of parts of the diplomatics discourse, are supervised Conditional random fields (CRF) models trained on a hand-annotated corpus of medieval charters. (*Corpus Burgundiae Medii Aevi* or CBMA).

The main Named Entity Recognition model has proven to be robust in its application to widely varying corpora in size, chronology and origin. The secondary model detecting parts of the diplomatic discourse, although less efficient, remains valid as a structuring tool. At the moment both can be used for indexing and studying a wide variety of diplomatics sources, thus saving huge human efforts.

We have developed different solutions to overcome the gap between model's dependence on its original training-set and its ability to be applied to other corpora. Similarly, various corrections and additions were made to the golden-corpus from several historical and linguistic analysis concerning writing phenomena in charters, which greatly helped to improve the initial performance.

In a later step we applied our automatic tools in the recognition of names of people, places and parts of the diplomatics discourse on thousands of charters from the CBMA corpus in order to study different questions concerning historical science and diplomatics. These studies concern the semi-automatic dating of a non-dated cartulary ; the evolution of the spatial vocabulary in the charters of the central Middle Ages and the indexing of charters from their scriptural modules, in particular formulae of the charter protocols. This studies has a twofold purpose : on the one hand have shown different strategies for abstracting and adapting to the automatic processing well-known methods of research in history ; on the other hand, seek to provide us tools with an applicative framework to obtain relevant knowledge to the historical science using massive processing.

No quiso desayunarse don Quijote, porque, como está dicho, dio en sustentarse de sabrosas memorias. Tornaron a su comenzado camino del Puerto Lápice, y a obra de las tres del día le descubrieron.

—Aquí —dijo en viéndole don Quijote— podemos, hermano Sancho Panza, meter las manos hasta los codos en esto que llaman aventuras.

Don Quijote de la Mancha, 1º parte, VIII

Le pregunté qué sabia de la *Odisea*. La practica del griego le era penosa ; tuve que repetir la pregunta. « Muy poco - dijo -. Menos que el rapsoda más pobre. Ya habrán pasado mil cien años desde que la inventé.»

Borges, El inmortal

Remerciements

Habiendo llegado al final de este pequeño camino debo reconocer el mérito y la ayuda de quienes han permanecido cerca. A mis directores, Pierre Chastang y Xavier Tannier por su infatigable amabilidad y rigor científico. Os agradezco vuestro compromiso constante y vuestra más que infinita paciencia.

A los miembros del jurado que han aceptado participar del tribunal de esta tesis. Será un grato placer rencontraros para esta lectura.

A Delphine, que habla perfectamente mi franco-español y ha corregido esta tesis con varias noches en blanco encima. Tu apoyo ha sido mayor del que esperaba y nunca menor del que necesitaba.

También a Fifi que me ha aportado siempre una luz en las oscuridades del latin medieval. Y a quien debo reconocer sus dotes premonitorias.

A todos mis amigos del Colegio de España. A José Luis y a Paco que han sido los mejores compañeros que alguien como yo podría encontrar y a quienes me une una amistad indeleble. A Juanma, Belén, Aitor, Tony, Estefanía, Felipe, Vane, exiliados en París, amigos, cuya inteligencia y carisma hicieron más llevaderos los tiempos difíciles.

A todos, gracias en grado superlativo.

Table des matières

Table des figures

Liste des tableaux

Glossaire

L'étoile (*) indique les définitions extraites de l'édition du Vocabulaire international de Diplomatique (VID) : ORTÍ, M. Milagros Cárcel (ed.). *Vocabulaire international de la diplomatique*. Universitat de València, 1997.

Acte écrit * : Écrit consignant l'accomplissement d'un acte juridique. Comme le document produit un effet de droit, les actes servent en tant que titres de propriété foncière ou de droit. (VID, p.21)

Annotation des parties du discours ou POS (Part-of-speech) tagging : Traitement lexical consistant à annoter les catégories grammaticales qui concourent sur chaque mot de la phrase. Cela correspond normalement aux parties du discours et aux caractères morpho-syntaxiques. Cet annotation est faite à l'aide de l'ordinateur en utilisant des modèles formés sur des corpus arborés.

Annotation sémantique (Entity Linking) : Tâche consistant à relier les entités d'un texte avec des contenus sémantiques portant l'identification de l'entité, une description et des metadonnées qui permettent la connecter à une ontologie.

Apprentissage supervisé : L'apprentissage automatique est dit supervisé lorsqu'on fournit à l'algorithme les réponses attendues sous forme de variables annotées. L'algorithme, après l'entraînement, doit être capable de généraliser son apprentissage sur des données non annotées.

Boîte noire : Dans le domaine des humanités numériques l'effet de boîte noire fait référence aux barrières interprétatives imposées par les algorithmes lorsqu'ils sont appliqués à la transformations des textes. On considère que leurs résultats, bien que performants, sont peu perméables à la critique et à la médiation scientifique et qu'ils ne permettent pas de bien cerner l'impact (et les distorsions) que les outils automatiques provoquent sur les résultats rendus.

Bootstrapping : Les méthodes dites de "bootstrapping" (traduit comme autoamorçage) consistent à échantillonner à maintes reprises les données obtenues à partir d'un échantillon initial. Ainsi, on utilise un petit ensemble d'échantillons annotés à la main afin de former un modèle qui, bien que médiocre, offre une annotation automatique facile à corriger, augmentant ainsi la disponibilité des données annotées.

Cartulaire * : Un cartulaire (lat. : c(h)artularium) est un recueil de copies de ses propres documents, établi par une personne physique ou morale, qui transcrit ou fait

transcrire dans un volume des titres relatifs à ses biens et à ses droits et des documents concernant son histoire ou son administration, pour en assurer la conservation et en faciliter la consultation. (VID, p. 36)

Chartrier * : Ensemble des chartes conservées par une personne physique ou morale – le plus souvent, un seigneur, une institution ecclésiastique, une ville – pour faire la preuve de ses droits ou conserver la mémoire de son histoire. (VID, p.27)

Corpus arboré (Treebank) ou corpus décoré : Corpus textuel portant, pour chaque phrase, l'annotation (*parsed corpus*) de sa structure syntaxique, sous la forme d'un arbre où les nœuds et feuilles représentent les relations de dépendance syntaxique entre les différentes occurrences de la phrase.

Désambiguïsations des entités : La désambiguïsations de l'entité (ou normalisation, ou encore liage référentiel) fait référence à la récupération des informations permettant d'identifier les entités nommées, ce qui se fait normalement en connexion avec des bases de données externes ou en exploitant massivement les sources internes afin d'avoir une identification relative.

Diplomatique : Science qui étudie la tradition, la forme et l'élaboration des actes écrits ; s'intéressant surtout aux enjeux liés à l'authenticité de l'acte, sa datation et sa composition. (Dans la science historique, les *diplômes* correspondent aux documents émanant d'une autorité publique.)

Distance d'édition : Mesure qui détermine le nombre de modifications (suppressions, remplacements ou insertions) nécessaires pour transformer une chaîne de caractères en une autre. Le coût d'édition de cette opération sert, par extension, pour déterminer le niveau de ressemblance entre deux mots ou phrases.

EAD (encoded archival description) et MARC (machine readable cataloging) sont deux standards utilisés pour la description d'information archivistique et bibliographique dans des catalogues numériques.

Eschatocole ou protocole final : Cadre formel final d'un acte écrit.

Faux-lemme * : Mot qui ne doit son existence qu'à une erreur de lecture (de scribe ou d'éditeur) ou à une faute d'impression. (VID, p.52)

Formulaire * : Recueil de formules destinées à servir de modèles aux rédacteurs des actes et, éventuellement, à contribuer à la formation des rédacteurs eux-mêmes. (VID, p. 37)

Formule * : Phrase, proposition ou groupes de mots dont use le rédacteur d'un acte pour exprimer chacune des clauses ou chacun des éléments formels de l'acte selon les usages diplomatiques ou juridictionnels. Les formules peuvent subir différents agencements et variations selon les typologies documentaires, les actions juridiques, l'identité des rédacteurs, bénéficiaires et auteurs, etc. (VID, p.54)

Lacunarité des données (data sparsity) : On parle de *data sparsity* lorsque un échantillon de données est composé d'un large nombre de variables et il n'existe pas suffisament d'observations pour chaque variable. Ce phenomène est assez commun dans les travaux de traitement automatique des langues (TAL) car si le vocabulaire d'une langue est composé de milliers d'unités, un texte n'en comporte normalement que quelques centaines et un domaine précis emploie un vocabulaire plus restreint encore ; autrement dit, la distribution est assez éparse.

Langage de marques (markup language) : Langage spécifiquement employé pour indiquer le format, le style et les principales structures d'un texte utilisant des balises. Les langages les plus connus sont XML et HTML.

Lemmatisation : Traitement lexical consistant à réduire les formes flexionnelles à leur forme canonique qui sont les lemmes ou lexèmes repérés dans les dictionnaires.

N-gramme : Séquence de N-items extraits d'une chaîne textuelle. Le n-gramme peut ainsi correspondre à une séquence de caractères, phonèmes ou mots, le plus souvent. L'utilisation de ces sous-séquences est assez habituelle dans les études statistiques de langue parce qu'elles permettent d'étudier le contexte immédiat des mots et de prédire les séquences, comme dans un modèle de langue, par exemple.

Notice * : Écrit dans lequel est consignée la substance d'un acte ou d'un fait juridique, soit par le destinataire ou le bénéficiaire lui-même, soit par un tiers, en vue d'en conserver la mémoire. (VID, p.96)

OCR (Optical character recognition) : Ensemble de techniques informatiques destinées à transformer les fichiers d'image d'un texte en un texte brut.

Parties du discours diplomatique : Différentes parties constitutives de l'acte écrit dont l'agencement, la structure et l'ordre forment le discours de l'acte.

Précision et rappel : Mesures traditionnelles de la performance des modèles. Le rappel exprime la sensibilité du système au moment de fournir des réponses possibles aux recherches. De son côté, la précision détermine le niveau de correction dans les résultats récupérés.

Protocole ou protocole initial : Cadre formel initial d'un acte écrit.

Recueil original : Édition diplomatique des acte émanant d'une même chancellerie ou d'un même auteur. Les recueils concernent normalement une même personne juridique ou suivent une organisation cohérente par région, action juridique ou chronologie. Le recueil originel s'oppose au recueil factice parce que ce dernier est un assemblage effectué par un propriétaire ancien, un érudit ou par le propriétaire actuel de ces documents.

Sac de mots : Modèle de représentation d'un texte comme l'ensemble des mots qui le composent sans soucis de propriétés ou contexte (ordre, fonction), autrement dit, c'est un vecteur de la fréquence d'apparition des mots.

SIG (système d'information géographique) : Système d'information qui organise, stocke, gère, et affiche différentes données géographiques et qui par extension permet la représentation et l'analyse cartographique de ces données.

Surapprentissage ou surajustement (overfitting) : Effet de surentraîner un algorithme sur un ensemble particulier de données, ce qui a comme conséquence une grande précision sur cet ensemble de données, car l'algorithme a appris les structures "par cœur", mais une mauvaise capacité de généralisation sur des données externes, même si celles-ci sont similaires à celles utilisées lors de l'apprentissage.

TEI (Textual Encoding Initiative) : Ensemble de modèles et lignes directrices qui visent à standardiser la description des propriétés et attributs d'un document en utilisant un langage de marques, notamment en XML.

Teneur de l'acte juridique : Composition formelle de l'acte qui se ramène aux trois groupes qui correspondent au cadre formel initial (protocole), à la matière centrale (texte) et au cadre final (eschatocole) de d'un acte.

Texte brut (*raw text*) : Format de texte qui représente seulement les caractères et ne porte aucune information ou représentation graphique.

TF-IDF (Term frequency – Inverse document frequency) : Indicateurs de l'importance d'un mot dans un documents prenant en compte sa fréquence ou rareté. TF représente la fréquence d'apparition d'un mot dans un corpus alors que IDF représente le nombre de documents dans le corpus portant le mot en question.

Validation-croisée : Technique consistant à valider la robustesse des modèles sur des échantillons de données non utilisées dans l'entraînement. Normalement un re-échantillonage (re-samplig) des données formant différents échantillons d'entraînement et de test sur les mêmes données, sur différents itérations, est suffisant pour bien valider un modèle.

Introduction

0.1 Le tournant numérique

Le recours aux outils numériques peut s'avérer très avantageux pour le traitement des grandes collections de textes. En particulier, ils permettent de réaliser une analyse des documents à une échelle globale, jusqu'au corpus entier, en parcourant des milliers de textes en un temps restreint. Le fait de pouvoir accomplir ce travail d'une façon automatique ou semi-automatique en quelques heures rend les outils numériques presque indispensables dans la recherche sur des questions impliquant de larges secteurs d'un corpus. Les modèles d'exploitation des données proposés dans le cadre scientifique du *data mining*, du *big data*, ainsi que la lecture à distance (*distance reading*) ont établi une méthodologie de travail assez claire à partir de l'extraction et de la formalisation, en peu de temps, des structures sous-jacentes aux grandes collections de textes.

Ce nouveau paradigme, qui privilégie une formalisation immédiate de l'information en formant des groupes, des classes, des modèles, des réseaux, des infographies, etc. s'avère ainsi très utile pour fournir des éléments scientifiques pour l'étude d'une question qui concerne tout un corpus ou parfois même plusieurs, mais elle peut rencontrer de nombreux obstacles lorsque elle est confrontée aux irrégularités et aux phénomènes linguistiques propres au langage naturel et au discours écrit, spécialement lorsqu'il provient d'états anciens de la langue. S'il est vrai que le recours aux outils automatiques nous offre un appareil structuré qui rend la masse textuelle plus facile à gérer, il n'est pas moins certain que l'usage de ces outils exige de nombreuses heuristiques à partir desquelles on peut, d'un côté, connecter le résultat algorithmique aux besoins et aux questions pertinentes pour les sciences sociales ; et de l'autre côté, mettre en place des manières d'interpréter et d'intégrer ces résultats dans les cadres scientifiques propres à ces disciplines.

Dans le domaine des sciences sociales, le recours aux outils numériques dans l'analyse de textes est une pratique devenue courante ces dernières années. Dans les études de critique littéraire, par exemple, elle est bien établie depuis au moins les années 1980 [1]. La facilité pour obtenir des statistiques, des lignes d'évolution d'un terme dans un corpus ou pour rendre possible une analyse factorielle rapide des corrélations entre plusieurs termes, mots et lemmes, ce qui jusqu'ici avait été réalisé à la main, a contribué à formaliser une approche dont les résultats se sont montrés très encourageants [2]. Depuis lors, sur cette niche d'études se sont développées les bases des

1. S Ramsay. "Special Section : Reconceiving Text Analysis : Toward an Algorithmic Criticism". In : *Literary and Linguistic Computing* 18.2 (2003), p. 167-174

2. Susan Hockey. *Electronic Texts in the Humanities : Principles and Practice*. en. OUP Oxford,

nouvelles pratiques proposées par une partie croissante de la communauté scientifique [3], rassemblées sous le terme générique des "humanités numériques" [4].

À la "boîte à outils" des travaux déjà classiques provenant du domaine littéraire comme la stylométrie, la détection des topiques et l'attribution d'auteurs, d'autres ont été ajoutés lorsque l'intérêt pour ce genre d'approches s'est étendu aux sciences sociales, apportant pour chacune des problématiques spécifiques. La nécessité d'abstraire les méthodes de recherche utilisées dans les humanités et de reproduire numériquement la formulation des questions qui résident au cœur de chaque discipline oblige à utiliser des outils de plus en plus spécialisés provenant des études du traitement automatique des langues ou de l'apprentissage artificiel. Le déclencheur de cette évolution doit être cherché dans au moins trois changements-clé des dernières années : un accès massif aux hautes capacités d'informatisation ; une disponibilité croissante des corpus numériques touchant toutes les sciences sociales et un intérêt soutenu pour les travaux interdisciplinaires de la part d'une communauté qui est capable d'utiliser à son profit toute une série d'outils autrefois considérés comme l'apanage des informaticiens [5].

Dans les études historiques, ces approches commencent à gagner un espace propre concernant en particulier la visualisation de réseaux sociaux, la reconstruction cartographique et les enquêtes sur les champs sémantiques des termes [6]. L'analyse quantitative qui opère comme source de ces constructions coïncide avec une partie importante du travail de l'historien lorsqu'il essaye de préciser les éléments et mécanismes de l'élaboration d'un document. De ce fait, les réseaux peuvent concentrer l'information-clé répertoriée depuis les niveaux lexicaux et gagner en généralité jusqu'à arriver à décrire la composition même d'une société ou l'organisation spatiale d'une ville, ayant comme cœur - ou plutôt comme nœuds - les personnes et les lieux. De manière similaire, les études de nature texto-métrique qui s'appuient sur une analyse morphosyntaxique ont commencé à gagner une certaine popularité. Elles mesurent

nov. 2000

3. Susan Schreibman et al. *A Companion to Digital Humanities*. John Wiley & Sons, 2008

4. L'ancien terme *humanities computing* qui définit plutôt l'intérêt de fournir un support technique et de rapprocher quelques outils informatiques pour les humanistes a perdu sa pertinence. Le terme actuel fait référence à une "boîte à outils" proche des méthodologies des humanités dont l'application fait « attention à la complexité, aux spécificités du milieu, au contexte historique, la critique et l'interprétation des résultats » Jeffrey Schnapp et al. "Digital humanities manifesto 2.0". In : *Hentet* 10 (2009), p. 2016

5. Voir la réflexion à propos de ce tournant numérique exposée dans David Lazer et al. "Computational social science". In : *Science* 323.5915 (2009), p. 721-723 " et dans la nouvelle édition de Susan Schreibman et al. *A new companion to digital humanities*. John Wiley & Sons, 2015

6. Voir à ce propos quelques travaux fondateurs qui témoignent de l'intérêt précoce des historiens français, notamment des médiévistes, pour la mobilisation des outils informatiques : Jean-Philippe Genet. "L'informatique au service de la prosopographie : Prosop". In : *Mélanges de l'École française de Rome* 100.1 (1988), p. 247-263 ; Alain Guerreau. "Analyse factorielle et analyses statistiques classiques : le cas des ordres mendiants dans la France médiévale". In : *Annales. Histoire, Sciences Sociales*. T. 36. 5. Cambridge University Press. 1981, p. 869-912 ; Marion Crehange et Lucie Fossier. "Essai d'exploitation sur ordinateur des sources diplomatiques médiévales". In : *Annales* 25.1 (1970), p. 249-284. Beaucoup de ces travaux ont également paru dans une revue pionnière, *Le médiéviste et l'ordinateur*, publiée entre 1978 et 2003 et accessible en ligne sur le site de persee.fr

le poids de chaque mot et de chaque chaîne de mots dans le but de caractériser un document et par extension le style d'un auteur ou la nature d'un modèle scripturaire. Cela permet de comparer des documents et de dégager des informations précieuses à propos des usages stylistiques, des vocabulaires, des formulaires, etc. et ainsi d'aider à confirmer certaines hypothèses dont chaque détail à retenir exigerait un long feuilletage des sources.

Cependant, la revue de la littérature que nous avons menée montre que la plupart des outils et méthodes d'exploration de corpus qui sont utilisés dans l'environnement numérique sont rarement mobilisés chez les historiens, et le panorama montre peu de signes en faveur d'un changement dans les années à venir. Si les outils sont nouveaux, favoriser une analyse des structures de sources à partir des données quantifiables est loin d'être une proposition nouvelle. Elle a été l'objet d'une longue série de débats dans les années 1970, ayant mené à un certain *status quo* qui est à nouveau contesté. Les méthodes statistiques d'analyse des données et la théorie de la modélisation adoptées en sociologie et en économie proposaient un cadre scientifique presque obligatoire pour ces disciplines, si elles voulaient être considérées comme des *sciences*. L'effet immédiat chez les historiens a été une polarisation des méthodologies de travail entre une analyse focalisée sur l'évènementiel et une étude des structures. La transformation des documents en tant que productions textuelles et sociales, en données et matrices numériques, considérées comme mécaniques et imperméables, apparaissait comme inacceptable pour une discipline, l'histoire, très attachée à l'exercice critique comme élément fondamental de ses méthodes.

Cette question est souvent réduite à une confrontation stérile entre méthodes quantitatives et qualitatives où on considère la récupération massive de preuves comme l'apanage des premières et l'argument critique de la source comme l'apanage des deuxièmes. En réalité la question est plus simple : comment introduire effectivement de nouveaux outils qui soient compatibles avec les questions pertinentes pour l'histoire ? Les outils statistiques ont laissé leur place à des outils basés sur l'algorithmique, mais ce sont les mêmes causes qui expliquent la modeste adoption tant des premiers que des seconds. En effet, en appliquant des méthodes qui délivrent des observations générales à grande échelle, on peut craindre que les méthodes utilisées par les études historiques s'en ressentent, car l'un de leur centres d'intérêt relève de l'analyse du contexte et des phénomènes particuliers présents dans des artefacts textuels par définition lacunaires. Et en même temps, l'application de méthodes d'étude assez détaillées nécessite d'une compréhension complète des chaînes logiques de la méthode historique, ce qui pourrait ne pas être disponible dans les méthodes de nature quantitative, soit parce que l'outil utilisé est opaque dans sa manière de produire un résultat, soit parce que l'historien est incapable d'élucider une interprétation des profondeurs énigmatiques de l'algorithme.

0.1.1 Les outils non adaptés.

Il peut s'avérer initialement ardu de définir la ligne permettant de connecter les outils de traitement automatique avec les questions pertinentes pour l'historien. Les travaux menés sur les corpus historiques relèvent de défis assez complexes, dus à la multiplication progressive des registres scripturaires et des états de la langue,

augmentant le nombre de situations linguistiques à observer. Jusqu'à la dernière décennie, la plupart des outils de traitement des langues étaient développés à partir d'échantillons de journaux en anglais du XXe siècle. Après les numérisations massives des années 2000, d'autres collections de textes, plus variées, ont attiré l'attention des chercheurs. Pour bien entamer ces recherches, les outils disponibles n'étaient pas suffisants parce qu'ils n'étaient pas capables d'apporter des solutions efficientes aux nouveaux problèmes soulevés par les changements dans la chronologie, le discours ou la langue. La littérature a fait peu de progrès dans ce sens, ce qui, combiné au manque d'outils généralistes qui pourraient offrir une bonne performance devant ces changements, a obligé les chercheurs à élaborer des modèles *ad hoc*, nécessitant souvent de lourdes adaptations des outils existants. Ils ont ainsi généré des myriades de modèles de traitement automatique attachés aux problèmes spécifiques relevés dans certains corpus. La situation est encore plus critique dans le domaine des études historiques puisqu'elles se focalisent, dans leurs démarches heuristiques sur les différents facteurs qui nourrissent les problèmes rencontrés, comme l'irrégularité dans le discours, les variations dans le vocabulaire ou les graphies, les anachronismes, les phénomènes scripturaires spécifiques et le chevauchement des usages lexicaux.

D'autre part, si on considère comme également indispensables à la recherche historique la valeur d'un corpus en tant que collection organique de textes, comme la sérialisation [7], le contexte documentaire et la représentativité, on dégage une nouvelle problématique, puisqu'on n'a pas accordé une importance similaire à ces éléments dans la recherche autour des outils automatiques. La recherche en histoire distingue deux types de corpus : ceux rassemblés par l'institution productrice ou les archives anciennes, normalement organiques, et ceux formés par différents choix éditoriaux à l'époque moderne (recueils factices). La valeur de la sérialisation donc diffère fortement, mais elle n'est pas prise en compte par les outils, ce qui constitue un sérieux défaut pour la recherche.

S'il est cohérent d'entraîner des outils automatiques avec plusieurs échantillons de journaux des trois dernières décennies puisqu'ils présentent un ensemble réduit et contrôlé des modifications linguistiques, un tel exercice est plus difficile lorsque l'on utilise des échantillons de cartulaires du XIIe [8] siècle ou des romans du XVIIIe siècle, par exemple, parce que chaque pièce peut présenter des phénomènes linguistiques uniques et des états assez hétérogènes du discours. On peut combler les lacunes des sources modernes en regroupant des textes similaires dans un arc temporel ample ; mais il serait fallacieux d'essayer de faire de même à partir de sources historiques ou de textes comportant des états anciens de la langue, qui de surcroît font normalement partie d'une série dont l'ordre et la disposition sont aussi des sources d'information. La production de modèles d'entraînement automatique, qui a besoin de jeux de textes

7. Voir une discussion autour de la sémiotique du corpus dans : Wagih Azzam et al. "Les manuscrits littéraires français : pour une sémiotique du recueil médiéval". In : *Revue belge de philologie et d'histoire* 83.3 (2005), p. 639-669

8. "Un cartulaire (lat. : c(h)artularium) est un recueil de copies de ses propres documents, établi par une personne physique ou morale, qui, dans un volume ou plus rarement dans un rouleau, transcrit ou fait transcrire intégralement ou parfois en extraits, des titres relatifs à ses biens et à ses droits et des documents concernant son histoire ou son administration". M. Milagros Cárcel Ortí. *Vocabulaire international de la diplomatique*. T. 28. Universitat de València, 1997, p. 35

relativement vastes, est par conséquent limitée aux corpus assez larges ayant une grande cohérence interne. Quiconque travaille avec des sources historiques sait que ce genre de corpus sont rares.

En considérant ces diverses exigences, l'application de l'approche automatique à la recherche historique nécessite par principe deux adaptations inspirées par notre domaine d'intérêt, et dont le développement est un défi majeur :

1. En premier lieu, des outils de traitement linguistique dont l'application soit adaptée aux discours historiques, aux pratiques de l'écrit et aux langues anciennes, autrement dit, aux artefacts écrits qui demeurent dépendants du contexte social et historique ; à défaut, des outils dont la robustesse ait été éprouvée sur la typologie documentaire concernée ;
2. Des bases de données nourries par de longues séries documentaires et lisibles par l'ordinateur (*machine-readables*) à partir desquelles il s'avère possible d'entraîner des modèles afin d'extraire une lecture par-delà une analyse statistique des unités textuelles, ou en tout cas plus significative que celle-ci.

0.1.2 La numérisation des sources historiques

La valeur accordée à la lisibilité d'un document par la machine est significative et de fait c'est un élément qui apporte une différence substantielle dans la performance des outils générés. Numériser un texte le rend interrogeable en tant que chaîne de caractères, mais il est impossible d'en déduire automatiquement des aspects pertinents comme la morphologie, les relations sémantiques, la signification ou le sens des mots. Il faut que ces aspects soient indiqués par un lecteur humain avant de le faire apprendre à la machine. Un travail qui doit aussi s'occuper des caractères externes du document normalement indiqués dans les metadonnées associées, qui doivent aussi être structurées et normalisées. Il existe des approches qui essaient de substituer ce travail humain mais aujourd'hui les meilleures performances sont obtenues à partir des corpus annotés, c'est-à-dire des corpus portant des informations linguistiques supplémentaires codifiées, et suffisamment larges pour pouvoir assurer un niveau acceptable de robustesse devant des données nouvelles, non familières.

De plus en plus de corpus historiques sont disponibles pour servir comme corpus d'entraînement des outils informatiques, mais très peu d'entre eux offrent des données annotées. Cette situation est problématique car jusqu'à présent la plupart de ce qu'a été publié numériquement est une simple reproduction du contenu des éditions imprimées, spécialement celles du siècle dernier. Dans le cas particulier des documents médiévaux, à ce jour, plus d'un demi-million d'entre eux ont été publiés numériquement. Ce qui est un nombre élevé si on considère que la plupart des sources médiévales sont manuscrites et que la technologie de reconnaissance optique n'y est pas encore applicable [9]. Malheureusement, la tendance autrefois en vigueur était de qualifier de "numérique" tout contenu hébergé par un site en ligne, dont le texte était simplement référencé par le

9. Voir à ce sujet les résultats de la dernière compétition pour la classification de l'écriture manuscrite médiévale ICDAR2017 : Florence Cloppet et al. "ICDAR2017 Competition on the Classification of Medieval Handwritings in Latin Script". In : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*. T. 1. IEEE. 2017, p. 1371-1376

code. De ce fait aujourd'hui la plupart des dites "éditions numériques", qui incluent des corpus de tout nature (travaux d'éditions érudite, collections de manuscrits, littérature et en général tout publication de nature textuelle) demeurent largement dépendantes des éditions imprimées - sur papier le plus souvent - et s'ils sont numériquement accessibles, ils ne sont pas *machine-readables*.

La première conséquence en est qu'actuellement une partie assez importante des documents disponibles pour la recherche en histoire, et en sciences sociales en général, est constituée par une masse chaque fois plus considérable de textes "bruts" [10]. Le manque des données structurées empêche de mobiliser les méthodes d'entraînement et d'exploitation les plus performantes. Cette situation a de surcroît généré un grand décalage entre l'information disponible et la capacité de lecture active par la communauté des chercheurs. Ceci est un problème encore plus sérieux dans les domaines avec des productions constantes, comme le journalisme ou les sciences politiques, parce qu'il y a littéralement plus de documents que ce qu'un chercheur peut aborder.

Certes, on a assisté dans la dernière décennie au boom des éditions en langages de marques sous le paradigme de la TEI (*Text Encoding Initiative*) [11]. Elles ont constitué un premier palier, solide dans certains cas, avec pour objectif général de produire des documents enrichis et interrogeables par la machine. De même, des bases de données ont été publiées par les grandes bibliothèques, en privilégiant les formats d'archivage *ad hoc* tels que EAD et MARC [12]. Mais dans un cas comme dans l'autre ce n'est pas normalement le contenu du document mais plutôt ses caractères extérieurs qui ont été l'objet d'une annotation, dont les modèles demeurent très attachés à ceux utilisées par les éditions philologiques savantes. En tant que reproductions numériques des éditions papier, elles demeurent peu exploitables par des méthodes de traitement massif.

La structuration de ces documents, dont l'annotation est un des éléments centraux, est un travail laborieux lorsqu'il est fait à la main. Fournir à ces documents un schéma, grâce à une annotation dans un langage codifié, qui explicite et représente ses caractères internes - par exemple les caractères sémantiques et syntaxiques du texte, comme dans le cas de cette thèse - est un travail complexe, pas seulement parce qu'il nécessite le concours d'une grande expertise mais également parce que, considérant la taille actuelle des corpus de milliers d'items, il ne peut pas être adopté à grande échelle. Un travail de telle nature et de telle ampleur ne peut s'envisager qu'avec l'aide d'un traitement informatisé.

Néanmoins, l'automatisation de la structuration de données se confronte à deux principaux problèmes :

10. Texte brut est une traduction de *raw text*, format de texte qui représente seulement les caractères et ne porte aucune information ou représentation graphique.

11. Un langage de marques (*markup language*) est un terme spécifiquement employé pour indiquer le format, le style et les principales structures d'un texte utilisant des balises. Les langages les plus connus sont XML et HTML. La TEI fait référence à un ensemble de modèles et lignes directrices qui visent à standardiser la description des propriétés et attributs d'un document en utilisant un langage de marques.

12. EAD (*encoded archival description*) et MARC (*machine readable cataloging*) sont deux standards utilisés pour la description d'information bibliographique et archivistique dans des catalogues numériques.

1. Le manque de corpus annotés à la main par des spécialistes, condition *sine qua non* pour pouvoir démarrer un entraînement ayant comme cœur un algorithme qui mettra en relation le corpus avec une série de règles et de patrons qui déterminent les traits textuels à privilégier lors de l'apprentissage ;
2. Le manque de flexibilité des modèles automatiques lorsqu'ils sont appliqués à des corpus provenant de domaines différents de celui d'origine, voire à des corpus formellement similaires mais différents en taille, dans leur chronologie ou tradition textuelle. En effet, une très bonne performance, obtenue à partir d'un modèle entraîné sur un corpus vaste, peut générer un outil très attaché à ses particularités textuelles, et le rendre peu exportable.

Cela nous amène à une situation où l'application des outils automatiques pourrait être doublement contrariée : d'un côté, l'annotation automatique se trouve restreinte par les faibles performances des outils disponibles ; de l'autre, la production d'outils plus performants s'avère compliquée du fait du manque de corpus annotés qui puissent servir comme des bases d'entraînement. À ce blocage s'ajoute un troisième problème plus circonstanciel : les outils produits avec des corpus annotés, même s'ils représentent un avancement significatif dans les niveaux élémentaires d'analyse, ne peuvent pas offrir un résultat complet si on les applique à des corpus présentant des dissemblances par rapport à leur corpus d'origine.

Le traitement automatique des langues peut être ainsi comparé à un effet de levier (*leverage*) où des solutions adaptées sont recherchées en utilisant le peu de corpus annotés disponibles afin de multiplier le profit - ici, la connaissance - que l'on peut tirer des plus vastes corpus en texte "brut". Indubitablement les solutions à toutes ces contraintes ont besoin d'une heuristique très bien définie pour, d'un côté, trouver des mécanismes satisfaisants afin de multiplier le profit des annotations, ce qui nécessite de proposer de multiples arrangements lors de l'entraînement des outils avec l'objectif de modérer l'attachement naturel des outils à leur corpus de base. Et, d'un autre côté, de fournir un environnement d'évaluation qui permette de détecter le minimum d'annotation nécessaire pour produire des outils acceptables, ce que rendra plus agile l'obtention de corpus annotés automatiquement.

Ces deux contraintes seront traitées *in extenso* dans la deuxième partie de cette thèse afin de montrer le processus d'élaboration d'un modèle automatique de structuration présentant une robustesse élevée sur des documents formellement similaires et une haute performance sur des documents proches mais plus hétérogènes. Il s'agira également de montrer que dans un corpus suffisamment large, annoter un ensemble représentatif de documents afin d'entraîner un modèle, peut être suffisant pour avoir en quelques heures une annotation automatique assez acceptable de tout le reste du corpus ainsi que d'autres corpus similaires, dont l'annotation manuelle complète aurait pris des années.

0.1.3 L'opacité de l'algorithme

Dans les travaux se fondant sur un algorithme, on peut considérer qu'une bonne performance, c'est-à-dire une sortie (*output*) automatique d'une qualité proche d'une entrée (*input*) donnée, est un résultat efficient ; mais les manières de générer de la

connaissance dans les sciences sociales peuvent s'en ressentir. Les outils sont finalistes, dans le sens où on peut observer précisément l'étape finale du traitement opéré, alors que le processus pour y arriver et les conditionnements imposés par les méthodes suivies ne sont pas si transparents. Si on ouvrait les modèles d'automatisation, on découvrirait de gigantesques matrices de chiffres, résultats du calcul du poids et des combinaisons de chaque trait et observation analysée. Ce décalage peut provoquer une inadaptation par rapport aux méthodes épistémologiques employées par les humanités, où le processus impliqué pour arriver à un résultat est aussi, voire plus, important que ce dernier. Au rapport technique normalement suffisant pour expliquer un bon résultat, il faut en ajouter un autre, qui inclut depuis une optique historienne, le "pourquoi et comment" de ce résultat, ce qui en plus va nous fournir des éléments heuristiques pour mieux adapter un outil à nos questions et évaluer l'impact ou le biais apporté par les méthodes choisies dans les résultats qu'ils nous offrent.

D'autre part, en suivant le même raisonnement, les outils de Traitement automatique des langues (TAL) entraînés au début à partir de corpus autres que ceux de textes anciens, comme on l'a mentionné plus haut, doivent s'adapter aux types de données fournies par la culture textuelle et graphique du manuscrit. Mais il faut aussi considérer que du côté de la recherche linguistique et statistique la réponse aux questions et les résultats sont proposés suivant les termes de son domaine, c'est-à-dire sous la forme de tableaux, étiquettes, chiffres, modèles, matrices, rappelant ainsi l'énorme besoin, à peine envisagé dans les travaux, d'établir un cadre d'interprétation dédié. On n'insistera jamais trop sur le fait que l'utilisation des outils automatisés n'a de sens que dans un dispositif de recherche qui vise à fournir un accès plus contrôlé et plus rapide à l'information. Les listes de noms récupérés, leur fréquence dans un corpus et le rapport sémantique de leurs interactions sont subordonnés à des questions pertinentes et posées en langage naturel : quelle est la nature de la relation entre un lieu et son entourage ? Quels changements expérimente un paysage et quel est le vocabulaire employé pour le décrire ? Que peut-on savoir de la vie sociale de telle ou telle personne ? En bref, il s'agit de trouver un compromis intellectuel entre, d'un côté, l'étude d'objets ou de représentations d'objets dont le contenu est indissociable de leur contexte social, culturel et sémiotique - catégories d'ailleurs difficiles à quantifier - et d'autre part, les résultats de l'analyse proposée par les nouveaux outils qui s'appuient sur une modélisation de régularités, de stéréotypes et de catégories linguistiques.

De fait, au vu de ce panorama, il n'est pas étonnant que les contraintes générées par ce tournant numérique chez les chercheurs en sciences sociales soient nombreuses. Les mécanismes dits obscurs de l'obtention de résultats à partir d'un traitement algorithmique ont servi à relancer les critiques à propos de la "boîte noire" et du positivisme, spécialement dans sa variante quantitative [13]. En outre, la faible perméabilité au criticisme et à toute question dialectique ont aussi généré du débat.

13. La littérature autour de la transparence des outils numériques dans la génération de la connaissance pour les humanités est abondante ; quelques travaux recensant les points principaux : Johannes Passmann et Asher Boersma. "Unknowing algorithms : On transparency of unopenable black boxes". In : (2017) ; Tal Hassner et al. "Digital Palaeography : New Machines and Old Texts (Dagstuhl Seminar 14302)". In : *Dagstuhl Reports*. T. 4. 7. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2014 ; F Gibbs et T Owens. *Building Better Digital Humanities Tools : Toward broader audiences and user-centered designs. Digital Humanit. Q. 6 (2)(2012)*

En fait, l'utilisation même de ces outils et des documents numériques a éveillé, et à juste titre, la suspicion parmi les chercheurs en humanités parce qu'ils travaillent dans un environnement numérique où les résultats obtenus peuvent être reproductibles, transférables et apparaissent univoques. 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 ces nouvelles pratiques.

Néanmoins, et comme on le verra dans cette thèse, cette problématique à multiples facettes en présente deux qu'il faut affronter urgemment. D'un côté le compromis invoqué dans les dernières années qui fait appel à l'utilité de ces techniques comme un coadjuvant de la recherche et en tout cas comme une manière de formaliser l'ensemble de la phénoménologie textuelle avant d'être une méthode de réflexion et de production de la connaissance en elle-même. Et d'un autre côté, la proposition et le développement d'une herméneutique robuste qui puisse offrir un cadre d'intégration des résultats offerts par l'algorithmique dans le système de validation scientifique des humanités. Après tout, l'usage de ces techniques permet de recourir à des moyens puissants afin d'obtenir des preuves et d'explorer des pistes de recherche solides, en enquêtant sur des quantités surhumaines d'information afin de mieux envisager une certaine hypothèse qui reste quant à elle attachée au domaine des sciences sociales.

0.2 Tâches de recherche

0.2.1 Les entités nommées

Comme on l'a mentionné, un texte non structuré peut difficilement être l'objet d'une analyse automatique ou semi-automatique dans le but d'identifier des nouvelles données et il faut le doter d'une représentation logique des différents niveaux d'information qu'il contient. Cette structuration des textes peut connaître différents degrés de finesse coïncidant avec les niveaux d'analyse grammaticale. Ainsi nous pouvons appliquer des outils sur l'extraction des données morphologiques (*stemming, lemmatisation*) ou syntaxiques (*pos-tagging, chunking*), comme pour déterminer la forme, l'ordre et la constitution d'une phrase, étape incontournable pour aborder des analyses plus complexes dans les niveaux syntaxiques (*word sens disambiguation, entity linking*) qui peuvent présenter les mots classés selon leur pertinence informative ou les relations de dépendance entre les mots d'une phrase. De même, des approches plus récentes s'intéressent à des questions relatives à la pragmatique comme le sens contextuel d'un mot ou l'analyse de l'opinion. Dans cette thèse on va retenir deux de ces approches qui ont montré leur efficience pour accélérer la récupération de données scientifiques dans de larges bases de données : la reconnaissance des entités nommées et la reconnaissance des parties du discours.

Tâche incontournable dans les systèmes modernes de récupération de l'information, la reconnaissance des entités nommées est un des principaux éléments de la structuration des textes. Elle permet de repérer tous les éléments d'un texte qui font référence à des éléments physiques et réels et de les classer selon un index de catégories. La détection des noms de personnes, de lieux, d'institutions et de manière plus large, de

dates, noms des produits ou même des maladies est une opération essentielle (*sub-task*) et un enjeu spécifique dans les analyses grammaticales automatiques car les entités nommées ne font habituellement pas partie du vocabulaire particulier d'une langue, elles ne sont donc pas récupérées par les outils d'analyse générale et il est nécessaire de les reconnaître dans un processus séparé.

La difficulté à les reconnaître fait en même temps leur intérêt car, en raison de leur spécificité, les entités nommées agissent comme des références absolues dans un texte, autrement dit, comme des éléments du langage qui font référence à une entité unique et concrète. Le sens et la signification d'un texte convergent vers le réseau des entités nommées qu'il contient car elles se trouvent normalement au centre de l'expression linguistique. En fait, dans le domaine qui nous concerne elles se rapportent directement aux questions élémentaires de la recherche historique : qui, où et quand. De plus, lorsqu'une entité nommée est mentionnée dans plusieurs documents elle devient un point de connexion avec l'information référentielle provenant d'autres séries d'un même corpus, voire d'autres corpus. C'est pourquoi l'identification et le classement de ces unités d'information s'avèrent des étapes incontournables pour comprendre et indexer un document et ouvrir des canaux dynamiques afin de mieux interroger une base de données.

Comme on l'a évoqué plus haut, la reconnaissance des entités nommées est une tâche de recherche du traitement automatique des langues (TAL), domaine dont l'objectif est d'améliorer les performances des ordinateurs - programmés dans un langage artificiel - dans l'exploitation des grandes quantités de texte en langage naturel. De ce domaine font partie plusieurs des nouvelles techniques qui ont commencé à être intégrées à l'étude des humanités comme la reconnaissance des caractères (OCR [14]), la lemmatisation et l'étiquetage morpho-syntaxique, comme d'autres, plus complexes, comme la désambiguïsation lexicale ou le plongement de mots (*word embeddings*).

La base théorique de la plupart de ces techniques se trouve dans le distributionalisme américain qui propose de concevoir le texte comme une série de mots interconnectés. Dans ce courant, l'extraction et l'analyse des co-occurrences, des "sacs de mots" (*word-bag* [15]), de la distance entre les mots et, en général, la mesure de toutes les relations de type contextuel sont privilégiées. Ce qui permet avec une certaine facilité de fournir une vision générale des réalités sous-jacentes et des relations dans les séries d'éléments qui constituent un texte. Dans ce panorama, la contribution décisive des entités nommées est l'identification d'éléments ayant une signification spécifique qui opèrent comme sujets ou objets des actions. Ainsi, une fois qu'une entité est identifiée, elle peut constituer, selon l'approche utilisée, un nœud dans un dense réseau spatial et social ; un point de connexion entre différents documents ou collections de documents et enfin, un élément indexé à l'intérieur d'une masse textuelle peu différenciée pour l'ordinateur.

14. OCR (*Optical character recognition*), définit l'ensemble de techniques informatiques destinées à transformer les fichiers d'image d'un texte en un texte brut.

15. Le *sac de mots* est un modèle de représentation d'un texte comme l'ensemble des mots qui le composent sans soucis de propriétés ou contexte (ordre, fonction), autrement dit, c'est un vecteur de la fréquence d'apparition des mots.

0.2.2 La détection des parties du discours diplomatique

De son côté, l'analyse du discours n'a pas attiré la même attention dans les études du traitement automatique des langues, parce que l'on dispose d'une quantité assez limitée de textes annotés avec cette information. De manière succincte, la reconnaissance des parties du discours détecte la séquence d'énoncés ou d'actes scripturaires utilisés ou mis en place lors de la rédaction d'un document. Chaque domaine a développé des modèles scripturaires particuliers qui déterminent la configuration de pratiques sociales et intellectuelles pour transmettre de l'information. Ainsi, la découverte des relations sémantiques établies entre les différentes séquences d'objets et d'énoncés formant un modèle permet de définir la caractérisation d'un style scripturaire aux fondements de la définition des traditions et des typologies scripturaires. En tant que pratiques sociales et intellectuelles, la récupération et la comparaison automatique des séquences peut aider la recherche historique concernant certains événements dont l'élucidation est complexe, comme la circulation des modèles scripturaires et des formules, les modes et les changements dans le vocabulaire ou l'évolution des pratiques de l'écrit [16].

L'élaboration d'un outil capable de récupérer les parties du discours acquiert une pertinence particulière dans le cas des actes écrits où est consigné l'accomplissement d'un acte de nature juridique ou administrative parce qu'ils sont normalement rédigés suivant un formulaire. Malgré leur stéréotypie, ces structures peuvent subir un nombre élevé de variations, mais, en général, il existe un modèle récurrent sous la forme classique pour les actes du Moyen Âge de la succession de trois parties : *protocole, texte* et *eschatocole* [17]. À l'intérieur de cette structure tripartite peuvent se dérouler de multiples séquences d'énoncés destinées à donner une teneur et une validité au document à travers l'adaptation des diverses typologies de modèles dont chaque détail, comme la manière de dater l'acte, la position d'une citation biblique, le choix d'une certaine formule ou la disposition des clauses, peut être l'objet d'observations parce qu'ils sont révélateurs des positions idéologiques ou politiques prises par l'auteur et plus généralement par l'institution rédactrice.

Les approches destinées à détecter les parties du discours ne se montrent pas plus complexes que celles destinées à récupérer les entités nommées, et, en fait, on peut se servir dans les deux cas des mêmes outils de base et des mêmes algorithmes. Mais l'heuristique déployée dans la détection des entités nommées peut être assez distante de celle employée pour les parties du discours car elle part d'une approche au niveau de la phrase (*phrase-level*) et non au niveau du mot (*word-level*). Le nombre d'étiquettes à évaluer lors d'un entraînement automatique peut être sensiblement plus large dans le cas de l'analyse des parties du discours. Il faut développer des stratégies nouvelles qui

16. Le concept de pratiques de l'écrit comprend l'étude des pratiques et usages autour de l'écriture et de l'écrit pendant le Moyen Age, ce qui a à voir avec la chaîne de l'élaboration d'un document, sa conservation et son archivage mais aussi avec la fonction du document écrit en tant qu'élément d'usage juridique, politique et social dans une société de rapports fondamentalement oraux. Voir : Étienne Anheim et Pierre Chastang. "Les pratiques de l'écrit dans les sociétés médiévales (VIe-XIIIe siècle)". In : *Médiévales. Langues, Textes, Histoire* 56 (2009), p. 5-10

17. La teneur d'un acte juridique se ramène à ces trois groupes qui correspondent au cadre formel initial (protocole), à la matière centrale (texte) et au cadre final (eschatocole). Voir : Ortí, *Vocabulaire international de la diplomatique*, p. 179, 188

prennent en considération la structure de larges chaînes de mots ou qui analysent la relation entre ensembles de mots assez distants, multipliant ainsi la possibilité d'avoir une erreur dans la détection. En fait, certains éléments, comme la rigidité des structures discursives, l'utilisation d'un vocabulaire dont l'évolution est lente, et la répétition des formules et des clauses, peuvent présenter un atout au moment d'entraîner un outil à cette finalité. Cela permet à l'algorithme de faire face à un nombre plus réduit de phénomènes et de généraliser la prédiction de certains éléments. Mais, à l'opposé cela peut également conduire à une incapacité à reconnaître des changements parfois très discrets (sur généralisation) que peut subir un acte dont les détails, comme on l'a déjà dit, constituent cependant une information précieuse.

0.3 CBMA, CDLM et corpus structurés

La réponse de la communauté des humanistes numériques au manque de corpus structurés a été diverse ; elle s'est surtout concentrée sur le développement d'outils utilisant des approches symboliques ou des méthodes non supervisées, c'est-à-dire d'outils qui n'ont pas besoin de corpus annotés. Mais dans certains cas les chercheurs ont aussi entamé la difficile mission d'annoter de larges corpus qui pourraient servir de base à ces traitements. À l'heure actuelle on ne compte pas plus d'une demi-douzaine de corpus historiques annotés, dont l'un des plus larges et des mieux organisé est le *Corpus Burgundiae Medii Aevi* (CBMA) qui compte vingt-neuf mille actes, soit la quasi-totalité de la production diplomatique connue de la Bourgogne médiévale, spécialement entre le Xe et le XIIIe siècle [18]. Il comprend notamment l'un des corpus les plus importants de la chrétienté occidentale, celui des chartes de l'abbaye de Cluny [19]. Ce recueil d'actes privés connu dans l'historiographie médiévale compte environ cinq mille cinq cent chartes, dont la majorité, environ quatre mille cinq cents, a été annotée à la main par le groupe de travail du CBMA. Dans l'annotation sont signalées et classées les entités personnelles et spatiales présentes dans le texte, et la totalité du corpus est éditée dans un format tabulaire qui précise aussi différentes informations-clé de chaque document, comme le type d'acte, la date, l'origine, l'affaire conclue, etc., soit un total de 56 paramètres de description.

La qualité du corpus CBMA, plus les corrections et ajouts que nous avons opéré, nous a permis de tester différentes solutions afin de combler les principales lacunes rencontrées dans l'entraînement d'outils sur d'autres corpus historiques. En particulier, la taille et l'extension chronologique de cet ensemble de chartes annotées a été suffisant pour nous permettre de générer un modèle de reconnaissance automatique des entités nommées entraîné sur un répertoire complet de phénomènes scripturaires et de situations linguistiques. Dans le modèle le plus performant, le taux de reconnaissance s'avère très élevé (au-delà de 95 %) sur l'ensemble des documents bourguignons. La variété et la cohérence interne du corpus nous a aussi permis de varier les échelles

18. Eliana Magnani. "Un corpus structuré et hétérogène de textes latins médiévaux (Bourgogne, Ve-XVe siècle)". In : *Bulletin du CERCOR-Centre Européen de recherches sur les congrégations et ordres religieux* 41 (2017), p. 59-65

19. BRUEL, Alexandre ; BERNARD, Auguste Joseph. Recueil des chartes de l'abbaye de Cluny : 802-954. Impr. Nat., 1876.

et les paramètres d'entraînement en formant des sous-corpus de différentes tailles, chronologies et origines afin de mieux observer la robustesse de l'outil dans différents scenarii sans trop diminuer le taux de reconnaissance. De même, nous avons déployé un dispositif d'évaluation afin de tester les outils générés sur des corpus étrangers à la Bourgogne, une opération obligatoire pour déterminer la flexibilité de l'outil.

Certaines spécificités du corpus clunisien ont été un atout pour la modélisation, à savoir :

1. les structures assez régulières promues par les formulaires lors de la rédaction des chartes ;
2. la stabilité conceptuelle d'un vocabulaire restreint et assez précis dans la définition de l'espace et des relations sociales ;
3. et une relative stéréotypie dans les étapes de la formation des noms de personnes et des lieux.

Mais ces spécificités, alors même qu'elles facilitent la modélisation, peuvent aussi poser de sérieux problèmes méthodologiques quant à la possibilité d'étendre le modèle à d'autres cartulaires. En particulier, le risque du surentraînement (*overfitting*) lié à des structures répétitives, ce qui donnerait un modèle peu robuste car très enclin à homogénéiser les résultats ; mais aussi, la présence d'un vocabulaire relativement limité auquel une institution est parfois très attachée sans qu'il soit nécessairement le même vocabulaire utilisé ailleurs ; et enfin, l'existence d'états de langue et de discours qui se montrent souvent très spécifiques d'une région, d'une époque ou d'une institution, ce qui est particulièrement problématique dans le cas du latin médiéval car la variabilité peut faire rapidement tomber les outils de base travaillant avec la lemmatisation et l'extraction des informations morpho-syntaxiques.

Concernant l'élaboration d'un dispositif de reconnaissance des parties du discours diplomatique, un deuxième corpus a été utilisé : le *Codice diplomatico della Lombardia Medievale* (CDLM). Ce corpus est un recueil de chartes provenant principalement des scribes et greffiers lombards monastiques, complété avec certains registres de la chancellerie civile. Achevé en 2006, ce projet est la suite et la réédition en format numérique des éditions du XIXe siècle dont le *Codex Diplomaticus Longobardiae*[20] constitue un des modèles. En tant qu'édition numérique, les 5200 documents qui forment l'ensemble du recueil sont les témoins d'un moment de transition entre les premières éditions critiques en format numérique et les éditions en langage de marques. Elle adopte le format XML tout en se conformant aux règles d'édition de la diplomatique traditionnelle, suivant ainsi la conviction, très actuelle, que les éditions des textes historiques doivent dépasser la simple reproduction des objets textuels. Cette édition XML est lourde pour les systèmes de traitement massif car elle superpose sur un même mot ou une même phrase plusieurs étiquettes pour répondre à un balisage des caractères externes et internes du texte, et ce en suivant le standard TEI, très répandu dans les éditions critiques numériques. De ce fait, elle a besoin pour son exploitation à grande échelle de plusieurs adaptations et nettoyages. Tout cela est cependant compensé par la qualité de l'annotation de certains aspects, absents des autres corpus,

20. Giulio Porro. *Codex diplomaticus Longobardiae, Augustae Taurinorum.* Regio Typographeo, 1873

et dont l'accomplissement a nécessité de longs efforts et une grande expertise. En réalité, il s'agit du seul corpus médiéval annoté qui procure une information concernant les parties du discours diplomatique.

Le CDLM présente des lignes générales différentes de celles du CBMA. Alors que le CBMA concentre pour l'essentiel des documents des Xe et XIe siècles, le CDML présente une accumulation croissante à partir de la deuxième moitié du XIe, qui culmine à la deuxième moitié du XIIe siècle : 45 % des chartes datent de cette période. En outre, les séries documentaires qu'il contient, documents produits par une même institution, ne sont pas très vastes. Provenant de plusieurs archives locales, ces séries ont été rassemblées dans l'intérêt de former une collection, autrement dit, un recueil d'éditeur contenant les documents produits dans une zone géographique déterminée. Le projet du CBMA est similaire mais il inclut des cartulaires, des recueils originaux, et des recueils factices assez vastes produits en grande partie par une même institution. Ainsi, alors que le CBMA inclut des séries comprenant des milliers de documents, plus de la moitié des séries reliées par le CDLM est constituée par quelques dizaines de documents seulement, et ses plus longues séries ne dépassent pas de 300 actes.

Sur le plan technique, cela ne signifie pas qu'un outil formé avec ce corpus devra nécessairement faire face à un niveau d'hétérogénéité plus élevé. Au contraire, compte tenu de la faiblesse de certaines séries et de la forte concentration sur une courte période, la variabilité pourrait avoir un impact limité sur les façons de prédire les parties du discours diplomatique. La vraie hétérogénéité se trouve en réalité à l'intérieur des séries qui peuvent présenter des documents assez différents dans leur composition formelle. Il est compliqué de détailler les décisions prises par l'algorithme mais en observant le modèle qui en résulte, on constate que la performance est médiocre sur certaines parties qui ne sont utilisées que dans quelques séries ou dans des chartes très formalisées dont le nombre est limité. Par contre, la performance est très bonne (supérieure à 85 %) dans la récupération des parties qui composent le modèle diplomatique "standard", si on peut l'appeler ainsi, et qui, par conséquent, est mieux représenté dans la plupart des séries (ce qui inclut par ex. invocations, notifications, adresses, suscriptions, dispositifs, dates).

0.4 Organisation de la thèse

Cette thèse est organisée en trois parties :

La première partie est divisée dans deux chapitres. Dans le premier chapitre on exposera une révision détaillée de la littérature scientifique autour des outils de traitement automatique de la langue, et spécifiquement de ceux développés et appliqués sur des corpus historiques et focalisés sur les entités nommées. Une vue d'ensemble de leurs atouts et des verrous scientifiques y sera proposée. Dans le deuxième chapitre on donnera un aperçu des principales approches techniques mobilisées dans la recherche appliquée aux corpus historiques et on fera une révision des autres techniques de traitement automatique des langues qui seront aussi utilisées dans notre étude du corpus.

La deuxième partie se trouve divisée en deux chapitres. Dans le premier chapitre

sont abordés des enjeux concernant l'organisation du corpus, en particulier sa représentativité par rapport à une réalité scripturale beaucoup plus large, et la question de la surreprésentation de certains styles et formes documentaires étant donné que la majorité du corpus a été recueillie par une seule institution. Puis, on s'intéressera à d'autres enjeux, concernant la pertinence de l'approche utilisée lors de l'entraînement tout en mesurant l'impact du déterminisme technique de l'outil sur les résultats ; et d'autres moins techniques qu'on peut regrouper dans une "boîte de solutions heuristiques". Dans le deuxième chapitre, nous nous occuperons du développement technique des deux modèles automatiques proposés dans cette thèse. Y seront expliquées toutes les questions relatives aux principes du traitement automatique du langage, les algorithmes et les solutions techniques. Dans ce chapitre seront également présentées les corrections et ajouts que nous avons opérés sur les corpus annotés comme les résultats en termes de robustesse de notre modèle suite à différentes étapes d'évaluation des outils.

La troisième partie comprend trois études de nature historique qui correspondent à la mise en œuvre du dispositif d'étude à propos des chartes bourguignonnes. La base de ce dispositif sont les textes enrichis avec les entités nommées et parties du discours, tant ceux déjà annotés par le groupe du CBMA que d'autres annotés automatiquement avec les outils développés ici, réunissant un total de neuf mille actes, ce qui inclut l'intégralité des actes clunisiens et de dix autres cartulaires, spécialement de ceux qui proviennent du *pagus* de Mâcon et de ses alentours et qui contiennent des actes datés des IXe et XIIe siècles. Les études proposées sur cet ensemble documentaire visent à mettre en œuvre différentes solutions heuristiques et à développer un cadre interprétatif qui puissent servir comme méthodologie de travail intégrant l'algorithmique dans l'étude en profondeur des corpus historiques. Ces trois études visent à développer certaines solutions pour obtenir rapidement des observations générales à partir des vastes collections documentaires. Chacune privilégie la mobilisation de l'un des trois caractères extraits automatiquement dans ce travail : les noms de personnes, les parties du discours et les noms géographiques.

1. Dans *la première étude*, nous avons entamé la datation semi-automatique d'un cartulaire quasiment dépourvu d'indications chronologiques, celui du monastère de Paray-le-Monial, ancien prieuré de Cluny. À cet effet nous avons développé une méthode qui génère des matrices de datation sur la base des noms de personnes qui apparaissent dans les actes de mutation foncière. Le principe qui sous-tend cette méthode est d'utiliser la mention de ces personnes dans différents cartulaires de la région pour essayer de récupérer toutes les données de nature chronologique qui nous permettant de dater un document ou en tout cas de proposer une date à fourchette serrée.
2. Dans la *deuxième étude*, nous proposons une méthode de récupération et de classement automatique des formules qui intègrent les parties du discours des actes du Recueil de Cluny. A partir des résultats de notre modèle de reconnaissance des parties du discours et en privilégiant une analyse centrée sur les protocoles, notamment sur les invocations, nous essayons différentes solutions techniques pour définir des formulations courantes et pour identifier les variations par rapport à celles-ci. Puis, nous enquêtons sur le rapport

existant entre la mobilisation d'une certaine formule ou d'une de ses variantes et l'influence de facteurs autres comme la qualité des personnages, le lieu de rédaction de l'acte, le type d'affaire conclu, la tradition scripturaire, etc. Finalement proposons d'offrir une vision globale, statistique, de toutes les solutions rédactionnelles utilisées par les scribes autour des parties du discours mobilisées dans les protocoles des actes.

3. La *troisième étude* présente une recherche autour des termes de description de l'espace, notamment dans le *pagus* de Mâcon, utilisés par les scribes dans les formules de description et d'inventaire fonciers. Ici nous analysons les entités géographiques et leurs co-occurrences qui constituent le vocabulaire couramment utilisé dans la description topo-spatiale. Nous proposons une vision sémantique, et chrono-spatiale autour des termes les plus mobilisés et des relations établies entre eux à l'intérieur des formules de localisation. Ceci est complété par quelques reconstructions cartographiques qui s'avèrent importantes pour bien comprendre les mutations des cadres territoriaux et des systèmes de découpage du paysage.

Chapitre 1

État de L'art

1.1 La reconnaissance et identification des entités nommées.

La reconnaissance des entités nommés est devenue dans les dernières années une tâche capitale dans la création des textes structurés permettant d'entamer des recherches dans le domaine du traitement automatique des langues et plus généralement dans la récupération de l'information. De ce fait, la plupart des applications visant des traitements complexes comme l'extraction de l'information, la découverte de connaissances ou l'analyse sémantique, incorporent dans leurs bases des modules qui assurent une bonne reconnaissance des entités nommées. Dans la littérature cette tâche de recherche est traditionnellement divisée en trois sous-tâches bien identifiées :

1. Le repérage des entités susceptibles d'être identifiées comme nommées, normalement des substantifs, et très souvent des noms propres ;
2. La classification des entités récupérées d'accord a des catégories prédéfinies et selon la réalité conceptuelle plus proche à laquelle elles font référence.
3. La désambiguïsation de l'entité (ou normalisation, ou encore liage référentiel) autrement dit, la récupération de ses informations identificatoires, ce qui se fait normalement en connexion avec des bases de données externes.

Mais, que sont exactement les entités nommées ? Leur définition est de plus en plus complexe au fur et à mesure que l'analyse massive de données a été insérée comme un élément fondamental de la recherche dans différentes disciplines. Dans les années 1980, dans le sillage des premières avancées en apprentissage automatique, différents programmes de recherche ont été favorisés dans le but de faire comprendre des textes à la machine. Notamment, le département américain de la Défense a financé une série de conférences (*Message Understanding Conference* ou MUC) visant à évaluer les techniques d'extraction du sens dans les messages militaires [21]. Rapidement, ces

21. Ralph Grishman et Beth Sundheim. "Message Understanding Conference-6". In : *Proceedings of the 16th conference on Computational linguistics* -. 1996 ; Damien Nouvel et al. *Named Entities for Computational Linguistics*. 2016

recherches ont été axées sur la détection de blocs de texte plus importants que d'autres, puis sur des entités dans le texte qui semblaient essentiels pour comprendre le message puisqu'elles canalisent toute la signification du texte. Ce sont ces travaux qui ont lancé les premières lignes directrices pour la reconnaissance des entités nommées.

Au début, les entités nommées définissaient des objets réels dont la catégorisation coïncidait généralement avec des noms propres. L'article fondateur [22] proposait ainsi une double catégorisation des entités nommées : ENAMEX (*entity name expression*, comme les noms de personnes, de lieux et d'organisations) et NUMEX *(numeric expressions*, comme les dates, quantités, pourcentages) auxquelles s'est ajouté après TIMEX (*time expressions*), formant la triade traditionnelle qui couvre la plupart des mentions nominales.

NUMEX et TIMEX sont encore l'objet de travaux dans les langues peu dotées mais il s'agit d'une tâche bien moins complexe que ENAMEX [23]. Étant donné que leurs entités sont fortement liées à un nombre restreint de mots et de symboles (numéros, mois, saisons, symboles de monnaie, etc.) elles ont pu être rapidement récupérées avec des patrons relativement simples et des approches basées sur de règles faites à la main (*hand-crafted rules*). En revanche, ENAMEX, puisque ses entités présentent des réalisations beaucoup plus irrégulières et des associations moins fréquentes, présente des défis dont le traitement a mobilisé des approches statistiques et basés sur l'apprentissage automatique (*machine-learning based*, voir section 3.4).

Stricto sensu, ENAMEX sont les seules entités qui doivent être considérés comme nommées et elles suffisent normalement pour la plupart des études [24], mais plusieurs débats ont remis en cause cette affirmation. Si ENAMEX peut suffire à la plupart des recherches, par exemple en sciences sociales, elle peut se montrer inefficace pour saisir le sens d'un texte lorsqu'il s'agit par exemple d'un rapport médical, d'un essai chimique ou d'un texte philosophique, textes avec une très basse incidence de noms de personnes, lieux et institutions. ENAMEX ne peut pas non plus collecter d'autres entités dont la nature est sans aucun doute nommée comme les acronymes, les abréviations, les coréférences, les métonymies. Cela a suscité à juste titre des classifications très fines des entités de la part de disciplines très diverses ouvrant ainsi à une multitude de sous-catégorisassions : pour les noms de villes, de régions, de pays, de métiers des personnes, d'entreprises, d'institutions juridiques, politiques, etc. [25]. De nouvelles catégories, en dehors la triade traditionnelle, ont ainsi été proposées afin de prendre en compte des instances laissées de côté comme les noms de maladies, marques enregistrées, molécules, processus physiques, œuvres d'art, etc. [26] arrivant

22. Grishman et Sundheim, "Message Understanding Conference-6"

23. David D Palmer et David S Day. "A statistical profile of the named entity task". In : *Fifth Conference on Applied Natural Language Processing*. 1997

24. Vikas Yadav et Steven Bethard. "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models". In : *Proceedings of the 27th International Conference on Computational Linguistics*. (2018), p. 2145-2158

25. Claudio Giuliano. "Fine-grained classification of named entities exploiting latent semantic kernels". In : *Proceedings of the Thirteenth Conference on Computational Natural Language Learning - CoNLL '09*. 2009

26. David Nadeau et Satoshi Sekine. "A survey of named entity recognition and classification". In : *Benjamins Current Topics*. 2009, p. 3-28

dans certaines études à systèmes proposant des centaines de catégories [27].

De ce fait, l'ouverture vers une grande diversité de cas à prendre en compte a contribué à assouplir les limites initiales du terme entité nommée, mais a aussi créé une généralisation excessive du concept. Formant différentes couches d'un concept extensible, les entités désignent les noms propres mais également toutes les instances catégorielles qui signalent un réfèrent pour incorporer, dans certains cas, la notion très ample des substantifs communs. Devant cette diversité, les disciplines traitant techniquement avec des entités nommées ont établi des définitions sélectives, afin de favoriser la recherche de certains traits considérés comme plus pertinents pour chacune d'entre elles.

Dans cette pléthore de définitions, privilégions trois d'entre elles qui proposent respectivement une optique linguistique, informatique et philosophique, :

— "*Although there is no standard definition we can say that NEs are particular types of lexical units which refer to an entity of the real world in certain specific domains, including human, social, political, economic or geographical, and have a name (typically a proper name or an acronym)*" [28]

— "*Entité nommée est la notion utilisée en TAL pour désigner les éléments discursifs monoréférentiels qui coïncident en partie avec les noms propres et qui suivent des patrons syntaxiques déterminés*" [29]

— "*The word 'Named' aims to restrict [Entities] to only those entities for which one or many rigid designators, as defined by S. Kripke* [30]*, stands for the referent (Kripke : "a designator* d *of an object* x *is rigid if it designates x with respect to all possible worlds where x exists, and never designates an object other than x with respect to any possible world*" [31]

Particular types, éléments monoréferentiels et *rigids designators* font, tous les trois, référence à une qualité exclusive des entités nommées : leur fonction comme référence unique pour un objet déterminé et comme instance particulière d'une classe d'objets. Cette indépendance et cette contingence lexicale des entités nommées sont précisément ce qui rend impossible de les répertorier dans des dictionnaires des formes d'une langue et par extension ce qui rend impossible de les récupérer automatiquement avec les outils généralistes qui travaillent à des niveaux morphosyntaxiques [32]. C'est cette caractéristique qui a fait gagner en popularité les outils de reconnaissance des entités nommées ou NER (*named entities recognition*) comme complément indispensable des analyses automatiques de textes. Puisque les entités sont des éléments fondamentaux pour la compréhension d'un texte, si leur détection n'est pas assurée il devient alors assez compliqué d'arriver à des résultats précis dans la récupération automatique de l'information.

27. Satoshi Sekine et Chikashi Nobata. "Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy". In : *LREC* (2004), p. 1977-1980.

28. Définition des entités nommées dans le ESTER1, MEU04

29. Montserrat Rangel Vicente. "La glose comme outil de désambiguïsation référentielle des noms propres purs". In : *Corela. Cognition, représentation, langage* HS-2 (2005)

30. Saul Kripke. "Identity and necessity". In : *Perspectives in the Philosophy of Language* (1971), p. 93-126

31. Nadeau et Sekine, "A survey of named entity recognition and classification"

32. Eszter Simon. "Approaches to hungarian named entity recognition". In : (2013)

La possibilité de récupérer les entités nommées à partir de répertoires de formes comme les dictionnaires onomastiques, index géographiques, ou les collections de patronymes, a été aussi très explorée mais elle se montre problématique parce que dans certaines langues ou états de langue une même entité nommée peut apparaître écrite de plusieurs manières à cause de la déclinaison, de la fusion graphique, des erreurs d'écriture, de l'adaptation d'une écriture phonétique, etc. [33] D'autre part, étant donné que les dictionnaires ne sont pas exhaustifs, les formes non répertoriées ne sont pas reconnues et un très grand nombre de règles et d'exceptions est nécessaire pour adapter cette reconnaissance à d'autres corpus. D'autres phénomènes récurrents parmi les entités nommées comme l'imbrication, le chevauchement, la coréférence et la métonymie sont autant d'obstacles à l'exhaustivité de leur reconnaissance. Enfin, comme on verra (voir partie 3.4), et spécialement devant des textes avec une plus grande variabilité, il est beaucoup plus efficient de développer un système capable de reconnaître une entité par sa position syntaxique que pour sa morphologie, dont la variété peut être très large, même si cet exercice nécessite la mobilisation de ressources plus complexes.

		LOC			ORG		
PERS					PERS		LOC
Hugonis	de	Breza	donat	monasterio	Sancto	Petro	Cluniacensis
NAM	PRE	NAM	VBE	SUB	QLF	NAM	NAM

TABLE 1.1 – Entités ENAMEX observées dans des differents niveaux d'imbrication. Dans l'exemple l'entité personnelle (*Hugo de Breza*) inclut un toponyme comme deuxième partie du nom (*Breza*) ; de même on peut considérer que le monastère de Cluny agisse comme une entité juridique dont le nom est composé par le nom d'un saint (*Petrus*) et un toponyme (*Cluniacus*).

1.2 La reconnaissance des entités nommées et leur classification dans les sciences sociales.

1.2.1 Les bibliothèques

Dans les sciences sociales, les travaux appliquant la REN (reconnaissance d'entités nommées) se sont multipliés à la même vitesse que les bases de données de textes en format numérique et ont permis de produire une bonne littérature autour de la question. Les bibliothèques, premiers producteurs de textes numérisés, commencent a y trouver une forme accessible et peu onéreuse pour ajouter des méta-données

33. Rodrigo AGERRI et German RIGAU. "Robust multilingual Named Entity Recognition with shallow semi-supervised features". In : *Artif. Intell.* 238 (2016), p. 63-82

concernant le contenu d'un document [34]. Cela a permis d'enrichir leurs collections avec des liens vers d'autres bases de données, afin de rendre plus facile la recherche d'un document à partir de quelques données qu'il contient et pas seulement par ce qui l'identifie. Ainsi différentes collections contenant des référents uniques comme des listes de personnes, d'auteurs, de centres de production, etc. sont normalement inclues dans les descriptions des objets documentaires sous les formats RDF ou OWL afin que ces données puissent être partagées et réutilisées par différentes applications, ce qui constitue l'un des principes du Web (ou toile) sémantique [35].

1.2.2 Les journaux

À partir du partenariat entre bibliothèques et chercheurs ont vu le jour dans la dernière décennie d'autres travaux touchant les sciences sociales. Lorsque la numérisation des grands corpus de revues et surtout de journaux anciens a été mis en place, le nombre de travaux a redoublé. De fait, de nombreux travaux ont été développés en se servant de ces corpus ; initialement en anglais, bien qu'on ait observé dans les dernières années une augmentation des corpus contenant des textes dans d'autres langues comme l'espagnol [36], l'allemand [37], le français [38] ou l'arabe [39] qui portent des contraintes linguistiques spécifiques. Les journaux anciens sont des sources historiques susceptibles d'être numérisées. Ils sont facilement disponibles, offrent un portrait quotidien de la réalité et physiquement les typographies modernes d'imprimerie sont acceptées par les outils actuels d'OCR, et même si l'on peut rencontrer un taux élevé d'erreurs dans la reconnaissance optique [40], la littérature propose déjà des guides de numérisation qui peuvent compenser en grande partie ce problème [41].

Techniquement, les travaux sur les journaux sont une porte d'entrée pour le travail portant sur les discours historiques. Ils présentent un état de la langue modérément anachronique, une diversité conceptuelle plus élevée que la langue moderne et nous montrent un nombre important de problèmes liés à leur numérisation [42], d'où on

34. Max De Wilde. "Semantic enrichment of a multilingual archive with linked open data". In : *Digital Humanities Quarterly* (2017)

35. Tobias Blanke et Conny Kristel. "Integrating Holocaust Research". In : *International Journal of Humanities and Arts Computing* 7.1-2 (2013), p. 41-57

36. Xavier Carreras et al. "Named entity recognition for Catalan using Spanish resources". In : *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - EACL '03*. 2003

37. Manaal Faruqui et al. "Training and Evaluating a German Named Entity Recognizer with Semantic Generalization". In : *KONVENS* (2010), p. 129-133

38. Andoni Azpeitia et al. "NERC-fr : Supervised Named Entity Recognition for French". In : *Lecture Notes in Computer Science*. 2014, p. 158-165

39. Ali Elsebai et Farid Meziane. "Extracting person names from Arabic newspapers". In : *2011 International Conference on Innovations in Information Technology*. 2011

40. Mika Koistinen et al. "How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine". In : *Proc. of LTC* (2017), p. 279-283

41. Chirag Patel et al. "Optical Character Recognition by Open source OCR Tool Tesseract : A Case Study". In : *Int. J. Comput. Appl. Technol.* 55.10 (2012), p. 50-56

42. Il est souvent évoqué le concept de *dirty OCR* qui fait référence à la sortie imparfaite des logiciels OCR sur des textes, normalement anciens (avant 1950), présentant des multiples défauts

peut dégager certaines assertions utiles pour des analyses sur d'autres documents, notamment :

1. Les modèles obtiennent en général un taux plus élevé de reconnaissance pour les lieux que pour les personnes, en raison d'une moindre diversité contextuelle des premières [43].
2. Normalement, les sources les plus anciennes (les plus anciens journaux datant de la fin du XVIIIe siècle) produisent des résultats moins bons en raison d'une plus grande variation dans l'écriture des noms, de mauvaises conditions physiques de conservation du papier et de la nécessité d'utiliser des sources de connaissances plus complexes pour clarifier la signification de nombreux termes [44].
3. Une partie importante des opérations dans le pré-traitement des sources consiste à normaliser les textes après l'OCR afin d'éliminer les erreurs de reconnaissance et le bruit occasionné par les appareils textuels originels.

1.2.3 Mémoires, lettres, rapports

Une dimension plus profonde de ces problèmes est constatée dans les travaux ayant comme corpus des lettres, mémoires [45], inventaires d'archive [46], rapports parlementaires anciens [47] ou rapports archéologiques [48]. Ces corpus obligent à se confronter à certains problèmes complexes de la langue naturelle et à un vocabulaire qui peut aller du registre familier au registre technique ou au jargon, mais toujours utilisé dans un environnement linguistique stable lié à la prévalence d'un style régulier, correct et formel. Ces corpus contiennent une grande proportion d'entités nommées, ce qui indique une densité d'information élevée [49]. Mais dans la plupart de ces travaux les performances obtenues, à condition d'avoir réalisé une bonne normalisation des textes, ne sont pas très différentes de celles obtenues dans les travaux sur les journaux anciens.

---

qui la rend inacceptable ou nécessitant d'un grand nettoyage. Voir son impact dans le traitement de textes historiques en : Mark J Hill et Simon Hengchen. "Quantifying the impact of dirty OCR on historical text analysis : Eighteenth Century Collections Online as a case study". In : *Digital Scholarship in the Humanities* (2019)

43. Caroline Barrière. "Searching for Named Entities". In : *Natural Language Understanding in a Semantic Web Context*. 2016, p. 23-38

44. K. Kettunen et al. "Old Content and Modern Tools-Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771-1910". In : *arXiv preprint arXiv :1611.02839* (2016) ; A. Erdmann et al. "Challenges and solutions for Latin named entity recognition". In : *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)* (2016), p. 85-93

45. Samet Atdag et Vincent Labatut. "A comparison of named entity recognition tools applied to biographical texts". In : *2nd International Conference on Systems and Computer Science*. 2013

46. Kate Byrne. "Nested Named Entity Recognition in Historical Archive Text". In : *International Conference on Semantic Computing (ICSC 2007)*. 2007

47. Grover, C., Givon, S., Tobin, R., & Ball, J. "Named Entity Recognition for Digitised Historical Texts". In : *LREC.* (2008)

48. Julian Richards. "Text Mining in Archaeology : Extracting Information from Archaeological Reports". In : *Mathematics and Archaeology*. 2015, p. 240-254

49. Isabelle Augenstein et al. "Generalisation in named entity recognition : A quantitative analysis". In : *Comput. Speech Lang.* 44 (2017), p. 61-83

Les problèmes demeurent les mêmes : une reconnaissance plus difficile des noms de personnes et une mauvaise maîtrise de la variabilité sur les corpus datant d'avant le XIXe siècle.

En outre, les tentatives pour réaliser des reconnaissances croisées en mobilisant modèles et jeux de test formés et extraits de collections de textes provenant de différents domaines ont échoué. Les entraînements rendent les modèles très attachés à leur domaine d'origine, compliquant ainsi tout effort pour générer des outils extensibles. Pour y remédier, certaines solutions apportant des sources supplémentaires, normalement des ontologies et bases de données externes, essayant de mieux distribuer la variété des entités nommées observées dans le jeu de test et le jeu d'entraînement, ont permis d'améliorer la performance de reconnaissance, sans toutefois la faire monter jusqu'à un niveau acceptable, supérieur à 80 %.

Un apport important de ces travaux réside dans la littérature autour des représentations des entités qui mettent dans un graphe ou une infographie toutes les relations et occurrences qu'elle fait émerger [50]. Il s'agit d'un travail pertinent étant donné que dans les sources évoquées on privilégie la description claire d'un processus et par extension des personnes, institutions et lieux y participant, ce qui facilite une reconstruction intégrale des connections existant entre d'un côté les entités nommées et de l'autre les concepts clés et les évènements qui les associent.

1.2.4 Les romans

Cette possibilité de produire des réseaux graphiques a été aussi appliquée à des corpus de romans. Dans les romans, qui constituent également de larges corpus faciles à numériser, le groupe des personnages est normalement réduit et ils partagent de fortes relations déroulées au fil d'événements exhaustivement décrits dans le récit. Cette particularité a permis de proposer des réseaux sociaux ego-centrés assez complexes [51]. Ces travaux combinés avec les réseaux produits à partir de l'étude des événements ont permis de construire une vision générale d'un roman ayant comme base les listes d'entités nommées récupérées. Ce genre de constructions permet aussi d'intégrer certaines perspectives classiques de la littérature comparée portant par exemple sur les dynamiques discursives, les représentations sociales et des relations de pouvoir entre les littératures provenant de différents genres, langues et périodes chronologiques.

D'ailleurs, la performance de la reconnaissance appliquée à ce genre littéraire, au moins dans les romans écrits à partir du XVIIIe siècle, est relativement élevée, grâce en partie au nombre très bas d'entités nommées si on les compare avec les journaux [52].

50. Andrea K. Thomer and Nicholas M. Weber. "Using Named Entity Recognition as a Classification Heuristic". In : *iConference 2014 Proceedings*. 2014 ; Sunghwan Mac Kim et Steve Cassidy. "Finding names in trove : named entity recognition for Australian historical newspapers". In : *Proceedings of the Australasian Language Technology Association Workshop 2015* (2015), p. 57-65

51. Mariona Coll Ardanuy et Caroline Sporleder. "Clustering of novels represented as social networks". In : *LiLT (Linguistic Issues in Language Technology)* 12 (2015) ; David K Elson et al. "Extracting social networks from literary fiction". In : *Association for Computational Linguistics. En Proceedings of the 48th annual meeting of the association for computational linguistics.* (2010), p. 138-147

52. Cyril Bornet et Frédéric Kaplan. "A Simple Set of Rules for Characters and Place Recognition in French Novels". In : *Frontiers in Digital Humanities* 4 (2017)

Cependant, en raison des nombreuses spécificités linguistiques qu'elles présentent, puisque chaque roman adopte un style unique et ses entités sont assez répétitives, on obtient de basses performances si on essaye d'utiliser les modèles de reconnaissance entraînes sur des journaux ou sur des textes d'autres genres littéraires, ce qui renforce l'opinion concernant la difficulté d'utiliser et adapter des modèles REN sur des corpus variés.

1.2.5 D'autres genres littéraires

Il n'est pas surprenant que les travaux sur d'autres genres littéraires soient assez rares. Ainsi, par exemple, on ne peut trouver que deux exemples de travaux sur le théâtre et la poésie [53] en raison de la très faible perméabilité de ces textes à la systématisation due à une phénoménologie linguistique plus complexe, mais aussi à une faible importance des entités nommées comme éléments clés pour la compréhension du texte. De plus, étant donné que dans ces cas il s'agit d'éléments textuels assez courts (la poésie) ou d'une succession de registres personnels (le théâtre), la reconnaissance se montre d'une complexité inabordable.

1.2.6 Deux enjeux clé

Le problème du croisement de modèles et plus généralement du transfert de connaissance (*transfer learning*) d'un modèle vers un autre n'a été pas négligé dans la littérature. Des modèles entraînés sur des courriels ont ainsi pu être appliqués à la littérature scientifique [54] ; il en est de même pour des modèles entraînes sur des fils d'actualité et appliqués à des courriels ou à Twitter [55] ; et des modèles appliqués à des journaux après avoir été entraînes sur Wikipedia [56], etc. Tous ces travaux ont permis de confirmer qu'il existe une forte dépendance du modèle à son corpus d'origine. Bien qu'il existe un niveau acceptable de reconnaissance parmi les textes proches, la performance chute si le genre ou le domaine diffèrent.

Une problématique semblable est soulevée dans les travaux essayant d'utiliser un modèle sur des textes dans une langue autre que celle des textes du jeu d'entraînement [57]. La plupart des modèles sont développés sur des corpus en anglais car le nombre de corpus annotés dans d'autres langues est encore faible. Ce problème, conjugué avec celui de la dépendance du domaine, oblige à développer des modèles

53. IV Foley et J John. "Poetry : Identification, Entity Recognition, and Retrieval". In : (2019)

54. Maynard, D., Tablan, V., Ursu, C., Cunningham, H., & Wilks, Y. "Named entity recognition from diverse text types. En 2001. p. 257-274". In : *Recent Advances in Natural Language Processing 2001 Conference* (2001), p. 257-274

55. Alan Ritter et al. "Named entity recognition in tweets : an experimental study". In : *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics. 2011, p. 1524-1534

56. Kettunen et al., "Old Content and Modern Tools-Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771-1910"

57. Ralf Steinberger et Bruno Pouliquen. "Cross-lingual Named Entity Recognition". In : *Benjamins Current Topics*. 2009, p. 137-164 ; Chen-Tse Tsai et al. "Cross-Lingual Named Entity Recognition via Wikification". In : *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. 2016

assez spécifiques pour chaque travail dans une langue autre que l'anglais qui comptent sur peu des ressources. Avec l'introduction de méthodes semi-supervisées (voir section 1.5.1) et pour économiser des efforts, certaines solutions ont été proposées. Il s'agit généralement de l'annotation à la main d'échantillons contenant une grande variété d'entités ou de l'adaptation de corpus annotés provenant des domaines proches de celui, à partir duquel l'annotation est auto-peuplée [58]. D'autres solutions peuvent avoir recours à des dictionnaires domaine-indépendants [59], aux métadonnées dans Wikipedia [60] ou, plus récemment, aux représentations de mots [61] ou aux plongements lexicaux [62] car la REN est aussi une tâche d'étiquetage de la séquence (*sequence-tagging*), dans laquelle il convient d'essayer d'extraire le sens des mots dans leur contexte.

En revanche, le problème de la reconnaissance des entités nommées dans les sources datant d'avant le XVIIIe siècle a été peu abordé dans la littérature. Les raisons apparaissent rapidement parce qu'ici se conjuguent à un niveau élevé tous les problèmes vus jusqu'à présent : les corpus ne sont pas très nombreux et leur numérisation depuis les éditions modernes peut présenter de sévères erreurs. Les éditions sont très lacunaires et en fait, même dans le cas des larges collections, elles peuvent être peu organiques et peu représentatives. En outre, la langue et spécialement l'état de la langue est problématique. Dans les récits anciens, il faut confronter différents états de l'évolution des langues dans leur processus de vernacularisation et, dans les cas des sources médiévales, il faut aussi travailler sur le latin médiéval dans tous ses différents niveaux discursifs [63]. Il existe aussi une intense variabilité du vocabulaire, des erreurs d'écriture, de syntaxe et une influence marquée du style particulier du scripteur [64]. Il n'est pas étonnant que les outils pour travailler sur ce sujet soient encore rares, de même que les bases mobilisables, comme les index géographiques (*gazetteers*), les collections de noms ou les ontologies ; ce qui peut rendre très compliqué tout analyse automatique.

Actuellement, l'accent est surtout mis sur la construction des corpus de référence (corpus contenant des annotations manuelles) et les outils de base comme les

58. Boliang Zhang et al. "Name Tagging for Low-resource Incident Languages based on Expectation-driven Learning". In : *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. 2016 ; Maud. Ehrmann et al. "Diachronic evaluation of NER systems on old newspapers". In : *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)* (2016), p. 97-107

59. Massimiliano Ciaramita et Yasemin Altun. "Named-entity recognition in novel domains with external lexical knowledge". In : *Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing.* (2005)

60. Joel Nothman et al. "Learning multilingual named entity recognition from Wikipedia". In : *Artif. Intell.* 194 (2013), p. 151-175

61. Rami Al-Rfou et al. "Polyglot : Distributed word representations for multilingual nlp". In : *arXiv preprint arXiv :1307.1662* (2013)

62. Cicero dos Santos et Victor Guimarães. "Boosting Named Entity Recognition with Neural Character Embeddings". In : *Proceedings of the Fifth Named Entity Workshop*. 2015

63. Michael W. Herren. "Latin and the vernacular languages. Medieval Latin : An Introduction and Bibliographical Guide". In : (1996), p. 122-130

64. Maud Ehrmann et al. "Building a multilingual named entity-annotated corpus using annotation projection". In : *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011* (2011), p. 118-124

lemmatiseurs, corpus arborés (*treebanks*) et vocabulaires formalisés[65]. Ainsi, les premiers travaux concernant les textes anciens et médiévaux ont été fait sur la littérature classique et médiévale en latin et sont focalisés sur les études autour des entités géographiques[66]. Dans ce contexte la Patrologie latine, le *Codex Thomisticus*, les collections d'auteurs classiques de certains projets et les bases de textes littéraires, notamment *Perseus* et le *projet Gutenberg*, ont été indispensables comme sources fiables de corpus. Comme dans le cas de l'analyse des lettres et des rapports scientifiques, le latin classique et scolastique oblige à se confronter avec certaines irrégularités du discours (notamment chez les écrivains médiévaux), toutefois contrebalancées par une haute correction dans le style et un attachement à la norme et aux modèles. Le problème fondamental réside dans le caractère lacunaire des sources (*data sparsity*) qui opère une transmission incomplète des styles, modes et traditions scripturaires. Une version encore plus normée du discours se trouve dans les sources médiévales à valeur juridique ; les actes rédigés suivant un modèle formulaire, se trouvent très attachés à une forme rédactionnelle bien définie, exigence qui leur confère une valeur comme preuve légale. Mais ils sont en même temps, sujets à une mise par écrit parfois assez personnalisée du scribe qui rédige souvent à l'aide de sa mémoire faisant à l'occasion intervenir son goût pour certaines formules et son inventivité.

1.3 La reconnaissance des entités nommées en Histoire

Tandis qu'une partie des corpus utilisés dans la reconnaissance des entités nommées sont de nature historique, la recherche réalisée ne l'est pas toujours. Elle est le plus souvent conduite par des disciplines autres que l'histoire, et chaque discipline essayant de configurer la recherche scientifique dans ses termes et autour de ses centres intérêts. Dans l'état actuel de l'art, on peut avoir l'impression que la plupart des travaux mobilisent des techniques provenant des études littéraires. En réalité, chaque discipline impose son influence. Dans les études littéraires, comme on l'a vu plus haut, il existe un intérêt très marqué pour les réseaux égocentrés et pour la construction d'infographies qui présentent l'aperçu complet d'un texte. Les travaux sur les journaux privilégient des réseaux construits sur les événements ou sur les personnages connus et visent à en extraire des informations sémantiques afin de mieux parcourir leurs fonds d'archives. De même, l'intérêt particulier pour les constructions de routes et d'itinéraires géographiques est très présent dans les travaux sur les sources gréco-latines. En outre, le fait qu'il s'agisse de sources modernes ou déjà indexées - dans le cas des cartes et des itinéraires - avec un discours historique qui n'exige pas trop d'adaptations, a beaucoup à voir avec la production de ce type de ressources.

65. Natalia Korchagina. "Building a Gold Standard for Temporal Entity Extraction from Medieval German Texts". In : *Conference on Language Technologies & Digital Humanities Ljubljana* (2016) ; Marco Budassi et Marco Passarotti. "Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon". In : *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. 2016

66. Mike Kestemont et Jeroen De Gussem. "Integrated sequence tagging for medieval Latin using deep representation learning". In : *arXiv preprint arXiv :1603.01597* (2016)

Le domaine à proprement parler historique ne demeure pas totalement à l'écart des recherches, mais il présente certaines particularités. Pour l'instant, les travaux cherchant à appliquer des techniques REN aux textes historiques autres que littéraires sont vraiment rares, mais il existe quelques travaux de nature historique dans le domaine de la *digital history* où des bases textuelles portant une annotation et une désambiguïsation correcte ont été utilisées. Dans les dernières années, la production d'interfaces pour des cartes numériques est devenue sans doute l'un des travaux préférés. Il est rare, en histoire, que des informations concernant la distribution spatiale, normalement réservée à l'archéologie, soient disponibles. Donc, la possibilité de visualiser un espace cartographié à partir de listes exhaustives et contextualisées de noms de lieux a attiré rapidement l'attention. Ce tournant n'est pas que numérique, il est très lié à l'intérêt porté à l'espace et au vocabulaire de l'espace qui a parcouru la recherche historique dans les dernières années. Les travaux portent principalement sur deux centres d'intérêt : les reconstructions de cartographie ancienne et la production d'infographies cartographiées sur des espaces contemporains.

Dans le premier cas, la plupart des travaux ont été réalisés à partir des index géographiques et en se référant à des bases de connaissance externes, notamment Wikipedia et Perseus. Les index géographiques (*gazetteers*) ne sont pas vus dans la recherche comme de simples listes de toponymes ; ils sont des répertoires bien indexés portant des informations précieuses complétant les cartes. Avec la domestication des outils d'information géographique, de nombreux projets de recherche ont envisagé une géolocalisation des lieux, c'est-à-dire, la conversion d'informations spatiales indirectes en informations précises et transférables sur une carte. À cet effet, ils ont combiné les coordonnées des listes de noms indexés qui apparaissent dans les cartes anciennes avec les réseaux de données modernes afin de produire une actualisation cartographique des scenarios anciens. À partir de cela, les cartes anciennes peuvent servir comme support d'une certaine narration historique [67] dans le but de reconstruire des routes et chemins anciens [68] , des entités politiques et des circonscriptions administratives [69] ou de grands mouvements migratoires, commerciaux, etc. Dans le même sens, il faut mentionner les magnifiques éditions numériques de cartes médiévales, qui sont en réalité des encyclopédies spatiales portant des interprétations cartographiques gréco-latines et arabes. Quelques-uns comme les *mappaemundi* sont devenues des interfaces très riches d'exploration de la toponymie historique et de la littérature latine et médiévale sur les voyages et les pèlerinages [70]. D'autres projets, plus récents, envisagent aussi la reconstruction des villes anciennes à partir des registres [71] en s'orientant vers la

67. R. Mostern et I. Johnson. "From named place to naming event : creating gazetteers for history". In : *Int. J. Geogr. Inf. Sci.* 22.10 (2008), p. 1091-1108

68. Elijah Meeks et Karl Grossner. "ORBIS : An interactive scholarly work on the Roman world". In : *Journal of Digital Humanities* 1.3 (2012), p. 1-3

69. Rainer Simon. "Towards semi-automatic annotation of toponyms on old maps". In : *e-Perimetron* 9.3 (2014), p. 105-128

70. Voir par exemple, The Virtual Mappa Project incluant les mappaemundi des abbayes anglaises : http ://sims.digitalmappa.org ; et tous les projets visant à cartographier les récits de voyages hébergés dans http ://globalmiddleages.org

71. Christina M. Fitzgerald. "Mapping the Medieval City : Space, Place and Identity in Chester c. 1200–1600". In : *J. Hist. Geogr.* 46 (2014), p. 133-134

reconstitution de la vie sociopolitique de l'époque.

La mise en place des outils d'information géographique a mieux fonctionné dans les travaux concernant les reconstructions infographiques d'espaces contemporains, réalisées dans la plupart des cas à partir de l'exploitation des archives urbaines des villes nord-américaines des XIXe et XXe siècles. Les systèmes de SIG [72] ont besoin d'informations assez précises et bien structurées, ce qui n'est pas le cas dans les sources anciennes qui peuvent présenter des points géographiques mal déterminés ou erronés [73]. En fait, les systèmes SIG ont dépassé la fonction d'outil géographique et se sont transformés en fournisseurs de bases de données permettant de transformer rapidement de grandes quantités d'informations en interfaces d'analyse et de question-réponse. De là sont nés quelques-uns des projets les plus intéressants cartographiant l'évolution urbaine [74], la circulation des informations, les relations entre les communautés rurales, les révolutions industrielles (Globalization Projet 1789-1861), qui transcrivent sur les plans de ville et les cartes des cadastres les entités nommées (notamment odonymes et patronymes) récupérées à partir des registres, des recensements, des journaux et des bilans politiques participant à des reconstructions événementielles.

Dans d'autres travaux, en dehors de ceux déjà commentés, l'intérêt a plutôt porté sur la possibilité d'activer des recherches alternatives dans les bases de données des documents historiques. Dans les moteurs traditionnels, la recherche se fait sur les métadonnées d'identification, basées principalement sur les anciens guides d'archivage. Mais avec l'incorporation des entités nommées, on peut offrir une recherche basée sur les noms de personnes et de lieux et, dans certains cas, récupérer et comparer leurs contextes immédiats d'apparition. Cette fonctionnalité est complétée par certaines statistiques lexicales qui peuvent aider à récupérer des documents plus importants. Néanmoins, il n'est pas encore commun d'avoir les textes en accès libre pour appliquer des méthodes de traitement massif et en réalité, dans certains cas, ce que l'on récupère c'est simplement l'image du texte et ses métadonnées descriptives.

Les travaux de *digital history* ont profité de l'expertise autour des entités nommées et des techniques développées dans les projets de littérature numérique. Mais ces travaux demeurent encore très loin des sources historiques les plus utilisées dans la recherche, c'est-à-dire les archives manuscrites. Les raisons en sont multiples : la reconnaissance automatique de caractères sur les manuscrits est pour l'instant au point mort, bien qu'il y ait eu des progrès significatifs et quelques plateformes pour aider à la transcription documentaire [75] ; les sources supplémentaires, c'est-à-dire les éditions critiques ou philologiques des manuscrits, susceptibles d'être numérisées,

72. SIG (*système d'information géographique*) décrit un système d'information qui organise, stocke, gère, et affiche différentes données géographiques et qui par extension permet la représentation et l'analyse cartographique de ces données.

73. Donald A. DeBats. "A Tale of Two Cities : Using Tax Records to Develop GIS Files for Mapping and Understanding Nineteenth-Century U.S. Cities". In : *Historical Methods : A Journal of Quantitative and Interdisciplinary History* 41.1 (2008), p. 17-38

74. Donald. A. DeBats et I. N. Gregory. "Introduction to Historical GIS and the Study of Urban History". In : *Soc. Sci. Hist.* 35.4 (2011), p. 455-463

75. Nous parlons ici de deux plateformes/logiciels qui ont gagné en popularité dans les dernières années pour le HTR (*handwritten texte recognition*) : Transkribus (https ://transkribus.eu/Transkribus/) et Himanis (https ://www.himanis.org/)

ne représentent qu'une infime partie de l'univers manuscrit conservé et ne sont pas favorisées par les numérisations massives en raison de leur faible intérêt pour le reste de la communauté. Cela sans oublier que l'extraction d'informations structurées à partir des sources manuscrites anciennes est confrontée à une très grande diversité linguistique, à l'absence d'un standard orthographique et au manque de ressources externes, exigeant ainsi une longue heuristique pour faire tourner les algorithmes et arriver à former des modèles. Étant donné que la plupart de ces projets s'appuie sur des financements publics et privés, l'investissement dans des projets qui ne présentent pas de résultats immédiats ou qui nécessitent des outils très exclusifs, n'est pas une priorité.

1.4 Outils du traitement automatique de la langue et corpus annotés.

L'une des réponses les plus énergiques de la communauté numérique concernant les corpus historiques a été de produire des outils de base et des corpus annotés à la main permettant de fonder un premier niveau de traitement automatique robuste pour les domaines et les langues dont la "boîte à outils" est restreinte (langues peu dotées) : typiquement il s'agit de langues pour lesquelles il n'avait pas été développé d'outils de traitement automatique jusqu'ici, telles que le latin ou d'autres langues anciennes.

Tout traitement automatique a besoin de trois éléments clés :

1. Des corpus annotés à la main. Comme on le verra, toutes les approches ont besoin de ces corpus prétraités, que ce soit comme base d'entraînement des modèles ou comme corpus d'évaluation - ou "test" - de la performance. Ils représentent aussi un bon premier niveau pour des techniques de *bootstrapping* essayant l'utilisation des échantillons de données supervisées pour guider un apprentissage non-supervisé. Les larges corpus médiévaux mis à disposition par les projets CBMA [76], CDLM [77] et SRCMF [78] en sont un bon exemple. Plusieurs équipes ont annoté à la main plusieurs attributs, tels que les caractéristiques morphologiques, les descriptions sémantiques, les relations syntaxiques, les entités nommées, etc.
2. Des outils de pré-traitement afin de réaliser une analyse dédiée sur une langue qui peut être adaptée à un état de langue particulier : *lemmatiseurs, treebanks, parsers, taggers*, etc. Depuis un premier corpus arboré (*dependency-treebank*) proposé pour le grec et le latin classique [79], d'autres comme L'*Index Thomisticus* et *Omnes* [80] offrent des outils pré-entraînés sur la littérature scolastique ; celui

76. http ://www.cbma-project.eu/
77. http ://www.lombardiabeniculturali.it/cdlm/
78. http ://srcmf.org/
79. David Bamman et Gregory Crane. "The Ancient Greek and Latin Dependency Treebanks". In : *Language Technology for Cultural Heritage*. 2011, p. 79-98
80. Barbara Mcgillivray et al. "The Index Thomisticus Treebank Project : Annotation, Parsing and Valency Lexicon". In : *TAL, 2009* 50.2 (), p. 103-127

du PROIEL [81] offre une extension vers quelques langues indo-européennes ; *Omnia* pour le médiolatin [82] ; Pandora [83] pour les variantes vernaculaires [84].

3. Des bases de connaissance supplémentaires : index géographiques, dictionnaires, thesaurus, collections, etc. Depuis un premier relevé statistique des œuvres latines par LASLA [85] d'autres - listes d'autorités, répertoires géographiques tels que proposés par Pelagios [86], Trimegistos [87], Perseus [88] - sont très utilisés dans de nombreux projets traitant avec la littérature classique et médiévale. Du même que le récent catalogue des gens dans les sources anciennes du Herodotus Projet. [89]

En outre, le nombre croissant de plates-formes créées pour faciliter l'analyse morphologique de nouveaux corpus, tels que Collatinus [90], Lemlat [91], GutenTag [92] et CLARIN [93], est remarquable. Ces plateformes offrent un outil d'extraction rapide d'une gamme complète d'attributs linguistiques. Enfin, certaines initiatives utilisent des techniques de traitement automatique du langage naturel (TAL), telles que Chartex sur les chartes médiévales [94], CompHistSem [95] sur la littérature latine tardive et Manuscripts Online [96], qui fournit un moteur de recherche permettant de répertorier les noms et les lieux mentionnés dans des livres imprimés et Trimegistos qui a appliqué la reconnaissance d'entités nommées avec succès aux sources en papyrus et épigraphiques grecques et latines afin de mieux indexer ses bases de données.

Tous ces efforts sont complétés par la disponibilité croissante de bibliothèques logicielles, y compris un accès facile aux techniques classiques telles que l'étiquetage,

81. https ://proiel.github.io/ ; Dag T T Haug et Marius Jøhndal. "Creating a parallel treebank of the old Indo-European Bible translations". In : *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)* (2008), p. 27-34

82. http ://glossaria.eu/outils/lemmatisation/

83. https ://github.com/hipster-philology/pandora

84. Récemment un universal treebank pour le latin a été publié en *Universal Dependencies* (https ://universaldependencies.org/treebanks/la-comparison.html) contenant ces trois latin treebanks (Latin Dependency Treebank 2.0 du Perseus Project, Latin-ITT de l'Index Thomisticus la section latine du PROIEL)

85. http ://web.philo.ulg.ac.be/lasla/ ; Joseph Denooz. "L'ordinateur et le latin, Techniques et méthodes". In : *Revue de l'Organisation Internationale pour l'Etude des Langues Anciennes par Ordinateur* (1978), p. 1-36

86. https ://github.com/pelagios ; Leif Isaksen et al. "Pelagios and the emerging graph of ancient world data". In : *Proceedings of the 2014 ACM conference on Web science - WebSci '14*. 2014

87. https ://www.trismegistos.org ; Mark Depauw et Tom Gheldof. "Trismegistos : An Interdisciplinary Platform for Ancient World Texts and Related Information". In : *Communications in Computer and Information Science*. 2014, p. 40-52

88. David A Smith et Gregory Crane. "Disambiguating Geographic Names in a Historical Digital Library". In : *Lecture Notes in Computer Science*. 2001, p. 127-136

89. https ://u.osu.edu/herodotos/ ; Erdmann et al., "Challenges and solutions for Latin named entity recognition"

90. https ://outils.biblissima.fr/en/collatinus-web/

91. http ://www.lemlat3.eu/

92. https ://gutentag.sdsu.edu/

93. https ://www.clarin.eu/

94. www.chartex.org

95. www.comphistsem.org

96. https ://www.manuscriptsonline.org

l'analyse, le découpage et la résolution des entités nommées en combinaison avec des outils d'annotation qui travaillent avec des lexiques tels que TreeTagger [97] et Lapos. Ils contribuent à l'intégration de la NER dans les projets de milieu de gamme. Stanford CoreNLP [98], Freeling, Natural Language Toolkit [99] et Scikit-learn [100] sont parmi les ensembles d'outils et de bibliothèques les plus utilisés suivant des méthodes telles que l'entropie maximale, des machines à vecteurs de support (SVM) [101], les réseaux de néurones (ANN) ou Champs aléatoires conditionnels (CRF) [102], la dernière étant la principale technique utilisée dans cette thèse.

1.5 Méthodes pour la reconnaissance automatique des entités nommées

1.5.1 Trois concepts clé.

Trois concepts doivent être visités avant de parler des approches classiques d'automatisation des entités nommées : le style d'annotation, la qualité du corpus et les mesures de performance :

Le schéma d'annotation

Les entités ENAMEX sont normalement balisées de manière assez simple : pour chacune on définit s'il s'agit de PERS (*person*), LOC (*location*), ORG (*organisation*), selon leur définition dans le Conll2002 standard corpus [103]. Mais afin de définir les frontières d'une entité, c'est-à-dire préciser si un mot fait partie d'une entité, la taille de cette entité, et, le cas échéant, le chevauchement d'entités, il est fréquent d'utiliser le style BIO [104] ou une de ses variantes. BIO est une abréviation de début (*Beginning*), intérieur (*Inside*) et extérieur (*Outside*), qui indique la manière correcte d'interpréter une entité composée par plus d'un mot (*multi-token*) comme le début d'une entité (B), le deuxième membre ou continuation d'une entité (I) et l'absence d'entité (O).

$$w_i = \left\{ \begin{array}{ll} \text{B-entity} & \text{si } w_i \text{ est le début d'une entité} \\ \text{I-entity} & \text{si } w_i \text{ est la continuation d'une entité} \\ \text{O-entity} & \text{O si il n'y a pas d'entité} \end{array} \right\} \tag{1.1}$$

Standardisation des corpus

97. https ://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
98. https ://stanfordnlp.github.io/CoreNLP/
99. http ://nlp.lsi.upc.edu/freeling/node/1
100. https ://scikit-learn.org/stable/
101. Thorsten Joachims. "Text categorization with support vector machines : Learning with many relevant features". In : *European conference on machine learning*. Springer. 1998, p. 137-142
102. John Lafferty et al. "Conditional random fields : Probabilistic models for segmenting and labeling sequence data". In : (2001)
103. Conll (conference natural language processing)
104. L. A. Ramshaw et M. P. Marcus. "Text Chunking Using Transformation-Based Learning". In : *Text, Speech and Language Technology*. 1999, p. 157-176

La littérature la plus récente fait une distinction entre deux types de corpus : le *golden standard corpus* et *silver standard corpus*[105]. Le *golden corpus* est le corpus de référence qui porte l'annotation souhaitée qui doit produire un modèle automatique. Autrement dit, c'est le corpus annoté à la main par plusieurs annotateurs et ensuite révisé par un spécialiste de la matière. Il s'agit de corpus longs à produire puisqu'ils demandent un grand investissement d'effort humain, ce qui explique qu'ils soient les plus demandés pour les approches statistiques.

Le *silver corpus* est quant à lui produit automatiquement à partir d'un texte sans annotation. L'annotation proposée est forcément imparfaite (la moyenne des meilleures modèles REN est autour de 85 %), mais elle est produite automatiquement en quelques heures. Avec les progrès des modèles REN, dans les étapes d'évaluation des modèles, ils sont chaque fois plus présents afin des comparer l'écart entre les performances offertes par les modèles *silver* et *golden* ou comme modèle principal qui, combiné avec des bases de connaissance extérieures, peut offrir des résultats similaires à ceux obtenus avec un *golden corpus*[106].

Mesures d'évaluation

Pour évaluer la performance d'un modèle on utilise les mesures de rappel, précision et f-mesure. Elles sont introduites dans MUC[107]. Étant donné une série d'entités à identifier N_{tags} un modèle proposera une série d'étiquettes correctes $N_{correct}$ et une série d'étiquettes incorrectes N_{wrong}, alors :

$$recall = \frac{Ncorrect}{Ntags} \tag{1.2}$$

$$precision = \frac{Ncorrect}{Ncorrect + Nwrong} \tag{1.3}$$

Le rappel détermine la capacité de récupération des entités pertinentes et susceptibles d'être des entités nommées. Le rappel exprime ainsi la sensibilité du système au moment de fournir des réponses possibles aux recherches. De son côté, la précision détermine le niveau de correction dans les résultats récupérés. La précision est ainsi liée au nombre réel d'entités classées complètement.

Puisque les deux mesures évaluent des caractéristiques différentes et parfois antagonistes (réduction du bruit et du silence), normalement les résultats finaux sont exprimés par une moyenne harmonique :

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{1.4}$$

105. Ning Kang et al. "Training text chunkers on a silver standard corpus : can silver replace gold ?" en. In : *BMC Bioinformatics* 13 (jan. 2012), p. 17

106. Michele Filannino et Marilena Di Bari. "Gold standard vs. silver standard : the case of dependency parsing for Italian". In : *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*

107. Grishman et Sundheim, "Message Understanding Conference-6"

Les résultats individuels de comparaison entre l'annotation à la main et l'annotation automatique peuvent être exprimés selon une terminologie plus fine : vrais positifs (tp), vrais négatifs (tn), faux positifs (fp) et faux négatifs (fn). Les termes positif et négatif font référence à la prédiction du modèle et les termes vrai et faux indiquent si cette prédiction est correcte ou pas selon l'annotation fournie. [108]

1.5.2 Approches supervisées et méthodes symboliques

Les approches classiques des systèmes de reconnaissance des entités nommées (REN) peuvent s'organiser en deux grandes branches : apprentissage supervisé et méthodes symboliques.

Dans le cas de l'apprentissage supervisé, on fournit à l'algorithme plusieurs exemples contenant l'annotation attendue à la sortie. L'algorithme doit modéliser la valeur de chaque étiquette en ajustant des paramètres à chaque itération, de manière à ce que le modèle final puisse produire une annotation automatique la plus proche possible de l'annotation manuelle fournie. Une partie importante de la recherche est dédiée à générer de modèles capables de réduire l'écart entre la dépendance du modèle à son corpus d'origine et sa capacité d'être appliqué à d'autres corpus (généralisation).

Par contre, les méthodes symboliques n'utilisent pas de données annotées, autrement dit, les exemples sont fournis à l'algorithme sans préciser la "sortie" souhaitée. Mais on fournit à l'algorithme un ensemble de règles, grammaires ou bases de données, afin de faire émerger des données pertinentes. Il s'agit alors d'une approche plus facile à mettre en œuvre, mais dont les résultats sont moins performants que dans le cas de l'apprentissage supervisé. Ce type de modélisation concerne surtout les approches à partir de règles (*rule-based*), de dictionnaires (*dictionnary-based*) et de statistiques par regroupement (*clusterisation*).

1.5.3 Méthodes symboliques

L'idée centrale de l'approche par des règles est de définir toutes les conditions que doit remplir un mot pour être considéré comme entité nommée. Les conditions varient fortement d'un corpus à l'autre. Il est nécessaire d'avoir une connaissance très précise des entités afin de couvrir tous les scénarios linguistiques possibles d'apparition (capitalisation, catégorie morphosyntaxique, ordre dans la phrase, préfixes, co-occurrences et d'autres régularités) ainsi que de sa structure (simples, complexes, modificateurs, accidents, fonctions, etc.). Le modèle est ainsi formé par l'ensemble organisé des règles conditionnelles, normalement construit en utilisant des expressions régulières.

La modélisation par des dictionnaires permet d'introduire de larges répertoires contenant des listes d'anthroponymes et des index géographiques, c'est-à-dire, des formes déjà validées d'entités nommées. Une fois les entités candidates détectées dans le texte, la classification se fait normalement par similarité avec celles contenues dans

108. Christopher D. Manning et Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. en. MIT Press, 1999

les dictionnaires. Cette méthode présente une importante limite parce que les formes qui ne sont pas présentes dans les dictionnaires ne seront inclues dans aucune catégorie de la taxonomie des entités. Afin de pallier ce défaut, dans les systèmes les plus performants une combinaison entre ensembles de règles et dictionnaires de formes peuvent offrir des performances assez acceptables sur de petits corpus.

Les méthodes symboliques statistiques consistent le plus souvent à regrouper des mots par affinités (*clustering*). À partir d'un texte prétraité, on programme l'algorithme pour trouver les tendances cachées (*hidden patterns*) sur lesquelles s'organise un texte, c'est-à-dire, les tendances qui règlent sa distribution. La méthode détermine les balises selon une maximisation des traits contextuels des mots. Des mots qui statistiquement partagent un même contexte font normalement partie d'un même groupe. Étant donné que le nombre de catégories n'a pas été préalablement indiqué, ces modèles peuvent annoter un corpus avec un large nombre de catégories au-delà du classique ENAMEX.

En général les résultats que l'on peut observer dans la plupart des travaux reposant sur des approches symboliques nous conduisent à trois conclusions :

1. Les solutions heuristiques privilégiant une approche combinée *ruled-based* et *dictionnary-based* montrent des performances plus élevées que celles reposant sur l'un seul d'entre eux ;
2. Dans les approches par des dictionnaires, on peut obtenir une précision très élevée, mais au détriment du rappel. En effet, un système entraîné sur un dictionnaire est très performant sur les formes identifiées mais il n'est pas bien formé dans la reconnaissance des formes qui ne sont pas recensés ;
3. En outre, comme les règles sont définies en fonction des spécificités de la langue et du domaine du texte, les modèles à base de règles sont difficilement exportable vers un corpus d'une langue ou d'un domaine différent.

Les approches symboliques sont encore très utilisées dans la recherche parce qu'elles n'ont pas besoin de données annotées. Compte tenu de la pénurie des corpus annotés et de la relative disponibilité des dictionnaires de noms, index géographiques et répertoires complexes de règles d'extraction, spécialement dans le domaine des études littéraires et historiques, ces modèles proposent une solution accessible pour structurer des textes. En plus, ils s'adaptent bien à la recherche en humanités car leurs résultats sont plus facilement interprétables par les experts qui peuvent modifier les règles d'extraction à chaque étage. Ces méthodes ont démontré leur efficacité dans l'annotation rapide, surtout de corpus très formalisés comme les registres administratifs ou les rapports médicaux.

1.5.4 L'apprentissage supervisé

Dans les approches statistiques supervisées, l'annotation aborde le problème de la classification des séquences de mots et sépare les sous-chaînes positives (entités nommées) de celles négatives (autres catégories). La tâche de l'algorithme sera de proposer une séparation similaire dans un texte nouveau afin de déterminer quels mots correspondent à une entité nommée, puis déterminer leurs frontières et leur fournir une

étiquette précisant sa typologie. Ici deux grands types de classificateurs sont utilisés : les génératifs et les discriminants. Si on considère qu'on a une séquence $(X_1...X_n)$ à laquelle on veut assigner un groupe d'étiquettes $(Y_1...Y_n)$, on a deux façons de le faire :

1. On calcule la séquence X_i qui correspond à chaque étiquette Y_i et on parle alors d'un classificateur *génératif* ;
2. On prédit l'étiquette Y_i pour chaque séquence X_i et on parle d'un classificateur *discriminant*.

Les génératifs apprennent la probabilité de distribution conjointe des séquences X et d'étiquettes Y ou $P(X, Y)$ et ils déterminent ensuite la probabilité de Y_i étant donné X_i ou $P(Y_i|X_i)$. Les discriminants par contre calculent directement $P(Y|X)$. Ainsi, les génératives *génèrent* un modèle représentatif de chaque classe Y en prenant l'ensemble des séquences X_i, alors que les discriminants doivent *discriminer* la meilleure classe Y_i pour une séquence X_i en calculant les frontières entre les classes.

Le classificateur génératif le plus utilisé est le *Hidden Markov Model* (HMM). Un modèle HMM doit détecter le meilleur état pour une séquence, mais un état et une séquence sont toujours influencées par l'état et la séquence antérieurs (propriété de Markov). Donc on doit détecter la meilleure séquence d'états pour une séquence d'observations. HMM désigne alors les catégories Y à partir des observables X.

Parmi les classificateurs discriminants, l'un des plus connus est le Conditional Random Fields (CRF), qui est utilisé dans cette thèse. À différence des HMM qui modèlent à la fois la probabilité de la séquence d'états et des étiquettes, les discriminants doivent modéliser la probabilité conditionnelle $P(Y|X)$ d'une séquence aléatoire d'étiquettes Y étant donné une autre séquence d'observations X. Les séquences X dans le CRF peuvent être des séquences multidimensionnelles (constituées par n propriétés), ce qui permet d'intégrer d'autres sources d'information disponibles. CRF désigne alors les catégories Y à partir des propriétés (*features*) des séquences X. Ces propriétés, comme on l'a déjà vu, correspondent normalement aux traits internes et contextuels des mots (POS-tag, capitalisation, suffixes, co-occurrences, position dans la phrase, etc.)

Approches hybrides

Alors que les méthodes utilisant des données annotées montrent des performances très élevées, mais que la *clusterisation* sur texte brut des méthodes non supervisées peut aussi offrir des résultats intéressants, au cours des dernières années quelques approches hybrides ont gagné en popularité. L'idée de base était de privilégier une approche non supervisée mais en commençant la modélisation sur un ensemble représentatif de données annotées. L'algorithme est fourni avec quelques groupes catégorisés (*clusters*) comme point de départ à partir desquels on essaye de maximiser les traits contextuels de tous les intégrants de chaque groupe afin de modéliser la distribution d'une étiquette et de chercher ensuite des candidats dans le texte qui apparaissent dans le même contexte. La puissance de cette méthode se trouve dans l'itération (*bootstrapping*) ; à chaque visite dans le corpus, le cluster s'est agrandi et son contexte est renforcé. Comme dans les approches statistiques non supervisées,

le modèle peut construire des catégories autres que celles d'origine et fournir des sous-catégories.

Les groupes originels, appelés *learning seeds* dans la littérature, peuvent être définis à partir des étiquettes, des règles ou des listes liées à un dictionnaire, ce qui permet d'avoir des données annotées automatiquement selon diverses observations et caractéristiques en profitant de certains caractères stéréotypés (dont on a vu l'exploitation dans les méthodes *rule-based*) des contextes d'apparition des entités nommées [109].

1.6 La désambiguïsation des entités nommées

La récupération de listes d'entités nommées dans un texte peut, dans certains cas, s'avérer un travail totalement inutile si ces entités ne sont pas correctement identifiées. La désignation d'entités nommées peut conduire à l'erreur de penser que l'on récupère des personnes alors qu'en réalité on récupère des noms propres ; elles peuvent devenir des noms de personnes ou de lieux lorsqu'elles sont associées à des données d'identification. Ceci peut donner lieu à un grand nombre de situations nécessitant plus ou moins d'efforts et la mobilisation de plus ou moins de ressources pour être réglées. Un nom peut être partagé par des centaines de personnes dans une base de données de journaux ; par quelques dizaines dans une base de données littéraires ou n'apparaître qu'une seule fois, comme cela est habituel dans les textes médiévaux. Le nom d'une ville peut avoir une douzaine d'homonymes à travers le monde (comme dans les cas de Paris, Tolède ou Valence) ou être un *hápax* correspondant à une ville inconnue par erreur de copiste. Plus encore, selon la langue ou le domaine, un même nom peut présenter une large variabilité graphique rendant plus lourde et désorganisée son exploitation. La tâche consistant à désambiguïser un nom, autrement dit, à spécifier l'identité d'une entité nommée et à définir une forme canonique pour chacune d'elles, passe par des méthodes de résolution différentes.

Annotation sémantique (*Entity Linking*)

Une des propositions les plus courantes vient des domaines des bibliothèques qui encourage à utiliser les bases de données d'auteurs, d'autorités ou les ontologies afin de connecter les listes d'entités récupérées avec des références uniques. Ce genre de bases de données a été cultivé depuis au moins deux décennies par les bibliothèques en associant les auteurs aux livres, les livres aux personnages ou aux sujets afin de mieux les classifier. On les retrouve dans la base de plusieurs systèmes publics de références. Existent actuellement quelques projets assez importants qui reposent sur l'aspiration de ce contenu structuré dans le but de produire des bases de connaissance dans un format lisible par l'ordinateur (normalement en SPARQL) dont le contenu puisse être

109. Miguel Won et al. "ensemble named entity recognition (ner) : evaluating ner Tools in the identification of Place names in historical corpora". In : *Frontiers in Digital Humanities* 5 (2018), p. 2 ; James Curran et Stephen Clark. "Language independent NER using a maximum entropy tagger". In : *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. 2003, p. 164-167

consulté et partagé facilement. Quelques projets assez connus comme Dbpedia [110], FreeBase [111] or Yago [112] extraient des contenus à différentes échelles de trois des plus grandes bases de données publiques : Wikipedia (à partir des info-boîtes, tables de données présentant les informations importantes sur un sujet), Wornet (lexiques et dictionnaires en anglais) et Geonames (index géographiques) afin de connecter entités et informations uniques d'identification.

Ces initiatives suivent souvent les principes du Web sémantique, des *linked data* (la plupart des référentes utilisent les URI's, identificateurs univoques) et de l'interopérabilité des donnés (les triplettes RDF sont le format d'échange le plus utilisé) qui deviennent essentielles dans la tâche de la désambiguïsation parce qu'ils facilitent le partage des données descriptives stockées dans des bases de données. Ce travail règle deux problèmes : l'existence de multiples dénominations pour une même entité qui ainsi peuvent être regroupées sous un même référent, et la difficulté pour former des graphiques de relations entre entités qui ainsi peuvent être connectées plus facilement dans une même base de données ou dans plusieurs [113].

Celui-ci est un travail qui en tout cas dépasse la tâche de la reconnaissance des entités nommées, mais par lequel les listes d'entités nommées sont un requis indispensable. En fait, dans une partie importante des travaux de désambiguïsation la *routine* classique se déroule comme suit :

1. La détection des entités nommées dans le texte avec une première classification de type catégorique ;
2. La sélection des candidates les plus proches dans la base de connaissance pour chaque entité nommée de la liste ;
3. La convergence entre le meilleur candidat et la meilleure entité étant donné la similarité dans leur contexte d'apparition.

Malgré un nombre important de travaux prenant cette direction, c'est un domaine de la recherche qui a trouvé rapidement ses limites. Le groupe des bases de connaissances les plus utilisées se sert principalement de Wikipédia comme base principale - de fait, le processus est souvent appelé wikification [114] - et de ressources lexicales en anglais. De surcroît, la nature encyclopédique de Wikipédia met un filtre important dans le nombre d'entités recensées, et se limite aux conceptions modernes, aux personnages connus ou aux évènements remarquables [115]. Elle est indispensable dans une première désambiguïsation des listes provenant de la littérature ou des journaux, mais elle se montre complètement inefficace dans les cas d'entités peu connues, retrouvées dans un faible nombre de sources ou en général dans la littérature

110. Jens Lehmann et al. "DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia". In : *Semantic Web* 6.2 (2015), p. 167-195

111. Kurt Bollacker et al. "Freebase : a collaboratively created graph database for structuring human knowledge". In : *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM. 2008, p. 1247-1250

112. Fabian M. Suchanek et al. "Yago : a core of semantic knowledge". In : *Proceedings of the 16th international conference on World Wide Web*. ACM. 2007, p. 697-706

113. Nouvel et al., *Named Entities for Computational Linguistics*

114. Tsai et al., "Cross-Lingual Named Entity Recognition via Wikification"

115. Zhang et al., "Name Tagging for Low-resource Incident Languages based on Expectation-driven Learning"

non recensée. Malgré quelques propositions, les solutions pour offrir un niveau primaire de désambiguïsation dans ce genre de sources, qui constituent la plus grande partie des domaines comme l'histoire, sont encore manquants.

D'ailleurs, on voit un nombre nettement supérieur de travaux d'*entity linking* sur les noms géographiques que sur les noms de personnes. Le travail sur ces entités est techniquement plus accessible, compte tenu du nombre inférieur de candidates à choisir. En général, le nombre total d'entités existantes, même s'il est assez élevé (*Geonames* par exemple recense 10 millions de noms), est forcément plus petit que l'autre. Dans plusieurs cas, les travaux sur des textes anciens se nourrissent d'importants thesaurus des noms historiques (voire ici *Perseus, Pelagius, Getty*) dont la désambiguïsation a été complétée après un soigneux travail d'érudition. On peut trouver toute une pléiade d'heuristiques pour désambiguïser une entité, mais les projets de désambiguïsation automatique restent rares [116].

Concernant les noms de personnes, les rares travaux qui existent sont attachés aux domaines littéraire et journalistique. Les méthodologies peuvent être diverses, en faisant appel aux techniques exploitant des dictionnaires des formes, des propriétés contextuelles ou en appliquant des méthodes plus complexes comme le *word sens disambiguation*. En tout cas, il s'agit de corpus complets, avec un nombre réduit de variants et au sein desquels on peut établir l'ensemble de relations entre les personnages, ce qui est impossible dans les corpus lacunaires.

116. Claire Grover et al. "Use of the Edinburgh geoparser for georeferencing digitized historical collections". en. In : *Philos. Trans. A Math. Phys. Eng. Sci.* 368.1925 (août 2010), p. 3875-3889

Chapitre 2

Corpus et transformation numérique

2.1 Les éditions numériques

L'un des objectifs principaux des éditions numériques est de produire une édition plus performante que celle des formats imprimés et de s'affranchir de la contrainte physique imposée par le papier [117]. Les éditions érudites peuvent fonctionner comme une base de données structurée, en ajoutant des éléments sur le contenu du texte édité, mais les contraintes physiques du codex en limitent la lisibilité : un volume incluant trop d'information serait ingérable [118] et la circulation entre les données est fortement contrainte par la matérialité du support. Avec l'arrivée de l'espace numérique - presque - infini et des indexations massives au moyen de métadonnées qui établissent des ponts sémantiques entre des mots et d'énormes quantités de contenu, cette limite semble avoir été dépassée. Les versions dématérialisées des éditions érudites, qui se fondent en réalité sur une photographie de l'édition imprimée, ont été transformées en source primaire privilégiée par plusieurs projets d'édition numérique.

La production d'éditions numériques à partir de sources dématérialisées est un processus de très longue haleine qui commence donc par un travail classique d'inventaire, de sélection et d'extraction d'anciennes éditions, suivi d'une série de travaux relevant de l'ecdotique [119]. Il est courant que le processus soit entravé par la perte de la tradition manuscrite ou par le manque d'organicité d'un corpus édité, ce qui oblige à former des collections rassemblant des textes selon des choix éditoriaux [120]. L'étape suivante comprend généralement la dématérialisation

117. Elena Pierazzo. "A rationale of digital documentary editions". In : *Literary and linguistic computing* 26.4 (2011), p. 463-477

118. Un très bon exemple seraient les éditions dites génétiques qui se occupent de l'étude de l'avant-texte (les manuscrites préparatoires) voir Almuth Grésillon. *Eléments de critique génétique. Lire les manuscrits modernes : Lire les manuscrits modernes*. Cnrs, 2016

119. Les travaux de critique textuelle ou de critique de restitution ne sont pas étrangers aux éditions numériques, mais on voit différentes positions à propos de la nature éditoriale des textes résultants : Ray Siemens et al. "Toward modeling the social edition : An approach to understanding the electronic scholarly edition in the context of new and emerging social media". In : *Literary and Linguistic Computing* 27.4 (2012), p. 445-461) ; Donald Francis McKenzie. *Bibliography and the Sociology of Texts*. Cambridge University Press, 1999

120. On peut ici chez les philologues ajouter lex travaux de généalogie des manuscripts à partir d'un

moyennant un traitement graphique dont la première phase est la récupération de l'image des éditions (ou des manuscrits) au moyen d'un scanner, plus rarement par la photographie [121], puis la transformation de l'image en fichiers-texte issus de l'ocerisation [122] ou de la saisie manuelle [123]. Les moteurs OCR les plus modernes offrent un résultat assez bon pour un texte imprimé, mais le travail se complique lorsque le texte contient des sections paratextuelles complexes, des textes dans plusieurs langues ou des textes présentant un état particulier de la langue, comme c'est le cas dans les éditions savantes. Dans certains rares cas, il peut être plus économique de saisir le texte à la main que de corriger le résultat donné par l'OCR. Un bon travail de sélection, de transformation et de correction des erreurs, parfois assez nombreuses, engendrées par les outils de prétraitement, permet d'obtenir une version numérique "propre" de l'édition savante, rendant ainsi possible la dématérialisation de l'appareil critique. Cet appareil critique déjà formalisé constitue souvent le cœur des précieuses métadonnées liées à la structure et au contenu qui forment l'épine dorsale de l'édition numérique moderne [124].

Jusqu'ici, cependant, cette matrice de données est assez semblable à la copie numérique d'un livre [125]. Et les éditions *ex novo* étant encore rares, il convient de se demander quelle est la contribution réelle d'éditions numériques qui se fondent sur des éditions papier préexistantes ? Nous pouvons identifier principalement deux apports : la présence d'outils d'exploration et de moteurs de récupération. Si le livre est limité par sa forme matérielle, l'espace infini du numérique permet de mettre à la disposition du lecteur différentes couches concentriques de contenu — commentaires, suppléments, références intra et intertextuelles, etc. —, révélant différentes connexions qui constituent une amélioration significative de l'expérience lecture, tant pour les amateurs que pour les scientifiques [126]. Parallèlement, l'exploration du texte en tant que séquence de caractères contextualisée est favorisée par une multitude de nouvelles techniques et outils permettant de cerner, classer, visualiser et même prédire de nouvelles informations à partir d'un texte, qu'il soit structuré ou non [127]. De nouvelles méthodes d'exploitation et de récupération des informations, à grande échelle, allant

stemma codicum.

121. À propos des problèmes liés à la dématérialisation du texte voir : Jacques Anis. *Texte et ordinateur : les mutations du lire-écrire.* université de Paris X-Nanterre, 1993

122. Traitement du texte par OCR : optical character recognition

123. Avec les progrès des techniques de reconnaissance optique, ce travail est déjà en désuétude

124. Des définitions du vocabulaire d'usage courant dans les éditions numériques peuvent être trouvés dans Kenneth Price. "Edition, project, database, archive, thematic research collection : What's in a name ?" In : *Faculty Publications–Department of English* (2009), p. 69

125. Il s'agit effectivement de la différence entre une "digital" édition (numérique) et une "digitized" edition (numérisée)". Dot Porter. "Medievalists and the scholarly digital edition". In : *Scholarly Editing* 34 (2013), p. 1-26

126. Pierazzo, "A rationale of digital documentary editions"; Paul Spence. "La investigación humanística en la era digital : mundo académico y nuevos públicos". In : *Humanidades Digitales : una aproximación transdisciplinar.* SIELAE. 2014, p. 117-131 les couches de contenu dépendent normalement d'outils de visualisation, ce qui posse des problemes d'interprétation, voir Johanna Drucker. "Humanities approaches to graphical display". In : *Digital Humanities Quarterly* 5.1 (2011), p. 1-21

127. Voir à ce sujet John Burrows. "Textual Analysis." In : *A Companion to Digital Humanities* (2004), p. 323-347

des statistiques massives aux détails, sont incorporées dans les éditions numériques au moyen de métadonnées ; elles peuvent - ou devraient - entraîner de multiples changements dans les pratiques de recherche en Humanités appliquées aux corpus.

L'un des facteurs décisifs dans ce cas est la disponibilité des outils pour un corpus ou collection donnée. Si des informations utiles qui ne peuvent être produites à la main, telles que la fréquence des termes, le nombre d'entités différentes attestées dans un corpus ou la récupération de n-grammes et de cooccurrences [128], peuvent être engendrées par un algorithme simple, des techniques plus complexes telles que la lemmatisation de mots, la résolution de coréférences ou la désambiguïsation du sens peuvent ne pas être disponibles - ou pertinentes - pour toutes les éditions. Dans une édition à usage général, les premiers se montrent plus utiles et pertinents que les seconds, qui acquièrent plus de signification dans un travail scientifique ayant comme objectif la résolution de problèmes de recherche concernant la langue écrite. Au niveau structurel, une édition philologique peut accorder beaucoup d'importance à des phénomènes linguistiques et à des figures rhétoriques n'ayant guère d'intérêt en dehors de la discipline, et mettre beaucoup moins l'accent sur les entités nommées ou l'étude des cooccurrences, d'une importance capitale pour les études historiques.

Par ailleurs, la disponibilité des outils est d'habitude conditionnée par l'accessibilité des textes au sein de la communauté de chercheurs. Cela dit, un instrument de recherche qui exploite des documents historiques est par définition complexe puisqu'il opère sur des contenus sujets ayant une très grande variabilité, de sorte que sa « boîte à outils » est généralement limitée aux structures élémentaires définies par des annotations faites à la main ou semi-automatisées. Cependant, une édition de textes modernes peut être linguistiquement plus accessible et comporter une gamme plus complète de structures annotées et d'outils d'exploration, qui peuvent même interagir avec des questions en langage naturel.

2.1.1 Le *Corpus des Chartae Burgundiae Medii Aevi* (CBMA)

Dans ce panorama, le corpus CBMA - qui correspond en réalité à une collection éditoriale - est l'une des meilleures éditions numériques d'un ensemble de textes médiévaux, parfait exemple d'une transformation numérique réussie, avec l'ajout d'annotations reprenant des aspects rarement formalisés dans ce genre de travail. L'édition sous forme de tableau que nous avons utilisée dans notre travail de thèse contient tous les détails disponibles dans l'appareil critique des éditions érudites : dates, titres, annotations, commentaires, bibliographie, lieux de production, genre, etc. auxquels d'autres, nouveaux, ont été ajoutés : entités nommées, termes clés, type documentaire détaillé, coordonnées, qui sont difficiles à obtenir, car ils nécessitent un long travail manuel et une très bonne capacité de lecture pour être extraits.

128. Un n-gramme est une séquence de N-mots. Ex. "*Iulius Caesar*" (bigramme), "*infandum regina iubes*" (trigramme). L'utilisation de ces sous-séquences est assez habituelle dans les études statistiques de langue parce qu'elles permettent d'étudier le contexte immédiat des mots et de prédire les séquences (modèle de langue). Il est bien probable qu'étant donné "*Iulius*" sa co-ocurrence (les mots intégrant la même sous-séquence) soit "*Caesar*", mais il l'est encore plus qu'étant donné "*infandum regina*" elle soit suivi par "*iubes*" puisque la phrase est exclusive d'Énnée.

Initié il y a plus de 10 ans, le CBMA est une édition encore en construction, intégrant régulièrement de nouveaux documents ; la dernière mise à jour date de septembre 2018 [129]. Initialement considéré comme un corpus diplomatique formé de la documentation de Cluny et Cîteaux, le CBMA a progressivement incorporé la quasi-totalité des documents diplomatiques bourguignons médiévaux dans un vaste arc temporel allant du IXe siècle à la fin du XIVe siècle [130]. La numérisation continue des nombreuses éditions de cartulaires produites depuis le XIXe siècle et de collections des originaux bourguignons, parmi lesquelles figurent quelques recueils célèbres de l'historiographie médiévale comme les cartulaires de Cluny, Saint Vincent de Mâcon et le Cartulaire général de l'Yonne, ainsi que certaines éditions de cartulaires mineurs, constituent la base du corpus.

La base de données diplomatique contient environ 23 000 documents, parmi lesquels un sous-corpus de 5 300 articles a été l'objet d'une annotation plus complexe. Ce sous-corpus, que nous mobilisons informatiquement, est principalement composé de chartes privées produites dans les abbayes clunisiennes et, pour une minorité d'entre elles, dans les abbayes cisterciennes. Les documents qu'il contient, provenant de près d'une centaine de petites localités de Bourgogne, sont extraites de dix cartulaires différents [131]. Cinq ont fait l'objet d'une étude par le passé : les cartulaires A, B et C de l'abbaye de Cluny (79 % du total), le cartulaire de Saint-Vincent de Mâcon, le cartulaire du prieuré de Jully-les-Nonnains et le cartulaire de l'abbaye cistercienne de Vauluisant (voir table 2.1)

Dans une deuxième étape d'extension de l'édition, une fois les actes épuisés, le corpus accueille et traite également des documents qui apportent une plus grande hétérogénéité à l'ensemble : chroniques, textes normatifs et hagiographiques, ce qui porte le nombre à près de trente mille documents [132]. Il s'agit d'un mouvement naturel et pertinent dans la mesure où les éditions hagiographiques sont parfois beaucoup plus abondantes que les éditions d'actes et leurs éditions numériques plus disponibles. Les textes hagiographiques qui constituent une ressource complémentaire à la recherche historique ajoutent au corpus un niveau pertinent d'hétérogénéité qui peut ainsi offrir une documentation liée tant à la gestion économique qu'à la vie institutionnelle et

129. `http://www.cbma-project.eu/`

130. Marie-José Gasse-Grandjean. "Les « Chartae Burgundiae Medii Aevi » (CBMA) et le numérique". In : *Francia* 40 (2011), p. 255-263 ; Magnani, "Un corpus structuré et hétérogène de textes latins médiévaux (Bourgogne, Ve-XVe siècle)"

131. RAGUT M.C., Cartulaire de Saint-Vincent de Mâcon : connu sous le nom de Livre enchaîné, Mâcon, Protat, 1864 ; BERNARD A., BRUEL A., Recueil des chartes de l'abbaye de Cluny. Tome 1 : 802-954, Paris, Imprimerie nationale, 1876 ; CHARMASSE A. de, Chartes de l'abbaye de Corbigny, Autun, 1889 ; GUIGUE M.-C., Cartulaire de l'église collégiale Notre-Dame de Beaujeu, Lyon, 1864 ; DESJARDINS G., Cartulaire de l'abbaye de Conques en Rouergue, Paris, 1879 ; CANAT de CHIZY P., Cartulaire du prieuré de Saint-Marcel-lès-Chalon, Chalon-sur-Saône, Marceau, 1894 ; REY C., L'entreprise archivistique de Jean de Cirey, abbé de Cîteaux (1476-1501). Le dossier documentaire de la seigneurie de Villars en Côte-d'Or, 2009 ; CATEL A., LECOMTE M., Chartes & documents de l'abbaye cistercienne de Preuilly, publiés et mis en ordre avec introduction, notes et tables, Paris, 1927 ; DUBA W. O., The cartulary of Vauluisant : a critical edition, 1994

132. Eliana Magnani. "Les CBMA en corpus structuré. Atelier 2. Le corpus hagiographique bourguignon. Débats et recherches. LaMOP-Sorbonne, 19 juin 2018". In : *Bulletin du centre d'études médiévales d'Auxerre/ BUCEMA* (2018)

Éditeur	N°chartes	pourcentage
Bernard et Bruel	4101	78.8 %
C. Ragut	635	12.2 %
Charmasse	26	0.5 %
Guigue	38	0.7 %
Canat de Chizy	115	2.2 %
Catel et Lecomte	32	0.6 %
C. Rey	43	0.8 %
Desjardins	10	0.2 %
Duba W.O	182	3.5 %
Miscelane	25	0.5 %

TABLE 2.1 – Nombre de documents par édition dans le corpus CBMA annoté

spirituelle des communautés [133].

Ces cartulaires et collections ont été édités, aux XIXe et XXe siècles, selon différentes normes éditoriales diplomatiques et philologiques ; la figure 2.1 montre comment ces actes sont répartis dans le temps. Les textes dactylographiés numérisés dans les éditions modernes sont la principale source des collections textuelles disponibles, où des éléments tels que la capitalisation, la ponctuation et le développement d'abréviations ont été ajoutés, permettant ainsi de moderniser les sources originales pour en faciliter la lecture. Le texte brut a été stocké dans une base de données dynamique et une équipe d'experts historiens et philologues a annoté manuellement les entités nommées personnelles et géographiques. En raison d'un manque de temps et de ressources, les entités juridiques et institutionnelles n'ont pas été identifiées.

Comme mentionné précédemment dans la partie 1.1, les entités nommées sont un élément fondamental de la structuration du contenu, notamment parce qu'elles ne sont pas récupérables avec les outils classiques du traitement de texte [134]. Ce n'est donc pas un hasard si les responsables du CBMA ont passé des mois à annoter manuellement toutes les entités nommées du corpus de Cluny, car il s'agit d'un corpus suffisamment complexe, du point de vue de la taille et de l'hétérogénéité pour permettre des pratiques de recherche panoramiques et massives jusqu'à présent interdites à l'historien. Annoter les entités nommées est une étape préalable fondamentale qui permet d'appliquer un ensemble plus large d'outils travaillant à des niveaux plus complexes lors de la récupération de nouvelles informations sur le contenu [135].

L'édition réalisée par le CBMA est également un excellent exemple de la diversité des approches qu'une édition numérique peut adopter. Le corpus, dont le téléchargement est gratuit, est disponible en trois formats : la version CSV (valeurs séparées par des virgules) sous la forme d'un tableau, utilisée dans cette thèse, est la

133. À propos du corpus hagiographique clunisien voir les études de Patrick HENRIET. "Chronique de quelques morts annoncées : Les saints abbés clunisiens (X e-XII e seècles)". In : *Médiévales* (1996), p. 93-108 ; et dans le CBMA : MAGNANI, "Un corpus structuré et hétérogène de textes latins médiévaux (Bourgogne, Ve-XVe siècle)"

134. SIMON, "Approaches to hungarian named entity recognition"

135. Simon OVERELL et Stefan RÜGER. "Using co-occurrence models for placename disambiguation". In : *International Journal of Geographical Information Science* 22.3 (2008), p. 265-287

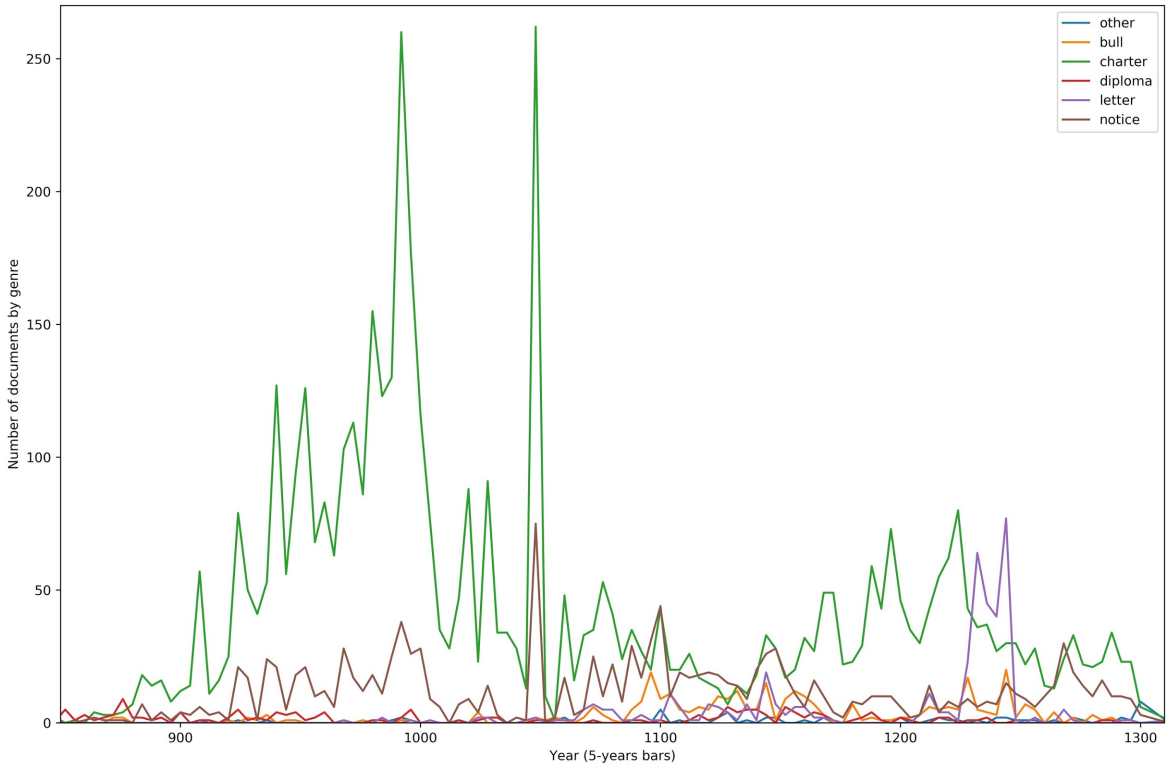

FIGURE 2.1 – Répartition dans le temps du corpus CBMA par type d'acte

plus facile à manipuler car elle possède la matrice de base en texte brut, une version au format propriétaire (Filmaker, qui sert de base de données intermédiaire pour la lecture et le traitement), une version de plate-forme, exploitée par Phylologic qui permet des recherches en tant que base de données relationnelle et connectée avec le logiciel TXT, contenant le corpus lemmatisé, dont le but est de permettre des recherches de type sémantique. Il ne s'agit pas de trois versions du corpus, mais de trois formats contenant différents niveaux d'informations destinés à différents segments de la recherche : traitement du langage naturel, édition numérique et textométrie.

2.1.2 Le corpus clunisien.

Le corpus de Cluny occupe une place centrale dans nos travaux tant pour le développement de modèles que dans les études de cas. Le fonds Cluny semble gigantesque par rapport aux autres fonds médiévaux numérisés : il contient plus de 5 500 items, auxquels il faut ajouter un important corpus de documents supplémentaires : comptes, censiers, coutumiers qui témoignent de l'intérêt de l'ordre pour compiler à la fois des documents de gestion économique et spirituelle. En comparaison, d'autres fonds européens d'importance et de chronologie similaires sont beaucoup plus succincts. Si l'on prend en compte quelques exemples dans leurs versions numériques, le corpus d'Île-de-France compte environ 1 200 éléments ; le corpus des *rolls* anglais, 1

400 ; le Becerro Galicano, 750 [136] ; le CDLM, 4 200. D'autres projets peuvent présenter des quantités proches du CBMA – 28 000 documents –, comme par exemple la *Diplomata Belgica* [137] qui héberge 30 000 documents du VIIe au XIIIe siècle, le fameux corpus *Diplomatarium Norvegicum* [138] qui contient 20 000 chartes pour tout le pays entre 1050 et 1590 et le DEEDS [139] autour de 35 000 actes pour les abbayes du sud de l'Angleterre [140].

Ces corpus de grande taille cités sont très généralistes, couvrant tout un pays à diverses époques. Il est encore très compliqué de trouver un corpus qui, comme celui de Cluny, est concentré sur une seule région et qui présente un arc temporel restreint puisqu'il s'étend du IXe siècle à la fin du XIIIe siècle, et que 80 % de ses documents étant concentrés entre 975 et 1120. Un tel corpus suggère une grande cohérence interne et un très haut niveau de représentativité. S'il est toutefois évident que le corpus actuel n'est qu'une partie de l'ensemble originellement compilé, et qu'un fragment de la production totale des chartes, il est difficile de comprendre les logiques de la sélection appliquée à l'origine aux actes existants, ainsi que les différentes modifications apportées par le temps à travers la tradition archivistique.

Les actes de Cluny ont été édités en 6 volumes, publiés lors d'une entreprise de longue haleine d'abord par Auguste Bernard, puis par Alexandre Bruel entre 1876 et 1920 [141]. C'est une édition dans laquelle les textes ont été extraits principalement des copies des cartulaires réalisées par Lambert de Barive au service du Cabinet des chartes, entre 1770 et 1790 [142]. A. Bernard commence l'édition en 1856 à partir des exemplaires conservés à la Bibliothèque nationale et de certains groupes d'originaux de la Bibliothèque de Saône-et-Loire. Ce n'est pas la première édition des actes de Cluny, car en 1656 avait était publié un petit recueil contenant des actes jugés très importants suivant les critères de l'époque : essentiellement des bulles et des privilèges [143]. Mais A. Bernart n'apprend cela qu'alors que les transcriptions des copies de Lambert de Barive sont presque finies puisque cette édition, la *Bibliotheca Cluniacensis*, publiée deux siècles avant avait presque disparue des bibliothèques. Les copies modernes des cartulaires ou d'extraits des cartulaires ne sont pas rares, mais fournissent régulièrement des versions abrégées ou corrigées des actes. Ce n'est pas le cas du cartulaire de Cluny où le copiste moderne reste non seulement fidèle aux chartes et cartulaires, mais fournit également de nombreuses copies d'originaux qui n'existent pas dans le cartulaire, n'ayant pas été jugés significatifs par les moines [144].

136. http ://www.ehu.eus/galicano/
137. https ://www.diplomata-belgica.be/
138. https ://www.dokpro.uio.no
139. https ://deeds.library.utoronto.ca/
140. Voir une description plus détaille de certains de ces corpus dans : Antonella Ambrosio et al. *Digital diplomatics : the computer as a tool for the diplomatist ?* Böhlau, 2014, p. 185-208
141. Alexandre Bruel et Auguste Joseph Bernard. *Recueil des chartes de l'abbaye de Cluny : 802-954*. T. 49. Impr. Nat., 1876
142. Sébastien Barret. "Un avocat au service du Cabinet des chartes : les travaux de Louis-Henri Lambert de Barive dans les archives de Cluny (v. 1770-v. 1790)". In : *Histoire et archives* 15 (2004), p. 29-64
143. Martin Marrier. *Bibliotheca cluniacensis*. Sumptibus Roberti Fovet Via Iacobaea, sub insigni Temporis et Occasionis, 1915
144. Voir au sujet des copies de L. de Barive et du fonds Moreau : Barret, "Un avocat au service

En conséquence dans le chartrier clunisien cohabitent une bonne quantité d'originaux et de copies [145], ce qui a permis à A. Bernard de réaliser un travail minutieux de collation avant la publication du premier volume. Dans les transcriptions, la plupart des erreurs sont typiques du travail d'un copiste : abréviations, oublis, sélections, très rarement des corrections. Dans quelques cas, l'édition de Bernard complète les omissions et les lapsus du cartulariste et du copiste, et, lorsque cela est possible, il annule les corrections qui altèrent l'original [146].

Les cartulaires sont arrivés en plusieurs livraisons à la Bibliothèque nationale pratiquement pendant tout le temps qu'a duré la publication des volumes préparés par A. Bernard et ils sont l'objet d'une description parallèle en catalogue par L. Delisle [147]. Les cartulaires arrivent dans leur intégralité, mais le chartrier primitif et surtout les collections d'originaux semblent affectées [148]. Compte tenu du nombre d'originaux dont L. de Barive semble disposer et du nombre d'items qui nous sont parvenus, la diminution de la taille du chartrier après les événements de la fin du XVIIIe siècle est importante. Les déménagements, les pillages et les altérations physiques sont généralement la cause de la disparition des fonds d'archives. Aux pertes naturelles des documents dans les différents chapitres avant le rassemblement des archives dans la tour nord de l'abbaye vers le XIIIe siècle, s'ajoutent les pillages pendant les guerres de religion et durant l'époque révolutionnaire.

Il semble que les archives gardées dans la tour nord de l'abbaye (Tour des privilèges) aient pu rester presque entières au moins jusqu'à la fin du XVIIIe siècle [149]. Lambert de Barive en témoigne, lui qui est chargé de copier ces archives à la demande du Cabinet des chartes, et il travaille sur une bonne quantité d'originaux entre 1770 et 1790 [150]. À la suite des évènements révolutionnaires, un grand nombre de volumes entrent dans des mains privées, comme c'est le cas pour de nombreuses archives ecclésiastiques. Le chartrier de Cluny, essentiellement les cartulaires, passe sous la garde de la Bibliothèque municipale de Cluny dans l'intention de constituer les futures Archives départementales de Saône-et-Loire, mais le mouvement final ne se produit jamais. Cet ensemble, et certaines des œuvres dispersées entre des mains privées, ont été transférés à la Bibliothèque nationale à partir des années 1880. Alors même que les premiers volumes de l'édition de Bernard et Bruel ont été publiés, des pièces et des

du Cabinet des chartes : les travaux de Louis-Henri Lambert de Barive dans les archives de Cluny (v. 1770-v. 1790)" ; Madeleine Oursel-Quarre. "A propos du chartrier de Cluny". In : *Annales de Bourgogne Dijon.* T. 50. 198. 1978, p. 103-107

145. Dominique Iogna-Prat. "La confection des cartulaires et l'historiographie à Cluny (XIe–XIIe siecles)"'. In : *Les cartulaires. Actes de la Table ronde, op. cit* (1993), p. 27-44

146. Bruel et Bernard, *Recueil des chartes de l'abbaye de Cluny : 802-954*, tome I, Avant-propos ; Jean Richard. "La publication des chartes de Cluny". In : *A Cluni : congrès* (1950) ; Harmut Atsma et Jean Vezin. "Autour des actes privés du chartrier de Cluny (Xe-XIe siècles)". In : *Bibliothèque de l'Ecole des Chartes* 155.1 (1997), p. 470-471

147. Léopold Delisle. *Le cabinet des manuscrits de la bibliothèque nationale : étude sur la formation de ce dépôt [...] avant l'invention de l'imprimerie.* T. 1. Imprimerie nationale, 1868, p. 563-564

148. Richard, "La publication des chartes de Cluny"

149. Sébastien Barret. "La mémoire et l'écrit : l'abbaye de Cluny et ses archives (Xe-XVIIIe siècle)". In : *Bulletin du centre d'études médiévales d'Auxerre/ BUCEMA* 13 (2009), p. 387-390

150. Atsma et Vezin, "Autour des actes privés du chartrier de Cluny (Xe-XIe siècles)" ; Richard, "La publication des chartes de Cluny"

séries continuent à arriver, notamment le cartulaire E. Ces livres, volumes, séries voire manuscrits isolés, parties d'une collection organique aujourd'hui perdue, constituent le premier fonds catalogué de Cluny.

Le chartrier ne souffre pas seulement des découpages et des filtres imposés par son hasardeuse histoire, mais également d'autres, de nature éditoriale. La première et la plus évidente est la clôture de l'édition des actes à l'an 1300, décision prise par A. Bernard qui destinait les chartes postérieures au XIIIe siècle au volume 7 de son édition. Bien que les chartes datées des XIVe-XVIIIe siècles soient minoritaires et le latin résiduel, l'ensemble constitue une importante série d'actes et de chartes en moyen français concernant diverses institutions.

Un second filtre est imposé par la décision du comité de publication du Recueil de ne pas rééditer les chartes figurant déjà dans la *Bibliotheca*[151] et dans le *Bullarium*[152]. En effet, ces deux éditions qui contiennent principalement des privilèges et des bulles ont été publiées au XVIIe siècle et elles ont fait l'objet de rééditions contemporaines à celle de Bernard et Bruel. Dans ces cas, l'édition du Recueil effectue un renvoi bibliographique aux deux éditions mentionnées.

Enfin, la mort d'A. Bruel a empêché la réalisation du septième et dernier volume qui devait contenir une série de suppléments indexés : le pouillé, les tables de concordance, la bibliographie et surtout les index toponymiques et onomastiques. Malgré les multiples reprises ce volume n'a jamais vu le jour et il n'a été que partiellement réalisé, dans les années 70 à l'Université de Münster. L'absence de ces supports d'indexation laisse l'édition orpheline d'instruments de recherche essentiels pour l'exploration de ce vaste corpus qui contient une énorme quantité de noms, de références internes et de séries documentaires[153].

2.1.3 Composition du corpus.

Le corpus se trouve donc constitué de trois groupes documentaires : les originaux, les cartulaires et les copies modernes, qui sont conservés pour la plupart dans quatre collections ou fonds d'archives : la Collection Bourgogne, le fonds de Nouvelles acquisitions latines et la Collection Moreau, toutes trois à la Bibliothèque nationale.

Originaux :

La Bibliothèque nationale garde la plupart des originaux. À côté de la collection Bourgogne, il existe d'autres petits ensembles repartis dans les différents fonds :

- Collection Bourgogne : constitué de 15 volumes qui contiennent des actes rédigés entre 813 et 1693. Du tome 75 au tome 85 : 613 pièces originales. Du 86 au tome 90, figurent des copies. S'y ajoutent diverses séries inégalement réparties dans les fonds de Manuscrits latins (LAT) et Nouvelles acquisitions latines (NAL) :

NAL

— 2272 : Collection de 102 lettres écrites par les abbés (1263-1561)

151. Marrier, *Bibliotheca cluniacensis*

152. Pierre Simon. *Bullarium sacri ordinis cluniacensis*. Lyon : Antonium Jullieron, 1680

153. Oursel-Quarre, "A propos du chartrier de Cluny" ; Atsma et Vezin, "Autour des actes privés du chartrier de Cluny (Xe-XIe siècles)" ; Maria Hillebrandt et al. "À la recherche de personnes perdues..." In : *Médiévales* (1991), p. 21-25

— 2154 : 88 chartes originales, Xe siècle.
— 2163 : 9 chartes originales, Xe siècle.
— 2265-2269 : 274 chartes ou pièces, la plupart originales, du XIIe au XVIIIe siècle.
— 2274 : Chartes provenant des archives de Cluny concernant le prieuré de Charité-sur-Loire du XIIe au XVIe siècle.

LAT
— 5461 : Collection de chartes du 1221 au 1455
— 8989 : privilèges accordés au Saint-Siège par les empereurs Frédéric Ier et Henri VI. XIIe siècle
— 17088 : 17 pièces sur parchemin, de 1079 à 1722
— 17715 : Chartes des archives de Cluny.

Copies :

La collection de copies par Lambert de Barive est intégrée dans les volumes 1 à 273 de la Collection Moreau de la Bibliothèque nationale. Elle est constituée d'environ cinq mille actes copiés vers 1790.

Il existe aussi, dans ce même établissement, d'autres copies concernant Cluny extraites du Grand Trésor, principalement des tableaux et des comptes, il s'agit des manuscrits latins 12823, 12740, 12768, 5214.

Les cartulaires

Les cartulaires divisés en lettres de A à E constituent la principale source de l'édition de A. Bernard et A. Bruel. L'historiographie montre clairement que les trois premiers (ms. NAL. 1497-1498 et 2262) font partie du même ensemble et sont conçus et commencés vers 1065-1080 [154]. Ils contiennent les actes produits sous les sept premiers abbés Cluny dans un arc temporel qui va de 825 (avant la fondation de l'abbaye) jusqu'à 1120. Alors que les cartulaires A et B contiennent des documents très similaires, essentiellement des actes privés, le cartulaire C (compilé vers 1095) recueille des documents émanant d'autorités, essentiellement des diplômes, privilèges et bulles, complétant ainsi le corpus des droits de l'abbaye [155]. Les deux autres, dénommés cartulaires D et E (ms NAL 2262 et LAT 5458) sont copiés l'un vers le milieu du XIIIe siècle, l'autre vers la fin de ce siècle, avec une intention d'archivage différente. D'une part, le volume se place dans la continuité de la compilation d'actes du cartulaire C, et, d'autre part, les moines visent à produire un cartulaire thématique indexé. Il comporte de ce fait des séries documentaires beaucoup plus variées contenant actes, diplômes, bulles et lettres.

• *Cartulaire A* : plus de 1300 chartes. Le cartulaire commandé probablement par Odilon garde des copies des chartes datées depuis le IXe siècle, même s'il a été réalisé dans le XIe

154. Iogna-Prat, "La confection des cartulaires et l'historiographie à Cluny (XIe–XIIe siecles)"' ; Harmut Atsma et Jean Vezin. "Gestion de la mémoire à l'époque de saint Hugues (1049-1109) : la genèse paléographique et codicologique du plus ancien cartulaire de l'abbaye de Cluny". In : *Histoire et archives* 7 (2000), p. 16-22 Isabelle Rosé. "Panorama de l'écrit diplomatique en Bourgogne : autour des cartulaires (XIe-XVIIIe siècles)". In : *Bulletin du centre d'études médiévales d'Auxerre/ BUCEMA* 11 (2007) ;

155. Barbara Rosenwein. *Cluny's immunities in the tenth and eleventh centuries : images and narratives*. Lit, 1998

siècle et fini au XIIe siècle (abbatiat de Pons).

D'après Bernard le cartulaire se trouve à la fois divisé en 4 sous-cartulaires selon l'abbé (voir figure 2.2) [156] :

- Bernon (910-927) . 29 actes. Folios 8 à 33.
- Odon (927-942). 189 actes. 4 folios avec des tableaux, puis folio 37 à 75.
- Aimard (942-954). 284 actes. 7 pages avec des tableaux. Folio de 82 à 143.
- Maïeul (954-995). 834 actes. Folios 162-307. Tableaux du folio 144 au 161.
- 6 pièces complémentaires : bulles, recensements, privilèges.

• *Cartulaire B* : 301 feuillets en parchemin. Rédigé probablement entre la fin du XIe et le début du XIIe siècle, et comprenant des suppléments du XIIIe et du XIVe siècle. Il se trouve également divisé selon les trois abbés de Cluny suivants :

- Odilon (994-1049). 805 aces. Folios 2-129
- Hugues (1049-1109). 736 actes. Folios 130-276
- Pons (1109-1122). 42 actes. Folios 277-291
- Il est remarquable que toutes les chartes du temps de Pierre le Vénérable soient manquantes.
- Un supplément de 35 numéros ajoutés aux feuillets blancs à la fin du XIIe siècle.

• *Cartulaire C* : 68 feuillets avec des actes datant entre 867-1095. Rédigé vers 1095 ou début du XIIe siècle.

- Supplément de 10 pièces ajoutées probablement au XIIIe siècle : items : 57, 58, 137, 138, 150-155

• *Cartulaire D* : 156 feuillets. Dernier quart du XIIIe siècle. 559 chartes au total.

- Bulles, diplômes et préceptes et lettres d'origine civile datant de 888 à 1298.

• *Cartulaire E* : : manuscrit 5458 de la Bibliothèque nationale. 285 feuillets répartis en 36 cahiers. Rédigé vers la fin du XIIIe siècle.

• *Cartulare E bis* : Copie incomplète du cartulaire E faite au XIVe siècle.

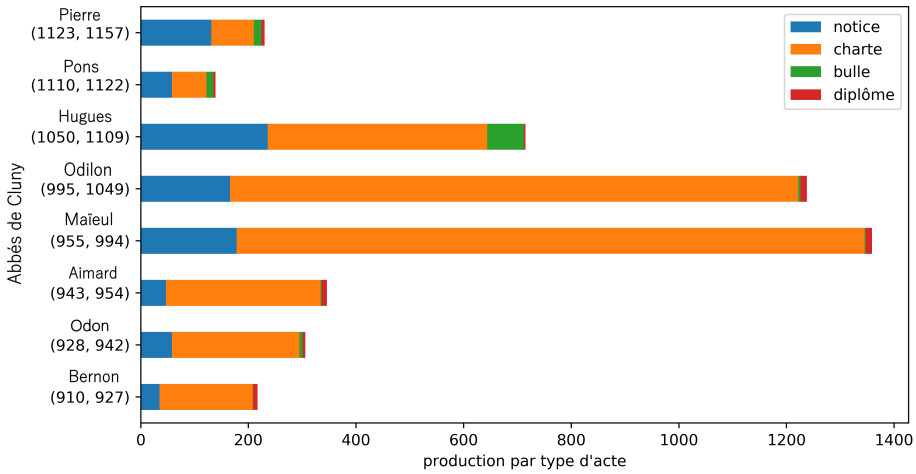

FIGURE 2.2 – Production d'actes dans les cartulaires de Cluny répartis par abbatiat. L'analyse statistique inclut les actes précisément datés et ceux datés dans une fourchette

156. BRUEL et BERNARD, *Recueil des chartes de l'abbaye de Cluny : 802-954*

2.1.4 Les typologies documentaires

La diplomatique a consacré d'intenses débats à définir ce qu'est un acte [157]. Les diplomatistes s'intéressent particulièrement à clarifier les mécanismes sous-tendant la construction d'un acte en tant qu'objet matériel et produit intellectuel. Les actes peuvent provenir d'une chancellerie royale ou pontificale, du notariat civil ou de la main d'un clerc dans un village ; ils peuvent comporter de longs préambules ou une simple invocation, des dizaines des témoins ou un seul, des *signa* ou des sceaux pendants, une date précise mentionnant lieu, mois, année, jour de la semaine ou aucun de ces éléments. Ils peuvent être rédigés dans un style subjectif, comme c'est le cas des chartes, ou objectif, dans le cas des notices. En tant que titre de propriété, ils peuvent avoir leur origine dans différents types d'action juridique : une donation, une vente, un échange, un procès ou tout à la fois. Enfin, ils peuvent être des originaux recueillis dans un chartrier ou insérés dans d'autres séries, ou être conservés en tant que copies contemporaines sous la forme de cartulaires, ou bien encore sous forme de copies modernes ou de regestes.

Puisque l'acte s'insère dans un carcan juridique, chaque réalisation dans sa structure s'avère le produit d'une négociation entre un formulaire proposé et une rédaction par le scribe où l'importance de la mémoire est primordiale. Il peut s'avérer compliqué d'établir un classement suivant la nature juridique de l'acte. Aux chartes privées peuvent être opposées les chartes publiques, mais une charte est-elle privée quand c'est une autorité publique qui l'a validée ? À l'opposé, une charte est-elle publique quand s'agit de l'affaire privée d'un roi, évêque ou pape ? Nous nous contentons ici du consensus défendu par quelques spécialistes [158], selon lequel les actes privés sont ceux qui portent des signes de validation autres que royaux, pontificaux ou d'une autorité urbaine. Nous allons différencier brièvement les types d'actes - et d'actions juridiques - trouvés dans notre corpus selon le classement fourni dans notre base de données numérique.

Les actes fonciers

L'acte - ou charte - privé sert d'instrument écrit qui valide l'existence d'un fait juridique, ce qui explique pourquoi il prend généralement la fonction d'un titre de propriété ou d'une preuve de droits sur un bien immobilier [159]. Les dons par des membres de l'aristocratie de biens fonciers à l'abbaye clunisienne sont les plus

157. voir à ce sujet Georges Tessier. *La diplomatique (3 e éd.)* 1966 ; Arthur Giry. *Manuel de diplomatique : Diplomes et chartes.-Chronologie technique.-Éléments critiques et parties constitutives de la teneur des chartes.-Les chancelleries.-Les actes privés.* T. 1. Paris, Hachette, 1894 ; Benoît-Michel Tock. "L'acte privé en France, VIIe siècle-milieu du Xe siècle". In : *Mélanges de l'école française de Rome* 111.2 (1999), p. 499-537

158. ibid. ; Laurent Morelle. "Pratiques médiévales de l'écrit documentaire". In : *Annuaire de l'École pratique des hautes études (EPHE), Section des sciences historiques et philologiques. Résumés des conférences et travaux* 139 (2008), p. 368-371 ; Alain de Boüard. *Manuel de diplomatique française et pontificale.* Picard, 1948, p. 12-35

159. Robert-Henri Bautier. "Olivier Guyotjeannin, Jacques Pycke et Benoît-Michel Tock.—Diplomatique médiévale. Turnhout, Brepols, 1992". In : *Cahiers de civilisation médiévale* 38.152 (1995), p. 285-310

abondants dans le corpus documentaire. Parmi ceux-ci, les donations *pro anima*, faites du vivant du donateur, *post obitum* ou au profit des membres de la famille, sont particulièrement nombreux. Les donateurs vivants peuvent imposer certaines restrictions à leur don ; il peut s'agir d'un don viager, ou d'un don partiel en échange d'une part d'usufruit, plus rarement d'un contrat précaire (*precaria*) [160]. Les ventes, les échanges ou les arrentements de terres passés entre un particulier et l'abbaye sont également courants, ceux entre deux propriétaires privés le sont moins [161]. Dans ce cas, le document est conservé, sous la forme de *munimina*, parce qu'il concerne l'abbaye directement ou indirectement. [162] et permettent d'établir l'historique de transferts d'une possession.

Ventes et échanges ne sont pas étrangers au corpus. L'abbaye peut avoir recours à l'achat de terrains afin d'accroître certains espaces lui appartenant. Dans certains cas, une vente peut être réalisée sous la forme d'une donation ou d'un échange et un transfert de multiples possessions peut être réalisé par une donation et une vente dans le même acte. Les actes d'échanges entre privés et l'abbaye sont nombreux et ils prennent habituellement la forme d'une notice et comportent des formules stéréotypées (*Placuit atque convenit inter .... commutare ; trado atque transfundo*), et dans le cas de ventes (*ego ... vendo / vendimus tibi / vobis ... et accipimus de vobis precium*).

Quelques actes classés comme fonciers traitent d'autres types de faits juridiques, par exemple, la renonciation à des droits sur la propriété ou de serfs en faveur de l'église, ce qui donne lieu à une *notitia werpitiones*. Si presque tous les actes portent sur des affaires foncières, on peut également trouver des actes portant sur d'autres biens ou instruments monétaires : prêts en argent, impignorations, loyers annuels de moulins, forges, outils, champs, obligations, confiscations, mises en gage, etc. La figure 2.3 montre la répartition des actes privés par type d'action juridique.

Les notices

La distinction entre charte et notice est habituellement portée par l'usage de la première personne ou de la troisième personne, développant des styles rédactionnels respectivement objectifs ou subjectifs. Mais dans certaines occasions une hybridation des deux styles est attestée dans les actes. La notice présente une version abrégée, et plus directe que la charte, laissant de côté les multiples clauses et la plupart de l'appareil protocolaire. Elle notifie une action juridique en adoptant l'outillage juridique minimal pour établir la validité de l'écrit dont le style direct, le protocole simple, un dispositif succinct et la présence d'une clause de sanction sont les éléments les plus communs. Il est par ailleurs naturel que les notices concernent un nombre élevé

160. « transfert qui mettaient une tenure à la disposition d'un laïc, qui abandonnait en échange une partie de ses biens, et s'acquittait chaque année d'un cens recognitif. », Michel Parisse, article « Précaire » du Dictionnaire du Moyen Âge, Claude Gauvard, Alain de Libera, Michel Zink (dir.), PUF, coll. « Quadrige », 2002

161. Bernard Vigneron. "La vente dans le Mâconnais du IX e au XIII e siècle". In : *Revue historique de droit français et étranger (1922-)* 36 (1959), p. 17-47

162. Didier Méhu. "Paix et communautés autour de l'abbaye de Cluny (Xe-XVe siècle)". Thèse de doct. Lyon 2, 1999, p. 17-41 ; Iogna-Prat, "La confection des cartulaires et l'historiographie à Cluny (XIe–XIIe siecles)"'

d'actes juridiques relevant d'autres affaires que la donation — ventes et échanges, ou bien encore déguerpissements—, car la notice ne transmet pas en général le cadre idéologique qui peut être mobilisé dans une donation [163].

Les actes de justice

Jusqu'à la fin du Xe siècle on peut trouver quelques actes et notices rédigées à propos de conflits entre propriétaires de terres ou entre Cluny et un tenancier. À partir de ce moment-là, ce genre de documents disparait pratiquement du corpus. En revanche, il existe des copies de procès civils et de conflits de juridiction dans la série H de la Collection Bourgogne, qui résultent probablement d'un registre beaucoup plus vaste [164]. D'autre part, des confirmations de droits fonciers, des renonciations de droits et de biens, la restitution de droits comme de prêts, et des engagements de paiement, apparaissent régulièrement tout au long du corpus de la fin du IXe siècle à la fin du XIIIe siècle. Comme dans le cas des ventes et des échanges, ce type d'actes peut s'avérer complexe à classer car l'affaire est réglée moyennant différentes aliénations ou acquittements sous la forme de donations, ventes, déguerpissements, reconnaissances de dettes, dont la rédaction apparaît ensuite dans l'acte.

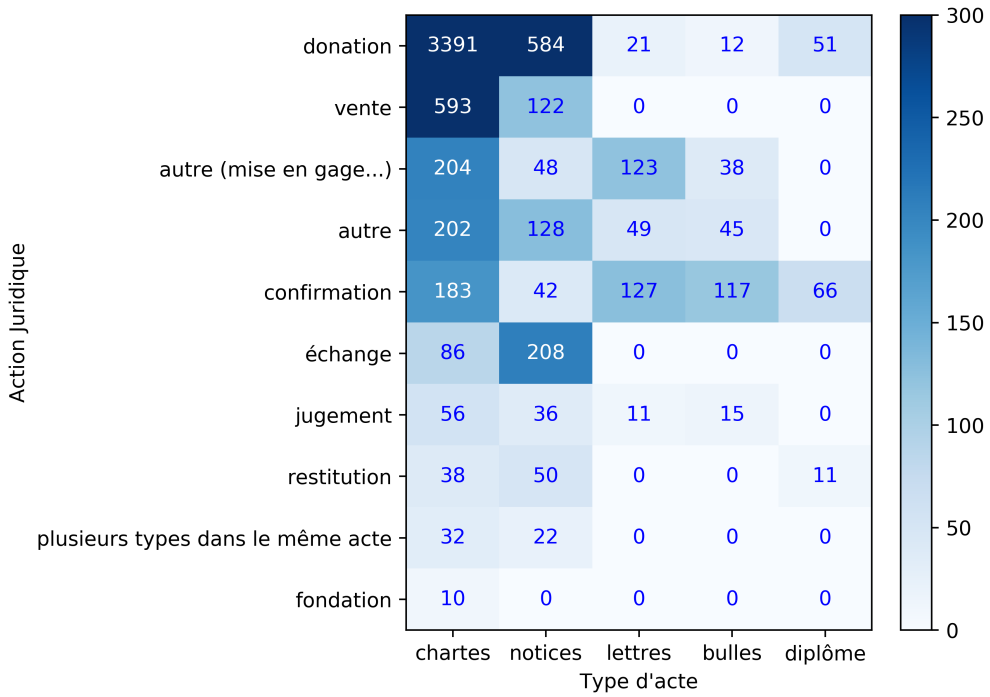

FIGURE 2.3 – Répartition des types d'acte selon l'action juridique

163. Laurent MORELLE. "Instrumentation et travail de l'acte : quelques réflexions sur l'écrit diplomatique en milieu monastique au xie siècle". In : *Médiévales. Langues, Textes, Histoire* 56 (2009), p. 41-74 ; BAUTIER, "Olivier Guyotjeannin, Jacques Pycke et Benoît-Michel Tock.—Diplomatique médiévale. Turnhout, Brepols, 1992"

164. MÉHU, "Paix et communautés autour de l'abbaye de Cluny (Xe-XVe siècle)"

Les diplômes

Les actes de confirmation des droits prennent la forme de préceptes lorsqu'ils émanent de l'autorité royale et de privilèges lorsqu'ils sont d'origine pontificale [165]. Ces deux types d'actes ont été rangés dans la catégorie de diplômes dans le classement du CBMA. Bien que le traitement intellectuel de ces actes soit semblable à celui des actes privés, le formalisme du discours exprimé dans la prolixité des formules introductives et dans l'utilisation d'un vocabulaire propre montre une plus grande complexité. Les préceptes justifient l'action royale dans le préambule et ouvrent le dispositif avec le verbe *concedo* préféré à *dono* qui est accompagné de clauses finales invoquant des notions rarement identifiables dans les actes privés : *iussio, perennisatio, auctoritas*. Les préceptes concernent pour la plupart des concessions à l'abbaye de terres dégagées de la fiscalité royale [166]. Ce type de concessions, comparées aux dons privés, sont normalement *totam cum omnibus* et impliquent le don de terres dans un rayon beaucoup plus étendu, dans certains cas une *villa* entière. Il est une pratique commune que la confirmation ou la restitution de donations ou immunités passées en cas de conflit de juridiction, donne lieu à l'émission d'un nouvel acte.

Les privilèges pontificaux concernent quant à eux les droits juridictionnels sur d'autres églises et bâtiments dont l'obéissance, la tutelle et la possession dépendent de l'abbaye. Le contrôle d'une église ou d'un monastère concerne également leurs biens et leurs terres, étendant ainsi l'emprise patrimoniale et spirituelle de l'abbaye sur un plus vaste territoire. Comme dans le cas des préceptes, les privilèges confirment généralement des dons antérieurs ou comportent l'autorisation de construire des édifices pieux.

Les préceptes royaux ont une périodicité très précise ; ils sont expédiés à peu près une fois tous les quatre ans à partir de l'an 813 (item 2 du cartulaire A) jusqu'à environ l'an 1000 où ils deviennent rares. Les privilèges pontificaux, eux, sont plus rares et tardifs, à l'exception des chartes de fondation, ils commencent à apparaître au milieu du Xe siècle et disparaissent pratiquement au début du XIIIe siècle.

Lettres ou bulles

Sous la catégorie lettres (aussi appelées *epistoles*), sont classes les missives souscrites par les autorités religieuses et les actes publics qui adoptaient une forme épistolaire. Lorsqu'elles adoptent une forme solennelle (*litterae solemnes*) elles sont alors dénommées comme bulles. Ces documents souscrits par les papes et évêques concernant Cluny ont pour objet un mandat, des instructions d'ordre ecclésiastique ou des réponses à des plaintes lorsque le conflit concerne l'autorité religieuse. Les lettres *pro monachis clunicensibus* sont fréquentes. Le pape intervient en faveur de l'abbaye dans des litiges juridiques, rappelant, confirmant et faisant respecter les privilèges accordés à Cluny. Les bulles émanant de la chancellerie pontificale sont normalement émises pour la confirmation des privilèges et possessions de Cluny ou de ses prieurés, ou pour le placement de certains monastères et églises sous la domination de l'abbaye.

165. Boüard, *Manuel de diplomatique française et pontificale* ; Bautier, "Olivier Guyotjeannin, Jacques Pycke et Benoît-Michel Tock.—Diplomatique médiévale. Turnhout, Brepols, 1992"

166. Tock, "L'acte privé en France, VIIe siècle-milieu du Xe siècle"

Certaines de ces lettres étaient attachées à des privilèges, notamment en ce qui concerne la confirmation des possessions et des droits. Enfin, dans d'autres cas, il s'agit de plaintes et de demandes d'intervention de divers ordres : spirituels, économiques ou dogmatiques.

Autres documents

Dans notre corpus, sous la rubrique "autres" sont regroupés certains documents qui, en dépit d'une tradition manuscrite similaire aux précédents, relèvent d'un traitement juridique différent et, dans la plupart des cas, il s'agit de documents uniques. On trouve par exemple les testaments des abbés de Cluny et quelques testaments d'individus léguant tous leurs biens à l'abbaye. Il existe des copies d'accords passés entre les autorités religieuses et l'abbaye concernant des conflits fonciers ou des droits (des *vidimus* sont également conservés), ainsi que des listes, probablement à usage interne pour la gestion économique des biens, des bâtiments et des produits vendus ; quelques recensements et quelques pancartes compilant le résumé d'actes.

2.1.5 Formalité et formule dans les actes du cartulaire.

L'importance des cartulaires dans l'histoire médiévale est bien connue, car ils rassemblent des transcriptions de chartes originales et des copies des divers types d'actes dont nous venons d'esquisser la typologie. Mais les cartulaires ne rassemblent pas l'exhaustivité des documents, et leur fonction n'est pas de compiler tous les actes faisant titre de droits. La confection du cartulaire inclut le plus souvent la sélection de certains documents et le refus d'autres. Les dynamiques de ce choix nous échappent habituellement, mais il semble acquis que ce processus d'assemblage relève en partie d'une intention mémorielle [167].

À partir des comparaisons entre les actes originaux et leurs versions compilées par les cartularistes, la recherche a montré que ces derniers restent en général très fidèles au contenu de l'acte, mais que par contre ils développent un travail consciencieux de correction de la langue et, dans quelques cas, d'abréviation - ou d'allongement - de descriptions et de formules prolixes, notamment dans les protocoles et les inventaires de biens-fonds.

Malgré cette transformation du document, la valeur du cartulaire comme source testimoniale n'est pas remise en cause. Ces actes sont des documents rédigés suivant des modèles, ce que nous appelons le formulaire, qui leur fournissent une structure plus ou moins stéréotypée. Le formulaire, par sa nature séquentielle et modulaire, intègre une chaîne d'énoncés destinés à donner validité aux actes et à formaliser des modèles qui prennent en compte la nature de l'action juridique.

Au moment de rédiger une charte, les scribes avaient devant eux, s'ils ne les connaissent pas déjà par cœur, des modèles nécessaires à la configuration l'acte. Toutefois, si le formulaire propose une structure fixe, les variations trouvées dans la

167. Au sujet des cartulaires-chronique on peut consulter : Olivier Guyotjeannin et al. *Les cartulaires : actes de la Table ronde organisée par l'École nationale des chartes et le GDR 121 du CNRS*. T. 39. École nationale des chartes, 1993 ; Robert Fossier. *Cartulaire chronique du prieuré Saint-Georges d'Hesdin*. T. 32. Éditions du Centre national de la recherche scientifique, 1988

rédaction de l'acte sont nombreuses, parfois très personnelles, car le scribe peut opérer des adaptations selon les circonstances qui peuvent être explicitées lors de la rédaction de l'acte.

Ainsi, le formulaire met en forme un discours dont la construction est liée au contexte de l'écriture, en présentant une forme assez structurée, fondée sur des énoncés répétitifs, mais qui autorisent un certain niveau de malléabilité. Dans cette structure, les noms de personnes, de lieux, d'institutions ecclésiastiques ou civiles, les dates, les titres, etc., constituent également des éléments spécifiques qui apportent une grande individualisation. Le formulaire structure l'expression écrite notariale et l'entité nommée spécifie et individualise le document. Néanmoins, les formes d'écriture des entités nommées peuvent se montrer assez hétérogènes. En dehors des contraints de la flexion latine, qui propose une version différente du nom selon sa fonction dans la phase, le manque de règles orthographiques - on écrit les noms de personnes et de lieux comme ils se prononcent ou comme on les a vu écrits. -, les abréviations et enfin les erreurs ou les actualisations dans la copie et la translation, peuvent conduire à une multiplication des versions graphiques d'une même entité. Les variantes possibles ne sont pas infinies, mais suffisamment nombreuses pour conduire, parfois, à sérieux problèmes d'identification.

Par conséquent, la récupération des modules constituant chaque document - les parties du discours - tout comme des éléments spécifiques qui lui confèrent son identité -les entités - s'avèrent indispensables pour tout dispositif de recherche visant à exploiter numériquement des bases de données à partir de cartulaires.

Ces deux caractères que nous venons d'évoquer concernant les actes peuvent être plus éclairés à partir de deux exemples de chartes. Daté entre 994 et 1007, cet acte de donation à l'abbaye de Cluny commence par un bref protocole (invocation et notification), suivi immédiatement par le dispositif indiquant l'objet et le but du don et les participants à l'acte [168].

> *In nomine Verbi incarnati. Cuncti noverint populi, quod ego Adalguis dono Domino Deo et sanctis apostolis ejus Petro et Paulo, ad loco Cluniaco, ubi domnus et reverentissimus Odilo abbas magis videtur preesse quam prodesse,*

168. CBMA 2609. "Au nom du Verbe incarné. Que tout le monde sache que moi, Adalgis, je donne au Seigneur Dieu et à ses saints apôtres, Pierre et Paul, en lieu et place de Cluny, où le maître et vénérable abbé Odilon semble plus gouverner que pour en tirer avantage, pour la rédemption de tous mes péchés, et pour que le Seigneur m'assiste dans le dernier jour du jugement.

Il y a donc ces choses dans le comté de Mâcon, dans la *villa* de Tasiaco : ceci est, en premier lieu, un curtilus (petit domaine) avec une maison, une vigne et un champ ; et il a des limites de trois côtés, d'un côté la terre de Rodulfus, d'un autre côte la terre de Seguinus, du troisième côte la terre des Francs et la voie publique.

Un autre champ ailleurs qui a des limites de quatre côtés, d'un côté la terre de Rodulfus, d'un autre d'un autre côté la terre de Fulchardus et d'un autre la voie publique. Et deux petits morceaux de prairies ; ils ont des limites de trois côtés, d'un côté la terre de Sanctus Quintinus, d'un autre côté la terre de Sanctus Martinus et de l'autre côté la voie publique.

Les recteurs de Cluny peuvent faire ce qu'ils veulent, à partir d'aujourd'hui, avec ce don. Cependant, si quelqu'un veut établir une dispute juridique contre ce don, laissez la colère de Dieu tomber sur lui et le submerger vivant dans l'enfer s'il ne revient pas à l'amendement.

Signum (Signum) Adalgis, femme, qui a ordonné que cet acte soit commis et a demandé de le valider. S (ignum) Ingelelmi, S. Arlei, S. un autre Arlei. S. Bernardi. S. Ebrardi."

> *pro redemptione omnium peccatorum meorum, ut Dominus dignetur auxiliare in extremi diem judicii.*

Ensuite vient une description détaillée des biens-fonds qui font l'objet de la donation. Les descriptions impliquent toujours une localisation moyennant une référence au système hiérarchique de l'organisation foncière, dans ce cas, *comitatus, villa, terra*, à l'intérieur desquels se trouvent enserrées les biens fonciers classés en parcelles spécialisées, ici *campum, vinea, curtile, pratum*, etc. Le scribe a aussi inclus une *terminatio*, c'est-à-dire, la détermination des limites des pièces de terre données par rapport au voisinage.

> *Sunt ergo ipsas res in comitatu Matisconensi, in villa Tasiaco : hoc est in primis unum curtile cum domo et vinea, et campum ; et habet fines de tres partes, de una terra Rodulfo, de alia Seguini, de tertia terra francorum, de quarta via publica. In alio loco alium campum ; habet fines de quattuor partes, de una parte terra Rodulfo, de alia terra Fulchardo, de tertia via publica. Et in aliis duobus locis duabus peciolas de prato ; habent fines de tres partes, de una parte terra Sancti Quintini, de alia Sancti Martini, de alia via publica.*

Le dispositif se referme par une *sanctio* contenant des clauses pénales destinées à renforcer et à garantir les actions juridiques :

> *De hanc donationem faciunt rectores de Cluniaco quicquid facere voluerint ab hodierno die. Si quis vero ullus homo donationem hanc contrariare voluerit aliquam litem, ira Dei incurrat super illum, et sit demersus in infernum vivus, si ad emendationem non venerit.*

Il existe enfin un eschatocole à deux parties avec la souscription du donateur et les signatures des témoins, contenant leurs noms et un "S." pour *signum* (signature).

> *S. Adalguis femina, qui hanc cartam fieri jussit et firmare rogavit. S. Ingelelmi. S. Arlei. S. item Arlei. S. Bernardi. S. Ebrardi.*

Ce modèle tripartite : protocole, dispositif et eschatocole est généralement bien identifiable dans les actes concernant les affaires les plus communes (donations, ventes, échanges, etc.). Comme on peut le constater l'appareil rédactionnel n'est pas très complexe ; le style n'est en général pas laconique, mais il est néanmoins assez direct et informatif. Ces actes ne comprennent habituellement pas de détails personnels ou narratifs, ce qui donne l'impression que nous sommes confrontés à une source répétitive et modulaire. Cette sécheresse est une nécessité dans les documents qui visent à devenir des preuves de droits.

Dans le cas des actes concernant des acteurs sociaux plus importants ou dans les cas des privilèges, ou les actes de chancellerie, même si on suit un modèle qui répond à une teneur similaire, le corpus de formules et les parties du discours peuvent se multiplier et sont l'occasion d'exprimer des motivations idéologiques en utilisant des citations d'autorité, passages bibliques ou commentaires d'ordre moral.

Par exemple, dans ce *praeceptum* de confirmation de donations de Robert II et de son fils Hugues à Cluny daté entre 1017 et 1025 le protocole présente deux parties : une *invocatio* - le patronage de la trinité - et une *subscriptio* - identité de l'auteur de l'acte - assez simples [169] :

> *In nomine sanctæ et individue Trinitatis. Rodbertus et Hugo filius suus, gratia Dei Francorum reges, omnibus sub nostro imperio militantibus, pacem et salutem.*

S'ensuit un long préambule qui, malgré son style moralisant, est un panégyrique où le rédacteur partage la charge idéologique avec la *potestas* royale. Dans ce préambule est invoquée une série de valeurs universelles (*dignitas, iustitia, charitas*) dont l'application peut être reconnue dans le contexte particulier de l'affaire en cours et qui servent de fondement à la validité du document :

> *Si precibus servorum Dei pro sancte Dei ecclesie statu necnon et eorum utilitatibus assensum prebemus, regiam in omnibus conservamus dignitatem. Quapropter necesse est ut quos intercessores pro nobis ad Deum premittimus, eorum petitionibus aurem accomodemus. Ita enim fiet ut et preces quas pro nobis effundunt ad aurem Dei omnipotentis ascendant, juxta illud Psalmographi : «Desiderium pauperie exaudivit Dominus, et petitiones cordis eorum audivit auris tua.» Idcirco omnium sancte Dei ecclesie fidelium nostrorumque noverit industria presentium scilicet et futurorum, quia venerabilis pater Odilo et fratres sub eo omnipotenti Deo militantes adierunt nostram regiam majestatem supplices, pro quibusdam rebus et potestatibus Sancti Petri nos interpellantes, quatinus ipsis regali auctorite faventes preceptum concederemus eorum justis petitionibus obaudientes :*

Finalement le dispositif s'ouvre avec le verbe *concedo* et continue par l'énumeration et la description des biens et des terres objets de la donation avec un style assez proche de celui utilisé dans les actes privés :

169. CBMA 3282. "Au nom de la sainte et indivisible Trinité. Rodbertus et son fils Hugo, rois des Francs par la grâce de Dieu, à tous ceux qui servent sous notre règne, la paix et la santé. Si nous offrons l'approbation pour les prières des serviteurs de Dieu pour l'état de la Sainte Église de Dieu et aussi pour ses bienfaits, nous préservons dans toute la dignité royale. Pour cette raison, il est nécessaire que nous écoutions les demandes de ceux que nous envoyons comme intercesseurs devant Dieu. Ainsi, il arrivera que les prières qu'ils prodiguent pour nous montent aux oreilles de Dieu tout-puissant, selon les paroles du psalmiste : "Le Seigneur a écouté le désir des pauvres, et ton oreille a entendu les demandes de son cœur" <cfr. Psaume 9 :38>.

C'est pour cette raison que tous les fidèles de la sainte église de Dieu et les nôtres, présents sans aucun doute et futurs, savent l'application avec laquelle le père vénérable Odilon et les frères qui servent sous lui au Dieu omnipotent se sont adressé suppliants à notre majesté royale et nous ont interpellés à propos de certaines choses et prérogatives de saint Pierre, de sorte que, favorisant ceux-ci avec l'autorité royale, nous avons accordé ce précepte prêtant l'oreille à leur justes demandes ; ainsi par notre autorité royale, nous leur accordons une petite abbaye dédiée aux saints martyrs Cosme et Damien, située à côté des murailles de la ville de Chalôn, qu'autrefois le comte Hugue, l'évêque Lambertus et son père Rodbertus, par la foi testamentaire, ont transmis au beatus Pierre et au monastère de Cluny. Aussi, nous accordons . . . "

regali itaque auctoritate concedimus eis quandam abbatiolam in honore beatorum martirum Cosme et Damiani dicatam, juxta mœnia Cabilonensis urbis sitam, quam olim comes Hugo et Lambertus episcopus et pater suus Rodbertus jure testamentario tradiderunt beato Petro et Cluniacensi monasterio. Concedimus item.....

2.1.6 L'utilité de la récupération des entités nommées dans les formulaires

D'un point de vue technique, les entités nommées sont finalement des éléments à usages multiples, susceptibles d'adopter différents rôles de nature historique. En effet, l'identification d'une entité nommée peut contribuer à la saisie d'un paysage régional complexe, qu'il s'agisse du nom d'un maître du sol, d'une propriété partagée, d'un nœud d'un réseau familial, ou d'un participant d'une longue série de relations et de transactions avec l'Église. Les anciens index de lieux et de personnes des éditions érudites de chartes, élaborés minutieusement, et à l'heure actuelle automatisables, sont indispensables pour la lecture globale d'un corpus parce ils apportent des repères temporels et spatiaux. La durée de vie d'un personnage connu, - ou même inconnu - peut permettre d'établir une date ou de serrer une date à fourchette dans un document non daté. Les mouvements de la propriété foncière peuvent être suivis à partir de la vérification sur les cartes des noms de villes, des lieux-dits et des cadres territoriaux hiérarchisés ; de même un nombre important de changements dans le discours formalisé peut être expliqué par la présence de tel ou tel personnage, la proximité de tel ou tel lieu ou l'orientation idéologique de telle institution.

Dans ce panorama, l'organisation de l'information disponible dans une charte doit nécessairement passer par des références spécifiques aux lieux et aux personnes en tant que concentrateurs de données mais surtout comme une aide à l'indexation du contenu de l'acte. L'ensemble des entités d'un corpus forment ainsi des pôles qui permettent de filtrer, classer et disposer les documents selon différentes valeurs et critères : par exemple tous les documents produits par un scribe dans un certain diocèse ; ou tous les documents datés de la régence de Lothaire concernant la *villa* de Varennas ; ou les actes de donation émis par un couple de *potentes* ; ou la mesure moyenne comparée d'un *pratum* ou d'une vigne dans deux donations ; ou les prix comparés de vente de terres entre *pagi* voisins ou dans leur évolution chronologique au sein d'un seul *pagus*, et ainsi de suite. De telle sorte que, dans un niveau élémentaire de la recherche dont l'objectif est de former des mailles fondamentales d'information, les entités nommées peuvent servir, autant que possible, à construire des objets plus complexes comme des tracés de *villae*, à restituer les cadres sociaux, juridiques et économiques de la possession de terres, ou bien encore à offrir une image plus détaillée de la vie économique dans les abbayes ou de la circulation de biens dans la région. Ce panorama constitue l'un des horizons majeurs de toute recherche dite numérique. Il laisse imaginer la puissance que possède la détection et la récupération d'entités nommées pour examiner rapidement et en détail de grandes bases de données issues de la documentation médiévale.

2.1.7 Autres cartulaires utilisés

Nous savons aujourd'hui que dans la Bourgogne médiévale, entre le XIe et le XVe siècle, environ 170 cartulaires ont été produits mais une partie importante est perdue [170]. Les cartulaires de Cluny, dont nous venons de dessiner la composition, ne sont certainement pas la règle et les cartulaires qui nous sont parvenus sont généralement beaucoup plus petits ou constitués de fragments en raison des multiples interventions dans leur tradition manuscrite. Dans d'autres cas les pièces autrefois compilées dans un cartulaire nous sont parvenues par des copies modernes des XVIIIe et XIXe siècles ou organisées dans des recueils factices de pièces jugées importantes au détriment d'autres considérées comme banales. Malgré tout, la quantité de documentation disponible est exorbitante si l'on considère qu'il s'agit de cartulaires régionaux, focalisés sur cinq diocèses : Mâcon, Autun, Langres, Châlon et Nevers, qui forment les centres des domaines de Cluny, Cîteaux et Clairvaux, trois institutions religieuses très prolifiques dans la production documentaire et qui deviennent rapidement des centres de culture et de production du savoir.

Cependant, seuls une cinquantaine de ces cartulaires ont fait l'objet d'une édition érudite, qui consiste parfois en une simple transcription, sans autre appareil critique que l'introduction historique des pièces. La plupart de ces éditions — quarante-trois — ont été incluses dans la base de données CBMA et sont disponibles dans leur intégralité. Notre étude sera étendue à quatre de ces éditions : Saint-Vincent de Mâcon, Marcigny-sur-Loire, Paray-le-Monial, Saint-Marcel-de-Châlon, et à une édition d'un recueil factice, celui de l'Yonne (voir figure 2.4). Les quatre premiers cartulaires appartiennent à l'orbite de Cluny et montrent des liens directs du point de vue de la documentation, élément capital pour la réalisation d'études conjointes. Le recueil d'actes concernant l'Yonne, bien que plus éloigné de l'orbite clunisienne, constitue un des ensembles documentaires les plus complets en dehors de celui de Cluny et peuvent offrir un portrait contrasté des évolutions du vocabulaire, de la conception de l'espace et des formes rédactionnelles. Au total, les documents de ces ensembles comptent environ 4 000, ce qui en fait un ensemble de taille similaire à celui analysé à Cluny.

En ce qui concerne les quatre cartulaires : Saint-Vincent de Mâcon est un cartulaire de chapitre cathédral. Le cartulaire rédigé vers la fin du XIIe siècle a été détruit lors des troubles huguenots de 1562, et il a été l'objet d'une édition érudite réalisée au XVIIIe siècle à partir de la copie d'une copie réalisée probablement au XVIe sur le livre enchaîné du trésor de la cathédrale [171]. La copie a été découverte dans les archives préfectorales de Saône-et-Loire par son éditeur moderne Camille Ragut au cours du XIXè siècle. Plus tard, un autre exemplaire du cartulaire, conservé dans la Bibliothèque impériale permettra de collationner quelques séries de chartes. L'édition est un recueil de 633 chartes s'échelonnant, comme pour les premiers cartulaires de Cluny, du Xe au XIIe siècle. Étant donné que les transactions sont effectuées dans le même temps et le même espace que celles mentionnées dans le cartulaire de Cluny, il est habituel de trouver, dans cet ensemble documentaire, les mêmes maîtres du sol. Cependant,

170. Rosé, "Panorama de l'écrit diplomatique en Bourgogne : autour des cartulaires (XIe-XVIIIe siècles)"

171. ibid. ; Camille Ragut et Théodore Chavot. *Cartulaire de Saint-Vincent de Mâcon : connu sous le nom de Livre enchaîné*. Impr. d'É. Protat, 1864

le cartulaire de Saint-Vincent, sans doute à cause de la perte totale des originaux et parce qu'il nous est transmis par la copie d'une copie, exhibe des sérieux problèmes dans sa chronologie. Cette problématique a été abordée par l'éditeur moderne, mais parfois avec un manque de rigueur scientifique.

Les cartulaires de Marcigny-sur-Loire et de Paray-le-Monial ont été compilés par deux prieurés bénédictins [172]. Isabelle Rosé a mis en évidence pour ces deux cas, ainsi que pour Saint-Marcel, une circulation institutionnelle des modèles d'organisation d'un cartulaire, ces trois cartulaires présentant des similitudes sur plus d'un plan. Marcigny est l'un des plus anciens cartulaires de la Bourgogne, écrit vers 1095, moment crucial dans les écrits clunisiens. I. Rosé a trouvé des similitudes substantielles entre la forme et l'organisation de ce cartulaire et ceux de Cluny, pointant notamment une même vocation à servir de recueil de gestion et de mémoire [173]. Une circonstance qui se répète dans le cas du cartulaire de Paray, commencé vers la fin du XIe siècle et qui s'apparente également au modèle établi par les premiers cartulaires de Cluny. Les deux cartulaires – Marcigny et Paray –proviennent, il est vrai, de monastères de fondation clunisienne, Marcigny étant une fondation personnelle de l'abbé Hugues (1055), et ils se trouvent tout deux dans l'orbite d'influence de l'abbaye-mère, dans un rayon de 50 km. Dans ces cartulaires et celui de Cluny, il existe en fait un nombre important d'acteurs ayant des propriétés dans toute la région allant de la Petit Grosne à la Loire. Les éditions intégrées dans le CBMA comprennent 307 documents rédigés entre 1044 et 1145 pour Marcigny, et 200 rédigés entre 977 et 1315, pour Paray, bien que la plupart des actes ne soient pas datés. Les événements de la Révolution ont conduit à la disparition des deux cartulaires, dont les éditions sont basées sur des copies modernes. Marcigny, comme l'indique le titre de l'édition "*Essai de reconstitution d'un manuscrit disparu*", est le produit des travaux de collecte et de reconstruction de deux des meilleurs connaisseurs de la Bourgogne médiévale, Jean Richard (1957) et Ulysse Chevalier (1851) à partir de copies modernes du cartulaire.

Le cartulaire de Saint-Marcel-lès-Chalon, bien que faisant partie de cette même orbite documentaire clunisienne, est un peu plus tardif et présente certaines différences. Cette abbaye devint dépendante de Cluny vers la fin du Xe siècle par donation comtale et le cartulaire a dû être rédigé vers 1120 [174]. Il se compose de 119 pièces avec des documents datant principalement des Xe et XIe siècles, avec quelques diplômes du XIe siècle. L'édition moderne est une publication *post mortem* (1851) réalisée à partir des papiers laissés par Chizy de Canat, qui l'a édité à partir de 6 copies différentes réalisées au XVIIIe siècle [175]. L'ordre suivi dans ce cartulaire n'est pas chronologique mais se fonde sur l'importance institutionnelle, ce qui constitue une innovation dont

172. Franz Neiske. "Les débuts du prieuré clunisien de Paray-le-Monial". In : *Paray-le-Monial actes du colloque* (1992), p. 1-12 ; Nicolas Reveyron. "Marcigny, Paray-le-Monial et la question de la chapelle mariale dans l'organisation spatiale des prieurés clunisiens au XIe–XIIe siècle". In : *Viator (English and Multilingual Edition)* 41 (2010), p. 63-94

173. Rosé, "Panorama de l'écrit diplomatique en Bourgogne : autour des cartulaires (XIe-XVIIIe siècles)"

174. Christian Sapin. "Saint-Marcel-lès-Chalon (Saône-et-Loire), église Saint-Marcel". In : *Bulletin du centre d'études médiévales d'Auxerre/ BUCEMA* 10 (2006)

175. Marcel Canat de Chizy et Paul Canat de Chizy. *Cartulaire du prieuré de Saint Marcel lès-Châlon*. L. Marceau, 1894

l'origine peut être reliée aux cartulaires D et E de Cluny. Le classement est donc effectué en fonction de la hiérarchie des autorités, en commençant par les rois et les papes et en terminant par l'aristocratie locale.

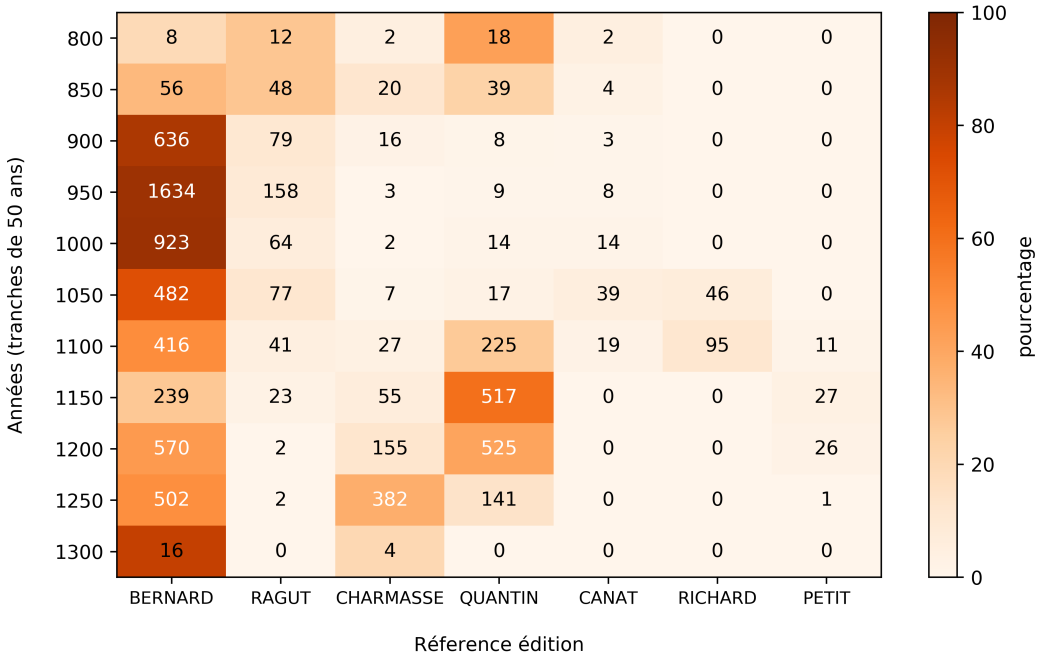

FIGURE 2.4 – Répartition chronologique des actes par édition dans notre corpus d'étude (la couleur indique le pourcentage que chaque groupe représente par rapport au total dans cette période)

Le recueil factice de l'Yonne regroupe une grande partie de toutes les chartes médiévales connues concernant ce département. La sélection est faite à partir de cartulaires d'origine monastique connus de différentes localités de l'Yonne - Auxerre et Sens avant tout. Son éditeur, Maximilien Quantin, travaille habituellement à partir de copies modernes et effectue une sélection de chartes, en retranscrivant celles qu'il considère pertinentes. Dans son édition sont inclus des actes par ailleurs édités dans le cadre de publication de cartulaires médiévaux, ou ayant été l'objet de travaux érudits. Parmi eux figure le cartulaire de l'abbaye Saint-Germain d'Auxerre, le cartulaire de l'abbaye cistercienne de Reigny, le cartulaire de l'abbaye de Molême, le cartulaire de Valvuisant, celui de l'archevêché de Sens. L'édition Quantin est l'une des rares qui contient un index topographique et onomastique et un tableau de classement en fonction de l'origine géographique de l'acte. Bien que l'ordre suivi par l'édition soit chronologique, les lacunes sont notoires dans de nombreuses séries, ce qui empêche de restaurer dans l'édition l'ordre original des actes. Le tome I, qui compte 394 actes rédigés entre le VIe siècle — des diplômes — et le XIIe siècle, présente des séries d'origine royale dont le nombre va decrescendo à partir du Xe siècle pour laisser place à de longues séries d'actes privés — donations des membres de l'aristocratie locale. Le tome II contient un total de 511 documents parmi lesquels actes, privilèges et bulles, 65 de la première moitié du XIe siècle et 446 du XIIe siècle.

2.2 Phénomènes de corpus

Le style d'écriture, la taille d'une charte, son propos peuvent varier, ce qui entraîne des changements dans les formulaires. Mais demeure un cadre formel qui offre une relative stabilité à toutes les relations établies entre clauses et entités nommées. Hormis dans les longs préambules et exposés, où l'on peut trouver, selon le type d'acte et ses participants, l'expression d'assertions de nature idéologique, les actes évitent l'inclusion de détails narratifs ou personnels. Cela peut donner l'impression que l'on est face à une source assez répétitive et construite de manière modulaire en utilisant des éléments distincts. Et si cela peut compliquer la lecture humaine - le formulaire masquant les spécificités du cas particulier - ces caractéristiques sont au contraire un point positif dans le traitement à grande échelle, parce que des hauts niveaux de régularité dans le discours limitent le nombre de possibilités offert à la décision automatique et en conséquence engendrent des résultats plus précis.

Or, si une relative stabilité dans les rapports entre formulaires et entités nommées est remarquée dans la production des actes et plus largement dans leur contexte immédiat d'écriture, certains phénomènes impactent directement les entités nommées et les usages linguistiques dépendent des changements liés au contexte historique et social de production de l'acte. C'est une banalité de le dire. L'acte, construit sur un format légal, n'est pas complètement perméable à ce genre de changements, mais les acteurs et auteurs du document, leurs intentions, leurs aspirations et leurs façons de s'exprimer peuvent y laisser une empreinte importante. Dans le même sens, la variabilité linguistique, les changements dans l'anthroponymie, l'évolution des juridictions spatiales et enfin les nouveaux rapports dans la hiérarchie sociale, qui impactent tous les formes des entités nommées et les formulations qui les contiennent, peuvent constituer de vrais défis pour la détection et la reconnaissance automatique dans les actes médiévaux. À tout cela il convient d'ajouter le contexte de conservation des documents et les vicissitudes de l'histoire des archives, qui peut contribuer à la surreprésentation de certaines structures et caractéristiques et à en cacher ou isoler d'autres, ce qui engendre un biais important pour le fonctionnement des modèles de récupération automatique.

Techniquement, effectuer des recherches à l'aide de REN (reconnaissance d'entités nommées) dans des corpus historiques n'est pas plus compliqué que de travailler avec des articles journalistiques ou des œuvres littéraires. Construire un modèle de reconnaissance dans l'un de ces trois domaines nécessite d'observer les principaux phénomènes pouvant affecter l'extraction des entités dont les contextes sociaux et historiques sont les plus importants. La littérature historique sur les cartulaires clunisiens ne manque pas, et une inspection bibliographique peut prévenir les principaux problèmes affectant la collection d'occurrences d'entités nommées. Quatre faits saillants doivent être signalés.

2.2.1 La révolution anthroponymique

Si l'on est entièrement plongé dans la récupération des entités nommées, tous les changements dans le processus de conformation des noms de personnes et dans l'évolution sociale de leur usage peuvent nous fournir des éléments précieux de nature

morphologique, concernant leur rôle comme identificateur social. Puisque le corpus bourguignon, et spécialement le corpus clunisien que nous analysons, va de la fin du IXe siècle au milieu du XIIIe siècle, nous allons en premier lieu nous confronter au processus de révolution anthroponymique en Europe occidentale qui commence au début du Xe siècle et se poursuit jusqu'au XIIIe siècle. Cette transformation est très bien décrite par des études désormais classiques [176]. De manière schématique, la dénomination par un nom unique, simplification de l'héritage romain utilisée depuis le début du Moyen Âge, a été remplacée vers la fin du Xe siècle par une nouvelle tradition utilisant un nom à deux composants ou plus. La construction de ce nom double ou triple présente différents rythmes d'évolution et de réalisation. Deux formes sont les plus fréquentes : le *nomem paternum* et les noms locatifs. Les réalisations des premiers consistent à ajouter au nom singulier une deuxième partie provenant du père ou dans un sens plus large, un nom apanage de la famille souvent par médiation d'une particule (par ex. *Hugues de Sinemuro, Letbaldus de Digonia*). Selon les régions et les traditions, cet ajout peut aussi consister en une déformation ou une adaptation du nom familier (par ex. *Rodericus Ferrandi*). Dans le cas des locatifs, il s'agit du nom d'un lieu, ville, région qui s'ajoute au nom individuel de la personne normalement par médiation de la particule "de", d'un génitif ou d'un ablatif. La relation entre la personne et le locatif peut être multiple : lieu ou région d'origine, lieu de provenance, lieu familier, lieu de propriété, etc. (par ex. *Rotbertus de Berziaco*)

Cette nouvelle tradition s'est consolidée tout au long des XIe et XIIe siècles au fur et à mesure que la production documentaire s'est consolidée et, que l'aristocratie a participé à cette dynamique scripturale. À partir du milieu du XIIe siècle, cette dénomination déjà stéréotypée a adopté des modes plus complexes faisant appel à d'autres traits dénominatifs : des références familières, des noms d'autres membres de la famille, des professions, des surnoms, des noms de fiefs, des microtoponymes et, dans d'autres cas mobilisant des périphrases, des triples noms de famille ou des locatif complexes (*Otonnis qui dicitur de Suso* ; *Grioulum filium Lafranci Caipeni*). Ce mouvement contribue à la diffusion, vers la fin du XIIe siècle, de noms personnels comportant de trois à six éléments.

La situation dans notre documentation, même si elle suit la dynamique anthroponymique générale, s'avère différente en certains points. Si presque la moitié de notre corpus date d'une période postérieure au Xe siècle, le nombre global d'entités personnelles à un seul élément est en revanche nettement supérieur au nombre de noms doubles ou d'entités complexes. Dans le formulaire, les parties plus solennelles peuvent présenter le donneur ou le vendeur avec son nom complet et, le cas échéant, ses titres, mais ces éléments ne sont pas conservés dans le reste du formulaire, qui présente une version abrégée du nom.

D'un autre côté, il existe différents niveaux d'adoption des noms doubles selon l'échelle sociale ; s'ils sont communs pour nommer les membres de l'aristocratie, ils le

176. À ce sujet voir les études de Monique Bourin et Pascal Chareille. *Genèse médiévale de l'anthroponymie moderne*. T. 2. Université de Tours, 1990 ; Patrice Beck et al. "Nommer au Moyen Âge : du surnom au patronyme". In : *Le patronyme. Histoire, anthropologie, société* (2001), p. 13-38 ; François Menant. "L'anthroponymie du monde rural". In : *Publications de l'École Française de Rome* 226.1 (1996), p. 349-363

sont moins pour dénommer les religieux, où la priorité est donnée à leur fonction ou charge. En outre, les entités nommées peuvent jouer d'autres rôles que des nominatifs : génitifs de pertinence, références de localisation, ou listes de souscriptions (avec *signa*), assez communs dans la documentation avant le XIe, et qui utilisent habituellement le nom simple. Ces listes contiennent souvent la plupart des entités personnelles d'un acte. Noms simples, noms doubles et noms complexes cohabitent parfaitement dans la documentation à partir du XIe siècle, bien que le nom simple commence à être de moins en moins utilisé. (voir la figure 2.5)

L'irrégularité croissante des noms représente un défi majeur pour notre modèle parce qu'elle oblige à définir des règles élastiques et à considérer une quantité croissante de détails. Même si les noms simples et les versions les plus communes des noms composés correspondent à la forme développée par l'immense majorité des entités personnelles, les noms complexes (environ 10 % de la totalité) amènent ce défi à son expression maximale étant donné qu'il s'agit des entités multiples qui peuvent introduire de forts bruits dans la régularité du système.

2.2.2 L'imbrication des entités nommées

En second lieu, un autre phénomène plutôt lié aux formes de dénomination dans les entités géographiques et juridiques a été rapidement repéré : l'imbrication et la superposition des entités nommées. Le mouvement général vers une plus grande complexité pour les noms de personnes, que nous venons d'esquisser, peut aussi être détecté pour les noms des lieux, mais avec un rythme particulier qui découle de trois situations documentaires principales :

(1) Il s'agit, tout d'abord, de l'apparition au début du Xe siècle du modèle formulaire de la *donatio pro anima*. La doctrine de l'aumône est le cadre idéologique dans lequel s'inscrivent les donations de propriétés à l'Église et les actes qui attestent le don suivent assez régulièrement ce modèle de charte [177]. Comme les cartulaires rassemblent des documents liés à la propriété ecclésiastique, et qu'une partie importante de ces propriétés a pour origine la donation privée, ce modèle de formulaire, dans ses différentes versions, se trouve constamment dans notre corpus. Avec le don, le donateur manifeste son intention de céder une propriété à un saint qui agit en tant qu'intermédiaire et non directement à l'Église. Cette propriété devient alors une partie des terres sous le *dominium* d'un saint, mais elle est gérée et possédée, par les effets juridiques de l'acte, par un monastère agissant en tant que personne morale. Comme l'ont montré les études d'Eliana Magnani et de Barbara Rosenwein [178], les motivations de ces donations peuvent être multiples.

Le plus intéressant dans ce panorama est le fait que ces entités, sous un même nom, jouent différentes fonctions. Dans les descriptions des biens-fonds les chartes utilisent de longues listes d'entités nommées mélangeant des éléments personnels, géographiques et institutionnels, par exemple :

177. Eliana Magnani. "Le don au moyen âge". In : *Revue du MAUSS* 1 (2002), p. 309-322 ; Anita Guerreau-Jalabert. "Caritas y don en la sociedad medieval occidental". In : *Hispania* 60 (2000), p. 27-62

178. Magnani, "Le don au moyen âge" ; Barbara Rosenwein. *To be the neighbor of Saint Peter : the social meaning of Cluny's property, 909-1049*. Cornell University Press, 2006

rebus [[Sancti Vincentii]$^{\text{PERS}}$ [Matiscensis]$^{\text{LOC}}$]$^{\text{ORG}}$;
a sero terra [[Sancta Maria]$^{\text{PERS}}$ de [Optimo Monte]$^{\text{LOC}}$]$^{\text{ORG}}$;
in honore [[Sancti Petri]$^{\text{PERS}}$]$^{\text{ORG}}$ constructam.

En outre les mêmes entités dans une même charte peuvent avoir différentes fonctions sans changer de forme :

donatio pro [sancti Petri]$^{\text{PERS}}$ ... ad terram [Sancti Petri]$^{\text{ORG}}$

Étant donné que les donations doivent décrire avec précision les limites des terres, la métonymie des frontières mélange dans un même format noms de personnes, lieux et institutions :

a mane [terra [Martino]$^{\text{PERS}}$]$^{\text{LOC}}$, a medio dia [terra [Arnulfo]$^{\text{PERS}}$]$^{\text{LOC}}$ presbiter ;
a cercio [terra [Sancti Vincencii]$^{\text{ORG}}$]$^{\text{LOC}}$

Il faut d'ailleurs remarquer que la construction nominale complexe (trois éléments ou plus) la plus utilisée est précisément celle qui mobilise un *sanctus* ou *beatus* parmi ses composants [179]

La complexité dans la classification des noms des institutions et des personnes est un problème général dans la reconnaissance des entités nommées parce qu'ils peuvent adopter des formes combinées ou imbriquées (par exemple : l'ambassade de France en Espagne ; le Paris Saint-Germain, etc.). Cependant, la rareté de ces cas pousse en général à les ignorer et à ne considérer que l'entité la plus large. Dans les cartulaires, cette question n'est en principe pas plus complexe, mais le fait que ce phénomène soit autant répété le transforme en un problème qu'il est indispensable de traiter.

(2) Un deuxième problème concerne directement les constructions de noms de personne utilisant des locatifs. Un nombre important de toponymes que l'on peut récupérer au sein du corpus, à partir du XIe siècle, apparaissent fusionnés dans des dénominations personnelles bipartites, agissant comme complément locatif. De telle sorte que la récupération de certaines dénominations personnelles, alors qu'elles sont imbriquées avec les toponymes, implique aussi la récupération des entités géographiques. Or, une récupération mélangée engendre un important problème de classification et d'ambiguïté. Comme dans le cas des entités personnelles, ce problème peut aussi subir un niveau croissant de complexité. Le nom double est assez stable tout au long des XIe et XIIe siècles et ces entités chevauchées ont généralement deux ou trois composants (ex. *Lambertus de Malliaco, Stephano de Cave Rupe*), mais, vers la fin du XIIe siècle, il n'est pas rare de trouver des entités avec quatre ou cinq composantes,

179. Il s'agit effectivement d'un intérêt pour délimiter le patrimoine ecclésiastique en le rapportant au « patrimoine des saints », diverses exemples de ce pratique peuvent être trouvés depuis le VIIIe siècle dans : Pierre Toubert. *Les structures du Latium médiéval : le Latium méridional et la Sabine du IXe siècle à la fin du XIIe siècle*. École française de Rome, 1973, p. 938-945

ce qui est dû en grande partie à la complexité des noms de lieux (ex. *Guillelmus de Sancti Stephano de Ponte*).

Il s'agit alors d'un problème technique difficile à régler car la désambiguïsation nécessite un modèle plus complexe capable de classer une occurrence en deux classes ou plus (nom de personne + nom de lieu), ce qui n'est pas toujours possible avec les classificateurs linguistiques actuels. Si nous ne prenons en compte que la plus grande entité — dans les cas cités, un nom personnel —, nous pourrions perdre des milliers de toponymes associés. Et, sauf dans les cas des microtoponymes assez difficiles à localiser, la majorité des toponymes associés à un nom de personne correspond à des endroits réels et identifiables. Une mauvaise détection signifierait une perte importante d'information.

(3) En troisième lieu, nous constatons une extension rapide d'un vocabulaire commun dans les actes pour décrire les terres, les propriétés et les biens, ce qui tend à fixer les descriptions spatiales à l'intérieur d'un modèle textuel stéréotypé par le formulaire. L'environnement textuel utilise et réutilise des concepts, des formules et des associations de mots disposés dans un discours contenant des parties généralement bien différenciées, ce qui est le principe d'un modèle [180]. D'un côté, cette caractéristique peut être un facteur essentiel de reconnaissance, puisque qu'elle crée de longues séries textuelles uniformisées dont les évolutions et les changements sont lents. Mais à l'opposé, ce contexte peut également devenir une source de « sur-entraînement » sur certains de ces vocabulaires et formules et conduire à une généralisation excessive sur les cooccurrences et en général sur le contexte linguistique de la phrase qui peut être parfois très répétitif.

Ce dernier problème est précisément l'une des causes principales de l'incapacité des modèles de reconnaissance utilisés en sciences sociales à s'adapter à d'autres textes que ceux-ci utilisés pour les entraîner. Puisqu'un style mobilise un vocabulaire limité et produit des combinaisons récurrentes, le schéma que le modèle apprendra sera forcément limité et spécifique, particulièrement s'il n'a été pas entraîné avec d'autres exemples stylistiquement hétérogènes, ce qui est généralement le cas. En conséquence, les structures nouvelles ou les éléments d'innovation ne seront pas bien reconnus parce que le modèle a appris "par cœur" les structures d'entrainement et n'a pas su généraliser. De nouveau nous avons, dans le cas des actes médiévaux, une version encore plus défigurée de ce problème car la répétition est courante et le vocabulaire bien plus réduit. Il faut pour régler ce problème bien définir les échelles d'entrainement et s'assurer d'un équilibre entre l'ampleur des séries et leur hétérogénéité.

180. Monique Bourin et Elisabeth Zadora-Rio. "Pratiques de l'espace : les apports comparés des données textuelles et archéologiques". In : *Actes des congrès de la Société des historiens médiévistes de l'enseignement supérieur public* 37.1 (2006), p. 39-55 ; Pierre Chastang. "Du locus au territorium. Quelques remarques sur l'évolution des catégories en usage dans le classement des cartulaires méridionaux au XIIe siècle". In : *Annales du Midi*. T. 119. 260. Privat. 2007, p. 457-474 ; Alain Guerreau. "Le champ sémantique de l'espace dans la vita de saint Maieul (Cluny, début du XIe siècle)". In : *Journal des savants* 2.1 (1997), p. 363-419

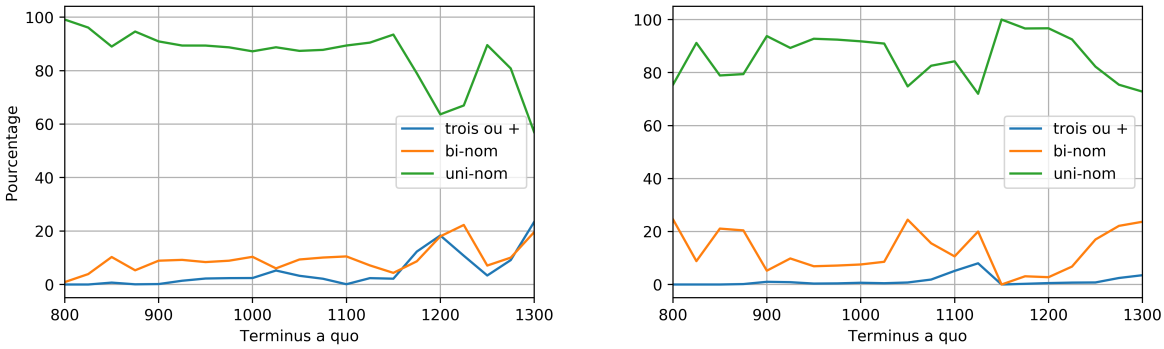

FIGURE 2.5 – Évolution du nombre de composants dans les noms de personnes (gauche) et lieux (droite). La période entre 1130-1160 est douteuse étant donné la quantité très faible de documents.

2.2.3 Le latin médiéval

Troisièmement, certaines particularités des états de la langue des documents peuvent altérer fortement la robustesse du modèle. Les documents intègrent deux grands pôles de transformation linguistique : d'un côté, le processus de vernacularisation de la langue qui amène le latin vers le français ; de l'autre, la personnalisation du document par son auteur à cause d'une méconnaissance du latin, et d'une certaine laxité dans les règles de son emploi. Nos documents étalés sur quatre siècles présentent des caractéristiques linguistiques très hétérogènes au regard de l'émancipation progressive des règles du latin classique [181]. À côté de ces évolutions, nous rencontrons différentes instances conformant la version du discours, celles imposés par les formulaires et la tradition : auxquelles s'ajoutent les interventions particulières de scribes et copistes.

En ce qui concerne l'abandon progressif de certaines contraintes de la grammaire latine, trois phénomènes morphosyntaxiques doivent être signalés car ils sont les plus périlleux : (i) la cohabitation d'un nombre élevé de prépositions remplaçant les déclinaisons (spécialement pour les génitifs et ablatifs) ; (ii) l'ordre plus souple de la phrase par rapport au latin classique, surtout dans les protocoles, ce qui permet une rédaction narrative et produit une inversion progressive de l'*ordo rectus* du latin ; (iii) le surcroît, d'origine phonétique, de versions déformées des mots les plus courants. Ces trois problèmes entament directement la robustesse des outils de base. Dans le cas du problème (i), la manifestation la plus dommageable est la présence d'une double règle superposée pour la distinction d'une même fonction ; le cas (ii) peut conduire à une mauvaise reconnaissance des dépendances établies par une entité et engendrer la formation d'associations incohérentes ou, à l'inverse, perdre les cooccurrences. Dans les deux cas, la subversion de la grammaire latine oblige à l'utilisation de règles assez souples pour reconnaître un texte qui peut apparaître syntaxiquement sectionné par des prépositions inattendues ou soudé par les déclinaisons. Le cas (iii) constitue la

181. Thomas BRUNNER. "Le passage aux langues vernaculaires dans les actes de la pratique en Occident". In : *Le Moyen Age* 115.1 (2009), p. 29-72 ; Marc VAN UYTFANGHE. *Le latin et les langues vernaculaires au Moyen Âge : un aperçu panoramique*. na, 2003, p. 2-18

cause principale des problèmes liés aux lemmatisations déficientes [182]. Un mot écrit de manières diverses peut devenir un piège linguistique qui risque d'appauvrir les résultats parce il est très compliqué déterminer au préalable les variations utilisées par un scripteur et toutes les versions graphiques possibles d'un même mot qu'elles sont susceptibles de produire.

Le sujet n'est pas d'exposer ici les difficultés de la reconnaissance des textes en langue latine classique, liées à une langue flexionnelle, mais de les mentionner rapidement, parce certaines d'entre elles sont présentes dans le latin médiéval. Il s'agit de la surabondance des suffixes : *-er, -ur, -am*, de l'ambiguïté que produisent certaines déclinaisons — entre ablatif, datif et nominatif — concernant le sens des mots, de l'irrégularité de certains verbes clés ou de l'existence de verbes à racine multiple (notamment *fero, eo, nolo*). Enfin, des problèmes d'ordre morphologique avec les modifications de *i* par *j*, *b* par *v*, *u* par *v*, *ae* par *e*, *uu* par *w*, etc.

Sur un autre plan, la rhétorique du discours et la correction stylistique peuvent se présenter à niveaux très différentes selon les circonstances de rédaction de l'acte. Il y aura forcément un niveau de correction et de soin plus élevé dans le latin de l'acte rédigé pour un pape ou un évêque que pour un simple laïc. L'inclusion des préambules et exposés dont la rédaction est moins dépendante de la formule, permet des espaces pour des rédactions rhétoriques. Les corrections et interventions au moment des compilations peuvent aussi supprimer, ajouter ou changer certains usages ou formules dans le but de donner une certaine solennité à un document intégré dans un nouveau contexte intertextuel. D'autre part, le niveau de connaissance du latin des scribes pèse nécessairement sur le résultat final. Quelques textes présentant des obscurités pourraient s'expliquer de cette manière, de même que l'énorme variabilité dans les façons d'écrire un mot ou de reproduire une formule. Le formulaire, dans sa nature conservatrice, peut ainsi parfois accumuler tout un répertoire de formules fossilisées qui sont toutefois incorrectes du point de vue linguistique.

C'est l'ensemble de ces éléments cumulés qui explique une grande partie des difficultés rencontrées lors de la construction d'un modèle automatique appliqué au latin médiéval des sources diplomatiques. La tâche est encore plus compliquée si l'on considère que les maigres ressources qui existent en traitement automatique des langues (TAL) pour le latin ont été construites à partir d'adaptations de textes littéraires écrits en latin classique ou en latin scolastique et visaient l'extraction de traits littéraires ou philologiques [183]. Les outils adaptés à la linguistique médiévale accusent un fort retard par rapport aux langues modernes ; en fait, le modèle proposé dans cette thèse tente de combler une lacune majeure dans cette recherche en TAL.

182. Geneviève Contamine. "Traitement des textes diplomatiques : les problèmes de la lemmatisation". In : *Publications de l'École Française de Rome* 31.1 (1977), p. 265-275 ; Kestemont et De Gussem, "Integrated sequence tagging for medieval Latin using deep representation learning"

183. Marco Passarotti. "From Syntax to Semantics. First Steps Towards Tectogrammatical Annotation of Latin". In : *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and humanities (LaTeCH)*. 2014, p. 100-109 ; David J. Birnbaum et al. "The Digital Middle Ages : An Introduction". In : *Speculum* 92.S1 (2017), S1-S38

2.2.4 La production dans les scriptoria

En quatrième et dernier lieu, deux remarques concernent la production des chartes dans les *scriptoria* bourguignons. La première concerne la confection des chartriers et la compilation de cartulaires, nettement dominée, jusqu'àu XIIIe siècle par les centres religieux du sud de la Bourgogne, notamment Cluny, aux dépends des centres du nord dont les niveaux de production de chartes similaires à celui de Cluny sont atteints vers la fin du XIIe siècle. La deuxième concerne l'évolution de la production des chartes à l'abbaye de Cluny, au sein de laquelle il est possible de distinguer trois périodes, liées à de profondes transformations de la réalité territoriale, de l'organisation de l'écriture et de la légalité des actes.

Comme l'ont montré des études assez récentes, la plupart des cartulaires bourguignons du XIe siècle sont d'origine bénédictine, parce qu'ils sont produits soit à l'abbaye de Cluny, soit dans l'une de ses dépendances. Cluny impose un fort contrôle intellectuel sur la dynamique de copie des chartes [184]. Cinq des six éditions de cartulaires mobilisées dans cette thèse ont cette origine et ils ont été privilégiés précisément parce qu'ils forment des suppléments du cartulaire clunisien. Le sixième, le recueil de l'Yonne a été composé à partir de chartriers produits dans les *scriptoria* du nord un siècle plus tard, lorsque d'autres ordres et institutions ont rejoint la deuxième vague de cartularisation sur la base d'un large réseau institutionnel établi dans le nord de la Bourgogne. Cette deuxième vague apporte des cartulaires libérés de la surveillance intellectuelle clunisienne, et comportant des traits plus hétérogènes (voir partie 3.2) : une diversité d'origine institutionnelle (prieurés, chapitres cathédraux, un cartulaire épiscopal) ; différents indices d'organisation (chronologique, topographique, par chancellerie) ; une variété des producteurs d'actes (aristocrates, ordres religieux, municipalités). Ce transfert de la production depuis les centres bénédictins vers les centres cisterciens, du sud au nord de la Bourgogne, ainsi que la participation de nouvelles institutions à la mise par écrit et en recueil de documents davantage liés à la gestion, n'est pas très bien cerné car la documentation clunisienne est bien moins repéré à partir de la deuxième moitié du XIIe siècle. Ainsi, dans la production de notre modèle une variété de recueils et de collections d'actes relativement large a été utilisée, bien que les séries les plus denses correspondent à celles produites au sein du réseau bénédictin, qui domine le panorama jusqu'àu XIIe siècle.

En outre, dans la confection du cartulaire clunisien qui constitue le cœur du modèle, différentes campagnes de copie et compilation peuvent être distinguées : 1063-1080, 1095, 1120 (cartulaire A, B et C) 1170-90 (cartulaires D et E) au fil desquels les lignes directrices de compilation semblent changer ainsi que les pratiques concernant la rédaction d'un acte. En général les chartes datant de la période antérieure à la fondation de l'abbaye jusqu'à la fin du Xe siècle (cartulaires A et B), concernent presque entièrement le patrimoine ecclésiastique et sont très attachés aux formulaires haut médiévaux et au motif juridique de l'acte, les variations et les adaptations étant faibles. [185]

184. Dominique Iogna-Prat. "La geste des origines dans l'historiographie clunisienne des XIe-XIIe siècles". In : *Revue bénédictine* 102.1-2 (1992), p. 135-191 ; Rosé, "Panorama de l'écrit diplomatique en Bourgogne : autour des cartulaires (XIe-XVIIIe siècles)"

185. Georges Duby. *La Societe aux Xle et Xlle siecles dans la region maconnaise*. A. Colin, 1953,

Ce panorama change à partir la deuxième moitié de XIe siècle. Une crise est perceptible dans la production des chartes. Le carcan juridique qui compose le formulaire autour la décennie 1030-1040 se desserre. Les chartes perdent la régularité qui les caractérisait, produit d'un fort attachement au formulaire, et elles sont en partie remplacées par des notices qui, comme on l'a vu, correspondent à des documents rédigés dans un style plus objectif, plus direct. À côté de cela, des changements profonds affectent le vocabulaire décrivant les réalités sociales et spatiales. Ce changement ne concerne pas seulement les documents, mais témoignent d'une évolution dans les structures spatiales et juridiques affectant la rédaction des chartes [186]. Le caractère général de ces changements en Europe occidentale, liés à la féodalité, a alimenté un débat historiographique ancien sur la mutation de l'an 1000 [187].

L'acte de donation en fait commence à disparaître du corpus clunisien vers 1040 et s'éteint définitivement à la fin du siècle. Avec l'introduction des cartulaires D, E et en partie du C qui reflètent l'intérêt de l'abbaye pour compiler ses actes concernant les relations avec le pouvoir public, l'expression formulaire des échanges fonciers disparaît et est remplacée par les cadres rédactionnelles plus libres des lettres qui dominent le panorama à partir de la deuxième moitié du XIIe siècle.

p. 9-18

186. François Bange. "L'ager et la villa : structures du paysage et du peuplement dans la région mâconnaise à la fin du Haut Moyen Age (IX e-XI e siècles)". In : *Annales. Histoire, Sciences Sociales.* T. 39. 3. Cambridge University Press. 1984, p. 529-569

187. Dominique Barthélemy. "La mutation féodale at-elle eu lieu ?(Note critique)". In : *Annales. Histoire, Sciences Sociales.* T. 47. 3. Cambridge University Press. 1992, p. 767-777

Chapitre 3

La modélisation informatique

3.1 Modélisation de la reconnaissance des entités nommées

En choisissant un corpus permettant d'entraîner l'algorithme à la reconnaissance d'éléments morpho-syntaxiques, nous devons rester très proches des normes définies dans la théorie du corpus, spécifiquement de celles relatives aux axes quantitatifs — quelle extension doit avoir un corpus ? — et qualitatifs — les documents sélectionnés sont-ils représentatifs ? —, tout en faisant l'appel aux spécificités d'une analyse qui privilégie les variations statistiques et le contexte immédiat des mots [188]. Le processus de formation du corpus ne nous concerne pas parce que nous prenons comme corpus l'ensemble de cartulaires et de recueils d'actes hérité, conçu intellectuellement comme un seul volume dont nous avons précisé, dans le chapitre 2, l'histoire et la composition. Néanmoins, le corpus originel et le corpus avec lequel nous avons entamé la construction du modèle automatique pour la reconnaissance des entités nommées ne coïncident pas nécessairement. Les problèmes que nous avons pointés exigent une réponse technique afin de contrôler les risques de surentraînement et de surgénéralisation que le modèle pourrait encourir. Ces risques sont liés à certaines des caractéristiques du corpus, notamment la forte dépendance aux institutions bénédictines, sa provenance régionale unique et le caractère stéréotypé du discours formulaire. Les mesures de contrôle passent par la formation de sous-corpus où l'on privilégie la présence de certains éléments homogènes — corpus de caractère spécifique — ainsi que de sous-corpus qui, par variation des échelles, contiennent des éléments hétérogènes — corpus de caractère général (voir partie 2.2).

D'ailleurs, s'il est vrai qu'une des questions les plus épineuses lorsqu'on travaille avec des états d'une langue disparue est le manque de locuteurs natifs et par extension de compétence linguistique complète de la part de l'analyste, cette compétence peut être partiellement remplacée par des instruments opérant à partir de dictionnaires,

188. À ce sujet quelques travaux de référence Anne O'Keeffe et Michael McCarthy. *The Routledge handbook of corpus linguistics*. Routledge, 2010, p. 345-359 ; Graeme Kennedy. *An introduction to corpus linguistics*. Routledge, 2014, p. 201-230 ; Tony McEnery et Andrew Hardie. *Corpus linguistics : Method, theory and practice*. Cambridge University Press, 2011

grammaires et classificateurs de séquences [189].

Par ailleurs, notre corpus est par définition une sélection limitée de documents qui transmet une image forcement incomplète du phénomène d'écriture dans la région ou même du processus de la rédaction des cartulaires. Essayer de construire un outil de vocation généraliste pour l'application sur d'autres corpus similaires se transforme ainsi en un défi important, spécialement si on part d'un ensemble de documents qui constituent un échantillon représentatif, certes, mais limité.

Finalement, le corpus en tant qu'objet « fermé » peut se révéler un univers linguistique singulier, puisque certains phénomènes liés à l'état de la langue n'apparaîtront que là, et feront de lui un référent textuel unique [190]. Ceci considéré, l'outil que l'on peut entraîner à partir d'un seul corpus n'a pas au premier regard une vocation généraliste. Cette situation oblige d'un côté à prouver jusqu'à quel point le corpus est suffisant pour construire la base d'un modèle fournissant un niveau correct d'adaptabilité et de l'autre côté à annoter des séries supplémentaires qui apportent soit des éléments détectés dans des corpus proches et mal représentés dans le corpus principal, soit des textes datés de périodes sans production ou avec une production très maigre. L'objectif est d'apporter à la machine un répertoire scriptural plus varié afin de combler certaines lacunes temporelles du corpus principal — Cluny en ce qui nous concerne —, mais aussi de tester si le modèle entraîné est robuste face à de nouveaux documents qui ont des caractéristiques un peu différentes.

3.1.1 Le pré-traitement du corpus

Modification et extension du corpus

Les modifications, adaptations et ajouts que nous avons effectués sur le corpus annoté ont donc été motivées par des questions relatives à la définition du corpus qui visent à réduire l'écart entre quantité et qualité documentaires en tant qu'exemples des phénomènes linguistiques. Cela implique en définitive de respecter des critères multiples : la représentativité phénoménologique du corpus, sa concentration et sa variabilité linguistique. Comptent également la multiplicité des typologies scripturales et des styles rédactionnels des actes, l'état de la langue et enfin le niveau d'erreurs et de fausses pistes que produisent les imperfections de la numérisation. Notre objectif est de développer un modèle basé sur un corpus régional opérant avec robustesse sur des corpus géographiquement et chronologiquement proches, mais aussi sur des corpus diplomatiques créés dans des espaces différents et appartenant à des traditions scripturales éloignées. Exposons donc les problématiques soulevées par le corpus clunisien.

189. Sophie Prévost. "Corpus informatisés de français médiéval : contraintes sur leur constitution et spécificités de leurs apports". In : *Corpus* 7 (2008), p. 35-64 ; Kennedy, *An introduction to corpus linguistics*

190. Prévost, "Corpus informatisés de français médiéval : contraintes sur leur constitution et spécificités de leurs apports" ; Christiane Marchello-Nizia. *Grammaticalisation et changement linguistique*. De Boeck-Duculot, 2006, p. 304

Représentativité

Un corpus de modélisation doit assurer stabilité et cohérence interne ; cela vaut pour toute recherche mobilisant des corpus. Un corpus très hétérogène ou répondant à des phénomènes multiples rend l'exploitation assez compliquée dans la mesure où les conclusions auront nécessairement un impact très nuancé. Dans cet état de la langue un corpus contient souvent des variations qu'on ne retrouvera pas ailleurs, si le corpus est suffisamment large ces variations peuvent être bien classées car on peut retrouver les structures et formules canoniques d'où elles proviennent. À un niveau strictement technique, cela constitue un atout. Puisque le modèle de reconnaissance généré est discriminant (voir partie 1.5.1), les règles provenant d'un seul corpus organique peuvent être plus efficaces que celles provenant d'une séquence de petits corpus variés ou de corpus divers mal échantillonnés qui pourraient fournir des observations bien plus variées. Mais, à l'opposé, l'utilisation d'un seul corpus peut affecter la capacité du modèle à reconnaître des entités dans des chartes externes au corpus, car celui-ci peut tendre à une uniformisation, conduisant à exclure partiellement voire totalement certains phénomènes et variations.

Par ailleurs, un corpus régional comme celui de Cluny peut manquer de représentativité par rapport à une réalité beaucoup plus vaste : en amont, la réalité de l'écrit diplomatique de la Bourgogne, en aval, une réalité plus large comprenant les centres producteurs de chartes de l'Europe occidentale. S'il est vrai que l'écriture formulaire apporte un certain niveau de ressemblance entre les chartes européennes, elle ne constitue que l'arrière-plan des modèles scripturaux qui sont influencés par des traditions particulières et par des états régionaux de la langue [191]. En revanche, un corpus provenant d'une institution au centre d'un réseau régional et qui garde de forts liens suprarégionaux, comme celui établi par les bénédictins, concentrera une pluralité de phénomènes supérieure à la moyenne du fait de son rôle dans la circulation des productions intellectuelles et dans la gestion des phénomènes économiques et sociaux.

Concentration

Le deuxième enjeu concerne la concentration documentaire liée au processus de cartularisation. Le cartulaire clunisien copie et accumule des actes principalement privés provenant des différentes mains et scriptoria de la Bourgogne. Les actes publics peuvent avoir une origine hors de la région, mais dans bien des cas ils ont été rédigés par le requérant et validées par l'autorité. À l'intérieur du corpus annoté, les chartes copiées proviennent de plus d'une dizaine de cartulaires et recueils factices — 80 % des actes sont de Cluny —, contenant, naturellement, des transactions juridiques de plus d'une centaine de petits espaces, se distribuant principalement dans cinq types d'actes : chartes, notices, diplômes, bulles et lettres et principalement cinq types d'actions juridiques : donation, vente, échange, confirmation et mise en gage.

191. La question entame directement avec celle de la « Urkundenlandschaft » introduite par H. Fichtenau pour définir l'existence de microrégions dont les actes présentaient des caractéristiques propres, tant du point de vue de la paléographie que de celui de la langue ou de la mobilisation des formulaires. Heinrich Fichtenau. "Das Urkundenwesen in Österreich vom 8. bis zum frühen 13". In : *Jahrhundert. MIÖG Ergbd* 23 (1971)

L'ensemble offre un panorama très ample, impliquant directement presque tout le sud de la Bourgogne, une partie du *pagus Lugdunensis* (Beaujolais) et indirectement tout le réseau monastique et les institutions publiques opérants dans la région. Cette amplitude géographique est pourtant plus limitée dans le plan diplomatique, car cette masse documentaire est produite et surveillée intellectuellement par deux institutions : Cluny et Saint-Vincent [192]. Cela ne concerne pas seulement la préférence de certains modèles formulaires au sein de son réseau, mais aussi un modèle de cartularisation imposant diverses filtres à la sélection des actes, comme certains distorsions éditoriales : correction de barbarismes, abréviation de certains actes, solennisation d'autres, etc.

Apparaît donc un double problème de perspective : d'une part, nous avons deux institutions centrales d'"accumulation" qui produisent et rassemblent les documents imposant des lignes directrices intellectuelles et éditoriales, tout en autorisant une certaine variabilité liée à l'origine multiple des documents et aux différentes traditions opérant dans chaque zone. D'autre part, nous avons une claire surreprésentation de ces deux institutions en tant que producteurs et récepteurs de documents par rapport à d'autres producteurs ou à des institutions moins intégrées, voire externes, au réseau bénédictin, ce qui réduit mécaniquement les styles et les phénomènes scripturaux mineurs.

Écriture formulaire

La troisième question concerne l'écriture formulaire dont nous avons déjà expliqué les caractéristiques (voir partie 2.1.5). Il faut ici rappeler qu'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 – dates, entités nommées, intitulation, signatures. La modélisation statistique s'appuie ainsi sur la reconnaissance de certains éléments structurels formant un tronc discursif commun associés à des éléments documentaires variés dont font partie les entités nommées. Les séquences textuelles sont prises en tant que produits de cette association d'éléments et les itérations, autrement dit, le calcul répétitif jusqu'à la satisfaction d'une condition, que l'algorithme opère sur ces séquences textuelles, ont pour objectif d'apprendre la distribution combinatoire de ces éléments et de générer un classement par probabilité d'apparition. De ce fait, dans des documents homogènes, le modèle peut proposer la balise la plus probable pour une combinaison requise, mais ce modèle pourrait être pris en défaut lors de son application à des documents provenant de traditions scripturales qui utilisent des formulaires différents et présentent des états de discours plus atypiques. Ils comportèrent alors nécessairement des distributions déformées par rapport aux paramètres d'apprentissage du modèle.

192. Voir aussi la question évoqué par N. Perreaux autour de la pénurie de production scripturaire dans le nord de la Bourgogne avant le XIIe siècle. Nicolas Perreaux. "L'écriture du monde (I).. Les chartes et les édifices comme vecteurs de la dynamique sociale dans l'Europe médiévale (viie-milieu du xive siècle)". In : *Bulletin du centre d'études médiévales d'Auxerre/ BUCEMA* 19.2 (2015)

L'état de la langue écrite

La quatrième question concerne l'état de la langue écrite. Dans les études de traitement automatique de la langue, la « boîte à outils » du latin est très pauvre, ce qui implique d'effectuer la modélisation à l'aide d'outils de traitement du langage indépendants de la langue, combinés à l'adaptation de certains outils développés pour le latin classique. Des études récentes ont porté sur la modélisation du latin, en produisant lemmatiseurs, dictionnaires et index géographiques, mais ceux-ci sont encore expérimentaux et demeurent focalisés sur la littérature en latin classique. Les variantes médiévales, qui résultent des états changeants de la langue imputables au processus de vernacularisation du latin, n'ont presque pas été abordées. Toute approche fondée sur un corpus médiéval particulier doit s'effectuer en prêtant une grande attention aux phénomènes singuliers, car il est fort probable qu'il recèle, pour certains d'entre eux, les seuls exemples existants. Toutefois, à la différence de ce qui arrive dans le domaine littéraire, les sources diplomatiques ne subissent pas de brusques changements dans le style, mais plutôt une lente émergence d'innovations dans laquelle l'individualité de l'écrivain se trouve entravée par la formule.

Numérisation

La cinquième et dernière question concerne la numérisation de sources. L'océrisation des éditions érudites, ou à défaut, la saisie manuelle, reste le moyen principal d'obtenir le texte brut qui constitue la base de travail. Les transcriptions originelles peuvent varier selon l'intérêt particulier de l'éditeur ; un philologue cherchera à fournir une grande quantité de détails graphiques trouvés dans le document et à préserver la forme originelle des mots, faisant plutôt une translitération, alors qu'un paléographe sera davantage enclin à présenter une version restituée et normalisée du texte et à fournir un cadre de références historiques. Selon le regard qui est porté sur lui, le contenu des éditions érudites offre, à travers le texte et l'appareil péri textuel, toute une série d'éléments de correction syntaxique, graphique et lexicale, comme la ponctuation, le développement des abréviations, la restitution des oublis du scribe, l'usage des majuscules, ainsi que des éléments diacritiques et para textuels — signes conventionnels indiquant les abréviations, les suppressions ou les modifications, les caractères spéciaux, les commentaires et les titres, etc. La normalisation du texte afin de faciliter la lecture humaine, facilite aussi la lecture par la machine et est un atout fondamental pendant la phase de modélisation. En même temps, l'excès de signes auxiliaires et de marqueurs de hiérarchie syntaxique peut rapidement engendrer un « bruit de fond », qui empêche un bon séquençage du texte affectant fortement la reconnaissance.

Afin d'étudier l'impact de toutes ces questions sur la qualité de notre modèle, nous avons effectué différentes expériences de validation croisée et d'évaluation de robustesse dont nous allons présenter à continuation les détails. Chacune d'elles s'est concentrée sur un seul aspect : la taille du corpus, les variations temporelles, les variations régionales.

3.1.2 Modifications effectuées sur l'annotation originelle et génération de sous-corpus

Dans plusieurs cas, que nous allons maintenant détailler, l'état de la numérisation mais surtout de l'annotation originelle limite la performance de notre modèle. Les corrections que nous avons opérées s'appuient sur l'observation minutieuse de plusieurs situations inhérentes aux formulations diplomatiques qui nous ont conduit à modifier ou enlever l'étiquette attachée aux entités.

Cette section décrit les premières étapes de ce processus, celles dédiées au pré-traitement et à la normalisation appliqués à tous nos corpus.

Suppression des éléments paratextuels

L'une des principales difficultés du travail effectué sur des textes édités par des érudits est l'excès du susmentionné de « bruit de fond » [193]. Les textes courants trouvés dans les éditions obligent 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, tout en étant extérieurs au texte lui-même.

> «Heldevini de Matriolis ([con]cedentis. [Æc]clesia [Vallis lucentis] nunc pos[sid]eat feodum ». [194]

> «Actum et datum Clun[iaci], anno Domini Mº CCº quadragesimo quinto, mense decenbri, in crastino beate Lucie. (Trace des trois sceaux.)». [195]

> «Ego denique domna maior qui hanc cartam fieri iussi ; legere audiui. & manu mea signum [roto]bus. {a} [roto] {b} Domnus aluarus testis, Domnus Munio testis, Semeno garciez testis / [signo : domna maioris SiGnvM] / {c}» [196]

Ces éléments paratextuels transforment les mots et altèrent les séquences, ce qui rend difficile l'obtention d'un texte segmenté (« tokenisé ») et lemmatisé, dont dépend l'application de l'algorithme lors de l'entrainement.

D'ailleurs un processus de normalisation de l'orthographe textuelle et du jeu de caractères est mis en place dans les scripts de tokenisation. On peut se heurter à un « mur de briques » si on essaie de gérer des données contenant des caractères extérieurs au jeu intégré à l'outil. Les transformations entre le format de codification originelle du corpus (Latin-1) et le format plus universel qu'utilisent la plupart des outils automatiques (UTF-8) n'est pas compliqué, mais dans le cas des caractères

193. Deux travaux qui traitent de questions similaires en corrigeant le "dirty OCR" : Tobias Blanke et al. "Information Extraction on Noisy Texts for Historical Research". In : *Digital Humanities* (2012) ; Thomas L Packer et al. "Extracting person names from diverse and noisy OCR text". In : *Proceedings of the fourth workshop on Analytics for noisy unstructured text data.* ACM. 2010, p. 19-26

194. CBMA 18415

195. CBMA 6313

196. CORHEN-0160

peu utilisés (æ œ, ô) une transformation dédiée peut être exigée. Cette normalisation diminue le taux d'erreur et facilite un usage universel du corpus.

Une fois les éléments paratextuels enlevés et les caractères normalisés, l'application des outils de base montre une amélioration significative des résultats. Le « toilettage » du texte réduit les taux d'erreur du système et il est facilement automatisable une fois qu'on a détecté l'ensemble des éléments problématiques. Toutefois, il a le défaut de souvent simplifier le texte. Les éléments supprimés, partie importante de toute édition critique, sont difficilement récupérables en raison de la complexité que présente le fait d'établir un registre de changements précis. Il a donc fallu définir des méthodes comparatives pour récupérer les données perdues dans le « toilettage » et ainsi restaurer les textes ayant subi un traitement automatisé.

Normalisation et segmentation

Une deuxième étape indispensable de la normalisation concerne différentes modifications que nous avons opérées sur l'annotation originelle afin de valider ou de mieux reconnaître des situations linguistiques qui, au préalable, présentent une résolution compliquée ou qui nécessitent une annotation plus précise. Les trois principales sont :

— Les entités imbriquées

— Les entités ayant un rôle « prédicatif »

— Les entités complexes périphrastiques.

En ce qui concerne les premières, l'une des pratiques les plus répandues dans les actes, et qui correspond à une réalité juridique et spatiale, est l'usage des noms de saints en tant qu'entités à la fois personnelles, juridiques et territoriales dont nous avons déjà vu quelques exemples (voir 2.2.2).

> « in *villa* Caucilla manso indominicato cum capella qui est in onore sancti Mauricii dedicata » [197]

> « Eldegrinus vivit, teneat et possideat, et pos suum discesso Sancti Petri perveniat » [198]

> « Et dono vobis vercariam unam in ipsa *villa*, que terminat a mane Sancti Vincentii et Belmontissa » [199]

> « in *pago* Matisconensi, in *vicaria* Sancti Pontii, in *villa* quae vocatur Burgundia » [200]

197. CBMA 1455
198. CBMA 1710
199. CBMA 387
200. CBMA 1994

Le plus cohérent serait d'annoter, dans la plupart des cas, ces entités comme des noms de institutions ou organisations (ORG), car elles jouent un rôle en tant qu'entité juridique qui agit sur différentes facettes de la réalité. Mais étant donné que cette catégorie n'a pas été prise en compte dans l'annotation initiale par le projet CBMA, elles ont été étiquetées à l'origine comme des personnes. Ici nous proposons une annotation différente, conforme à leur fonction : nous les avons annotées comme lieux. Si morphologiquement ces entités sont des noms de personnes, elles constituent des références génériques dont la valeur n'est pas celle de personnes. Elles n'ont d'ailleurs pas toujours d'existence historique. De plus, leur fonction de jalons spatiaux et de référents stables s'apparente davantage à celle d'une *villa* ou d'un lieu-dit. En réalité, quand elles font référence à un saint en tant que personne juridique, elles désignent ordinairement un bâtiment — monastère, église, chapelle, hôpital —, un espace intérieur ou une parcelle localisable. Si nous acceptions ce genre d'entités comme noms de personnes, cela conduirait le modèle à considérer un contexte clairement spatial comme étroitement lié à un contexte personnel, débouchant ainsi sur de nombreux faux positifs.

En revanche, nous avons totalement ôté la balise dans les cas, peu nombreux, où il est fait référence explicitement à la figure d'un saint, lorsqu'elle sert d'inspiration pour une formule ayant une portée morale ou qu'elle est liée à une citation biblique. Il en est de même dans les cas où une date est indiquée par la mention de la festivité d'un saint selon le calendrier [201].

> « et per singulos annos, in festivitate Sancti Martini , pro investitura, sextarios VIII de vino persolvam » [202]

> « secundum regulam Sancti Benedicti de semet ipsis post Odonem » [203]

> « ut pius Dominus, per intercessionem sancti Petri » [204]

Par ailleurs, dans les chartes foncières, il est assez courant de trouver une description de la forme, de l'orientation et de la surface des terres, objet de la donation, de la vente ou de l'échange, qui fait référence à sa place dans le parcellaire local. Il est compliqué de préciser si cela correspond à un bornage opérant dans le cadastre, ce qui n'est pas d'ailleurs probablement le cas (voir partie 6.3). Quoi qu'il en soit, il est évident que les biens fonciers sont décrits après une enquête sur le terrain et l'indication de leur morphologie se faisait en indiquant les limites avec les propriétés contigües qui, en l'absence de nom officiel, étaient évoquées par le nom d'un propriétaire [205].

201. On n'a pas modifié en revanche les trois personnages bibliques habituels des clauses de condamnation : Dathan, Abiron, Judas

202. CBMA 1937

203. CBMA 1698

204. CBMA 2046

205. Voir au sujet des cadastres et des techniques de bornage et arpentage : Pierre Portet. "Les techniques du bornage au moyen âge : de la pratique à la théorie". In : *Sfruttamento tutela e valorizzazione del territorio. dal diritto romano alla regolamentazione europea e internazionale.* T. 18. Jovene, Napoli. 2007, p-195 ; Michel Lauwers et Laurent Ripart. "Représentation et gestion de l'espace dans l'Occident médiéval". In : *Actes du colloque «Rome et l'État moderne européen :*

Normalement les scripteurs utilisent alors deux types de formules pour indiquer la mesure (*perticatio*) et l'orientation / limites (*terminatio*) :

> «...qui habet fines de uno latere/de tres partes/de quatuor partes...»

> «...qui terminat a mane...a cercio...a sero..a medio die...»

Le syntagme manifestement locatif peut se développer, selon le scripteur, sous la forme d'un accusatif de direction — préposition + accusatif — ou d'un ablatif de séparation sans préposition. Dans les deux cas, l'entité nommée au centre joue le rôle d'un complément adoptant la forme d'un génitif d'appartenance. Il n'est pas rare, par contre, de trouver des combinaisons alternant les deux cas. Effectivement dans les expressions de mouvement et de détermination des limites, le latin classique se sert de l'accusatif avec *in/ad* et de l'ablatif introduit par les prépositions *ab/ex*. Néanmoins, l'ablatif, en particulier après la période des IIIe au VIe siècles, pouvait être préféré pour exprimer l'idée locative en supprimant la préposition, comme dans certains usages du latin classique. Dans le latin médiéval, et en rapport direct avec la formation latine des scripteurs, on trouve une coexistence des deux formes :

> « terminat...de una parte terra Arnaldi, de altera terra sancti Petri, a meridie rivo currente » [206]

> « unum campum terminat a mane terra Francorum, a medio dia similiter, a sero Sancti Stephani et silva insimul » [207]

> « ista [vinea] habet fines a mane ad terram Immonis, a medio die ad terram Sancti Petri » [208]

> « terminat a mane increpito, a medio die et a sero ad terram Sancti Petri, a cercio ad terram Bernardi » [209]

Au niveau morphosyntaxique, cet usage ne complique pas la reconnaissance automatique, mais considéré dans un sens strictement contextuel, il engendre deux problèmes : d'un côté, la présence d'entités nommées personnelles jouant un rôle locatif et de l'autre, la présence de cooccurrences typiquement associées à la description de lieux et bien-fonds (*terra, manus, parte, serus*, etc.) accompagnant des noms de personne. Dans l'annotation originelle, ces entités ont été considérées comme des entités personnelles, puis nous les avons changées en entités de type géographique. Comme dans le cas précédant de *sanctus/Beatus*, il s'agit ici d'une accumulation de

une comparaison typologique». In : J.-P. Genêt (dir.), Rome et l'État moderne européen. Rome : Collection de l'École française de Rome. T. 377. 2007, p. 115-171 ; Jean-Loup Abbé. "Arpenter et borner les terroirs de l'Europe méridionale au Moyen Âge : savoir et savoir-faire". In : *Annie Rousselle (éd.), Monde rural et histoire des sciences en Méditerranée. Du bon sens à la logique, Perpignan, Presses universitaires de Perpignan* (1998), p. 51-62

206. CBMA 2116
207. CBMA 505
208. CBMA 1606
209. CBMA 1526

fonctions assignées à la même entité. Nonobstant, les entités présentes se comportent comme des références de lieu plutôt que comme des indicateurs de personne, et leur utilisation n'a de sens qu'à l'intérieur d'une reconstruction parcellaire. Il serait assez compliqué de les mobiliser pour d'autres réseaux que ceux formés par les distributions spatiales des terres et de leurs formes d'appropriation. En outre, en changeant la balise, on s'assure de ne pas apporter de « bruit de fond » supplémentaire au modèle.

Finalement, comme nous l'avons vu, à partir du XIIe siècle, la dénomination personnelle peut adopter une forme composée, incluant des éléments qui ne sont pas strictement nominatifs, ce qui constitue un défi important pour le modèle. Dans ces cas, nombreux à partir de la décennie 1160, le choix le plus pertinent a été de modifier l'annotation originelle afin d'éviter une mauvaise reconnaissance :

> « Joannes nepos ejus et Theobaldus filius ipsius Hugonis » [210]
>
> « Milo, filius defuncti Henrici Cambellani » [211]
>
> « Galterus Sapiens filius Renaudi de Plaseto, Jaquetus et Grivellus fratres ejus » [212]
>
> « Bernardo qui Parvus cognominatur » [213]
>
> « Andegauensis Comes Gaufredus, cognomine Martellus » [214]
>
> « Hildeburga cognomine Martiniana » [215]
>
> « Iohanne dicto de sancto Symphoriano » [216]
>
> « silvam Sancta Maria, que vulgo dicitur Boerecia » [217]

Ce genre d'occurrences, facilement repérables car elles sont introduites par un nombre réduit de verbes au passif (*cognominatur, dicitur, appelatur, nominatur*, etc.), sont une expression de la variété de formes nominatives qui émergent en Europe à partir de la fin du XIe siècle. Ce phénomène peut être aussi observé dans l'appellation de certains lieux, notamment de certains lieux-dits. Les compléments dans ces noms complexes ont surtout une utilité documentaire ; ils permettent de distinguer et de bien identifier le personnage participant à l'acte ou la localisation du bien objet de l'acte. Dans le cas des noms de personnes, il peut s'agir d'un deuxième nom familier, ayant une valeur d'usage, d'un surnom à la manière des *cognomina* romains, ou d'un locatif d'origine ou d'appartenance.

210. CBMA 18428
211. CBMA 17897
212. CBMA 18393
213. CBMA 632
214. CBMA 15833
215. CBMA 14547
216. CBMA 830
217. CBMA 1494

Cela dit, cette catégorie peut être divisée en deux types : ceux qui correspond a des dénominations transitoires ajoutant une entité afin de préciser l'identité et ceux qui ajoutent un nom propre à l'individu ou au lieu. Dans le premier cas nous avons gardé l'annotation originelle qui distingue les deux personnes (ou lieux), même si la deuxième a été invoquée uniquement pour distinguer la première. Par exemple, *Theobaldus filius ipsius Hugonis* ou *Iohannes dicto de sancto Symphoriano* n'est pas le nom de la personne mais une construction transitoire qui peut toutefois servir à la récupération des coréférences et à la désambiguïsation des individus, mais qui pris dans la simple détection des entités nommées peut apporter des entités inutilement complexes. Dans la même logique, les noms comportant un surnom ou locatif ont été considérés comme une seule entité et la balise sur la deuxième entité a été enlevée. Dans *Hildeburga cognomine Martiniana*, les deux entités qualifient la même personne bien qu'elles forment une unité périphrastique doivent être récupérées dans leur intégrité par le modèle.

Dans d'autres cas, l'annotateur a introduit des entités complexes en incluant une co-occurrence qui fonctionnent, c'est vrai, comme élément précisant l'identification, par exemple : *Dalmatius Miles, Hugo Camerarius, Walterius Pistor*, etc. La deuxième partie du nom s'agit de l'indication de l'office ou dignité de la personne : dans ces cas on a opéré une séparation des entités. Dans certains cas, plus ou moins repérables, par ex. *Bernardus Grossus, Symon Parvus*, le « sobriquet », a devenu effectivement une partie du nom transmis aux enfants.

Enfin, dans le tableau CSV (*comma separated values*) contenant le corpus nous avons opéré un changement concernant la virgule. Dans le format originel le séparateur d'entités est une virgule mais certaines entités contenant des virgules (ex. *Gauscerannus, cognomento Taunel*) celle-ci devenait ambiguë. Nous l'avons donc remplacée par dièse comme séparateur d'entités (#).

3.2 L'entrainement du modèle

Dans cette partie nous allons décrire les étapes techniques de formation du modèle de reconnaissance automatique qui suivent une approche traditionnelle tripartite :

1. création des ensembles d'apprentissage et de test ;
2. validation croisée (cross-validation)
3. évaluation de la performance.

Les deux premières parties visent traditionnellement à réduire le plus possible l'attachement du modèle entraîné à son corpus d'origine. Effectivement, lors de l'apprentissage, le modèle est ajusté à un certain corpus d'entraînement afin qu'il puisse prédire l'apparition de structures similaires dans des textes nouveaux, mais parfois cet ajustement est trop serré ou trop peu. Les bonnes pratiques dans l'entraînement cherchent un équilibre entre un modèle qui ne soit pas trop entraîné sur certaines structures et un modèle qui soit capable de trouver un schéma cohérent dans le corpus. Ces deux situations problématiques appelées dans la littérature *overfitting* et

underfitting rendent des modèles inutilisables car ils ne s'adaptent pas à des nouveaux textes [218].

Dans le premier cas, l'algorithme est nourri avec une grande quantité de textes très hétérogènes ou portant trop de variables avec un nombre réduit d'observations. En conséquence, il n'est pas capable de reconnaître un schéma qui justifie la variété. Ce modèle alors apprend les structures « par cœur » et fait des prédictions correctes s'il trouve de nouveau la même structure — par exemple dans le corpus test —, mais il est difficilement applicable à d'autres textes que ceux du corpus d'origine. Dans le cas contraire, plus rare, l'algorithme est nourri avec un petit corpus très peu varié ou avec petits échantillons de textes, donnant comme résultat un modèle pauvre et attaché à un schéma trop généraliste (voir figure 3.1). Dans ce cas, la capacité d'apprentissage du modèle est sous-utilisée, mais il est possible de l'améliorer, contrairement à la situation d'*overfitting*.

Afin de réduire ces difficultés, ou à tout le moins de constater leur impact, on applique deux procédures. D'abord, une division du corpus de référence (c'est-à-dire, déjà annoté par des experts humains) en deux segments : l'un utilisé pour l'entrainement et l'autre utilisé pour tester la robustesse du modèle généré. Cette comparaison va nous donner un premier résultat indicatif de la qualité d'ajustement du modèle. En deuxième lieu, on introduit différentes variations dans le corpus d'entrainement : variations dans la taille des ensembles, dans les dimensions et les attributs, afin de faire des tests croisés (*cross-validation*) entre un ensemble contre tous les autres et vice-versa. L'objectif étant d'obtenir un classement des résultats qui puisse nous aider à identifier des problèmes spécifiques et d'y remédier pour améliorer le résultat global.

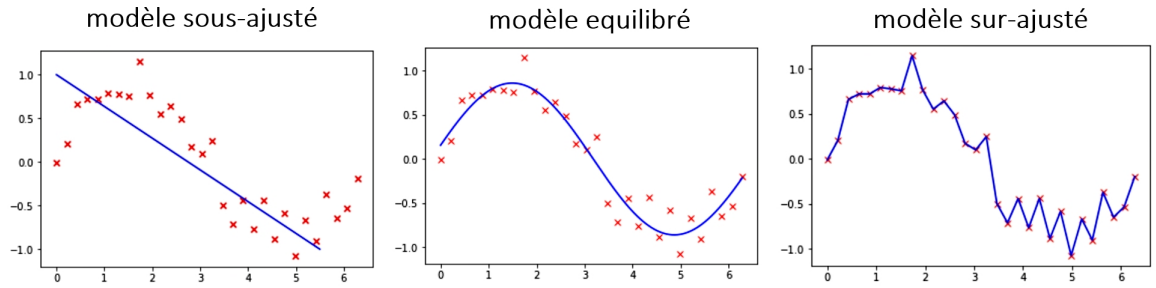

FIGURE 3.1 – Différents états d'ajustement d'un modèle à ses données.

3.2.1 Division des données

Les méthodes de TAL divisent le corpus principal en deux parties : l'une consacrée à l'entraînement du modèle et l'autre à tester sa robustesse et sa validité. Dans notre corpus annoté de 5 300 unités, nous définissons un corpus d'entraînement de 4 000 unités – plus d'un million de mots – et un corpus-test de 1 300 documents. Dans

218. Voir à ce propos le glossaire de termes de Kohavi & Provost : : `http://ai.stanford.edu/~ronnyk/glossary.html` et le manuel de Ripley : Brian D RIPLEY. *Pattern recognition and neural networks*. Cambridge university press, 2007, p. 354-360

ce dernier, il est de bonne pratique de garder un petit ensemble, dans notre cas 300 documents, afin de former un corpus de validation qui permet de contrôler le biais du modèle tout au long de l'entraînement. Finalement, les résultats chiffrés que nous présentons sont donc calculés sur 1 000 documents

Après chaque itération sur l'ensemble d'apprentissage, le modèle appris est appliqué à l'ensemble du validation et le processus est arrêté lorsque le résultat de la validation ne s'améliore plus, ce qu'on désigne par « arrêt précoce » (*early stopping*). Puis on teste la performance en demandant au modèle d'annoter automatiquement l'ensemble du corpus-test qui était réservé en vue de cette l'évaluation et qui est composé par des documents qui ne participaient pas à la phase d'apprentissage afin d'éviter tout biais.

Ventilation chronologique

Un aspect complexe, qui soulève directement la question de la représentativité, concerne cette division du corpus dont le but est de former les ensembles d'entraînement et de test nécessaires aux différentes étapes de construction du modèle. Idéalement, les deux parties doivent respecter une distribution interne identique, c'est-à-dire qu'elles doivent comporter un ensemble similaire de phénomènes linguistiques, malgré des dimensions différentes. Il s'agit d'une condition nécessaire pour former des modèles entraînés sur l'ensemble des attributs d'un corpus et pour valider sa robustesse en utilisant un corpus réduit (*corpus test*) qui soit représentatif du corpus d'origine. Normalement, dans des corpus vastes et organiques, une division aléatoire ou pseudo-aléatoire est suffisante car la possibilité de trouver des évènements similaires dans tous les ensembles est très élevée.

Cependant, si on privilégie une division aléatoire, l'homogénéité de la répartition pourrait ne pas être assurée dans un corpus tel que le CBMA qui présente d'importantes asymétries, liées à des pratiques et à des usages textuels fortement associés à des époques, à des copistes et à des institutions bien individualisés. On ne peut pas calculer a priori la variété des phénomènes et leurs moments précis d'apparition, hiatus et abandon, ce qui serait très utile pour bien équilibrer les deux ensembles. Il faut donc, au moment de comparer la distribution des corpus d'entraînement et des corpus test, privilégier deux aspects généraux : la variété typologique et la dimension documentaire. En effet, les statistiques générales montrent des décennies de grande variété documentaire et des décennies de grande pénurie ainsi que des moments de surreprésentation d'un type particulier d'acte. Afin d'éviter des creux chronologiques, une double ventilation a été testée : pseudo-aléatoire, en utilisant une clé de distribution et chronologique, par tranches de 25 ans.

(i) L'approche pseudo-aléatoire génère deux ensembles dont la distribution est complètement imprévisible et l'ordre difficilement discernable. L'approche à partir d'une « graine aléatoire » (*random seed*) propose une distribution aléatoire déterministe, car on donne une première valeur à partir de laquelle les séquences aléatoires sont générées. Cela veut dire que si on donne deux fois la même graine on va obtenir deux fois la même chaîne.

(ii) Dans l'approche par tranches on divise le corpus par groupes chronologiques chaque 25 ans et on réalise à l'intérieur de chaque groupe une division aléatoire en deux ensembles respectant le ratio 4 :1.

Cependant, comme on peut l'observer dans la figure 3.2, la différence statistique entre ces distributions montre à peine des différences remarquables, ce qui nous amène à penser qu'aucune n'est plus pertinente que l'autre. La distribution temporelle et typologique des trois jeux de données s'avère très similaire à la distribution de l'ensemble du corpus. Cela dit, la distribution aléatoire a été retenue afin de préserver les valeurs originelles dans la distribution du corpus. Il peut aussi être intéressant, afin que les expérimentations ultérieures soient reproductibles, de garder la clé de distribution.

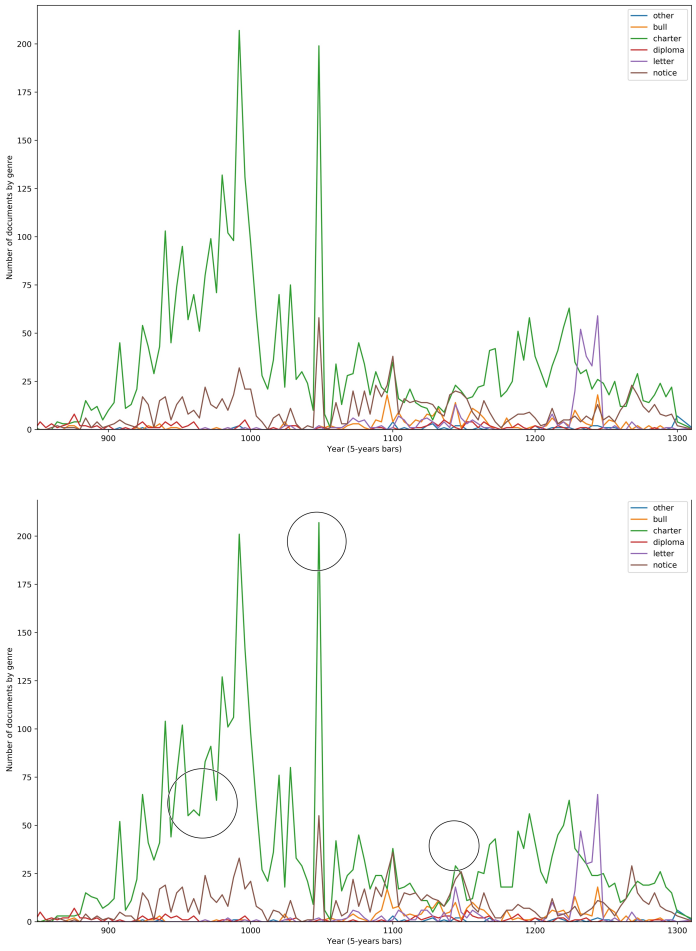

FIGURE 3.2 – Ventilation chronologique aléatoire du corpus (en haut) vs répartition par tranches à 25 ans (en bas). Les cercles indiquent les différences les plus remarquables

3.2.2 Validation croisée et formation de sous-corpus

Partie indispensable de la validation d'un modèle, la validation croisée est faite en répliquant de manière itérative la même comparaison training/test, mais en utilisant des ensembles d'entrainement réduits. En bref, on forme de petits ensembles en variant la composition de l'ensemble, en ce qui concerne la taille, la dimension, le nombre d'observations ou en faisant un regroupement par affinités, qui dans notre

cas correspond aux critères de la chronologie, des origines et des types. Ensuite, des modèles sont entraînés avec les sous-ensembles et on teste leur robustesse face aux autres ensembles d'entrainement ou face au corpus-test originel.

Cette ventilation maîtrisée du corpus peut constituer une réponse aux questions posées par la concentration et la représentativité du corpus. En favorisant une bonne vérification du niveau du surajustement du modèle, ainsi qu'une localisation des sous-ensembles les plus problématiques, elle favorise l'extension du modèle à d'autres corpus et valide notre approche générale. Par conséquent, afin d'obtenir un modèle robuste et d'étudier l'impact des facteurs hétérogènes sur le corpus, il a fallu appliquer trois sous-divisions dans le CBMA et construire un corpus test composé par des documents provenant d'autres régions que la Bourgogne.

a. Lors de la première expérience, on a pratiqué une validation croisée imbriquée (*nested cross validation*). Tout le corpus a été divisé en 10 parties ($K_1 ... K_{10}$) d'environ 500 documents. Ensuite on a entrainé dix modèles en progression arithmétique (K_1, K_1+K_2, $K_1+K_2+K_3$). Pour chaque modèle généré, les mêmes protocoles de validation que ceux du modèle général ont été appliqués. L'objectif était de trouver le meilleur équilibre entre l'efficacité et la taille du sous-ensemble d'entraînement, afin de développer un modèle moins dépendant du corpus d'origine, plus robuste sur des corpus variés et moins exigeant en termes de ressources informatiques (Figure 3.3). Cela permet également d'estimer la quantité d'annotations manuelles nécessaires pour parvenir à un bon niveau de performance.

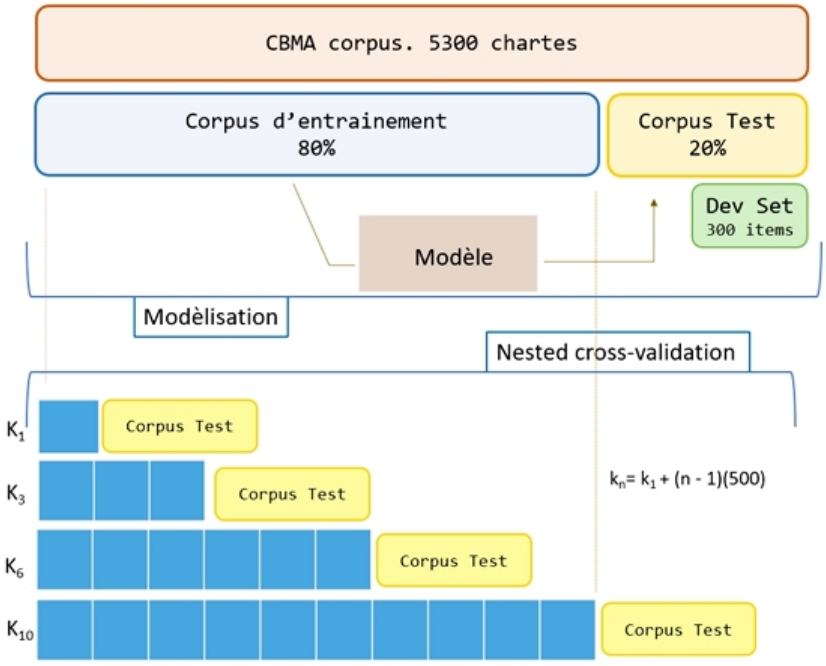

FIGURE 3.3 – Modélisations à partir des ensembles d'entrainement et test et validation croisée.

b. Suivant des paramètres similaires, 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 d'entraînement et de corpus test afin de réaliser des comparaisons *1 vs all*

et vice-versa. Cette expérience pourrait être critiquée dans la mesure où il est assez commun de trouver des chartes non datées, mal datées, mais surtout datées selon une fourchette. Ce problème a été corrigé en plaçant une même charte à l'intérieur de deux sous-corpus, si sa date estimée enjambait deux siècles – par exemple 980-1020. Comme dans le cas précèdent, une validation croisée entre les quatre ensembles a permis de faire des comparaisons plus précises et de vérifier l'effet de la variabilité ainsi que de tester la robustesse du modèle sur différentes unités chronologiques, puis de valider ou non l'application du modèle à une plage temporelle plus large.

c. L'une des observations les plus préoccupantes pendant la ventilation du corpus était la présence de périodes et zones avec très peu voire aucune présence documentaire dans le corpus annoté. Cette pénurie intrinsèque au corpus clunisien n'est pas forcement la règle pour d'autres corpus proches. Afin de réduire l'impact de ce manque de données, nous avons introduit une nouvelle modification dans le corpus originellement annoté en ajoutant un sous-corpus de 400 documents supplémentaires balisé à la main afin de couvrir les « zones grises » des IXe, XIIe et XIIIe siècles. Puisque le corpus est très dense pour la période qui va de 940 à 1090, il n'a pas été jugé nécessaire de couvrir les petits hiatus trouvés durant cet intervalle de temps. L'objectif final était d'éviter de perdre, à cause de ces lacunes chronologiques, certaines variétés scripturales et d'apporter des documents de la même époque en remplacement de ceux perdus soit par sélection, soit à cause d'avatars historiques.

À cet effet nous avons sélectionné des actes provenant d'autres diocèses que celui de Mâcon, à savoir, l'Yonne, Dijon, Autun, Nevers et Langres disponibles dans notre corpus soit sous la forme de cartulaires soit sous la forme de recueils factices [219]. Nous avons privilégié des actes produits depuis la fin du XIe siècle jusqu'à la première moitié du XIIIe siècle, lorsque la conservation du corpus clunisien fut prise en défaut. De plus nous avons annoté 20 actes de l'edition de la *Bibliotheca cluniacensis* dont les documents à l'origine appartenant au chartrier clunisien avaient été détachés et ne sont pas inclus dans l'édition de A. Bernard. Pour l'annotation de ce jeu de 400 documents nous avons appliqué les mêmes protocoles et le nouveau standard d'annotation proposé après les modifications sur le corpus originellement annoté.

d. Enfin, le test final réside dans la validation de la performance sur d'autres

219. CHARMASSE A. de, Cartulaire de l'Eglise d'Autun, 1978 (35 actes) ; LESPINASSE René de, Cartulaire du prieuré de la Charité-sur-Loire (Nièvre), 1887 (35 actes) ; QUANTIN Maximilien, Cartulaire général de l'Yonne, 1873 (70 actes) ; P. JUENIN, Nouvelle Histoire de l'abbaye royale et collégiale de Saint-Filibert et de la ville de Tournus, 1733 (25 actes) ; CHARRAULT abbé L., La chartreuse de Bellary (1209-1793), 1908, (20 actes) ; LESPINASSE René de, Les chartes de Saint-Etienne de Nevers, 1907 (20 actes) ; LALORE Ch., Chartes de l'abbaye de Mores, 1873 (30 actes) ; DUBY G., Recueil des pancartes de l'abbaye de la Ferté-sur-Grosne : 1113-1178, 1953 (15 actes) ; RICHARD J., Le Cartulaire de Marcigny-sur-Loire : 1045-1144, 1957 (15 actes) ; PROU M., VIDIER A., Recueil des chartes de l'abbaye de Saint-Benoît-sur-Loire, 1907 (25 actes) ; CHEVRIER G., CHAUME M., Chartes et documents de Saint-Bénigne de Dijon, prieurés et dépendances : des origines à 1300, 1943 (30 actes) ; MARILIER J. (abbé), Chartes et documents concernant l'abbaye de Cîteaux, 1098-1182, 1961 (15 actes) ; BERTHOUMEAU L., Du vol et de sa répression en Bourgogne sous l'ancien droit et Chartes de l'abbaye de Saint-Etienne de Dijon, de 1260 à 1270, 1914 (15 actes) ; LAURENT J., Cartulaires de l'abbaye de Molesmes, ancien diocèse de Langres, 916-1250, 1907 (30 actes) ; COTTIN H., Chartes de l'abbaye Saint-Etienne de Dijon, de 1291 à 1300, 1910 (20 actes) ; MARRIER M., Bibliotheca cluniacensis, 1614 (20 actes).

corpus que ceux mobilisés pendant la formation du modèle. À cette fin ont été annotés quatre sous-corpus formés par de chartes provenant de quatre régions avec une intense production documentaire (table 3.1), dont les sources sont :

1. le corpus des cartulaires d'Île de France (XIIe-XIIIe siècles) publié par l'École des Chartes [220] ;
2. le corpus DEEDS (Xe-XIIIe siècles) réunissant des chartes d'origine ecclésiastiques des abbayes du sud de l'Angleterre [221] ;
3. le CDML (Codice diplomatico della Lombardie Medievale, XIe-XIIe siècles) contenant des chartes lombardes de chancellerie et d'origine ecclésiastique [222] ;
4. le corpus CODEA (CHARTA, Xe-XVe siècles) composés par des chartes médiévales castillanes dont un des objectif est de former un corpus pour l'étude de la formation du castillan médiéval [223].

Ces corpus de petite taille – 70 à 100 chartes – ont été annotés à la main suivant les mêmes protocoles que le CBMA, après avoir été constitués selon les mêmes critères chronologiques et typologiques que le corpus principal. Tout cela vise à produire un cadre de validation de la robustesse de notre modèle, avec des documents similaires du point de vue typologique, chronologique et diplomatique, mais provenant d'espaces extérieurs à la Bourgogne.

Siècle/ Corpus	ANGLO	CASTILE	ILEFRANCE	LOMBARDY	corpus originel	corpus modifié	ensemble_extra 400 docs
10th	10	10	12	10	2292	3230	12
11th	24	11	22	16	1510	2050	27
12th	24	15	53	22	816	860	182
13th	12	14	63	2	638	730	149
Nº Tokens	11110	15616	41608	12441	1096095	1096095	104330
Nº Entities	1326	1841	3594	1222	84752	84752	8263

TABLE 3.1 – Nombre de chartes par siècle et nombre de « tokens » et d'entités nommées dans les principaux corpus et les corpus test européens.

3.3 Modèle et algorithme

Sur le plan formel, un texte est une composition relationnelle et séquentielle de signes interdépendants. De ce point de vue lexical, les entités nommées correspondent à une catégorie bien distincte du reste des mots. Elles ne sont pas recensées dans les dictionnaires et en conséquence elles ne sont pas facilement idéntifiables, mais elles sont un composant à part entière de la phrase et jouent un rôle syntaxiquement très

220. http://elec.enc.sorbonne.fr/cartulaires/
221. https://deeds.library.utoronto.ca/
222. http://www.lombardiabeniculturali.it/cdlm/
223. http://www.corpuscodea.es/corpus/consultas.php

similaire aux noms propres. Puisque le défi principal de la reconnaissance des entités nommées est de diviser un texte entre les mots constitutifs de ces entités et ceux qui entrent dans une autre catégorie, puis de baliser chaque entité, une analyse qui identifie un grand nombre d'attributs de chaque mot s'avère fondamentale. La caractérisation de chaque mot, en tant qu'unité sémantique, permet de construire un index de ses propriétés intrinsèques (*features*), ce qui constitue l'un des piliers de l'opération de classification.

Un autre pilier de l'analyse correspond à la caractérisation des mots en tant qu'éléments relationnels formant des séquences. Le but ultime des modèles de reconnaissance automatique est de bien prédire la séquence d'étiquettes adaptée à la séquence d'observations. La reconnaissance des entités nommées est conditionnée à une bonne analyse de leurs contextes d'apparition dans la phrase. La fonction remplie par une entité, comme tous les autres mots de la séquence, est précisée par le rôle que joue chaque élément dans la séquence.

Cela dit, l'algorithme doit modéliser, en s'appuyant sur des séquences multidimensionnelles, des données qui contiennent une matrice d'attributs pour chaque mot et une matrice des différents états relationnels de chaque mot dans la séquence.

Parmi les différentes méthodes qui peuvent être adoptées pour modéliser cela, la technique des champs aléatoires conditionnels (ou CRF pour *Conditional Random Fields*) semble appropriée [224]. Celle-ci accepte des modélisations multidimensionnelles, considérant le poids des attributs dans l'étiquetage de la séquence. De plus, cette technique modélise directement le problème de prédiction standard et analyse plusieurs relations de la séquence, sans considérer des dépendances trop strictes — se limiter aux mots adjacents — ce qui est plus adapté à la réalité linguistique de notre corpus dans lequel deux mots reliés peuvent ne pas être situés de manière adjacente dans la séquence.

La modélisation de toutes les relations possibles entre les variables en question, de tous les états et 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 de chaque séquence d'étiquettes d'être correcte selon certaines observations, ce qui est généralement suffisant pour déterminer la classe d'une entité.

3.3.1 Matrice de données

Pour appliquer cette forme d'apprentissage automatique supervisé, chaque mot d'une phrase doit être considéré comme un *token* [225] dont il est indispensable d'expliciter certaines propriétés. L'ensemble du corpus a été converti en un format tabulaire, fournissant des informations lexicales, syntaxiques et morphologiques de

224. Lafferty et al., "Conditional random fields : Probabilistic models for segmenting and labeling sequence data" ; Hanna M Wallach. "Conditional random fields : An introduction". In : *Technical Reports (CIS)* (2004), p. 22

225. On parle souvent de « tokens » pour éviter de leur donner une valeur linguistique trop marquée, car « mot » a une définition linguistique et la "tokenisation" ne conduit pas forcément à des mots au sens strict.

chaque *token*. Ainsi, le corpus entier devient une base de données dans la mesure où chaque mot est reproduit dans un tableau à sept colonnes comme suit :

— TOKEN (mot d'origine)
— POS (catégorie morphosyntaxique – *Part-of-speech*)
— LEMMA (forme sans déclinaison du mot)
— CASE (indique si la première lettre est en majuscule ou en minuscule)
— SUFFIX (trois derniers caractères de chaque mot)
— ENTITÉ (appartenance ou non à la catégorie des entités nommées – une colonne pour les noms de personnes, une pour les noms de lieux)

Les trois premières colonnes donnent au modèle un premier niveau de catégorisation car elles ajoutent les informations grammaticales et morphologiques du texte. Ils contiennent la version segmentée – réduite à des unités indivisibles – de chaque mot, la catégorie morphosyntaxique (*Part-of-speech*), obtenue à partir d'une version de TreeTagger [226] développée par le groupe Omnia en 2013 [227] et le lemme, 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 un suffixe figé, formé des trois dernières lettres, où figure la déclinaison en latin déterminant la fonction grammaticale du mot.

Comme dans la plupart des modèles traitant d'un problème de classification, nous utilisons le format BIO pour représenter les entités nommées, ce que correspond aux deux dernières colonnes. Pour rappel (voir partie 2.2) : B, I et O représentent le début (B-entité, *Begin*), la poursuite (I-entité, *Inside*) ou l'absence (O, *Outside*) d'une entité.

Nous avons considéré le problème en deux étapes : la première étape extrayait les noms de personne, la seconde étape les noms de lieux. Un seul classificateur extrayant conjointement les noms de personnes et de lieux personnels n'a pas pu être implémenté, car le corpus contient de nombreuses entités qui se chevauchent. C'est pourquoi les dernières colonnes de la table 3.2 répertorient les classes au format BIO. Cependant étant donné que l'imbrication porte presque toujours sur un nom de personne contenant un nom de lieu, et non l'inverse, les deux étapes permettent au modèle des personnes de se servir de l'information prédite sur les lieux.

La conception du modèle passe ensuite par une phase de définition des caractéristiques que l'algorithme utilisera pour réaliser ses calculs. Comme l'illustre la table 3.2, nous utilisons, pour chaque *token* (ligne) :

— La valeur de surface du token lui-même (TOKEN), ainsi que des deux tokens précédents et des deux tokens suivants, soit une fenêtre glissante de 5 tokens.
— La catégorie morphosyntaxique (POS) et le lemma (LEMMA) du token
— La capitalisation des mots
— L'information concernant l'annotation.

L'algorithme L-BFGS, fourni par Wapiti [228], permet une optimisation de la sélection des données par CRF et ainsi l'usage de la mémoire vive de l'ordinateur. Il

226. Helmut Schmid. "Treetagger| a language independent part-of-speech tagger". In : *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart* 43 (1995), p. 28
227. `http://glossaria.eu/outils/lemmatisation/`
228. L'outil se trouve disponible dans `https://wapiti.limsi.fr/`

TOKEN	POS	LEMMA	CASE	SUFFIX	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 %[-1,0]	NAM	-	UPPER	aco	I-PERS	B-LOC
perpendens %x[-2,0]	VBE	perpendeo	LOWER	ens		
,	PON	,	LOWER	-		

TABLE 3.2 – Example d'entrainement pour la séquence *Quod ego Hugo de Berziaco perpendens*. La zone grise indique une seule observation (concernant le mot "de") qui combine toutes les caractéristiques de tous les colonnes dans une fenêtre de 5 tokens dans une fenêtre de 5 tokens (2 tokens avant et deux tokens après le token observé)

s'agit d'un outil d'étiquetage de séquences développée dans le laboratoire Limsi-CNRS qui est capable de travailler sur des données multi-labélisées avec des millions de caractéristiques [229].

L'algorithme - et par extension la méthode CRF - fonctionne alors en formant un modèle discriminant et en recherchant la meilleure option d'état à partir d'un corpus d'apprentissage contenant des observations sur des états et des attributs balisés. À partir d'une série d'observations étiquetées, il construit une interprétation et détermine l'étiquette la plus probable pour une nouvelle séquence inédite.

3.4 Résultats des expérimentations sur les modèles

La figure 3.4 résume le processus d'apprentissage : le corpus de 5 000 documents annotés subit les divers processus détaillés ci-dessus, afin d'obtenir notre modèle de reconnaissance des entités nommées. Vient ensuite une étape de validation appliquant le modèle sur l'ensemble réservé au test et en effectuant une validation croisée. Puis, on compare les résultats obtenus avec le modèle principal et les sous-modèles à ceux fournis grâce à l'annotation manuelle dans l'ensemble test. Tous ont été évalués avec les mesures traditionnelles de précision, de rappel et de F1-mesure (voir partie 2.3) avec deux différentes configurations implémentées dans l'outil BratEval [230] : La correspondance exacte (ou EM pour *Exact match*), correspond à un vrai positif, c'est-à-dire à une entité nommée correctement catégorisée et dont les sont limites parfaitement identiques à celles annotées dans le golden-corpus. La correspondance partielle (ou PM pour *Partial match*) correspond à un vrai positif dont l'entité extraite ne partage toutefois qu'une correspondance partielle avec l'entité annotée dans le corpus de référence. Cette catégorie regroupe essentiellement des entités composées de plusieurs éléments, ce qui représente entre 15 % et 20 % du total. Une fois le modèle validé, on peut l'appliquer aux documents non annotés (voir figure 4.4).

229. Thomas LAVERGNE et al. "Practical Very Large Scale CRFs". In : *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Uppsala, Sweden : Association for Computational Linguistics, 2010, p. 504-513. URL : http://www.aclweb.org/anthology/P10-1052

230. Le script BratEval est disponible librement dans : https ://bitbucket.org/nicta_biomed/brateval

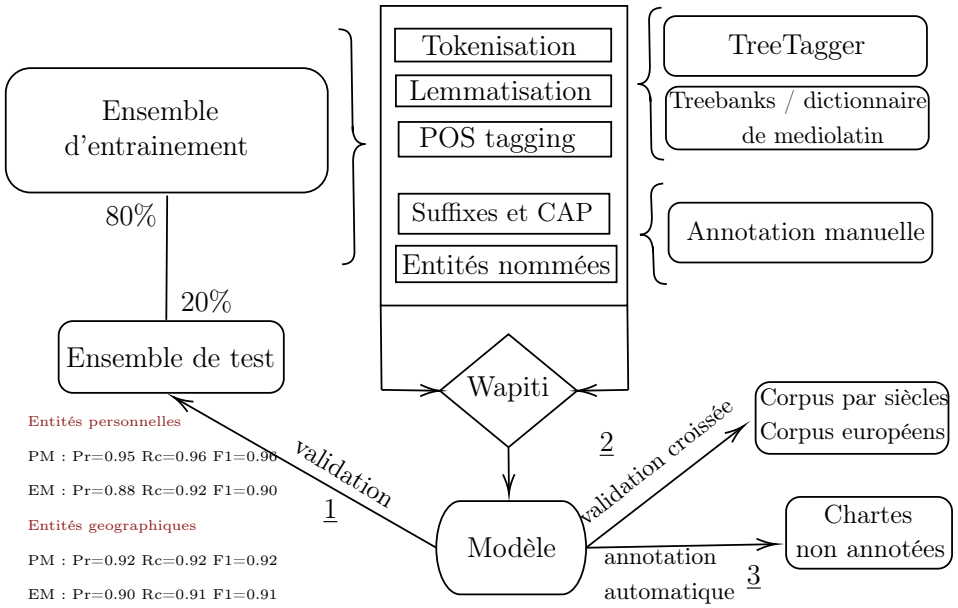

FIGURE 3.4 – Résumé du processus de modélisation de la reconnaissance des entités nommées

3.4.1 Modèle général

Nous avons d'abord produit le modèle général formé à partir des ensembles principaux du CBMA, soit pour mémoire 4 000 documents de corpus d'entraînement, 1 000 documents de corpus-test et 300 documents de corpus de développement. Ce modèle atteint dans le cas des entités personnelles un *exact match* sur la F1-mesure de 0,95 et *partial match* de 0,96 ; concernant les entités de lieu il atteint un *exact match* de 0,91 et un *partial match* 0,92 (voir table 3.3).

Nom de personne	PR	RC	F1	TP	FP	FN
B-PERS	0.95	0.96	0.96			
I-PERS	0.88	0.92	0.90			
Partial match	0.95	0.97	0.96	12965	615	291
Exact match	0.93	0.96	0.95	12729	851	529
Nom de lieu						
B-LOC	0.91	0.93	0.92			
I-LOC	0.81	0.80	0.80			
Partial match	0.92	0.92	0.92	7171	590	550
Exact match	0.90	0.91	0.91	7035	726	681

TABLE 3.3 – Meilleur ratio de reconnaissance en termes de précision et recall et selon les paramètres de l'outil Brateval : TP (true positive), FP (false positive) et FN (false negative)

Si l'on observe les expériences d'évaluation dans tous les autres ensembles de taille progressive, elles montrent que précision et rappel sont élevés, ce qui écarte le risque d'avoir un modèle sous-ajusté (voir partie 4.2). De plus, le fait d'avoir de très bons résultats sur un corpus test suffit pour écarter, sur ce corpus, le risque d'un modèle surajusté. La différence entre les résultats des *exact* et *partial matches* ne dépasse pas

deux points, ce qui confirme que la confusion entre catégories d'entités de la part du modèle est rare et que la détection des limites des entités nommés (I-PERS et I-LOC) sont satisfaisantes, cette dernière tâche étant souvent ardue dans les travaux de TAL.

En outre, dans les analyses visant à déterminer le meilleur rapport performance/taille, on observe, comme prévu, une performance nettement supérieure entre le modèle k_{10} et le modèle k_1 (plus de 15 points de différence). La moyenne peut être retrouvée dans k_5 (2 500 documents) qui montre un résultat de 90 points, seulement 3 points au-dessus de la meilleure performance alors que le modèle a été entraîne avec la moitié de documents. La plus mauvaise performance est observée à partir du modèle k_3, au-dessous duquel les résultats chutent jusqu'à 79 % en k_2 et 74 % en k_1, les rendant presque négligeables.

Dans une deuxième expérimentation (tableau 3.4), nous avons ajouté à chacun des k-modèles le corpus supplémentaire de 400 documents annoté à la main pour éviter la disparition de styles mineurs et minimiser les zones grises. L'ajout de ce corpus montre de discrètes améliorations dans les performances de k_1 à k_3, qui s'intensifient à partir de k_4 et qui ajoutent une amélioration significative en k_{10} (2 points de plus). Si on recalcule les résultats après cet ajout, k_4 (1500 + 400 docs) présente la performance la plus équilibrée entre taille et robustesse, 5 points au-dessous de la meilleure performance alors qu'il n'est entraîné qu'avec 38 % de l'ensemble principal d'apprentissage (plus l'ensemble de 400 docs). Ceci vient valider l'hypothèse qu'on peut proposer un modèle beaucoup moins coûteux en termes d'annotation et de ressources mobilisées sans perdre trop en efficacité.

Entité/		B-PERS			I-PERS			B-LOC			I-LOC		
Mesure/		Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1
k_{10}	4000+ 400	0.95	0.96	0.96	0.88	0.92	0.90	0.91	0.93	0.92	0.81	0.80	0.80
k_5	2500	0.90	0.91	0.91	0.86	0.90	0.88	0.89	0.90	0.90	0.80	0.79	0.80
k_4	1500+ 400	0.90	0.89	0.90	0.85	0.87	0.86	0.88	0.87	0.88	0.78	0.76	0.77
k_3	1500	0.86	0.84	0.85	0.81	0.82	0.81	0.86	0.82	0.84	0.73	0.70	0.71
k_2	1000	0.81	0.77	0.79	0.77	0.79	0.78	0.82	0.77	0.79	0.68	0.63	0.66
k_1	500	0.75	0.72	0.74	0.73	0.69	0.72	0.76	0.70	0.73	0.64	0.58	0.60

TABLE 3.4 – Validation croisée entre tous les sous-ensembles de corpus. PERS : personnes, LOC : lieux, PM : *partial match* ; EM : *exact match*

Si on fait une interprétation plus détaillée des résultats, on constate que nous avons obtenu, dans les modèles les plus performants, un taux élevé de *match partiel* (0.96 et 0.92) - c'est-à-dire une reconnaissance de l'entité même si elle demeure incomplète. Cela étant dit, notre modèle permet de détecter des entités nommées, quelle que soit leur taille, avec une efficacité d'au moins 96 % et de les classer correctement dans au moins 90 % des cas, puisque le score minimum d'*exact match* est de 91 %. De plus, la différence observée entre *partial match* et *exact match* n'est jamais supérieure à 2 points, ce qui implique que la reconnaissance des entités complexes avec au moins deux composants 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 *partial* et *exact matches* sont respectivement 4 points en dessous des résultats obtenus pour les noms de personne (91 % et 92 %).

3.4.2 Modèles par siècles

Les tableaux 3.5 (noms de personnes) et 3.6 (noms de lieux) montrent le résultat de nos expériences de validation croisée sur les modèles par siècle. Pour rappel, nous avons formé quatre sous-ensembles de documents triés par siècle (du Xe au XIIIe siècle), que nous avons ensuite modélisés puis testés. Les résultats croisés n'ont pas montré une trop grande hétérogénéité et ils sont proches de ceux obtenus avec des modèles de grande taille (entre 5 et 10 points de moins). Les modèles présentant la meilleure performance sont ceux des siècles centraux (XIe et XIIe siècles), et les plus faibles sont ceux des siècles périphériques (Xe et XIIIe siècles). Trois faits saillants paraissent ici remarquables :

1. le modèle du Xe siècle, entrainé à partir du plus grand ensemble de données, atteint la performance la plus défavorable lorsqu'il est appliqué aux autres chartes, en particulier aux ensembles du XIIe et XIIIe siècle ;
2. l'application des modèles du Xe, XIe et XIIe siècles sur l'ensemble du XIIIe siècle offre le moins bon rapport taille/performance ;
3. la moins bonne performance individuelle est obtenue en appliquant le modèle du Xe siècle sur les documents du XIIIe siècle.

Cette relative homogénéité entre les quatre résultats et la proximité avec ceux du modèle général suggèrent deux observations principales : (i) la régularité observée au fil des siècles prévaut sur les changements, tout drastiques qu'ils aient pu paraître dans la variété de la formulation diplomatique. Le schéma (*pattern*) que le modèle doit capturer doit donc être assez régulier dans sa morphologie mais comporter des pics de variété toutefois bien contrôlés, ce qui pourrait expliquer qu'on atteint une détection constante des phénomènes de dénomination lorsque l'on considère les modèles entrainés sur un seul siècle ; (ii) en conséquence, les problèmes de détection sont concentrés au niveau des attributs spécifiques et de ce que le modèle considère comme « bruit-de-fond ». Ce phénomène est probablement attribuable à des événements spécifiques pouvant correspondre à des dénominations irrégulières ou complexes. La moindre performance des modèles des Xe et XIIIe siècles renforce cette hypothèse et indique la présence de variations scripturales ou de changements dans la composition du nom plus importantes durant ces deux siècles qu'au cours des deux siècles centraux. Dans les actes du Xe siècle, le nom n'est pas encore complexe et l'apparition du nombre double est mieux visualisée dans les actes du siècle suivant, en conséquence le modèle formé n'arrive pas à bien saisir les différentes formes de former les noms attestés dans les siècles postérieurs. Par ailleurs, les documents clunisiens du XIIIe siècle, lorsque l'acte de donation a déjà disparu, correspondent principalement à des lettres et bulles dont les modèles formulaires, et par extension les façons de présenter les entités nommées, ne sont pas ceux de l'acte de mutation foncière dominante dans les siècles précédents.

3.4.3 Modèles européens

Dans la même veine, les tableaux 3.5 et 3.6 — correspondant aux noms de personnes et de lieux — présentent les résultats obtenus en appliquant nos différents modèles aux corpus « étrangers » décrits à la section 3.2.2. Nous avons appliqué à ces corpus régionaux tous les modèles générés jusqu'ici : d'une part les modèles les plus performants, exposés dans le passage b de 4.4.1 et que l'on peut appeler « modèles généraux CBMA », et, d'autre part, les quatre sous-modèles entraînés avec les documents provenant d'un même siècle dont nous venons d'analyser les résultats. L'application des modèles généraux aux chartes européennes donne de bons résultats, alors que l'application des modèles par siècle est moins robuste, mais présente tout de même une très bonne performance.

Les six résultats obtenus sont assez proches. Les modèles généraux offrent une meilleure performance en F1-mesure, environ 2 à 4 points supérieure aux modèles par siècle, et, encore une fois, nous constatons que le modèle formé sur 1 900 chartes (1500 + jeu de 400) montre une performance légèrement inférieure (1-2 points) par rapport au modèle de 4 400 chartes (4000 + jeu de 400). L'utilisation d'un plus petit ensemble de documents annotés pourrait donc être acceptable dans la modélisation. Chacun des modèles par siècle, atteint presque les mêmes résultats en étant appliqué aux différents corpus de chartes européennes, excepté celui du XIIIe siècle qui est plus hétérogène.

En ce qui concerne l'analyse des résultats fondés sur l'outil BratEval, le *partial match* dans PERS varie entre 94 % et 93 %, et l'*exact match* montre des résultats compris entre 88 % et 82 % pour les quatre corpus régionaux. Dans le cas d'entités LOC, le *partial match* atteint des résultats compris entre 85 % et 82 % et l'*exact match*, entre 81 % et 73 %. La différence entre *partial* et *exact match* est beaucoup plus prononcée (environ 6 - 7 points) que celle obtenue dans le premier modèle général (1 point) ce qui indique des problèmes plus sérieux dans la reconnaissance des entités composées, notamment géographiques.

En analysant uniquement les résultats numériques des modèles par siècle sur ces chartes, nous avons relevé de nombreuses similitudes avec les résultats obtenus lors de l'évaluation croisée des modèles par siècle entre eux :

(i) le modèle du Xe siècle, formé avec le plus grand nombre de documents, offre la plus faible efficacité en termes de reconnaissance sur les corpus étrangers. Nous avons également constaté une corrélation dans les résultats entre les modèles par siècle et les corpus de chartes européennes ;

(iii) Le corpus parisien (Île-de-France), composé principalement de chartes médiévales tardives (notamment de la fin du XIIe et du XIIIe siècle), offre des résultats légèrement plus élevés lorsqu'il est évalué avec le modèle du XIIe siècle et ces derniers s'accroissent lorsqu'il est évalué avec le modèle du XIIIe siècle, qui est pourtant le moins efficient sur les autres corpus européens.

En outre, nous pouvons nous permettre de diviser les corpus en deux selon les différences qui existent entre *partial* et *exact match*. D'un côté les corpus parisien et lombard, pour lesquels on observe des différences relativement modestes (6-7 points), et de l'autre, les corpus anglais et espagnol qui présentent des différences plus élevées, notamment pour les chartes ibériques. Ces différences proviennent notamment de problèmes dans la reconnaissance des entités complexes qui peuvent avoir différentes

causes : formes de dénomination personnelle périphrastique, moins bonne segmentation ou nettoyage du corpus, manque d'organicité dans la composition du corpus. En fait, si l'on considère ce dernier détail, le corpus espagnol est à l'origine un recueil sans vrai organicité, et le corpus anglais, plus organique, est un recueil éditorial composé d'autres recueils factices ou de cartulaires.

Pour résumer, les similarités des résultats obtenus à partir des six modèles ces résultats nous permettent de renforcer les hypothèses formulées concernent les modèles par siècle :

(i) les particularités des quatre corpus externes ont été traitées de manière globale sans diminution considérable de la performance générale ce qui est la preuve que les modèles eux-mêmes sont robustes et suffisamment valides pour être appliqués à des corpus variés. Par ailleurs, cela suggère qu'il existe des similitudes à la fois dans la composition du nom et dans sa détermination contextuelle beaucoup plus fortes que les différences suggérées par chaque tradition régionale.

(ii) les documents contiennent une longue série de caractéristiques qui résistent aux changements scripturaux intervenus entre les Xe et XIIIe siècles, et qui sont si tant de la taille que de l'origine des chartes ; qui sont mieux traitées par un modèle itératif basé sur le formulaire et sur des traditions scripturaires très bien établies, assurant ainsi un rapport de reconnaissance initialement acceptable ;

(iii) il existe des ensembles plus discrets de caractéristiques, qui s'ajoutent aux précédentes, déterminées par des spécificités locales et des modifications de la documentation qui constituent la principale source d'erreur.

TABLE 3.5 – Résultats en termes de précision (Pr), rappel (Rc), f1-mesure (f1) et Brateval sur les entités nommées personnelles. Les values en rouge indiquent la difference entre exact match (EM) et partial match (PM). Legend : TP (true positive), FP (false positive), FN (false negative)

Model/ Test		ENGLAND							CASTILE							ILE_FRANCE							LOMBARDY						
		TP	FP	FN	Pr	Rc	F1		TP	FP	FN	Pr	Rc	F1		TP	FP	FN	Pr	Rc	F1		TP	FP	FN	Pr	Rc	F1	
4400	PM	471	32	37	0,94	0,93	0,93	-10	860	57	78	0,94	0,92	0,93	-12	1645	139	81	0,92	0,95	0,94	-7	535	23	59	0,96	0,90	0,93	-8
	EM	427	76	104	0,85	0,80	0,83		757	160	184	0,83	0,80	0,81		1524	260	213	0,85	0,88	0,87		495	63	106	0,89	0,82	0,85	
1900	PM	442	24	89	0,95	0,83	0,89	-6	843	52	96	0,94	0,90	0,92	-12	1629	103	101	0,94	0,94	0,94	-6	541	21	59	0,96	0,90	0,93	-5
	EM	416	50	115	0,89	0,78	0,83		736	159	205	0,82	0,78	0,80		1530	202	207	0,88	0,88	0,88		509	53	92	0,91	0,85	0,88	
10th	PM	487	38	41	0,93	0,92	0,93	-7	862	85	65	0,91	0,93	0,92	-17	1618	298	116	0,84	0,93	0,89	-7	541	27	56	0,95	0,91	0,93	-6
	EM	453	72	78	0,86	0,85	0,86		704	243	237	0,74	0,75	0,75		1495	421	242	0,78	0,86	0,82		507	61	94	0,89	0,84	0,87	
11th	PM	497	34	33	0,94	0,94	0,94	-6	857	62	79	0,93	0,92	0,92	-15	1627	230	105	0,88	0,94	0,91	-6	542	22	56	0,96	0,91	0,93	-5
	EM	468	63	63	0,88	0,88	0,88		720	199	221	0,78	0,77	0,77		1526	331	211	0,82	0,88	0,85		514	50	87	0,91	0,86	0,88	
12th	PM	466	26	66	0,95	0,88	0,91	-8	765	42	176	0,95	0,81	0,88	-12	1596	123	136	0,93	0,92	0,92	-6	494	18	106	0,96	0,82	0,89	-5
	EM	423	69	108	0,86	0,80	0,83		666	141	275	0,83	0,71	0,76		1479	240	258	0,86	0,85	0,86		465	47	136	0,91	0,77	0,84	
13th	PM	397	22	133	0,95	0,75	0,84	-13	662	31	278	0,96	0,70	0,81	-12	1522	90	204	0,94	0,88	0,91	-8	452	19	146	0,96	0,76	0,85	-7
	EM	338	81	193	0,81	0,64	0,71		566	127	375	0,82	0,60	0,69		1388	224	349	0,86	0,80	0,83		419	52	182	0,89	0,70	0,78	

TABLE 3.6 – Resultats idéntiques sur les entités nommées de lieux

Model/ Test		ENGLAND							CASTILE							ILE_FRANCE							LOMBARDY						
		TP	FP	FN	Pr	Rc	F1		TP	FP	FN	Pr	Rc	F1		TP	FP	FN	Pr	Rc	F1		TP	FP	FN	Pr	Rc	F1	
4400	PM	338	24	122	0,93	0,73	0,82	-6	561	57	141	0,91	0,80	0,85	-12	1194	181	249	0,87	0,83	0,85	-6	264	19	80	0,93	0,77	0,84	-3
	EM	311	51	145	0,86	0,68	0,76		477	141	208	0,77	0,70	0,73		1111	264	326	0,81	0,77	0,79		252	31	91	0,89	0,73	0,81	
1900	PM	342	29	118	0,92	0,74	0,82	-5	567	58	144	0,91	0,80	0,85	-12	1206	191	238	0,86	0,84	0,85	-6	259	19	84	0,93	0,76	0,83	-4
	EM	317	54	139	0,85	0,70	0,77		477	148	208	0,76	0,70	0,73		1114	283	323	0,80	0,78	0,79		246	32	97	0,88	0,72	0,79	
10th	PM	238	21	222	0,92	0,52	0,66	-6	462	64	245	0,88	0,65	0,75	-12	992	200	458	0,83	0,68	0,75	-6	204	17	141	0,92	0,59	0,72	-7
	EM	214	45	242	0,83	0,47	0,60		381	145	304	0,72	0,56	0,63		901	291	536	0,76	0,63	0,69		184	37	159	0,83	0,54	0,65	
11th	PM	329	29	131	0,92	0,72	0,80	-6	565	72	154	0,89	0,79	0,83	-13	1184	231	266	0,84	0,82	0,83	-7	260	23	84	0,92	0,76	0,83	-5
	EM	301	57	155	0,84	0,66	0,74		464	173	221	0,73	0,68	0,70		1087	328	350	0,77	0,76	0,76		245	38	98	0,87	0,71	0,78	
12th	PM	319	30	138	0,91	0,70	0,79	-4	569	78	140	0,88	0,80	0,84	-11	1193	189	248	0,86	0,83	0,85	-6	219	19	126	0,92	0,63	0,75	-5
	EM	300	49	156	0,86	0,66	0,75		483	164	202	0,75	0,71	0,73		1120	262	317	0,81	0,78	0,79		204	34	139	0,86	0,59	0,70	
13th	PM	282	53	174	0,84	0,62	0,71	-7	476	56	229	0,89	0,68	0,77	-12	1162	142	286	0,89	0,80	0,84	-6	164	35	180	0,82	0,48	0,60	-4
	EM	310	25	149	0,93	0,68	0,78		398	134	287	0,75	0,58	0,65		1064	240	373	0,82	0,74	0,78		152	47	191	0,76	0,44	0,56	

3.5 Le modèle des parties du discours diplomatique

Sur la base du modèle CRF dont nous venons d'exposer les détails, nous avons entraîné un modèle secondaire pour la détection des parties du discours diplomatique. Comme nous l'avons déjà mentionné le corpus que nous avons utilisé pour cette modélisation (*Codice diplomatico della Lombardia medievale* ou CDLM) est un corpus d'éditeur réuni dans un grand projet entre 2000 et 2006 à partir de différentes éditions de chartriers et collections de chartes [231] provenant des réseaux monastiques et ecclésiastiques lombards, dont les actes ont été émis principalement entre la deuxième moitié du XIe siècle et la fin du XIIe siècle [232]. On peut trouver, intégrées dans le projet, des éditions critiques de chartes originellement publiées en papier (par exemple celles du monastère de San Pietro in Ciel d'Oro [233] et du monastère de Santa Maria di Morimondo [234]), mais une partie importante des chartes, spécialement celles des petits recueils, on été éditées *ex novo* pour cette édition numérique et n'existent pas ailleurs.

L'édition, qui compte 5175 documents, est très riche en actes relevant de la diplomatique notariale. L'institution notariale avait en effet été confirmée en Italie bien avant qu'en France où elle n'est réintroduite que vers la moitié du XIIe siècle. Chronologiquement la plupart de sa documentation appartient au moment de transition de la charte vers l'instrument public [235]. En conséquence, on retrouvera des actes notariés suivant des formulaires et formes d'instrumenter qui peuvent présenter certains dissemblances formelles par rapport à ce que nous sommes habitués à trouver dans les recueils clunisiens. En outre, l'édition contient un nombre très élevé de *muminima*, actes souvent stipulés entre privés conservés dans les archives des monastères afin de garder l'historique des possessions ou droits récemment acquis. Ainsi, on retrouve dans l'édition des actes qui concernent directement les monastères et d'autres les affaires contractuelles entre laïcs. L'édition du CDLM est d'ailleurs assez bien répartie entre différentes actions juridiques. On retrouve un nombre important d'actes privés souvent notariés - du fait que le clergé et l'aristocratie constituent la majorité des clients de notaires - concernant ventes, achats, achats-ventes et échanges, ainsi que des documents de nature judiciaire et de gestion. De même il existe un nombre important d'actes publics notariés - le notariat impérial et le notariat apostolique étaient aussi très bien établis- : confirmations impériales, royales et pontificales, actes épiscopaux, privilèges, lettres, etc. Par contre, les actes de donation tant privés que

231. Un chartrier est l'ensemble des chartes conservées par une personne physique ou morale -le plus souvent, un seigneur, une institution ecclésiastique, une ville pour faire la preuve de ses droits ou conserver la mémoire de son histoire ; une collection est un ensemble matériel de pièces réunies artificiellement par l'effort volontaire de l'auteur de ce rassemblement.Ortí, *Vocabulaire international de la diplomatique*, p. 27

232. La répartition du corpus, qui va du IXe au XIIIe siècle est en effet inégale : autour de 70% des documents sont situés entre le milieu du XIe à la fin XIIe siècle. Ils sont répartis dans plus de 60 éditions avec des séries courtes ; les plus longues comptent entre 200 et 400 documents (10 éditions), les plus courtes ne comptent qu'une vingtaine de chartes

233. Ezio Barbieri et al. *La carte del monastero di San Pietro in Ciel d'Oro di PAvia*. Fontes, 1984

234. Michele Ansani. *Le carte del monastero di Santa Maria di Morimondo, 2 vol (1010-1200)*. Spoleto, 1992

235. Les minutes sont encore rares à trouver, car l'édition s'arrête à la fin du XIIe siècle lorsque ils commencent se faire communs dans la pratique notariale.

publics, qui constituent le cœur du CBMA, sont peu nombreux et on les retrouvera surtout dans les documents du Xe et XIe siècle [236].

La version numérique du corpus

La version numérique du corpus CDLM est assez lourde à traiter car c'est une édition complexe qui inclut dans un même ficher XML des balises concernant les caractères externes du document (état de conservation, support de l'acte ou matière subjective), les caractères internes (langue, parties du discours) et même, bien que de manière incomplète, de l'information concernant les entités nommées de personnes, lieux et des liens de parenté. Les éditeurs ont suivi des lignes éditoriales proches des attentes d'une édition diplomatique traditionnelle, reproduisant ainsi le travail d'édition effectué auparavant. En conséquence, dans le fichier XML un même mot doit supporter souvent jusqu'à 3 étiquettes à la fois (sans compter les attributs), ce qui peut rendre son exploitation assez compliquée en termes d'analyse massive.

```
<!--
=====================================================================
DTD dei Codice diplomatico della Lombardia medievale (CDLM)
http://cdlm.unipv.it/edizioni/cdlm.dtd
=====================================================================
-->
#actes différents de la charte:
<!ELEMENT TENOR >
    <!ELEMENT MINUTA >
    <!ELEMENT REGESTUM >
    <!ELEMENT NOTITIA >

<!ELEMENT PROTOCOLLO #cadre formel initial
    (INVOCATIO | INTITULATIO | INSCRIPTIO | DTCRON | DTTOP |)*>

<!ELEMENT TESTO #Partie se rapportant directement à l'acte juridique.
    (EXORDIUM | NARRATIO | PROMULGATIO | DISPOSITIO | ESTIMATIO |
    RES | FORMULAE | CLAUSULAE | CORROBORATIO | SANCTIO |  CED)*>

<!ELEMENT ESCATOCOLLO #cadre formel final
    ( DATATIO | DTCRON | DTTOP | SMR | SMC | SMF | SMT | SME |
    SUBSCRIPTIO | RECOGNITIO | COMPLETIO | IT | ROGATIO | IUSSIO |
    TENOR-ADDITUM | CED)*>
=====================================================================
```

236. Les classements proposés dans l'édition sont très fines (par ex. "Pagina constitutionis Oberti Mediolanensis archiepiscopi", "Carta investiture causa iudicati inter vivos"), ce qui complique l'obtention de chiffres globales sur les typologies documentaires et les affaires conclues. Nous avons calculé, grossièrement, qu'il existent environ 3054 chartes, 621 breves, 174 libellus, 111 sententias, 66 préceptes, 38 privilèges, 36 diplomas dont 1525 ventes (*Carta venditionis*), 512 échanges (*Cartula comutacionis/permutationis*), 211 donations (*Carta donationis/offersionis*), 64 loyers (*Carta locationis*), 208 investitures [de fief] (*Cartula investiture*), 130 concessions (*cartula promissionis*), 96 fines (*finis cartula*), 21 lettres d'évêques et archevêques, etc.

Parmi ces éléments, ceux qui nous intéressent ici, les parties du discours, ont été balisées selon un cadre assez formel d'acte juridique [237]. Les actes du CDLM sont en général plus longs que ceux trouvés dans le CBMA (ils comptent environ le double de lignes que ceux du CBMA pour un nombre similaire de documents) et se rapportent à certains parties du discours qu'on ne voit, au moins jusqu'à l'avènement du XIIIe siècle, que rarement mobilisées dans la rédaction la plus courante, car ils correspondent à des actes notariés, à des chartes solennelles ou de chancellerie. La définition de type de document (DTD), énoncée ci-dessus, nous montre l'"arbre" des éléments considérés dans la composition de l'acte juridique [238] :

La teneur de l'acte privé est composée par trois groupes bien signalés dans la DTD : le protocole (le cadre formel initial, « Protocollo » dans la DTD), le texte (« Testo », partie centrale du document qui se rapporte à l'acte juridique) et l'eschatocole (« Escatocollo », cadre formel final). À l'intérieur de ce cadre formel peuvent être utilisés différents modules scripturaires (présentés comme éléments-fils dans la DTD), ce que nous appelons les parties du discours diplomatique, destinées à donner validité à l'acte et à reproduire, avec plus ou moins rigidité (voir le chapitre 5 pour une étude plus en détail des parties du discours) le modèle rédactionnel ou formulaire qui correspond à l'action juridique en cours.

L'allure du texte, autrement dit, l'aspect formel, général de sa rédaction, peut présenter des variations importantes selon les différentes circonstances qui doivent être exprimées dans les actes : l'action juridique, la qualité des participants, la chancellerie ou *scriptorium* d'origine, etc. En conséquence on trouvera des styles objectifs, d'autres subjectifs et même des formulations ampoulées selon la solennité de l'acte. Différents formulaires, styles de rédaction et formes d'instrumenter un acte mobilisent des parties bien déterminées du discours diplomatique. En conséquence, et ici se trouve une des premières difficultés, il existe des parties du discours très peu mobilisées, car correspondant à des pratiques restreintes (par ex. actes solennels, bulles, privilèges) et il existe d'autres parties qui sont mobilisées dans la quasi-totalité des actes. Le corpus CDLM témoigne bien de ces différences si on compte le nombre d'occurrences (entre parenthèses) de chaque partie :

Parties mobilisées dans des cadres très formels :

237. voir les lignes directrices de l'édition dans : Michele Ansani. "Edizione digitale di fonti diplomatiche : esperienze, modelli testuali, priorità". In : *Reti Medievali Rivista* 7.2 (2006), p. 1-1

238. Les parties indiqués dans la DTD utilisent la nomenclature de tradition allemande, alors que nous allons nous référer à celle de tradition française, placée entre parenthèses. Par ailleurs, les éditeurs ont mobilisé des éléments qui ne répondent pas aux définitions proposées dans le Vocabulaire international de la diplomatique afin de distinguer des sous-parties à l'intérieur d'autres, par exemple ceux utilisés pour distinguer les rôles ou rangs des souscripteurs : SMR (*signa manuum rogantium*), SMC (*signa manuum consentientum*), SMF (*signa manuum fideiussorum*), SMT (signa manuum testium), SME (*signa manuum estimatorum*), de la même manière "Completio" et "Datatio" regroupent certains éléments finaux, hors cadre : salutations, sceaux, validations, notes, etc

Recognitio [239], Corroboratio [240], Inscriptio [241], SMF [242], Iussio [243], SME [244], Estimatio [245]

Plus nombreuses, mais appartenant encore à un modèle d'acte assez formel :
Promulgatio [246], Sanctio [247], Narratio [248], SMC [249], Rogatio [250], Exordium [251], Tennor-Additum [252]

Parties appartenant aux modèles d'actes les plus communs :
Intitulatio [253], Invocatio [254], Subscriptio [255], SMR [256], SMT [257], Clausulae, DTTOP [258], Dispositio [259], Formulae [260], DTCRON [261].

Il faut ajouter que le corpus est quasi-entièrement composé de chartes, car d'autres typologies documentaires sont faiblement représentées :

Minuta (8), Notitia (25), Regestum (106)

239. Souscription de chancellerie faite par un officier qui déclare prendre la responsabilité de la pièce
240. Clause de corroboration de l'existence d'un acte
241. Élément indiquant le nom et éventuellement les titres et qualités de la personne (ou des personnes) à qui l'acte est adressé (adresse)
242. Signature de celui qui garde ou enregistre l'acte
243. Ordre écrit de procéder à l'établissement de l'acte écrit
244. Souscription de celui qui évalue l'acte
245. Partie où est indiqué les noms des personnes qui estiment et commandent l'acte juridique.
246. Formule par laquelle ce qui suit est porté à la connaissance (notification)
247. clauses destinées à assurer l'exécution de l'acte
248. Partie du texte par laquelle sont expliquées les circonstances du commandement de l'acte ou ses raisons (exposé)
249. Signature de la personne qui concède l'acte
250. Mention de l'ordre reçu d'écrire l'acte
251. Partie du texte par laquelle celui-ci est justifié de façon générale par des considérations juridiques, religieuses ou morales (préambule)
252. Élément à usages multiples qui porte sur des mentions hors de propos ou sur des gloses
253. Partie du protocole qui fait connaître le nom de l'auteur de l'acte écrit et sa titulature (suscription)
254. Formule de dévotion qui met l'acte sous le patronage divin ou d'un saint
255. Formules par lesquelles les parties, les témoins de l'acte juridique manifestent leur volonté personnelle, leur consentement ou leur présence.
256. Signature de la personne qui commande l'acte
257. Signatures des témoins de l'acte
258. Data Topica : date de lieu
259. Partie fondamentale du texte, par laquelle l'auteur manifeste sa volonté et fait naître l'acte juridique (dispositio)
260. La distinction entre Clausulae (clauses) et Formulae (formules) n'est pas du tout claire. Son usage est assez variable selon l'annotateur, néanmoins, on s'est aperçu que l'intention est d'annoter les clauses finales du dispositif, mais la typologie étant assez vaste, on rencontre des nombreuses erreurs dans l'annotation
261. Data Cronica : date de temps

Parties du dipl. discours	Freq	% du corpus	Longueur moyenne	Longueur médiane	Section charte
DTCRON	3766	80.5	13.2	13	A
Formulae	3429	73.3	41.5	43	B
Dispositio	4493	96	299.4	261	B
DTTOP	3789	81	4.2	4	A
Clausulae	3374	72.1	79.8	71	B
SMT	2165	46.3	16.3	14	C
SMR	2285	48.8	13.7	13	C
Subscriptio	1295	27.7	8.7	7	C
Invocatio	1444	30.9	3.6	3	A
Exordium	812	17.4	35.8	25	A
Rogatio	601	12.8	6.9	6	C
SMC	510	10.9	15.3	15	C
Narratio	365	7.8	218.4	138	A
Intitulatio	214	4.6	6.7	6	A
SME	181	3.9	16.7	16	C
Estimatio	226	4.8	82.5	93	C
Sanctio	200	4.3	39.5	30.5	C
Iussio	176	3.8	9.1	4.5	C
Inscriptio	126	2.7	15	13.5	A
SMF	127	2.7	9.2	9	C
Promulgatio	85	1.8	16.9	13	C
Corroboratio	90	1.9	16.2	17	C
Recognitio	40	0.9	9.4	9	C

TABLE 3.7 – Fréquence des parties du discours diplomatique dans le corpus annoté CDLM. "% du corpus" indique le pourcentage de chartes contenant la partie du discours signalée. "Section charte" indique chacune des trois parties majeures d'une charte : protocole(A), texte (B) et schatocole (C).

Observations sur le corpus

Après un examen attentif, le corpus présente diverses situations qui doivent être prises en compte si on envisage une modélisation de ses données. Nous allons signaler les 4 principales :

- (i) L'existence de plusieurs étiquettes à l'intérieur d'un même nœud, par exemple :

— 47 DTCRON dans la DATATIO
— 59 DTTOP dans la DATATIO
— 112 IUSSIO dans la COMPLETIO
— 8 NARRATIO dans le DISPOSITIO
— 45 SANCTIO dans le DISPOSITIO

Il s'agit de sous-parties à l'intérieur des modules les plus longs dont l'annotation se présente naturellement imbriquée. Les parties du discours ne sont pas simplement juxtaposées et certains documents peuvent présenter des structures assez complexes à distinguer, spécialement dans le cas des clauses finales comme c'est le cas pour <Sanctio> et <Iussio>. Par ailleurs <Datatio> n'existe pas en tant qu'élément du discours diplomatique, c'est une licence de la part des annotateurs de CDLM afin

de regrouper les formules chronologiques, parfois assez longues, présentes dans les eschatocoles ; il n'est donc pas étonnant d'y trouver diverses DTCRON et DTTOP.

- (ii) La relativement grande disparité entre certaines parties du discours diplomatique. Notamment le *dispositio* (dispositif) est 4 à 5 fois plus long que les unités suivantes, telles que *exordium, narratio, completio*. Le dispositif est la partie fondamentale du texte, par laquelle l'auteur manifeste sa volonté et fait naître l'acte juridique. Les parties suivantes sont relativement moins utilisées, on les voit parfois placées dans les protocoles afin d'ajouter justifications, considérations juridiques, religieuses, morales, etc. À l'autre extrême se trouvent des parties *mono-formulaires* ou *bi-formulaires*, dont la longueur est en général entre 5 et 20 mots : c'est le cas des parties très mobilisées dans les protocoles comme l'invocation, l'*inscriptio* (adresse) ou la *promulgatio* (notification) ainsi que d'autres vus régulièrement dans les eschatocoles : listes de témoins, dates de lieu et de temps.

- (iii) La composition quasi exclusive du corpus par des chartes, notamment de chartes notariées. Le corpus est bien renseigné à propos des typologies et origines des chartes et des affaires conclues. Sa diversité pourrait autoriser un outil modélisé à partir de celui-ci d'être applicable à un large spectre de chartes surtout celles produites à partir du XIe siècle ; par contre, son fort attachement à la charte compliquera ou empêchera singulièrement l'application du modèle sur d'autres typologies documentaires assez communes, notamment lettres et notices dont les solutions rédactionnelles répondent à des modèles différents que ceux de la charte ou à des chartes reposant sur des formulaires plus anciens.

- (iv) La perte de la rigidité du formulaire dans l'organisation interne des parties les plus longues. En effet, si ces parties démarrent en général de manière stéréotypée, les séquences successives montreront un arrangement plus souple de la phrase car elles doivent souvent exprimer les spécificités de l'affaire. Les scribes et notaires suivent de manière plus ou moins stricte les modèles, mais les formules peuvent subir de multiples arrangements d'ordre fonctionnel ou rhétorique. Le scribe suit le formulaire, non comme un cadre contraignant, mais comme un modèle auquel se rapporter.

Apprentissage

Contrairement aux entités nommées, pour les parties du discours notre modèle CRF vise à annoter les catégories d'une manière "plate", car la fréquence des séquences imbriquées ne permet - et ne justifie pas - pas un travail en deux étapes ; en conséquence, nous avons écarté du corpus de référence les annotations se présentant ainsi (Dans notre cas <Iussio> et <Clausulae>). Le corpus en format tabulaire à quatre colonnes (table 3.8) inclut les traits suivants que le modèle aura à utiliser pour la prédiction d'étiquettes :

- Le "token" (mot d'origine)
- Marquage morphosyntaxique (POS)
- Lemme (version sans flexion du mot)
- Catégories au format BIO du corpus de référence.

Le fichier de patrons (en annexe) analyse la séquence à partir d'une fenêtre contextuelle glissante de [-4, +4] par rapport à la position courante, générant uni-grammes et bi-grammes.

Token	**POS**	**Lemma**	**Bio format categories**
In %x[4,0]	PRE	in	B-INVOCATIO
Christi %x[3,0]	QLF	christus	I-INVOCATIO
nomine %x[2,0]	SUB	nomen	I-INVOCATIO
. %x[1,0]	SENT %x[1,1]	. %x[1,2]	I-INVOCATIO %x[1,3]
Die %x[0,0]	SUB %x[0,1]	dies %x[0,2]	B-DTCRON %x[0,3]
mercurii %x[-1,0]	SUB	mercurius	I-DTCRON
quinto %x[-2,0]	NUM	quintus	I-DTCRON
intrante %x[-3,0]	VBE	intro3	I-DTCRON
iernuario %x[-4,0]	QLF	ianuarius	I-DTCRON
,	PON	,	I-DTCRON
in	PRE	in	B-DTTOP
claustro	SUB	claustrum	I-DTTOP
officialium	QLF	officialis	I-DTTOP
Sancte	QLF	sanctus1	I-DTTOP
Brigide	NAM	-	I-DTTOP

TABLE 3.8 – Corpus d'apprentissage au format tabulaire pour le modèle des parties du discours diplomatique. Dans la figure la séquence "*In Christi nomine. Die mercurii quinto intrante iernuario, in claustro officialium Sancte Brigide*". Les zones en gris indiquent une observation sous la forme d'un bi-gramme pour la transition de l'Invocation à la Date de temps (*DTCRON*) combinée avec les différents traits extraits d'une fenêtre de 9 tokens (4 avant et 4 après l'observation centrale)

Contrairement aux entités nommées, l'information concernant la capitalisation des mots et les suffixes, l'analyse des séquences étant faite au niveau de la phrase, ne se montrent pas utiles. La modélisation porte alors sur le mot d'origine ("token"), les catégories morpho-syntaxiques et les étiquettes des catégories au format BIO. Le plus important est la modélisation prenant en compte les bi-grammes : en effet, il existe dans les documents un ordre d'usage assez bien déterminé, ce qui exige que l'analyse porte sur un nombre plus élevé d'uni-grammes et surtout d'observer les bi-grammes - les transitions - dans certains motifs afin de regarder à la fois la catégorie de la ligne courante et la catégorie de la ligne précédente. L'objectif est ici que notre modèle apprenne qu'une séquence <Promulgatio> (notification) est souvent suivie d'une séquence <Dispositio>, autrement dit, qu'un <I-Promulgatio> peut être suivi d'un <B-Dispositio> et qu'à l'opposé une séquence <Narratio> n'est jamais, ou presque, suivie d'une séquence <Iussio> ou <Subscriptio>.

De plus, afin de contrôler la longueur de certains parties, notamment, *dispositio* et *narratio*, nous les avons sectionnées de manière artificielle selon le nombre de phrases qui les composent (1ère phrase de la section <Dispositio-1>, 2ème phrase <Dispositio-2>, etc.). Cela évite le fort déséquilibre du système qui tendait à annoter les parties difficiles à reconnaître comme faisant partie du *dispositio* pour minimiser son taux d'erreur.

Plusieurs modèles ont été formés avec différents hyperparamètres afin de trouver le meilleur taux de généralisation et le meilleur rapport taille/performance, mais n'ont pas permis d'économies d'annotation. Le corpus de référence est composé de 4680 documents (2,5M de "tokens"). Le modèle le plus équilibré (voir résultats dans la table 4.8) est réparti comme suit : 3000 docs (65 %) pour le jeu d'entraînement, 300 pour le jeu de validation. La performance est en général 2-3 points plus basse en f1-mesure (ceci dépend de la section) qu'avec deux ensembles plus complets de 3700 documents (80 %) pour le jeu d'entraînement et 370 pour celui de validation.

Par ailleurs, l'apprentissage à partir de bi-grammes se montre assez exigeant en termes de ressources informatiques ; dans notre cas environ 4 heures d'entraînement sur 16 coeurs et 50Go de mémoire vive générant autour de 150M de traits (*features*).

Nous discutons ensuite les résultats de l'évaluation du meilleur modèle generé.

3.5.1 Évaluation du modèle

La table 3.9 compare les hypothèses du modèle avec la référence du jeu test. On observe une performance assez bonne (supérieure à 80 %) sur les parties du discours diplomatique dont nous avons un nombre substantiellement plus élevé d'exemples. Cela correspond aux parties les plus mobilisées dans les modèles les plus courants de chartes médiévales produites entre le Xe et le XIIe siècle. D'un côté nous observons que *dispositio* et *completio* qui correspondent globalement au texte et à l'eschatocole sont bien détectés. De l'autre côté, nous avons une bonne détection sur des étiquettes qui, bien que ne correspondant pas à des parties du discours diplomatique à proprement parler, distinguent le rôle des souscripteurs de l'acte (SMC, SME, SMR, SMT).

Par ailleurs il existe une performance assez acceptable (entre 70 % et 85 %) en *exact match* dans les parties constituées par une ou deux formules dont les formes courantes sont plus ou moins répétitives. C'est le cas notamment de *invocatio*(invocation), *subscriptio* (souscription), *inscriptio* (adresse), *promulgatio* (notification), et les dates de lieu (DTCRON) et de temps (DTTOP). Cela correspond aux parties les plus communes trouvées dans les protocoles, cadre initial de l'acte (ou dans les eschatocoles, selon la tradition et l'époque).

Par contre, la performance est médiocre dans les parties du discours les moins mobilisées et dont les exemples sont peu nombreux. Ici on peut distinguer deux sous-types : d'un côté *exordium* (préambule), *narratio* (exposé) et *intitulatio* (suscription) dont les débuts de la séquence sont bien détectés, mais la suite est très problématique, ce qui se traduit par le grand écart en f1-mesure entre *exact match* et *partial match* ; de l'autre côté : *rogatio* (rogation), *recognitio* (recognition de chancellerie), *corroboratio* (clause de corroboration) et SMF (signature de qui registre l'acte) qui montrent de mauvais résultats dans toutes les mesures. Les raisons semblent plus ou moins évidentes. Les éléments du premier sous-type introduisent, dans les protocoles, des éléments de justification ou des antécédents de l'affaire conclue ; elles sont normalement introduites de manière stéréotypée, mais présentent des détails plus particuliers par la suite. Le modèle arrive a bien saisir les débuts, mais la suite étant plus contingente, n'est pas bien repérée. Les éléments du deuxième sous-type sont très peu mobilisés avant le XIIIe siècle, car il s'agit de signes de validation et d'authenticité

appartenant à des actes de chancellerie ou plus tardivement du notariat public et le CDML comporte un nombre assez faible d'exemples, rendant leur reconnaissance très compliquée.

Partie / mesure		TP	FP	FN	Précision	Rappel	F1	dif
COMPLETIO	PM	674	2	4	0,9970	0,9941	0,9956	0,02
	EM	658	18	20	0,9734	0,9705	0,9719	
CORROBORATIO	PM	11	1	2	0,9167	0,8462	0,8800	0,40
	EM	6	6	7	0,5000	0,4615	0,4800	
DISPOSITIO	PM	2563	151	119	0,9444	0,9556	0,9500	0,09
	EM	2339	375	333	0,8618	0,8754	0,8685	
DTCRON	PM	808	14	9	0,9830	0,9890	0,9860	0,22
	EM	630	192	188	0,7664	0,7702	0,7683	
DTTOP	PM	694	17	17	0,9761	0,9761	0,9761	0,21
	EM	542	169	169	0,7623	0,7623	0,7623	
EXORDIUM	PM	103	4	14	0,9626	0,8803	0,9196	0,24
	EM	76	31	41	0,7103	0,6496	0,6786	
INSCRIPTIO	PM	4	0	9	1,0000	0,6087	0,7568	0,54
	EM	14	10	19	0,2857	0,1739	0,2162	
INTITULATIO	PM	30	0	6	1,0000	0,8333	0,9091	0,39
	EM	17	13	19	0,5667	0,4722	0,5152	
INVOCATIO	PM	238	4	6	0,9835	0,9754	0,9794	0,15
	EM	201	41	43	0,8306	0,8238	0,8272	
NARRATIO	PM	50	19	13	0,7246	0,7937	0,7576	0,42
	EM	22	47	41	0,3188	0,3492	0,3333	
PROMULGATIO	PM	29	2	4	0.9354	0.8787	0.9062	0,08
	EM	25	4	6	0.8620	0.8064	0.8333	
RECOGNITIO	PM	2	4	4	0,3333	0,3333	0,3333	0,16
	EM	1	5	5	0,1667	0,1667	0,1667	
ROGATIO	PM	78	24	33	0,7647	0,7027	0,7324	0,10
	EM	67	35	44	0,6569	0,6036	0,6291	
SMC	PM	82	7	12	0,9213	0,8723	0,8962	0,01
	EM	81	8	13	0,9101	0,8617	0,8852	
SME	PM	33	1	3	0,9706	0,9167	0,9429	0,14
	EM	28	6	8	0,8235	0,7778	0,8000	
SMF	PM	12	1	10	0,9231	0,5455	0,6857	0,17
	EM	9	4	13	0,6923	0,4091	0,5143	
SMR	PM	456	8	8	0,9828	0,9828	0,9828	0,08
	EM	421	43	43	0,9073	0,9073	0,9073	
SMT	PM	454	2	10	0,9956	0,9784	0,9870	0,06
	EM	428	28	40	0,9386	0,9145	0,9264	
SUBSCRIPTIO	PM	191	9	12	0,9550	0,9409	0,9479	0,12
	EM	154	46	72	0,7700	0,6814	0,7230	
Overall	PM	6520	270	295	0,9602	0,9567	0,9584	0,12
	EM	5709	1081	1124	0,8407	0,8355	0,8381	

TABLE 3.9 – Résultats de l'évaluation du modèle des parties du discours diplomatique sur le jeu de test en termes de *Exact Match* (EM) et *Partial Match* (PM) selon l'outil BratEval. Legende : précision (Pr), rappel (Rc), f1-mesure (f1) TP (true positive), FP (false positive), FN (false negative), dif : différence entre partial et exact match sur f1.

Évaluation sur un jeu test du CBMA

Comme dans le cas des modèles pour les entités nommées, nous avons annoté un jeu de référence avec de documents inconnus du modèle afin d'évaluer sa capacité de généralisation à des documents provenant d'autres corpus. Étant donné qu'il s'agit d'une annotation assez laborieuse, le jeu compte seulement 100 documents sélectionnés du corpus CBMA (Les résultats se trouvent dans table 3.10).

D'une manière générale ces résultats répliquent ceux obtenus sur le jeu de test, bien que la performance soit de quelques points inférieure dans presque tous les secteurs. Dans cette nouvelle évaluation, les parties les plus courantes et les parties contenant une ou deux formules sont bien repérées (entre 80% et 85%), ce qui inclut la plupart du protocole, la dispositio et les parties essentielles de l'eschatocole. Cela correspond aux parties du discours bien répandues tant dans le CDML que dans le CBMA. Par ailleurs, exordium et narratio, comme dans l'évaluation du jeu de test, ont leurs débuts bien repérés mais la détection complète est problématique. À l'opposé la performance sur les parties appartenant à un style de charte plus solennel ou provenant d'une chancellerie est médiocre ; de fait nous n'avons pas trouvé dans nos sondages plus de 5 chartes du CBMA portant *recognitio* et *corroboratio* employés dans les actes notariés ce qui traduit que le CBMA n'inclut que très peu de chartes le contenant et exclusivement des chartes publiques.

	Partial Match			Exact Match		
	Précision	Rappel	F1	Précision	Rappel	F1
COMPLETIO	0.969	0.780	0.864	0.843	0.818	0.830
DISPOSITIO	0.898	0.975	0.935	0.831	0.896	0.862
DTCRON	0.850	0.944	0.894	0.789	0.833	0.810
DTTOP	0.920	0.958	0.938	0.791	0.863	0.826
EXORDIUM	0.785	0.846	0.814	0.533	0.727	0.615
INSCRIPTIO	0.785	0.611	0.687	0.583	0.411	0.482
INTITULATIO	0.931	0.836	0.881	0.656	0.600	0.626
INVOCATIO	0.878	0.966	0.920	0.806	0.892	0.847
NARRATIO	0.692	0.750	0.720	0.416	0.416	0.416
PROMULGATIO	0.868	0.942	0.904	0.794	0.843	0.818
ROGATIO	0.666	0.800	0.727	0.571	0.666	0.615
SMC	0.885	0.861	0.873	0.781	0.735	0.757
SMF	0.833	0.625	0.714	0.666	0.500	0.571
SMT	0.897	0.897	0.897	0.801	0.777	0.789
SUBSCRIPTIO	0.927	0.888	0.907	0.743	0.679	0.709

TABLE 3.10 – Résultats de l'évaluation du modèle des parties du discours diplomatique sur le jeu de test de documents du CBMA en termes de *Exact Match* et *Partial match* selon l'outil BratEval.

En résumé notre modèle est capable de détecter globalement les trois groupes qui composent la teneur du texte : protocole, texte et eschatocole. Cependant ses capacités de récupération sur les parties qui se trouvent à l'intérieur des cadres formels montrent différentes performances selon la taille, la régularité d'usage et le style rédactionnel : la performance est bonne lorsque il s'agit de formules, moins bonne si elle comporte des détails assez spécifiques ou sur des parties peu mobilisées. Le modèle le plus usuel de charte peut être reconnu dans ses composants principaux car le modèle se

montre robuste dans la reconnaissance des cadres initiaux (voir notre étude autour des protocoles dans le chapitre 5) et finaux, comme du dispositif. Les types plus complexes et formels d'actes peuvent néanmoins se montrer bien plus compliqués à détecter.

Enfin, la reconnaissance sur le *dispositio* mérite un commentaire spécial. Le modèle est capable de repérer le début et la fin de cette partie fondamentale de l'acte, mais nous n'avons pas modélisé sur les parties trouvées à l'intérieur, notamment les clauses finales (prohibitives, dérogatives, de réserve, pénales, etc.) qui constituent une source d'information très riche pour l'historien. Ces clauses avaient été partiellement annotées sous l'étiquette <clausulae>, mais les clauses étant très variées, son annotation sous une même étiquette générait trop bruit pour le modèle, en conséquence elle a été ôtée pendant l'entraînement. Une modélisation prenant en compte cette information, après l'avoir corrigée et complétée, est toutefois envisageable.

3.6 Conclusion

Suivant les différentes évaluations que nous avons effectuées, notre modèle de reconnaissance des entités nommées offre une performance élevée et est assez robuste pour être appliqué à un vaste rayon documentaire en conservant une qualité satisfaisante. Nous l'avons appliqué sur des corpus fort variés en taille, chronologie et origine et il s'est montré valide et suffisamment performant pour permettre de semi-automatiser une étape primordiale de la structuration de grandes quantités de données provenant d'actes manuscrits, économisant ainsi beaucoup d'efforts humains.

Les différentes modifications et ajouts que nous avons effectué sur l'annotation originellement fournie partent de la constatation de plusieurs situations problématiques à l'égard d'une automatisation et nous permettent de proposer un standard d'annotation pour de futures modélisations. Nous avons ainsi corrigé massivement l'annotation originelle lorsqu'elle empêchait d'avoir une meilleure performance car proposait des contextes contradictoires au modèle. Ensuite nous avons apporté deux nouveaux jeux de documents annotés : un jeu supplémentaire de 400 documents annoté à la main couvrant les zones de pénurie documentaire de notre corpus principal et un jeux de 450 documents annotés manuellement provenant de corpus contenant des actes diplomatiques d'autres régions européennes. Ces deux ensembles peuvent à l'heure actuelle être utilisés comme des jeux de test.

Par ailleurs, notre modèle secondaire pour la reconnaissance des parties du discours diplomatique se montre robuste pour la détection des parties les plus courantes des actes et peut être mobilisé pour obtenir un premier niveau assez acceptable de structuration. Une reconnaissance plus fine pourrait s'avérer problématique, mais cela correspond à des types d'actes plus solennels et dont le nombre est assez réduit.

Enfin nous avons largement prouvé qu'un travail comme ceci dépend d'une pratique interdisciplinaire, car un standard de modélisation fait suite à l'analyse des différents états de la langue et du discours, qui constitue le cœur de la méthode, auquel doit s'ajouter l'observation attentive des spécificités proposées par le document médiéval. L'équilibre méthodologique aide à polir l'annotation sur laquelle l'algorithme est appliqué et produit des observations de nature historique qui aident à mieux déterminer l'application numérique.

Chapitre 4

Datation assistée par ordinateur

4.1 Introduction

Dans les pages qui suivent nous allons proposer une méthode de datation assistée par ordinateur des actions juridiques reflétées dans les actes d'un cartulaire médiéval lorsqu'elles ne sont pas datées. Pour expérimenter notre méthode de travail nous avons choisi un cartulaire appartenant à la base du CBMA, le cartulaire du monastère de Paray-le-Monial, compilé pendant le XIIe siècle et dont la quasi-totalité des actes ne sont pas datés. La méthode se fonde sur la construction automatique d'une matrice combinatoire de données présentant deux informations essentielles : d'une part, les noms des personnes figurant dans chaque document de notre cartulaire, ce qui a été obtenu grâce à notre modèle de récupération des entités nommées et, d'autre part, la liste des données chronologiques qui leur sont associées. Une fois que la présence de personnes communes à la fois dans des actes datés et dans des actes non datés est vérifiée, chaque personne remplissant ces 2 conditions a fait l'objet d'une recherche par approximation morphologique dans l'ensemble de la base diplomatique de la Bourgogne afin de collecter toutes les données chronologiques disponibles concernant cette personne. L'objectif est ici le transfert de ces données chronologiques, à partir des personnes du groupe d'actes datés (et accessoirement des bases de données externes) vers le groupe d'actes non datés.

Les sous-chapitres se concentrent sur cinq développements : une brève exposition du panorama scientifique autour de la datation automatique des chartes médiévales ; une analyse des caractéristiques et de l'histoire du cartulaire dont nous allons essayer de dater les actes ; l'exposition détaillée de trois exemples de datation réalisés grâce à notre matrice de données afin de valider la méthode proposée ; une discussion autour des éléments historiques et diplomatiques qui autorisent l'application de notre méthode ; et finalement, la présentation de notre séquence technique de travail (*pipeline*). Ce travail est complété par un annexe trouvé à la fin de la thèse : le jeu complet des actes non datés du cartulaire avec la chronologie restituée et un *index personarum* présentant les personnages les plus pertinents mobilisés ici.

4.2 Comment dater les actes d'un cartulaire ?

La datation semi-automatique des documents d'un corpus, et dans ce cas précis d'un cartulaire, est un travail très complexe qui met en jeu tant la connaissance approfondie du corpus que les outils informatiques capables de fournir de l'information structurée à partir de l'analyse des textes. La datation d'un document est une information capitale et un prérequis indispensable à tout travail historique parce qu'elle nous aide à mieux préciser le fil des actions signalées mais surtout à mieux interpréter les informations en les rapportant à un contexte scripturaire bien déterminé. Un corpus qui n'est pas daté ou dont les documents portent une date incomplète, ce qui est assez commun dans le monde médiéval, est parfaitement mobilisable comme preuve scientifique, mais tant que la date reste douteuse sa valeur comme preuve solide peut se voir sérieusement affectée.

Dans les actes bourguignons, l'absence de la date est assez commune avant le XIIe siècle et particulièrement au cours du XIe siècle. C'est la raison pour laquelle les dates d'une partie importante des documents de cette époque-ci ont été estimées grâce à la mention qui y est faite des abbés de Cluny ou des évêques, mobilisant un travail d'érudition classique concernant la chronologie : la fourchette de datation. Dans notre travail nous allons employer la même méthode, mais en mobilisant une masse bien plus ample d'informations. La datation par fourchette est un travail assez aride, qui s'effectue par le transfert de la datation depuis un personnage figurant dans un document au document lui-même. La chronologie de cette personne, elle, a le plus souvent été attestée par sa mention dans des documents datés préalablement. Ainsi, la datation se réalise par le transfert de l'information chronologique structurée d'un personnage vers une source non structurée et non datée mais où apparaît ce même personnage. L'efficacité de cette méthode dépend directement de la portée de l'information disponible et dans le cas des personnes qui apparaissent dans un texte médiéval, elle est très limitée parce que le nombre de personnes dont on connait la chronologie est très réduit. Rois, comtes, évêques, abbés ont souvent une chronologie bien définie, mais leur présence dans les documents étudiés est statistiquement basse et même infime si on les compare aux milliers de personnes qui apparaissent, par exemple, dans les actes d'un cartulaire de taille moyenne et pour lesquels nous n'avons habituellement aucun autre témoignage documentaire.

Si la probabilité de trouver une personne chronologiquement bien identifiée dans les chartes médiévales est faible, même lorsqu'on a la chance de la rencontrer, les dates fournies se montrent parfois peu utiles. Dans le cas du corpus bourguignon, ce sont des abbés qui sont régulièrement mentionnés dans les actes et souvent pris comme critère fondamental de datation. Néanmoins, comme on l'a vu (figure 2.1), les fourchettes chronologiques des trois abbés de Cluny des siècles centraux, Maieul, Odilon et Hugues est supérieure à 40 ans, et s'étend même sur 60 dans le cas de Hugues ; ainsi un classement chronologique des chartes à partir de ces données génère une vision altérée de la réalité transmise dans les actes qu'on a pu observer lorsqu'on a fait des graphiques généraux sur la chronologie et la répartition typologique des actes du cartulaire de l'abbaye de Cluny.

Une définition beaucoup plus précise des fourchettes chronologiques passe par

l'itération du même processus de transfert de la date, mais appliqué cette fois-ci sur toutes les catégories de personnages présents dans un document et retrouvés dans un autre. Il ne s'agit pas d'une nouveauté car le croisement des nombreux groupes de personnages souvent présentés en tant que témoins, confirmateurs, souscripteurs, etc. a été traditionnellement une riche source d'information chronologique pour l'historien. Si nous proposons de changer d'outil, ce processus ardu continue à imposer deux conditions : d'un côté, la récupération et la classification complète des entités nommées, et de l'autre, la récupération à partir de sources externes et internes des dates qui leur sont associées. Si on arrive à fournir une tranche chronologique pour chaque personne d'une certaine importance, ce qui dans un cartulaire équivaut à dire ceux qui apparaissent au moins deux fois dans la documentation, on pourrait avoir bien plus d'éléments informatifs pour déterminer la chronologie de chaque document.

Ces dernières années différents systèmes de détection automatique de la date des documents médiévaux ont été proposés, et nous avons déjà évoqué l'un de ceux-ci, le DEEDS, construit à partir la base de données des chartes des monastères anglais (voir partie 2.1). L'approche utilisée, appelée *matching word patterns* ("Modèle de correspondances de mots"), extrait des chartes datées toutes les unités sémantiques (groupes de mots fréquemment associés) puis en effectue un suivi chronologique sur elles qui vise à déterminer leur date d'apparition et de disparition. Ensuite, le même processus est réalisé sur les chartes non datées. Puis, on compare les unités sémantiques extraites de chaque groupe afin de transférer les chronologies lorsque l'on trouve de coïncidences.[262] Ce modèle, construit sur une très vaste base de données, arrive à dater avec une bonne précision les chartes non datées d'un même ensemble, avec une fourchette de l'ordre de 10 ans. Mais, étant très dépendent de l'existence d'un vaste groupe de chartes datées qui permet d'avoir plus de « motifs de mots » ou *word patterns*, il montre une performance peu satisfaisante lorsque les groupes de chartes datés sont plus réduits, l'écart dans ce cas pouvant dépasser les 50 ans. Une classification chronologique avec une approximation de plus de 50 ans peut être utile comme premier classificateur, mais elle est loin de régler le problème de la datation manquante.

Les actes d'un recueil documentaire tel qu'un cartulaire, notamment d'un cartulaire-chronique, peuvent présenter de nombreux éléments qui faussent le calcul automatique de la chronologie et qui rendent difficile ce genre de comparaisons massives par groupes de mots. Tout d'abord, il comprend une grande variété d'éléments anachroniques, ce qui n'est pas surprenant si on considère que la compilation d'un cartulaire est un travail de sélection toujours postérieur à la rédaction des actes, et, dans les cas des cartulaires-chroniques, il s'agit d'un groupe de documents de nature légale inséré dans un récit historique. Lors de la compilation d'un cartulaire, les responsables peuvent avoir ajouté des documents qui sont postérieurs aux documents initiaux, et effectué modifications, corrections, adaptations de style voire altérations volontaires et parfois même des forgeries. Dans d'autres cas un document peut être la notice abrégée d'une charte de la même date, ou à l'inverse il peut avoir subi un allongement par l'ajout des préambules et des clauses absentes dans une rédaction originellement objective ; il peut être la confirmation d'un document ancien ; il peut

262. Michael Gervers. *Dating undated medieval charters*. Boydell & Brewer Ltd, 2002, p. 30-72.

s'agir d'une insertion : transcription d'un document antérieur dans le corps d'un autre ; un *vidimus* ou *inspeximus* (ce qui inclut une copie littérale) d'un acte validé un demi-siècle avant ou la continuation, pas forcément immédiate, sous la forme de complément, d'une donation ancienne.

Le cas des notices est assez particulier et assez commun dans les cartulaires. Les notices sont normalement la copie d'actes rédigés à l'origine dans un style objectif, privilégiant la fixation de l'action juridique ; plus rarement il s'agit d'une copie rapide et abrégée d'une charte privilégiant la transmission de la *dispositio*. Dans un cas comme dans l'autre, elles négligent les caractères internes et les parties des actes plus formels et peuvent apparaître rédigées sans trop d'attention aux formes stéréotypées, présentant une plus grande souplesse et de là une plus grande irrégularité dans leurs composants linguistiques et diplomatiques. Tous ces éléments se retrouvent habituellement au moment de la compilation des actes, mais il ne faut pas négliger l'existence d'un nombre considérable de contributions personnelles de la part du scripteur de l'acte, dépendant de son niveau de connaissance du latin, de sa mémoire, son goût personnel au moment d'assembler le formulaire et enfin de son inventivité littéraire, éléments qui peuvent tous contribuer à la datation, mais peuvent également introduire des bruits forts dans un système de calcul automatique appliqué aux actes de cartulaire. En outre, les collections de documents originaux, qui pourraient atténuer ces problèmes dans une certaine mesure, ne sont pas très courantes. Ainsi, un même cartulaire peut héberger des documents contemporains assez différents selon les modes de compilation, les changements introduits par le compilateur et enfin le niveau de liberté adopté par le scripteur originel. Il existent des documents très similaires mais dont les rédactions sont séparées par des décennies et des documents complexes auxquels on devrait fournir deux dates ou plus car ils contiennent plusieurs actions juridiques, ce qui complique immensément tout exercice de datation auto ou semi-automatique, et jette le doute sur la datation d'un acte par son degré d'adaptation à un ensemble d'unités sémantiques et linguistiques [263].

Ce panorama n'est pas du tout étonnant car comme on l'a vu dans le chapitre 1.2 de cette thèse, les travaux relevant de l'apprentissage automatique sur des textes antérieurs au XVIIIe siècle, et spécialement sur les textes médiévaux, se tournent souvent vers des modèles *ad hoc*, offrant très peu d'applicabilité à l'extérieur du corpus d'origine. Cela ne signifie pas que les méthodes d'apprentissage ou de comparaison automatique soient dénuées d'intérêt, mais plutôt que pour l'instant ces techniques peuvent se montrer efficaces si elles sont appliquées à des états élémentaires ou intermédiaires de la recherche et non à des états avancés. L'interdisciplinarité, bien entendu, ne vise pas à abandonner les méthodes classiques mais à les soutenir par des outils informatiques dans les tâches les plus fastidieuses, ce qui a conduit ces dernières années à une exigence académique bien plus élevée en ce qui concerne les résultats.

La méthodologie de travail que nous proposons ici pour fournir une chronologie

263. Voir à ce sujet : Giry, *Manuel de diplomatique : Diplomes et chartes.-Chronologie technique.-Éléments critiques et parties constitutives de la teneur des chartes.-Les chancelleries.-Les actes privés*, p. 865-870 ; Véronique Gazeau. "Recherches autour de la datation des actes normands aux X e–XII e siècles". In : *Dating medieval undated charters* (2000), p. 61-79, p. 61-79 ; Olivier Guyotjeannin. "La diplomatique médiévale et l'élargissement de son champ". In : *La Gazette des archives* 172.1 (1996), p. 12-18, p. 12-18

aux actes dépend directement d'un outil d'apprentissage automatique capable de nous fournir les listes structurées des personnes, autrement dit, des entités nommées, mais elle se trouve insérée dans une heuristique assez bien établie dans la science historique.

4.3 Le cartulaire de Paray-le-Monial

Du groupe de cartulaires de l'orbite clunisienne que nous avons mobilisés, celui du monastère de Paray-le-Monial semble le plus pertinent pour cet exercice de datation. [264] Nous l'avons précédemment présenté dans ces grandes lignes au moment d'exposer notre base documentaire de travail (partie 2.1.7). Il s'agit d'un cartulaire de petite taille : environ 225 actes – nous ne comptons pas les items disparus – étaient présents dans le codex originel organisé en 114 feuillets de parchemin. Dix-huit documents du XIIIe siècle ont été ajoutés par l'éditeur moderne ; ils correspondent à des visites diocésaines et à des recensements. L'ensemble des actes se présente presque entièrement sous forme de notices de donations, de donations-vente ou de résolution de litiges ; il inclut quelques chartes de donation et de confirmation et une dizaine d'actes délivrés par l'autorité publique qu'elle soit royale, comtale ou provenant de l'évêché. Les notices présentent un style rédactionnel objectif, parfois laconique, fournissant des renseignements issus normalement d'une rédaction originale abrégée fixant la mémoire de l'action juridique ou d'originaux significativement abrégés.

Le cartulaire est une reconstitution issue du travail minutieux d'Ulysse Chevalier à partir d'un codex original perdu et d'au moins 5 copies fragmentaires, dont une faite par Lambert de Barive vers 1786 [265] du cartulaire originel ou de copies postérieures. Lambert de Barive estime la date de rédaction du cartulaire au XIIe siècle [266], mais Canat de Chizy, dans un travail à peine postérieur à l'édition de Chevalier, propose quant à lui une date de rédaction située à la fin du XIe siècle. [267] Dans deux articles publiés en 1992, Frank Neiske et Marie Hillebrandt revalident cette date [268] et placent la rédaction au temps du prieur Hugues (ca. 1080 – 1115) et après la mort de Hugues II, comte de Chalon (1079), dont la famille avait une participation privilégiée dans les affaires de Paray depuis la fondation du monastère en 977 par le comte Lambert (930-978). Cette participation de la famille est poursuivie après la donation de l'abbaye à Cluny, effectuée en 999 par son successeur le comte-évêque Hugues I de Chalon (987-1039).

Cela étant dit nous devons considérer un état de la tradition diplomatique des actes présents dans l'édition érudite du cartulaire qui se déroule en au moins quatre étapes :

1. Les chartes et notices originales ;
2. La compilation des actes dans le cartulaire ;
3. La copie du cartulaire faite par Lambert de Barive et les autres copistes

264. Ulysse Chevalier. *Cartulaire du prieuré de Paray-le-Monial, ordre de Saint-Benoît*. A. Picard, 1890.
265. (ibid.), Introduction, V-X
266. (ibid.), Introduction, VIII
267. M Canat de Chizy. *Origines du prieuré de Notre-Dame de Paray le Monial*. Saone-et-Loire, 1876, p. 122.
268. Neiske, "Les débuts du prieuré clunisien de Paray-le-Monial", p. 140.

4. Les versions reconstituées par Ulysse Chevalier comme base de la collection et la critique des copies conservées.

Les actes du cartulaire ne sont pas systématiquement datés et ce défaut dans la chronologie, qui jusqu'à présent n'avait pas été reconstituée, nous empêche de bien rattacher ce source à une tradition scripturaire. Dans tout le recueil on compte à peine onze actes dont la date est renseignée dans le texte [269]. De plus, huit de ces onze actes correspondent à des préceptes et diplômes datés après la deuxième moitié du XIIe siècle et ne sont donc pas vraiment connectés avec l'ensemble principal qui concerne les transactions foncières. Cela nous laisse un cartulaire presque entièrement vierge en ce qui concerne la chronologie de ses actes, puisque même si on peut rapidement fournir une chronologie basée sur les noms de abbés qui apparaissent dans une quarantaine d'actes, ou des comtes de Chalon qui font souvent des donations au monastère, la tranche est si vaste (40 ans de moyenne) qu'elle rend la datation très peu utile pour une recherche minutieuse.

Le cartulaire suit, par ailleurs, un modèle rédactionnel proche de celui suivi par le reste des cartulaires clunisiens, sous l'influence intellectuelle de l'abbaye-mère. En fait, le monastère de Paray, à l'origine fondation grâcieuse des comtes de Chalon, est transféré avec toutes ses propriétés à Cluny en tant que prieuré, à peine 30 ans après sa fondation (977). Cette influence clunisienne contribue à ce que les typologies, modèles, vocabulaires, formules et références spatiales et sociales soient très proches, voire identiques à celles trouvées dans les grands cartulaires bénédictins et que, par extension, la méthode développée ici puisse être adaptable aux larges ensembles de chartes non ou mal datées que l'on y trouve.

4.3.1 Le monastère de Paray-le-Monial

Les comtes de Chalon dotent dès sa fondation le monastère d'un important patrimoine, comme en témoignent de nombreuses notices du cartulaire. Les familles des *potentes* du Charollais et du Brionnais ne tardent pas à accroître ces dotations dans une mesure bien plus grande que ne le laissent voir les notices de l'époque, souvent laconiques, et dont le nombre est peu élevé au Xe siècle. Cet immense patrimoine qui assurait une vie économiquement confortable pour le monastère est transféré à Cluny en 999 dans une donation assez inattendue [270] qui contribuait à définir un équilibre entre Cluny, les comtes de Chalon et le roi. Sous la domination de Cluny, Paray devient un simple prieuré, statut qui contraste avec la richesse de son vaste patrimoine. Dans les premiers temps, Cluny touche à peine l'intégrité du patrimoine originel du monastère mais, conséquence de sa sujétion à l'ordre, Paray cesse d'être le destinataire privilégié des donations foncières dans la région. À peine une vingtaine de documents des premières décennies du XIe siècle concernant Paray sont conservés dans le corpus bourguignon et presque tous sont liés aux donations de la famille comtale qui continue à doter le monastère. Dans ce contexte, la stagnation patrimoniale et la perte d'indépendance dans la gestion ont pu été assimilées à une crise économique dont les

269. CBMA 7020, 7203, 7205, 7218, 7222, 7230, 7231, 7238-7244
270. CBMA 7230

symptômes sont très fortement perceptibles vers la fin du XIe siècle dans le recueil documentaire constitué par l'institution.

Il est incontestable que le cartulaire soit commencé à l'époque du prieur Hugues, personnage omniprésent, même lorsqu'il ne souscrit pas les actes du cartulaire. Hugues, membre d'une des plus importantes familles de la région, les *Buxol*, est le deuxième ou troisième fils issu du mariage entre *Aya* et *Artaldus de Buxol*, dont nous connaissons tous les descendants jusqu'à la troisième génération (voir ci-dessous la chronologie des *Buxol*). Selon nos estimations, il est né à la fin de la décennie 1030 ou au début de 1040 [271] et il est entré comme moine à Paray vers 1063. Il est bien possible qu'il ait été nommé prieur en remplacement de *Girbertus*, du temps du comte Hugues II de Chalon (et donc avant 1079, date de sa mort), mais nous ne disposons d'aucun acte où tous les deux soient présents. Dans la première notice qui témoigne de sa nomination au poste de prieur, nous le trouvons souscrivant une donation d'Adelaïde, la comtesse de Chalon, pendant la période la plus dure de l'interrègne et du conflit de succession entre les héritiers Gui de Thiers et Geoffrey de Donzy, survenu après la mort du comte Hugues II, c'est-à-dire vers 1080-1085 [272]. Nous considérons qu'il s'agit de la date clé de sa prise de fonction comme prieur de Paray. Comme le précise le cartulaire [273], Hugues est prieur au début de la compilation des actes et il est le prieur le plus mentionné dans les textes [274]. La date du début de la compilation du cartulaire n'est pas connue, mais elle ne doit pas être bien postérieure aux premiers événements où Hugues apparaît, c'est-à-dire autour de l'an 1090 (voir discussion en 5.2).

Hugues de Buxol demeure prieur au moins jusqu'en 1115, lorsque nous le trouvons souscrivant un document avec le duc de Lorraine Symon Ier (1115-1139) [275]. Son successeur dans la fonction, *Bernardus*, ne le reste que brièvement. Il apparaît dans deux actes [276] seulement (voir la chronologie des prieurs de Paray ci-dessous). En revanche, le suivant, *Artaldus*, se trouve déjà en poste en 1119, date presque certaine d'un jugement d'excommunication conservé dans le cartulaire, document unique dans son genre, et pour lequel on a proposé une chronologie à l'époque de l'élection du pape Calixte II dans l'abbaye de Cluny [277]. Néanmoins nous avons considéré pour Hugues un *terminus ante quem* fixé à l'an 1115, en l'absence d'une date plus certaine. Cela veut dire que Hugues de Buxol meurt à un âge avancé pour l'époque, environ 75 ans, qu'il serait entré au monastère de Paray-le-Monial ayant 20 à 25 ans, aurait été nommé prieur environ 15 ans plus tard, et aurait occupé cette fonction 35 ans, probablement jusqu'à sa mort. Évidemment, nous estimons ces dates d'après une première série d'indices chronologiques fournie par les actes de Paray, travail que nous allons expliquer

271. S'il est déjà né à l'époque d'une donation de son père datée, selon nous, vers 1030-1039 (CBMA 7112) il est probablement encore enfant. Son père lui-même est né au plus tôt dans la décennie de 1010 et mort dans la décennie de 1080 (CBMA 7113). Hugues entre certainement dans la vie religieuse vers 1063 (CBMA 7232) et il vit jusqu'à 1115 environ ; il est donc raisonnable d'envis*ager* qu'il soit né vers 1040.

272. CBMA 7103

273. CBMA 7028

274. Il y a dans le cartulaire 47 mentions pour le prieur Hugues et seulement 23 pour tous ses prédécesseurs et successeurs réunis.

275. CBMA 7131

276. CBMA 7206, 7225

277. Voir CBMA 7223 et le commentaire dans l'annexe.

dans le sous-chapitre qui concerne la construction de notre matrice de données.

Le prieur Hugues – accompagné de sa famille qui est très active dans les affaires de la région – est le personnage le plus important pour la détermination de la date des actes du cartulaire parce qu'il souscrit ou intervient dans une cinquantaine d'actes. Chacune de ses apparitions permet de dater le document dans la fourchette des 35 ans que nous venons d'estimer comme celle de ses fonctions de prieur. La période demeure large, et Hugues rencontre au moins deux générations de personnes au cours de cette période, ce qui va nous permettre d'essayer de réduire cette fourchette en la divisant en deux sous-périodes. Les dates obtenues pour certains de ces personnages à partir des cartulaires de Cluny, Marcigny, la Fierté-sur-Grosne et Saint-Vincent de Mâcon nous signalent assez bien les limites de ces deux générations ce qui, ajouté à d'autres phénomènes reflétés dans les actes, nous a conduit à distinguer deux périodes durant le temps où il était prieur : une première qui va de 1080 jusqu'à environ l'an 1100 et une seconde période depuis cette date jusqu'à l'an 1115, mais dont la portée peut s'étendre jusqu'à la fin de l'abbatiat de Pons à Cluny en 1123. Ainsi, la fourchette au sein de laquelle dater des documents est réduite à deux sous-périodes de vingt et quinze ans respectivement.

Dans les actes datés de la première période (1080-1100), nous trouvons essentiellement des donations à Paray, de trois types principaux : les donations *pro anima* et *pro remedio anime*, surtout par les contemporains du prieur Hugues (ceux nés dans la décennie de 1040-50), mais parfois au nom de leurs parents à leur décès ; des donations *pro susceptione* (entretien), lors de l'entrée dans la vie religieuse de membres des familles puissantes ; des arrangements qui donnent lieu à des donations après des litiges ou des *calumniae*, souvent liés à la délimitation de terres ou à l'exercice de droits fonciers. Les donations « par libéralité » sont peu nombreuses. L'activité de la génération de Hugues, dont les deux frères, *Girardus* et *Artaldus*, sont des très bons représentants, se trouve entremêlée avec les derniers gestes de la génération antérieure, et elle commence à être éclipsée au cours de la décennie 1090 par celle de ses enfants, nés au cours des décennies 1060-70 et dont l'activité s'étend jusqu'aux années 1120. La génération du prieur Hugues est surtout active depuis la fin de la décennie 1060, comme on le voit dans les quelques notices de cette époque qui sont conservées. Donc, au moment de la rédaction des actes de cette première période, cette génération, qui est déjà dans la quarantaine ou la cinquantaine, se trouve au plus fort de son activité.

Pour déterminer l'an 1100 comme année charnière entre ces deux périodes du priorat de Hugues, nous avons pris en compte trois critères. Le premier est la disparition physique de la génération de Hugues, ce qui est reflété par de nombreuses *donations pro anima*, ce qui commence à se produire aux alentours de l'an 1100. La génération suivante, qui fait son apparition vers la fin de la décennie 1090, accomplit à son tour ces donations et l'activité de la plupart de ses membres est attestée jusqu'à après la mort de Hugues, dans les derniers documents enregistrés dans le cartulaire, au cours de la décennie 1120. Aux alentours de l'an 1100, deux autres événements rapportés dans les chartes nous ont semblé significatifs : d'un côté, le début des croisades (1095-1096) qui déclenche des séries de donations *pro redemptione animarum*, des donations-ventes pour financer les voyages auprès de l'Église, mais aussi des litiges autour des *malae*

consuetudines qui grevaient les paysans[278] ; d'un autre côté, la stratégie agressive du prieuré pour consolider ses possessions et accroître son emprise sur des cellules territoriales à travers l'usage de la donation-vente qui devient le type de charte le plus utilisé après l'an 1100, dont les débuts timides sont toutefois perceptibles à la période antérieure.

La compilation du cartulaire étant aussi une entreprise personnelle de Hugues, elle débute à son époque, et plus de la moitié des actes peuvent être datés du temps de son priorat. Le prieur *Artaldus*, du temps de l'abbé Pons de Cluny (1109-1123), ajoute encore une dizaine d'actes, mais du temps de Pierre le Vénérable, entre 1124 et 1157, ne nous reste qu'une vingtaine d'actes. Nous observons dans ce dernier groupe une succession de *placita* entre Paray et quelques familles puissantes, notamment les Bourbon-Lancy, autour de litiges non résolus depuis des décennies. Deux litiges avec l'évêché d'Autun se terminent dans des termes peu avantageux pour le monastère de Paray. La donation *pro anima* disparaît pratiquement et la donation-vente, qui devient le cadre juridique et scriptural de solution des conflits, oblige le monastère à débourser des sommes importantes pour dédomm*ager* le donateur.

Finalement, neuf des dix documents conservés dans le cartulaire pour les époques postérieures à l'abbatiat de Pierre sont datés car il s'agit de chartes et *precepte* qui relèvent d'une forme diplomatique bien différente. La compilation du cartulaire est effectivement désactivée après l'époque de l'abbé Pons (1109-1123) et les quelques chartes et diplômes ajoutés ensuite, comme les cartulaires D et E de Cluny, relèvent d'un récit patrimonial et historique concentré sur les relations de l'établissement avec le pouvoir public, notamment royal[279]. Ainsi ces derniers documents transcrits reflètent principalement deux actions juridiques : l'arrêt des hostilités et des exactions fiscales de la part des comtes de Chalon sur Paray et Tolon, qui était passé sous le contrôle du prieuré à une époque très précoce ; et le serment et la confirmation par les comtes de Chalon, avec l'intervention royale, des fameuses cartae (chartes de communauté) sur les populations de ces lieux.

On ne peut pas affirmer si le monastère connaît un temps de plénitude à l'époque de la confection du cartulaire, pendant la deuxième période de l'abbé Hugues de Cluny et sous la direction du prieur Hugues de Buxol, puisque notre vision des affaires du monastère se trouve précisément médiée par le cartulaire, mais cette époque correspond sans doute à un moment de croissance patrimoniale, ce qui est probablement à mettre en relation avec une plus grande proximité entre Cluny et ses dépendances dans le Brionnais et le Charollais après l'élection de Hugues Semur (1049). La deuxième vague de donations, plus timide que la première, qui commence aux alentours du changement de siècle, et les efforts du monastère pour réunir ses propriétés pendant la deuxième partie du priorat de Hugues (1100-1115), ont rapidement échoué et la désactivation du cartulaire peu de temps après ne nous permet de suivre l'évolution de cette entreprise (que nous supposons arrêtée). La consolidation de l'abbaye de Cîteaux, fondée en 1098,

278. Sur la figure des "mauvais usages" au XIe siècle voir : Florian Mazel. *Encore les "mauvaises coutumes...Considérations sur l'Église et la seigneurie à partir de quelques actes des cartulaires de Saint-Victor de Marseille"*. 2010, p. 613-626

279. Bruel et Bernard, *Recueil des chartes de l'abbaye de Cluny : 802-954*, Préface, p. XXVIII-XXXVIII

et du monastère de la Ferté-sur-Grosne en 1113 ont détourné les flux de donations des familles auparavant adressées à Paray. Le monastère perd également la protection et la faveur des comtes de Chalon après la mort de Gui de Thiers (1113), alors qu'apparaissent les premiers mouvements communautaires au sein des populations environnant le monastère, qui cherchent une protection ducale et royale.

4.3.2 Classement et cotation du cartulaire.

Selon l'édition du XIXe siècle, le cartulaire original était constitué de 245 items [280], 13 d'entre eux n'ont été transmis par aucune copie. Cinq autres items sont illisibles ou très fragmentaires, et 24 autres sont tellement laconiques – entre 10 et 30 mots – qu'on peut à peine les considérer comme informatifs compte tenu du manque de contexte qui rend les éléments présents difficiles à identifier. On peut réaliser pour les quelques 200 documents restants une classification diplomatique classique, parfois compliquée il est vrai. Ce qui nous conduit à distinguer 12 chartes, 162 notices, 4 privilèges et 4 préceptes. Comme nous l'avons mentionné précédemment, seuls onze de ces actes mentionnent une date, dont huit ont une chronologie postérieure à 1147. Il existe aussi dans l'édition 16 documents administratifs – visites, recensements, comptes. Ce dernier groupe ne faisait pas partie du cartulaire originel : il s'agit d'un ajout éditorial. En fait, la plupart de ces derniers documents sont datés entre 1270 et 1342 et appartiennent à une tradition documentaire et intellectuelle clairement différente.

Une datation plus précise des actions reflétées dans les actes du cartulaire n'avait jusqu'ici jamais été proposée, et c'est ce que nous allons essayer de faire. Comme dans le cas de Cluny, on peut réaliser une estimation rapide en se référant à la présence des abbés de Cluny. Dans 48 documents un abbé est mentionné, et dans 31 des cas, il s'agit de l'abbé Hugues (1049-1109), dont nous avons déjà mentionné la longévité problématique pour la datation. Faisant une extrapolation rapide sur la base du vocabulaire et des séries, on pourrait estimer qu'au moins la moitié des documents originels a été rédigée à son époque, c'est-à-dire entre 1049 et 1109. Dans une dizaine de documents l'abbé Pons (1109-1122) est mentionné, permettant une datation entre 1109 et 1123, et dans quelques-uns sont aussi mentionnés les abbés de Cluny Pierre (1122-1157) et Odilon (994-1049).

Le cartulaire se trouve grossièrement divisé, comme indiqué initialement par Ulysse Chevalier et Canat de Chizy, en trois groupes documentaires : un groupe mémoriel, un groupe de gestion, c'est-à-dire les actes fonciers, et un dernier groupe relatif au pouvoir public, ce qui correspond globalement au début, au milieu et à la fin du cartulaire, selon l'ordre reconstitué par l'éditeur. L'historiographie actuelle autour des cartulaires-chroniques, comme celui-ci, leur accorde la fonction de construire une mémoire de l'institution religieuse, ce qui se traduit dans la confection d'un appareil documentaire soutenu, normalement au début, par un récit historiographique, dogmatique ou hagiographique suivi d'un large ensemble d'actes issus des différentes actions juridiques auxquelles l'institution a attribué un intérêt particulier parce qu'elles

280. Chevalier, *Cartulaire du prieuré de Paray-le-Monial, ordre de Saint-Benoît*, Introduction, V-X

sont un témoignage à la fois légal et historique de sa constitution patrimoniale [281]. Cette fonction, qui est parfaitement remplie dans le cas de Cluny, est aussi détectable dans le cartulaire de Paray, même s'il inclut à la fin des documents reflétant les relations avec l'autorité qu'elle soit royale, comtale ou ecclésiastique. U. Chevalier a aussi ajouté quelques documents assez intéressants concernant la gestion patrimoniale : copies des visites diocésaines et recensements.

D'ailleurs, comme les différentes copies du cartulaire l'indiquent, il se trouvait lui-même initialement subdivisé en parties numérotées en chiffres romains [282]. L'éditeur a essayé de restituer les différentes divisions du cartulaire mais est un tâche ardue car après les premières divisions la foliation n'apparait que de manière intermittente dans le cartulaire. Dans le Recueil de l'abbaye de Cluny au moins trois campagnes de copie sont détectables dans les premiers cartulaires. Est-ce le cas pour Paray ? Les divisions observées peuvent-elles correspondre à différents moments de compilation du cartulaire ? Nous pensons, à la lumière de la chronologie restituée que c'est le cas, mais en notant deux points : les deux premières divisions, qui portent une numérotation originale, semblent les seules à faire partie de l'entreprise d'origine, les autres sont en réalité des ajouts à différentes époques et la plupart des actes de la dernière division semblent avoir été ajoutés à l'époque de rédaction de l'acte ; d'ailleurs, même si les divisions présentent une certaine cohérence, la logique de compilation de chacune pouvait répondre tout simplement à des impératifs pratiques en absence d'un vrai modèle de compilation :

- Première division [283] : dans la foliation originelle cette division est numérotée de l'item I jusqu'à l'item CXVII. Les premiers items du cartulaire (CBMA 7017-CBMA 7058), spécialement ceux qui correspondent au récit mémoriel (les chartes de fondation et les chronologies des comtes de Chalon) ont probablement été rédigés au début de la confection du cartulaire. La vingtaine d'actes restants correspond au tournant du siècle et nous les avons principalement datés de la décennie de 1090-1100. Les documents de cette division ainsi que des trois divisions qui suivent ont été probablement compilés du temps du prieur Hugues parce qu'aucun acte ne dépasse sa chronologie (voir figure 4.1). D'ailleurs, même si on ne peut pas préciser la date de compilation du cartulaire, la forte concentration des premières divisions dans la décennie 1080-1090 semble nous indiquer que cette période, ou la décennie qui suit, est la plus probable pour le début de compilation ;

- Deuxième division [284] : cette vingtaine d'actes reprend une nouvelle comptabilisation de chapitres par des chiffres romains, de l'item LXXIX à l'item CXV. Il ne s'agit pas d'un sous-groupe interne à la première division sinon d'une nouvelle

281. Voir à ce sujet : Patrick Geary. "Entre gestion et gesta". In : *Ecole des chartes (éd), Les Cartulaires, Paris* (1993), p. 13-26, Guyotjeannin et al., *Les cartulaires : actes de la Table ronde organisée par l'École nationale des chartes et le GDR 121 du CNRS*, p. 13-27, Iogna-Prat, "La geste des origines dans l'historiographie clunisienne des XIe-XIIe siècles", p. 135-191

282. "Ce Cartulaire, grand in-4º en parchemin, couvert de même contient cent quatorze feuillets cottés en chiffres romains, dont plusieurs, surtout au commencement et vers la fin, sont lacérés et morcelés. L'écriture est d'environ l'an 1200, époque approchant des plus anciens cartulaires" (Chevalier, *Cartulaire du prieuré de Paray-le-Monial, ordre de Saint-Benoît*), Introduction, VIII

283. CBMA 7017-7082

284. CBMA 7083-7103

série par chapitres. Les actes de cette division sont très difficiles à dater, ils présentent une forme abrégée bien plus marquée que dans le premier groupe. Cet ensemble est probablement compilé dans les premières années du XIIe siècle ; il inclut quelques actes de cette époque, mais surtout des actes de la période antérieure (1080-1100) ;

- Troisième division [285] : ce petit groupe d'actes est introduit par une périphrase narrative : *In præcedenti narratione hujus operis.* Il s'agit probablement, comme pense l'éditeur, d'une addition postérieure, du temps de l'abbé Pons (1109-1122). En tout cas elle semble une addition faite au plus tôt dans la deuxième période du prieur Hugues (après l'an 1100) car elle concerne à d'importantes donations de la décennie de 1090 ;

- Quatrième division [286] : en CBMA 7112 commence une nouvelle division par chiffres romains ; les 6 premiers de manière consécutive mais ensuite la numérotation devient intermittente. La plupart de ces actes correspond à la deuxième période du priorat de Hugues, et les autres sont un mélange de toutes les époques antérieures depuis l'époque du comte-évêque Hugues (999-1039). Il semble que le compilateur essaie de compléter les titres fonciers des décennies antérieures à l'an 1080. On pourrait penser, mais sans certitude, que la compilation des actes de cette division est faite alors que le prieur Hugues est encore vivant puisqu'il apparaît régulièrement dans les actes.

- Cinquième division [287] : cette nouvelle division porte deux titres de section : *Incipiunt cartae de Tolon et Monti – Cap. I* (CBMA 7180) et *Incipiunt cartae Baronenses* (CBMA 7196). Le premier groupe inclut des actes de la deuxième période du prieur Hugues et, dans le deuxième, le compilateur se focalise sur la récupération des chartes de donation du temps de la fondation du monastère (977) jusqu'aux premières décennies du XIe siècle, donc de l'époque des comtes de Chalon, Lambert (†988) et son fils le comte-évêque Hugues (988-1039). Quelques documents sont déjà de l'époque du prieur *Artaldus* (vers 1118 -1125/30), ce qui peut laisser penser que cette partie de la compilation se fait sous sa direction.

- Sixième division [288] : à partir de CBMA 7217, la compilation du cartulaire commence à ralentir et les actes ont une chronologie plus hétérogène. On observe un « saut » vers l'époque de l'abbé Pierre, soit entre 1123-1157, même si on a encore quelques documents des époques antérieures. Commencent à apparaître les premiers documents datés dont nous avons déjà expliqué la nature, et en particulier le groupe final CBMA 7237-7244 qui inclut des préceptes et des privilèges probablement ajoutés de la date de leur réception : 1147, 1206, 1243.

Enfin, à partir de CBMA 7059, c'est-à-dire après le groupe d'actes correspondant au récit mémoriel du cartulaire, dans la première division, il est assez commun que la copie des documents se fasse par séries de trois, quatre ou cinq notices concernant le plus souvent les donations ou accords au sein d'une même famille. Des notices « doubles » sont également présentes, l'une fonctionnant comme complément de l'autre [289]. À d'autres occasions, et nous l'avons déjà exposé, différentes actions juridiques s'intègrent dans une même notice [290], ce qui impose de considérer deux ou trois dates pour chaque

285. CBMA 7104-7111
286. CBMA 7112-7179
287. CBMA 7180-7195 et CBMA 7196-7216
288. CBMA 7217- 7244
289. Par exemple : CBMA 7060-7061 ; CBMA 7173-7174
290. Par exemple CBMA 7103, 7127

document. Cette sérialisation qui montre une fouille intentionnelle dans les archives par le compilateur est faite probablement avec l'intention soit de rassembler les donations qui impliquent un donateur ou sa famille, soit celles qui impliquent une même terre, *villa* ou église, conformément à l'effort de Paray de réunir des propriétés dispersées entre plusieurs mains. Cet exercice de sérialisation regroupe le plus souvent des chartes non contemporaines, et l'intervalle qui les sépare peut aller de quelques années à plusieurs décennies. Malheureusement cette sérialisation, qui aurait pu ordonner l'ensemble du cartulaire, n'est pas systématique et nous trouvons des séries un peu partout sans que cela puisse être considéré comme un modèle de rédaction.

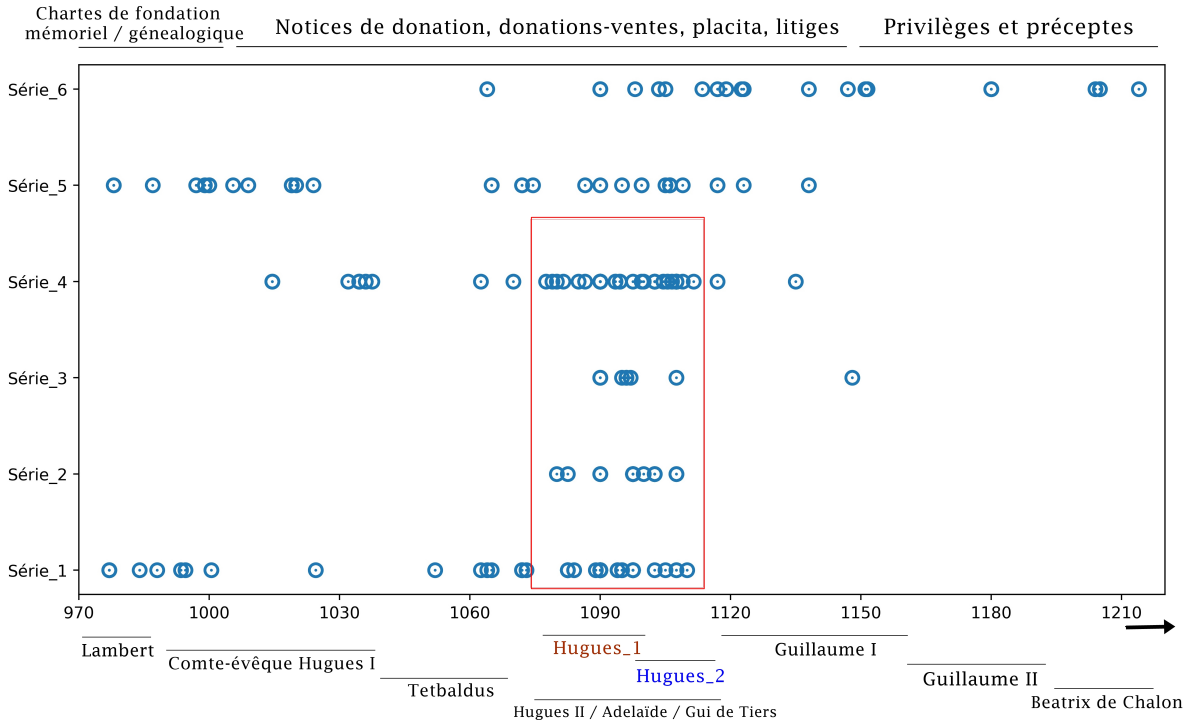

FIGURE 4.1 – Distribution chronologique restituée du cartulaire de Paray-le-Monial. Les séries correspondent aux divisions internes du cartulaire et elles peuvent répondre à des ajouts et à différents moments de compilation. En haut, est indiqué une division typologique générale des documents ; et en bas, les chronologies des comtes de Chalon et du prieur Hugues (le cadre en rouge indique les actes datés et probablement compilés du temps de son priorat).

4.4 Trois exemples de datation des actions juridiques dans les actes

Dans les exemples qui suivent nous allons commenter en détail la datation des actions juridiques de trois actes du cartulaire de Paray-le-Monial. Nous avons choisi ces trois exemples parce qu'ils sont représentatifs des avantages qu'offre l'usage de notre matrice comme outil d'assistance à la datation, et qu'ils permettent d'ilustrer comment elle permet de dépasser les erreurs de datation des techniques traditionnelles. En effet, le chercheur peut se confronter à trois types d'erreurs de datation difficilement

discernables par les méthodes traditionnelles, spécialement dans les actes des XIe et XIIe siècles : les erreurs de datation diffusées dans les éditions érudites ; les erreurs de datation à cause de l'homonymie et les erreurs de datation trouvées dans le document lui-même. La matrice nous dresse rapidement un panorama assez complet de toutes les chronologies des personnes mentionnées dans un document, et permet de proposer une date plausible. Ce repérage massif peut parfois conduire à des chronologies erronées à cause de la nature des actes (voir discussion dans la partie 5.5) mais ces erreurs sont facilement discernables. Finalement les critères de sélection chronologique demeurent l'apanage du chercheur puisqu'il est pour l'instant impossible d'avoir une datation, basée sur cette méthode, qu'on puisse qualifier d'auto- voire semi-automatique.

a) CBMA 7095 (acte non daté ; date proposée : 1090-1115)

« *Quidam miles, relinquens seculum et accipiens monachicum habitum, dedit Deo et huic loco unam vercheriam in* villa *de Prisiaco, quæ est sita juxta domum monachorum, et de duabus partibus terminat terra dominæ Stephanæ, de tercia terra Sancti Petri, de alia terra Sancti Grati. Dedit etiam ad ipsum locum omnem suam partem de terra quam habebat, cum fratre suo Humberto, in* villa *de Montet et in Chaloer : tali convenientia ut a natale sancti Martini in tres annos, si voluerit frater suus Humbertus, redimat ipse CXX. solidos ; si vero C. solidos in hoc tempore reddiderit, iterum monachi terram teneant, donec XX. solidos reddat. Quod si in istis tribus annis eam non redimerit, post hoc etiam in perpetuo eam teneant monachi et possideant, cum omni integritate sua, sive in silvis, pratis, vineis, aquis aquarumque decursibus, totum ad integrum usque ad inquirendum. S' Hugonis, Humberti. S' Hugonis de Saleniaco. S' Seguini de Colmines. S' Ansedei præpositi, Artaldi Ruil.* »

Notre matrice de données (tableau 4.1) prend en compte trois souscripteurs de l'acte : *Hugues de Saleniaco, Seguinus de Colmines* et *Artaldus Ruil*. Le premier, *Hugues de Saleniaco*, est mentionné comme souscripteur dans les actes de Cluny et Marcigny-sur-Loire entre 1087 et 1122. À Paray on le retrouve dans trois actes ; nous avons daté deux d'entre eux de la première période du prieur Hugues (1080-1100) et le troisième est daté dans une fourchette réduite (1094-1098) grâce à la présence[291] d'*Humbertus*, prieur de Marcigny entre ces dates.

Le deuxième souscripteur, *Seguinus de Colmines*, est mentionné dans quatre actes à Cluny et Marcigny ; trois sont datés dans le document entre 1106 et 1108 et un quatrième est daté dans une fourchette grâce à la chronologie de l'abbé Hugues (1049-1109). À partir de ces deux témoins nous pouvons déjà approcher une date probable pour le document : entre 1090 et 1115, c'est-à-dire la 2eme période du prieur Hugues ; cette date est soutenue par les autres souscripteurs qui apparaissent dans les actes cités et qui ont une forte activité documentaire principalement dans cette période.

Les dates du troisième souscripteur, *Artaldus Ruil*, semblent, au premier abord, problématiques. En effet, il est présent dans deux autres actes : l'un, en C2, daté de 1050, est très précoce par rapport aux périodes pré-citées ; l'autre, en C1, a une datation dans une fourchette très ample : 1049-1109. Notre matrice n'a pas récupéré de dates pour les souscripteurs qui accompagnent *Artaldus Ruil* dans ces deux actes

291. CBMA 7108

de l'abbaye de Cluny (CBMA 4689, 4751), mais les deux actes sont liés à la famille de *Berziaco* (Berzé-la-Ville), très puissante dans la région, et il n'est pas déraisonnable de penser qu'ils ont été rédigés dans la même période.

S'agit-il d'une erreur de datation de la part des éditeurs du cartulaire de Cluny ? Nous pensons que c'est le cas. L'acte en question, qui nous apporte la date de l'an 1050 [292] vient des copies qu'avait faites Lambert de Barive du chartrier original de l'abbaye de Cluny. Lambert de Barive, lui, avait daté le document de l'an 1100 environ, mais cette date avait été corrigée par Bernard et Bruel dans l'édition des actes de Cluny, argumentant des inconsistances dans le vocabulaire pour un document de l'an 1100. À la lumière de notre chronologie reposant sur deux autres souscripteurs, la datation de Lambert de Barive semble correcte et une nouvelle date doit être proposé pour cette charte du cartulaire de l'abbaye de Cluny, parce que *Artaldus Ruil* semble un personnage actif dans la documentation au plus tôt dans la décennie de 1080.

	Terminus a quo	Terminus ante quem	Numéro CBMA	Cartulaire/ Recueil	Version trouvée
A	Hugonis de Saleniaco				
A1	1090	1109	4439	abbaye de Cluny	Hugonis de Soloniaco
A2	1087	-	5052	abbaye de Cluny	Hugo de Saligiaco
A3	1106	-	5280	abbaye de Cluny	Hugo de Salenniaco
A4	1147	-	5563	abbaye de Cluny	Hugo de Siniciaco
A5	1088	-	11186	Marcigny-sur-Loire	Hugonis de Saloniaco
A6	1122	-	11410	Marcigny-sur-Loire	Hugonis de Saliniaco
A7	1080	1109	7064	Paray-le-Monial	Hugo de Saliniaco
A8	1080	1100	7036	Paray-le-Monial	Hugo de Saliniaco
A9	1094	1098	7108	Paray-le-Monial	Hugo Saliniaco
B	Seguini de Colmines				
B1	1049	1109	4655	abbaye de Cluny	Seguini de Culminas
B2	1106	-	5270	abbaye de Cluny	Seguino de Culminis
B3	1108	-	5304	abbaye de Cluny	Seguinus de Culminis
B4	1106	-	11281	Marcigny-sur-Loire	Seguino de Culminis
C	Artaldi Ruil				
C1	1049	1109	4689	abbaye de Cluny	Artaldus Ruilus
C2	1050	-	4751	abbaye de Cluny	Artaldus Ruil

TABLE 4.1 – Matrice chronologique pour CBMA 7095. Les lignes en gris montrent la version orthographique de chaque personnage présent dans le document. Nous cherchons dans la base du CBMA, par similitude des chaines de caractères, toutes les coïncidences pour chaque personne. S'il y a un résultat positif on récupère les données associées : dates, numéro, et cartulaire. La dernière colonne montre les différentes versions orthographiques du nom de cette personne trouvées dans les documents de référence.

b) CBMA 7168 (acte non daté ; date proposée : 1105 – 1113)

« *Domnus Lebaudus de Digonia faciebat multas querimonias et molestias contra locum et fratres Aureæ Vallis, de quadam silva quæ vocatur Rasneria, quam silvam dederat Adelaidis comitissa, filia Teobaldi comitis, Deo et ad luminariam hujus loci, et quandam capellam,*

292. CBMA 4751

dictam ad Sanctam Ecclesiam, justa eamdem silvam : dicebat enim dictus Lebaudus dictam silvam de suo esse casamento. Tandem vero in fine cœlitus inspiratus, sentiens sibi mortem vicinam, misit et vocavit priorem hujus loci, et dedit et quictavit quicquit ipse et sui in dicta silva cum terra pertinente habebat vel habere debuerat, ad luminariam ecclesiæ de Paredo, pro remedio animæ suæ, et laudavit donum quod de eadem silva domina Adalaidis fecerat. Dedit etiam idem Lebaudus quemdam hominem justa eamdem silvam manentem, nomine Alardum Fadi, et heredes ejus, sicut ipse et sui antecessores tenuerunt, in perpetuum pacifice possidenda. Testes et laudatores : domina Adalaidis et Wido de Tier, filius ejus ; uxor domni Lebaudi et filius ejus, Jocerannus de Copetre, Atto Buxul, H. de Ozoles, P. de Civin', H. de la Tor et alii multi.»

Dans ce deuxième exemple (tableau 4.2) notre matrice de données nous autorise à considérer six souscripteurs dans l'acte afin d'approcher une chronologie. Nous pouvons y trouver deux personnes de premier ordre : la comtesse Adelaïde et son enfant, Gui de Thiers, comte de Chalon (1085-1113). La chronologie de ces deux personnes nous renvoie principalement à la 1ere période du prieur Hugues (1080-1100). Si l'on connait la date de disparition de Gui de Thiers (1113), pour Adelaïde, fille de Tetbaldus, comte de Chalon, la chronologie est bien plus floue : elle est née vers la décennie de 1020 (son père est né vers 990/95) et il est peu probable qu'elle soit encore vivante après la décennie de 1100.

En ce qui concerne les quatre souscripteurs : *Letbaldus de Digonia, Iotcerannus de Copetra, Atto de Buxol* et *Petrus de Civiniaco*, on les identifie dans la matrice comme appartenant surtout aux décennies 1090 - 1110 et ils participent encore à quelques documents du temps de l'abbé de Cluny Pons (1109-1123). A partir de ceux-ci et de Gui de Thiers on peut déjà estimer une première chronologie, très satisfaisante : 1090-1113, mais nous pouvons la préciser davantage.

Le document que nous essayons de dater constitue la *donatio pro anima* - donc, très probablement rédigée à la fin de sa vie - de Letbaldus de Digonia, personnage récurrent de la 1ère période du prieur Hugues et dont la dernière apparition datée, selon la matrice, est de l'an 1105 dans le cartulaire de l'abbaye de Cluny (A3). Notre document doit être probablement postérieur à 1105, ce qui nous autorise à proposer une fourchette bien plus serrée (1105-1113). En tenant en compte la chronologie d'Adelaïde, qui en 1105 serait très âgée - 75 ans environ – ce qui est rare compte tenu de l'espérance de vie de l'époque - peut être devrions nous rester plus proches de 1105 que de 1113. Il n'est même pas déraisonnable de penser qu'Adelaïde ne participe pas à la donation et qu'elle est évoquée parce qu'elle participe aux antécédents de la donation. Mais tout cela reste trop hypothétique.

Il y a par ailleurs quelques remarques à faire sur les dates récupérées dans la matrice qui, au premier regard ne semblent pas si cohérentes avec la chronologie proposée, ce qui est représentatif de deux problèmes assez communs dans le repérage des noms de personnes : d'un côté les dates à fourchettes amples, et de l'autre l'homonymie dans les noms doubles de personnes.

Ainsi dans le document D1 la chronologie proposée par J. Richard à Marcigny pour *Atto de Buxol* (1065-1097) semble un peu décalée par rapport à la nôtre (1105-1113). *Atto de Buxol* est le neveu du prieur Hugues et il est peu probable que sa naissance soit antérieure à la décennie de 1060 (voir la chronologie des *Buxol* ci-dessous). Il est

mentionné dans trois actes de Paray à partir de la décennie de 1090. Il faut alors considérer que la date l'acte de Marcigny où *Atto de Buxol* est présent est bien plus proche du *terminus ante quem* (1097) que du *terminus a quo* (1065) proposés pour ce document, ce qui suggère une réévaluation de la date de cet acte.

	Terminus a quo	Terminus ante quem	Numéro CBMA	Cartulaire/ Recueil	Version trouvée
A	Lebaudus de Digonia				
A1	1090	1110	4488	abbaye de Cluny	Letbaudo de Digonia
A2	1090	1095	4489	abbaye de Cluny	Letbaudo de Digonia
A3	1100	1115	7082	Paray-le-Monial	Letbaldus de Digonia
A4	1080	1100	7171	Paray-le-Monial	Letbaldus de Digonia
A5	1080	1100	7175	Paray-le-Monial	Letbaldus de Digonia
A6	1097	1115	7183	Paray-le-Monial	Letbaldus de Digonia
B	Wido de Tier				
B1	1085	1100	7103	Paray-le-Monial	Widone de Tier
B2	1096	1100	7224	Paray-le-Monial	Wido de Tierno
C	Iocerannus de Copetre				
C1	1080	-	5000	abbaye de Cluny	Ioceranno de Coperia
C2	1083	-	5027	abbaye de Cluny	Iocerannus Coperia
C3	1089	-	5061	abbaye de Cluny	Ioceranni de Iopera
C4	1123	-	11319	Marcigny-sur-Loire	Iocerannus de Chopetra
C5	1130	-	11320	Marcigny-sur-Loire	Ioceranni de Chopetra
C6	1065	1079	7061	Paray-le-Monial	Iocerannus de Copetra
C7	1080	1100	7175	Paray-le-Monial	Iocerannus de Coperia
C8	1115	1123	7192	Paray-le-Monial	Iocerannus de Copere
C9	1147	1166	7220	Paray-le-Monial	Gausceranno de Copetra
C10	1115	1122	7223	Paray-le-Monial	Gauscerannus de Copetra
D	Atto Buxul				
D1	1065	1097	11227	Marcigny-sur-Loire	Atton
D2	1080	1100	7036	Paray-le-Monial	Atto de Buxol
D3	1090	1098	7044	Paray-le-Monial	Atto de Buxol
D4	1090	1110	7065	Paray-le-Monial	Atto filius eius
D5	1085	1100	7103	Paray-le-Monial	Atto Buxol
E	P. de Civin'				
E1	1105	-	5257	abbaye de Cluny	Petrus de Civiniaco
E2	1108	-	5304	abbaye de Cluny	Petrus de Civione
E3	1128	-	5436	abbaye de Cluny	Petrus de Civegnone
E4	1098	-	11279	Marcigny-sur-Loire	Petro de Civiniaco
E5	1098	-	11302	Marcigny-sur-Loire	Petrus de Sivignon
E6	1100	1115	7082	Paray-le-Monial	Petrus de Civinon
E7	1100	1113	7175	Paray-le-Monial	Petrus de Civinun
E8	1123	-	7192	Paray-le-Monial	Petrus de Civignun

TABLE 4.2 – Matrice chronologique pour CBMA 7168.

En ce qui concerne *Iotcerannus de Copetra*, souscripteur, sa chronologie est problématique parce que la matrice a récupéré, sous la même combinaison de nom et prénom, un père et son fils (ou son neveu). Le premier est un participant actif de la 1ère période du prieur Hugues (1080-1100) et il apparait dans trois actes datés de cette même période de l'abbaye de Cluny (C1, C2, C3). Le second est attesté dans différents actes entre 1123 - 1130 à Marcigny et 1146 - 1160 à Paray (C4, C5, C9, C10). S'il est possible d'envis*ager* qu'un individu adulte entre 1080 et 1100 soit encore vivant en 1123, il est impossible qu'il soit encore en vie en 1146 ; le plus logique est

donc de penser qu'il s'agit de deux personnes différentes : l'une vivant du temps du prieur Hugues et l'autre, un descendant en ligne directe, et actif dans les documents à la fin de l'époque de l'abbé de Cluny, Pons (1109-1123) et pendant l'abbatiat de Pierre (1123-1157).

c) CBMA 7192 (Acte daté de l'an 1149 ; date proposée : 1123)

« *Domnus Guichardus de Digonia, quadam die apud Trenorchium in quodam tirocinio percussus, apud Cluniacum se deportari fecit et coram fratribus suis, Letbaldo et Joceranno, et aliis amicis suis fecit testamentum. Et inter cetera dedit Deo et loco Cluniacensi, pro sepultura et pro remedio animæ suæ, duos homines, Humbertum et Bernardum Closers, fratres, in* villa *quæ Vetus Vinea dicitur, sicut ipse et antecessores sui tenuerunt, liberos et inmunes. Et hoc fratres ejus, Letbaldus et Jocerannus, in præsentia ejusdem et donni Hugonis abbatis et Artaldi prioris Cluniacensis, laudaverunt et concesserunt. Domnus vero Artaudus, prior Cluniacensis et Parede, dictos homines, de consensu et voluntate fratrum et abbatis Cluniacensis, dedit ad luminariam de Paredo et heredes eorum, sicut dictus G. dederat ad locum Cluniacensem, sine retentione aliqua. Testes ex hoc dono : Lebaudus et Jocerannus de Digonia, B. de Calvomonte, Jocerannus de Copere, Hugo Buxul, Jocerannus de Marcili, G. de Chasanes, Petrus de Civignun. Actum anno gratiæ* $M^{o}.C^{o}.XL^{o}.V^{o}.IIII^{o}.$»

Dans notre troisième exemple (tableau 4.3) nous allons nous servir de 9 personnes dans le but de proposer une date approximative. D'un côté, les personnages secondaires : *Guichardus de Digonia, Iotcerannus de Digonia, Iotcerannus de Copetra, Hugo Buxul, Iotcerannus de Marcile, Gaufridus de Cassanias* et *Petrus de Civignion*, et de l'autre, deux personnages du premier ordre : le prieur de Paray *Artaldus* et l'abbé Hugues de Cluny. Comme on peut l'observer dans la matrice, les personnages secondaires figurent surtout dans la 2eme période du prieur Hugues, de 1098 jusqu'à 1123. Ils font partie de ce que nous considérons d'être la génération qui suit celle du prieur Hugues (voir 5.2), constituée d'individus qui apparaissent vers le milieu de la décennie de 1090 et s'éteignent après la décennie de 1120.

En ce qui concerne les deux personnages principaux, comme on l'a déjà indiqué dans la chronologie des prieurs de Paray, ci-dessus, on sait de manière certaine que le prieur *Artaldus* se trouve à Paray vers 1118 jusqu'à environ 1125/1130, et qu'Hugues abbé de Cluny demeure en poste jusqu'à sa mort en 1109.

Nous notons ici une première incohérence : le prieur *Artaldus* est entré en fonction vers 1118, soit 10 ans après la mort de l'abbé Hugues, ce qui rend impossible leur participation commune à ce document. Ensuite, l'acte est effectivement daté dans l'eschatocole de l'an 1149, ce qui est une date trop tardive tant pour les souscripteurs et que pour les personnages principaux. En effet, notre matrice, s'appuyant sur les personnes secondaires et le prieur *Artaldus*, nous renvoie à la décennie de 1120 pour ce document, alors que la date présumée officielle du document le date de 25 à 30 ans plus tard, ce qui semble à première vue complètement invalider la proposition.

S'agit-il d'une erreur de la matrice ou, plus probablement, comme l'avait suggéré Canat de Chizy qui avait déjà souligné cette incohérence, d'un lapsus du copiste ou du scripteur ? En tout état de cause notre matrice est soutenue par des sources documentaires incontestables. Un dernier élément peut faire pencher la balance : en

ce qui concerne la présence de l'abbé Hugues : il serait plus plausible, et cohérent avec notre matrice, qu'il s'agisse non pas de l'abbé Hugues de Sémur, abbé de Cluny (1049-1109), mais de Hugues II, un abbé qui a passé quelques mois à ce poste en 1123, peut être à la suite de la démission de l'abbé Pons, avant que Pierre le Vénérable lui succède. Cette hypothèse est cohérente avec les éléments de certaines et si on fait confiance à notre matrice elle permet de dater le document à l'année 1123.

	Terminus a quo	Terminus ante quem	Numéro CBMA	Cartulaire/ Recueil	Version trouvée
A	Guichardus de Digonia				
A1	1228		6030		Guichardus de Digonia
B	Gauscerannus de Digonia				
B1	1150		11042	la Ferté-sur-Grosne	Jocerannus de Digonia
B2	1147	1156	7220	la Ferté-sur-Grosne	
B3	1147		7222	Paray-le-Monial	Gauscerannus de Digonia
C	Hugo de Buxol				
C1	1050		1730	abbaye de Cluny	Hugo de Bussol
C2	1112	1119	11284	Marcigny-sur-Loire	Hugonem de Buxolio
C3	1098	1112	11302	Marcigny-sur-Loire	Hugo de Bussul filius Girardi
C4	1100	1125	11403	Marcigny-sur-Loire	Hugo de Buxol
C5	1030	1039	7112	Paray-le-Monial	Ugonem de Buxol
C6	1090	1100	7171	Paray-le-Monial	Hugo de Busol, filius Artaldi
C7	1112	1115	7235	Paray-le-Monial	Hugonem de Buxolio
D	P. de Civin'				
D1	1105		5257	abbaye de Cluny	Petrus de Civiniaco
D2	1108		5304	abbaye de Cluny	Petrus de Civione
D3	1128		5436	abbaye de Cluny	Petrus de Civegnone
D4	1098		11279	Marcigny-sur-Loire	Petro de Civiniaco
D5	1098		11302	Marcigny-sur-Loire	Petrus de Sivignon
D6	1100	1115	7082	Paray-le-Monial	Petrus de Civinon
D7	1100	1113	7175	Paray-le-Monial	Petrus de Civinun
D8	1123		7192	Paray-le-Monial	Petrus de Civignun
E	Iocerannus de Copetre				
E1	1080		5000	abbaye de Cluny	Ioceranno de Coperia
E2	1083		5027	abbaye de Cluny	Iocerannus Coperia
E3	1089		5061	abbaye de Cluny	Ioceranni de Iopera
E4	1123		11319	Marcigny-sur-Loire	Iocerannus de Chopetra
E5	1130		11320	Marcigny-sur-Loire	Ioceranni de Chopetra
E6	1065	1079	7061	Paray-le-Monial	Iocerannus de Copetra
E7	1080	1100	7175	Paray-le-Monial	Iocerannus de Coperia
E8	1115	1123	7192	Paray-le-Monial	Iocerannus de Copere
E9	1115	1122	7223	Paray-le-Monial	Gauscerannus de Copetra
F	Gaufridus de Cassanias				
F1	1100		5229	abbaye de Cluny	Gaufredi de Cassanias
F2	1080	1100	7040	Paray-le-Monial	Gaufridus de Cassagnias
F3	1090	1098	7041	Paray-le-Monial	Gaufredus de Cassagnias
F4	1080	1100	7175	Paray-le-Monial	Gauffredus de Cassaneis
F5	1115	1122	7223	Paray-le-Monial	Gaufredus de Cassanias
G	Iotcerannus de Marcili				
G1	1090	1115	7038	Paray-le-Monial	Iocerannus Marcile

TABLE 4.3 – Matrice chronologique pour CBMA 7192.

4.5 Les entités nommées dans les cartulaires.

Si la variabilité intrinsèque aux documents des cartulaires limite la performance des méthodes d'apprentissage automatique pour le calcul de la date, d'autres éléments peuvent nous aider à resserrer les fourchettes de datation pour un document à partir du croisement des personnes. Parmi ces derniers, trois sont particulièrement importants : la confection des chartes et par extension des cartulaires sur des séries restreintes d'éléments ; la dimension sociale que prennent les affaires notifiées dans les chartes et l'existence de relations de proximité, sous forme de réseaux sociaux, spatiaux et de parenté, entre les personnes.

Avant tout, un cartulaire est un ensemble de chartes qui contient des séries d'un nombre limité d'éléments. Le groupe de chartes qui intègre un cartulaire est un groupe déjà filtré et sélectionné depuis un univers de chartes bien plus vaste : on conserve et copie les chartes qui concernent l'institution rédactrice et, parmi elles, celles qu'on pense être les plus importantes. [293] Les chartes qui concernent un transfert de patrimoine vers l'institution jouent ici un rôle de premier plan. Les bienfaiteurs publics constituent un groupe réduit, où figurent rois, papes et évêques. Les bienfaiteurs privés forment en revanche un groupe bien plus ample, mais presque monopolisé par l'aristocratie locale. Dans certaines régions, celui-ci peut être un groupe important numériquement mais qui demeure maitrisable et surtout organisé en un réseau social sur une base territoriale. Pourtant la vocation régionale des cartulaires, ayant comme centre l'institution, étant bien affirmée, on va le plus souvent trouver des réseaux de donateurs concentrés sur une unité géographique bien définie, unité qui d'ailleurs peut être assez spécifique dans sa division, mais bien délimitée dans son extension. À ce groupe des disposants s'ajoute celui formé par les individus présents pour donner leur validité juridique aux actes. Ainsi scripteurs et notaires, revêtus de l'autorité légale et juridictionnelle nécessaire, vont souvent apparaître à côte des souscripteurs pour clôturer le document. Tout cela se traduit par l'existence dans les cartulaires de listes restreintes d'entités nommées qui correspondent tant aux personnes qu'aux lieux qui participent majoritairement à la plupart des affaires notifiées dans les documents.

Un recueil de chartes, dans la mesure où il est étroitement lié aux intérêts fonciers de l'institution commanditaire, présente des listes d'entités corrélées, reflétant les liens existant dans la réalité sociale et spatiale. C'est précisément l'existence de ces connexions qui fait qu'une datation chronologique à partir des personnes « inconnues » peut être plus efficiente. La raison en est simple : la majorité des personnes connectées dans les documents sont contemporaines et le transfert de données chronologiques profite de ces connections pour aboutir à un transfert chronologie immédiat sur tout le réseau d'individus. Les personnes identifiées dans un seul acte daté permettent de transférer ce repère chronologique vers des documents non datés où ces mêmes personnes apparaissent et fournissent ainsi une borne chronologique pour d'autres personnes mentionnées dans ces documents.

Dans les actes copiés ou transférés, les personnes jouent principalement deux rôles : elles marquent l'origine de l'action juridique en tant que donateurs, vendeurs, etc.,

293. Iogna-Prat, "La confection des cartulaires et l'historiographie à Cluny (XIe–XIIe siecles)"', p. 27-44.

ce qui est normalement inscrit dans le protocole de l'acte ; ou elles participent à la validation juridique de l'action en tant que témoin ou *laudator*, ce qui apparaît dans l'eschatocole de l'acte. Aucun de ces noms n'est là par hasard, mais ils font partie d'une structure en réseau qu'on peut dire corporative. Dans la donation, qui est l'action la plus mobilisée dans les cartulaires de la Bourgogne, les membres d'une famille peuvent jouer les deux rôles. On donne en couple, entre frères (ou en frérèches), les enfants avec leurs parents, les parents avec leurs enfants. Les actes peuvent évoquer plusieurs motivations pour ce don : on donne pour le salut de l'âme des proches mourants, de ceux qui viennent de mourir et parfois pour celui des ancêtres. On donne pour la rédemption des péchés, pour l'entrée d'un fils au monastère ou d'une fille au couvent. On donne pour mettre fin à un litige ou pour compléter une autre affaire, une vente, un loyer, un échange. Cette donation en famille, surtout quand il s'agit d'une aliénation d'une part d'un patrimoine ancien, est approuvée par les parents (*laudatio parentum*) et peut porter le *placitum* de tous les proches qui participent à l'exploitation de la terre ou de la cellule territoriale objet de la donation. L'opposition manifeste à une donation de la part de l'un d'eux figure parfois sur le document. Si la donation suppose une participation monétaire, surtout après le XIe siècle, celle-ci est répartie entre les membres de la famille. Et si l'institution religieuse demeure le bénéficiaire matériel de la donation, la famille est la seule bénéficiaire spirituelle de la transaction.

Mais la donation, qui met en relation un individu privé, et par extension sa famille, avec un centre religieux, n'a pas seulement valeur d'aumône. Elle constitue aussi dans certains cas une tentative pour contrebalancer le pouvoir local en s'associant au pouvoir religieux. La dynamique du don a des implications bien plus profondes qui ne sont pas entièrement expliqués par la libéralité promue par l'aumône. Il faut se rappeler que dans plusieurs cas la donation-vente se produit parce que le vendeur, décidé à vendre, considère l'institution religieuse comme un acheteur prioritaire. Cette relation peut aller même au-delà d'une relation de propriété ou de copropriété dans le sens économique, et il est assez commun parmi les familles donatrices que certains individus renoncent à la vie séculière et soient reconnus comme membres des institutions religieuses, ce qui pouvait aussi donner lieu à une donation de soutien. [294] La relation privilégiée établie entre ces familles et le centre religieux est à l'origine d'une dynamique mutuellement bénéfique et de proximité entretenue par les générations suivantes.

La dimension sociale des chartes ne se limite aux proches. On voit régulièrement apparaître d'autres personnes appartenant à des familles de la région avec qui les disposants ont des alliances préexistantes. Souscrire un acte de donation en tant que témoin valide et renforce une relation sociale et familiale entre le signataire et le donateur, et traduit habituellement un lien plus approfondi dans d'autres sphères de la vie, notamment économique, au sein de l'espace régional [295]. Le témoignage ou la *laudatio* d'une action juridique accomplie peut aussi révéler le désir de donner plus de légitimité au document de la part du donateur, faisant participer dans l'acte

294. Joachim Wollasch. "Parenté noble et monachisme réformateur. Observations sur les «conversions» à la vie monastique aux XI e et XII e siècles". In : *Revue historique* 264.Fasc. 1 (535 (1980), p. 3-24, p. 3-24.

295. Voir à ce sujet : Martin Aurell. "La parenté en l'an mil". In : *Cahiers de civilisation médiévale* vol. 43 (2000), p. 125-142, p. 125-142 ;

différentes figures d'autorité ou de pouvoir. De manière similaire, les échanges fonciers entre personnes privées, sous la forme de vente, location ou comme moyen de mettre fin à des litiges, sont souvent précédés par des relations de contact social. Ainsi, par le transfert effectif de terres et de biens, la donation accroît et renforce le lien social et territorial entre certains groupes. Elle constitue une réponse aux initiatives de la doctrine de l'aumône, mais aussi une réponse à la nécessité de légitimer des relations sociales à travers un document écrit et d'exprimer ces liens à l'occasion d'un événement médiatisé par l'institution ecclésiale [296].

Le rôle joué par le nom personnel va donc au-delà de la simple déclinaison de l'identité personnelle et se présente comme un nœud dans un large réseau à partir duquel les effets d'une donation ou d'une donation-vente, passées à titre personnel, peut se ramifier jusqu'à atteindre, en amont tous les membres d'une famille, et en aval, un ensemble de familles liées à la transaction. La connexion ainsi établie entre deux familles est loin d'être le fruit du hasard puisque qu'elle correspond à l'expression de la structure lignagère de la société féodale. Elle se démultiplie tout au long d'un cartulaire, permettant de déterminer les différentes ramifications dans la structure sociale, ce qui facilite la diffusion des chronologies entre groupes de personnes identifiés.

Si les entités personnelles présentent un individu qui n'agit pas individuellement dans un document, les entités géographiques ne se comportent pas très différemment puisque nous sommes devant une structure foncière qui prend la *villa* comme l'unité fondamentale de distribution territoriale pendant les Xe et XIe siècle et qui constitue effectivement une cellule répartie entre plusieurs propriétaires et seigneurs, y compris les établissements ecclésiastiques. Elle-même se trouve insérée dans différentes mailles de distribution spatiale du paysage humanisé, le trinôme *pagus-ager-villa* jusqu'à la moitié du XIe siècle dans la Bourgogne et le tandem *paroisse-ecclesia* lors de la cristallisation des ressorts paroissiaux tout au long du XIIe siècle. Les différentes vagues de transfert de propriété mettent en marche un bouleversement spatial à l'intérieur des *villae* lorsque la terre et les biens dépendant d'elle commencent à être mobilisés, à certains moments massivement, à l'occasion par exemple de l'apparition d'un nouvel établissement religieux, des mains des laïcs à celles de l'institution ecclésiastique [297]. Il est assez courant de trouver des manses, des vignes, des champs, et même des églises dont la propriété est partagée entre plusieurs membres d'une même famille ou de plusieurs familles ; leur donation occasionne parfois des litiges à cause du bornage, du déguerpissement et de l'occupation abusive ou de l'opposition d'un membre à la donation. En certaines occasions l'institution religieuse est la seule propriétaire de toute une *villa* qui était à l'origine morcelée dont la propriété complète a duré des décennies avant d'être transférée aux mains des ecclésiastiques. [298]

Ainsi, il est évident devant ce panorama qu'existent différents types de relations

296. Voir à ce sujet : Rosenwein, *To be the neighbor of Saint Peter : the social meaning of Cluny's property, 909-1049*, p. 50-77 ; Georges Duby. *Qu'est-ce que la société féodale ?* Flammarion, 2002, p. 1459-1465, p. 1459-1465

297. Voir à ce sujet : Bange, "L'ager et la villa : structures du paysage et du peuplement dans la région mâconnaise à la fin du Haut Moyen Age (IX e-XI e siècles)" ; Lauwers et Ripart, "Représentation et gestion de l'espace dans l'Occident médiéval"

298. Rosenwein, *To be the neighbor of Saint Peter : the social meaning of Cluny's property, 909-1049*, p. 78-108.

privilégiées entre les entités qui peuvent nous aider à mieux opérer le transfert de la date. Des groupes relativement restreints de personnes sont associés de manière à peu près stable – sans que cette relation soit exclusive. Ces liens associatifs commencent dans la famille et s'étendent à des cercles sociaux régionaux, par des relations de parenté biologique ou artificielle. Face à cette situation, un premier niveau de structuration, comme celui fourni par la récupération automatique des entités nommées, est nettement insuffisant pour aborder l'extraction d'informations chronologiques utiles pour l'historien, et il est nécessaire de valider scientifiquement le résultat obtenu grâce à une méthode numérique. Ce premier résultat qui se présente sous forme d'une liste d'entités nommées doit être utilisé pour détecter les structures sous-jacentes qui déterminent les relations entre les entités, c'est-à-dire les réseaux sociaux et familiaux. L'étude des entités nommées présentes dans un cartulaire pourrait alors permettre une reconstruction au moins partielle de la chronologie d'apparition des membres de l'aristocratie participant aux affaires de l'abbaye ou de toute autre institution rédactrice d'un cartulaire.

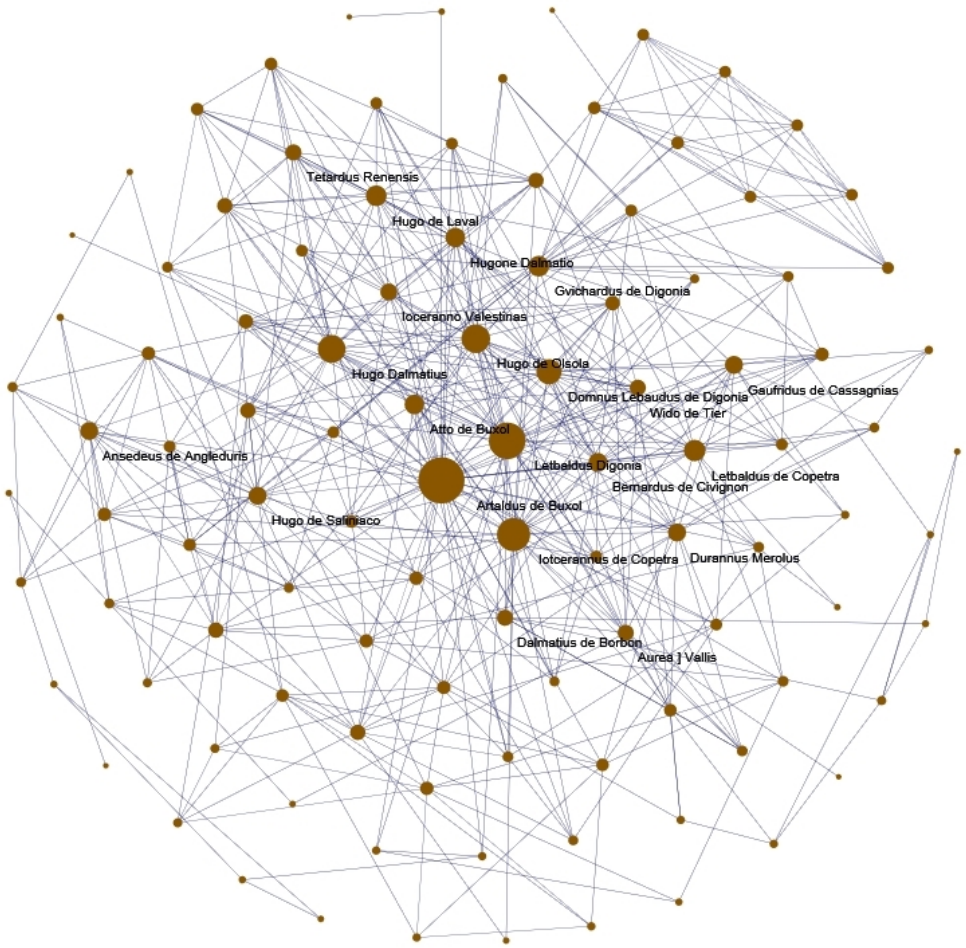

FIGURE 4.2 – Réseau social établi entre les différents chefs de familles représentés dans les actes du cartulaire de Paray-le-Monial.

Cela étant dit, nous devons ajouter au travail de détection et de classification des entités nommées la récupération des liens existants entre les membres d'une même famille et entre les différentes familles liées entre elles, puisque la datation chronologique et la configuration des réseaux spatiaux et sociaux sont, comme on l'a vu, intensément interdépendantes. Ces deux tâches ont été abordées dans la littérature concernant le traitement automatique des langues dont nous avons déjà parlé (partie 1.6) : la désambiguation des entités nommées et l'extraction des relations. L'objectif est ici double : bien identifier l'entité récupérée afin de restituer au maximum sa valeur comme *désignateur rigide*, et présenter l'ensemble des relations qu'elle a établi avec des entités contextuellement proches. Ces deux tâches sont particulièrement compliquées dans le cas des chartes médiévales, à cause des états de langue, de la très fréquente homonymie et, plus généralement de la sècheresse et du caractère lacunaire des sources. À cela s'ajoute un dernier problème : le manque de sources externes contenant de l'information vérifiée à laquelle comparer les listes des noms. Ici les seules sources disponibles pour comparaison sont les chartes elles-mêmes ou celles d'autres cartulaires proches chronologiquement et spatialement, présentant donc les mêmes limites.

Dans la partie qui suit cette première étude, que nous considérons comme la partie applicative de la thèse, nous allons détailler comment nous avons abordé ces deux questions en nous appuyant sur notre modèle de récupération d'entités nommées. Ensuite, nous étudierons quelles solutions techniques on peut envisager pour multiplier la récupération de l'information à partir des données initialement disponibles dans une logique qui dépasse la traditionnelle datation bijective des actes.

4.6 La datation assistée par ordinateur

4.6.1 Matrices chronologiques et *pipeline*

Nous venons d'expliquer, en faisant appel aux implications sociales sous-jacentes à rédaction des actions juridiques, les raisons pour lesquelles une datation par croisement de personnes, et dans une certaine mesure par croisement d'événements, doit être considérée comme une méthode privilégiée pour la datation assistée des chartes médiévales. Dans ce chapitre, nous nous proposons d'expliquer d'un côté les dispositions techniques nécessaires pour mener à bien ce processus en tirant parti des outils d'automatisation, et de l'autre d'exposer l'heuristique proposée pour surmonter les principaux problèmes que présente cette approche.

Dans ce récit, on va suivre une méthode, dite de pipeline, en cinq étapes qui correspondent à différents niveaux de disponibilité de l'information :

1. L'extraction, l'indexation et la canonisation des entités nommées, à partir du calcul de similarité entre elles, afin de configurer des dictionnaires de formes orthographiques de chaque personne et lieu ;
2. L'extraction et la classification des cooccurrences des entités afin de déterminer, d'un côté les titres, qualités, fonctions des personnes et, de l'autre, les liens de parenté entre entités ;
3. La récupération d'une première couche chronologique à partir des personnes les

plus connues détectées dans les chartes et qui sont citées dans d'autres sources : bases de données, listes généalogiques et nécrologiques ou autres cartulaires de la Bourgogne, notamment celui de Cluny ;

4. Une deuxième couche chronologique sera obtenue à partir des groupes de documents classés chronologiquement au cours de l'étape antérieure ; autrement dit, nous transférerons des données depuis le groupe daté vers le groupe non daté ;
5. Si les deux étapes précédentes sont soutenues par la recherche à partir du nom double, un troisième niveau d'information sera proposé à partir de l'exploitation du nom simple présents dans les réseaux familiaux, c'est-à-dire le cadre généalogique que nous avons reconstitué, dans la mesure du possible, au cours de l'étape précédente.

4.6.2 Les index personarum et locorum

Les deux premières étapes correspondent en somme à l'extraction et la formalisation des *index personarum et locorum* qui a constitué traditionnellement un travail fastidieux et de longue haleine pour les éditeurs de chartes. La localisation documentaire des personnes, lieux et biens (moulins, forges, bâtiments, etc.) était effectivement un travail qui nécessitait un dépouillement soigneux des textes par l'éditeur. De ce fait, ces index n'ont pas toujours été réalisés. Ils figurent souvent dans les éditions de petits cartulaires, mais le travail est resté inachevé pour le vaste corpus clunisien par exemple. Ce travail chronophage est aujourd'hui automatisable, comme nous allons le montrer. On peut considérer qu'il s'agit d'une tâche presque résolue sous réserve de deux conditions imposées au chercheur :

(i) L'extraction soigneuse de toutes les variantes orthographiques tant pour les noms de personnes que pour les noms géographiques, ce qui permet de relier toutes les formes attestées sous une seule forme canonique et de récupérer avec certitude toutes les relations établies entre deux acteurs et toutes les modifications subies par un espace ;

(ii) La détermination des cooccurrences à partir d'une extraction par fenêtres de mots, autrement dit des chaines plus longues que l'entité en soi, ce qui est essentiel afin de déterminer les titres, qualités, fonctions, charges, surnoms et appellations autres que porte un personnage ou un lieu. Cette information peut se révéler fondamentale au moment d'identifier un personnage et de déterminer sa chronologie étant donné le très haut niveau d'homonymie dans la dénomination des personnes, spécialement dans les chartes antérieures au XIe siècle.

Comme nous l'avons expliqué précédemment, il est indispensable, avant d'entamer l'extraction de couches chronologiques, d'effectuer la reconstruction de quelques arbres chrono-généalogiques des familles aristocratiques qui concentrent la plupart des relations sociales et de pouvoir dans une région. Déterminer les trois relations les plus communes exprimées dans les actes (mariage, lien de paternité et lien de fraternité) est fondamental tant pour déterminer une chronologie que pour établir un cadre familial de référence. Cela s'avère important parce que dans le cas des notices ou dans les souscriptions des chartes, les noms de familles peuvent être supprimés

et la seule manière d'identifier un personnage repose sur l'identification de noms de baptême attribués aux membres d'une certaine famille.

En théorie, le travail fait au cours de ces deux étapes pourrait nous montrer rapidement les traces générales d'une dynamique à l'échelle sociale et économique, puisqu'on a récupéré un réseau d'entités nommées, tant de noms de personnes que de lieux, bien connectées entre elles à partir des relations exprimées dans les actes. Une esquisse de ce réseau (figure 4.2) peut à cet état du travail nous renvoyer une image relativement proche de la réalité. Mais cela demeure encore une image incomplète que l'on peut affiner davantage une fois réglées les deux questions fondamentales qu'on a énoncées auparavant : la désambigüisation des entités et la détermination de la chronologie des actions juridiques.

4.6.3 Les couches chronologiques

Pour déterminer la chronologie à partir de l'extraction de dates disponibles dans d'autres sources d'information incluant les personnes du cartulaire de Paray, nous allons mobiliser en premier lieu des généalogies et des nécrologies pour les personnages les plus connus ou portant un titre. En deuxième lieu, nous utiliserons les chartes datées des autres cartulaires clunisiens, puisqu'un nombre important de personnes apparaissent dans plus d'un cartulaire. Étant donné qu'il s'agit d'un processus de retour d'information (*feedback*), les propres chartes de Paray, au fur et au mesure qu'elles sont datées, peuvent être utilisées pour en dater d'autres.

L'érudition se sert de trois éléments textuels pour dater un document : la présence de personnages datables, les changements dans le vocabulaire et l'évolution des formules discursives. Chaque recueil de chartes présente un état particulier dans chacun de ces trois domaines, ce qui facilite ou complique la datation de ces documents. Dans le cas de Paray, un examen rapide nous prévient de ces situations :

Nous détectons autour de 620 différentes personnes présentes dans les documents, tant comme donateurs ou vendeurs, que comme témoins et souscripteurs. Le nombre de témoins est bien plus élevé dans le cas des donations faites par des personnages importants. Nous sommes contraints à l'approximation concernant le nombre de personnes car dans certains cas il très difficile de distinguer deux personnes dont l'orthographe est très proche, mais la chronologie décalée, comme par exemple :

Hugo de Saligiaco (CBMA 5052) / *Hugonis de Saleniaco* (CBMA 7095)
Hugo Scabellus (CBMA 7109) / *Hugues de Scamellis* (CBMA 5302)

Le nombre de personnes entrant dans le calcul général semble élevé si on considère qu'il s'agit d'un cartulaire de 200 actes concernant un espace restreint et dans un arc chronologique limité à un demi-siècle pour la plupart des actes. Ceci est en partie généré par les chartes tardives (des XIIe-XIIIe siècles) qui comportent un nombre substantiellement plus élevé de noms de personnes. Dans les chartes du XIe siècle, on compte en moyenne 6 personnes présentes dans un acte. D'ailleurs, même si le nombre total de personnes est élevé, le nombre de celles qui jouent le rôle de concentrateurs, autrement dit, de « nœuds » dans le réseau documentaire est d'environ 28. Ces personnes apparaissent chacune dans au moins 5 actes et cumulent à elles seules plus

de la moitié du total des connections détectées entre les actes. Cette trentaine de personnages, comme peut voir dans la figure 4.2, appartiennent principalement à 10 familles : *Buxol, Copetra, Bourbon-Lancy, Digonia, Olsola, Angleduris, Centarberent, Cavazola* et *Saliniaco*. Leurs donations à Paray sont observées depuis la donation du monastère à Cluny par Hugues, comte de Chalon et évêque d'Auxerre (999), et connaissent un deuxième moment de plénitude précisément du temps du prieur Hugues et de la rédaction du cartulaire.

Pour la quasi-totalité de ce groupe, on ne connaît pas de dates biographiques. Celles que l'on connaît viennent normalement de quelques chartes datées du cartulaire clunisien où ils apparaissent parfois faisant donation ou jouant le rôle de témoin. Le seul cadre chronologique qu'on peut établir est de noter leur première et leur dernière apparition documentaire par rapport aux autres personnes. Mais comme on le verra plus loin, la présence de ce groupe concentrateur facilite la datation, parce qu'ils vont être associés, par les chartes, à une date ou série de dates. Il suffit de la présence d'un seul personnage datable souscrivant un document avec eux ou l'attestation de leur présence dans un document daté d'un autre cartulaire pour commencer à resserrer l'intervalle chronologique les concernant. De ce fait, si on arrive à les associer à un jeu de dates, même si elles sont approximatives, on peut les combiner pour écarter des chronologies impossibles et en valider d'autres qui sont quant à elles plausibles.

La difficulté rencontrée dans cette approche est la coexistence de deux ou plus générations de personnes dans les actes datés dans une fourchette. Comme on l'a déjà mentionné, la datation par l'abbé Hugues dans une fourchette de 60 ans nous met face à un ensemble de chartes dans lesquelles nous pouvons trouver jusqu'à trois générations de souscripteurs. Puisqu'on ne connait pas de dates précises, il est recommandé d'estimer la longévité des personnes, aussi approximative qu'elle puisse l'être. Les cas de l'abbé Hugues, qui a vécu 85 ans, et de l'abbé Odilon, 88 ans, sont vraiment exceptionnels. Les sources dont nous disposons ne nous autorisent pas à établir de mesures de longévité, mais si l'on se réfère aux quelques chiffres estimés par des études de paléo démographie en Bourgogne, l'espérance de vie à 25 ans (et non à la naissance en raison de l'importante mortalité infantile) parmi les aristocrates et les grands personnages peut certes dépasser les 65 ans, mais semble se situer autour de 60 ans de moyenne. Les calculs basés sur les documents fiscaux de la Bourgogne pour les XIIIe et XIVe siècles, parlent aussi d'un renouvellement total de la population tous les 25 ans [299]. Ainsi, sauf en cas de mort brutale comme celles des comtes Lambert ou Hugues II en Espagne, nous retenons le chiffre de 60 ans comme marqueur chronologique pour chaque personne, lorsque on n'est pas arrivé à documenter d'autres marques chronologiques.

Dater l'apparition et la disparition d'un personnage est souvent rendu plus difficile par l'homonymie. Pour cet exercice nous avons privilégié les personnages

299. Il y a très peu d'études dans ce domaine : Henri Dubois. "Population et fiscalité en Bourgogne à la fin du Moyen Âge". In : *Comptes rendus des séances de l'Académie des Inscriptions et Belles-Lettres* 128.4 (1984), p. 540-555 ; Robert Fossier. "La démographie médiévale : problèmes de méthode (Xe - XIIIe siècles)". In : *Annales de démographie historique*. JSTOR. 1975, p. 143-165 ; Robert Fossier. "Peuplement de la France du nord entre le Xe et le XVIE siècles". In : *Annales de démographie historique*. JSTOR. 1979, p. 59-99 ; J Houidaille. "Mortalité masculine dans les familles regnantes au Moyen Age". In : *Population* 27 (1972), p. 1131-1134

dénommés avec plus d'un élément, puisque travailler autour des dénominations simples, habituellement le nom de baptême, offre des résultats peu fiables. Afin de peaufiner les résultats chronologiques, il est possible de combiner les noms simples après l'extraction de la relation à laquelle ils sont attachés – normalement une filiation directe – et d'essayer de les retrouver associés dans d'autres chartes. Mais il s'agit d'un exercice complexe et peu rentable. En revanche, la récupération des noms doubles limite beaucoup le problème de l'homonymie étant donné la spécificité de la deuxième partie du nom liée à une famille, qu'il s'agisse d'un locatif d'origine ou d'appartenance ou d'un patronyme. Néanmoins, la transmission du nom et sa répétition intergénérationnelle au sein d'un groupe très réduit peut se révéler un problème majeur, spécialement dans l'analyse des groupes familiaux, parce que l'on va trouver des pères, fils, neveux, et grands-pères qui portent le même nom de baptême. Dans plusieurs cas, on ne peut affirmer avec certitude si le souscripteur est le père ou le fils, et s'ils souscrivent un document avec un troisième personnage, cela introduit du « bruit » dans les marqueurs chronologiques. Dans de rares cas, le rédacteur de la charte lui-même a surnommé un personnage *senex* ou *iuvenis* afin de les distinguer, et il arrive parfois que la filiation soit explicitée, mais on ne peut pas toujours récupérer automatiquement cette information.

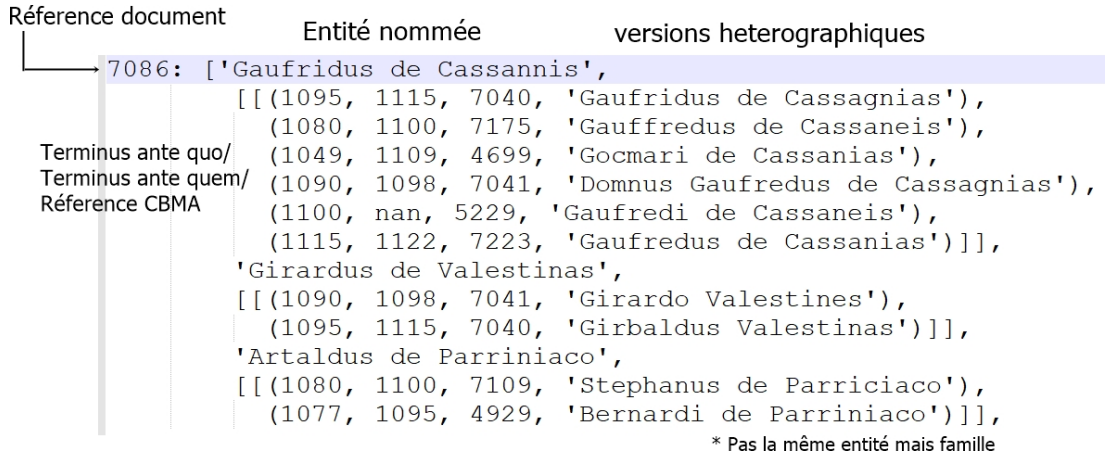

FIGURE 4.3 – Exemple de la collection de versions des entités nommées dans chaque document sous la forme d'un dictionnaire. (Dans ce cas CBMA 7086)

4.6.4 Distance de Levenshtein et Wikification

Pour l'étude de ce cartulaire en particulier, nous considérons l'étude automatisée de l'évolution du vocabulaire des chartes concernant les qualifications sociales et spatiales, et l'étude de l'évolution des parties du discours diplomatique, comme des données secondaires ; ceci pour plusieurs raisons. Tout d'abord, les deux-tiers des chartes recueillies sont concentrées chronologiquement sur une période relativement courte, autour de 60 ans (entre les décennies 1070 et 1120), qui ne présentent pas de très grandes évolutions scripturaires. Ensuite, le cartulaire de Paray n'est pas suffisamment important numériquement pour révéler clairement des changements qui puissent nous fournir des marqueurs chronologiques plus précis. Dans le cartulaire clunisien, où pour

la même période il y a plus de 600 chartes, nous pouvons les caractériser avec une plus grande précision à partir du vocabulaire utilisé. Cela ne veut dire pas qu'il n'existe pas de changements dans le vocabulaire des chartes entre les différentes périodes distinguées au sein du cartulaire de Paray – de fait, nous avons constitué une liste de 22 termes qui présentent des variations significatives –, mais ils apportent une information chronologique qui, dans ce cas, est très générale. C'est le cas par exemple de *miles, dominus, uxor, werpitio, oporteo, concedo, debeo, offero, laudo, elemosyna*, etc. Présents régulièrement dans les actes de la fin du XIe siècle, mais qui disparaissent presque entièrement dans les actes des périodes des abbés Pons et Pierre. De même, le vocabulaire spatial change avec la transformation de l'*ager* vers la première moitié du XIe siècle et la disparition progressive du *pagus* qu'on trouve rarement dans les chartes du XIIe siècle où il a déjà été remplacé par le diocèse et la paroisse, ce qui va de pair avec la disparition d'autres usages anciens de détermination spatiale qui passent par l'usage de *domus, pratum, mansus, pascus, incultus, silva, locus, dicitur*, etc.

Les variations orthographiques dans les noms peuvent-elles être prises comme critère de datation ? Ce n'est habituellement pas le cas. Il est vrai qu'il existe une évolution dans la façon d'écrire les noms de famille dont la raison tient souvent à la francisation des formes latines et à l'absence de règles d'écriture. En fait, il est commun de trouver deux ou trois versions graphiques d'un nom dans un même document. Par exemple le nom de famille *Buxolio* peut apparaître écrit comme *Buxol, Buxel, Busul, Buxelio* dans les chartes de Paray et de Cluny pendant le XIe siècle. En revanche les versions de ce même nom de famille observées dans les chartes de la fin du XIIe et XIIIe siècle : *Buziaco, Bussuil, Busy* semblent être plus en accord avec un vrai changement dans la manière de prononcer tant le nom de famille que la localité de provenance, et dans la manière de traduire l'oralité en signes graphiques. Mais il serait fallacieux de tirer des données chronologiques à partir de ces changements. En tout cas, il s'agit d'un élément qui n'apporte aucune utilité dans l'espace chronologique qui concerne le cartulaire de Paray. Ce phénomène d'hétérographie, loin d'être une aide, représente l'un des principaux défis lors de l'utilisation des listes d'entités nommées, puisqu'il exige de réunir sous une même entité toutes les variations graphiques trouvées. Dans le cas contraire, on risque d'altérer la réalité présentée par les chartes. Il faut ajouter aux variations graphiques d'ordre scripturaire celles d'ordre linguistique, c'est-à-dire les variations dues à la déclinaison latine et celles dues à la vernacularisation – peu usitée au XIe siècle – des versions latines des noms : *Albus* par *Blancus*, *Chastel* par *Castello*, *Rufus* par *Rubeus*, *Sinemuro* par *Setmur*, *Chauve* par *Calvus*, etc., qui peuvent apparaitre de manière contemporaine sous l'une ou l'autre forme. Ce travail de « canonisation » doit faire appel à des mesures vectorielles de similarité entre chaînes de mots, afin de réaliser des comparaisons massives entre toutes les entités récupérées de Paray, et celles d'autres cartulaires.

En deuxième lieu, nous sommes confrontés à un cartulaire composé en grande partie par des notices qui sont 9 fois plus nombreuses que les chartes et documents d'origine publique. La notice, comme on l'a précédemment observé (partie 2.1.4), présente un style objectif et une rédaction plus libre par rapport au carcan juridique mobilisé dans la charte ; elle ne suit pas un formalisme dans sa construction que nous pourrions situer dans une chronologie. Comme dans le cas de l'évolution du vocabulaire, certaines

parties des chartes, notamment les protocoles, présentent une série importante de changements, spécifiquement entre les documents de la fin du XIe siècle et ceux qui sont postérieurs, mais leur nombre n'est pas suffisamment élevé pour proposer un classement solide. Les notices offrent peu d'aide dans ce cas compte tenu du fait qu'elles prennent une forme diplomatique très objective et qu'elles se dispensent de la plupart des formulations d'une charte. Le laconisme d'un nombre important de notices est vraiment sévère, et dans certains cas il s'agit d'une simple note qui a pour seuls témoins le scripteur et le donateur [300]. Ce laconisme est presque une marque d'identité dans notre cartulaire et le travail de réduction des chartes du compilateur peut être clairement observé quand une même action juridique a produit des documents pour Cluny ou Marcigny et Paray [301]. L'absence d'éléments de contextualisation et l'omission d'un nombre important d'éléments de formalisation diplomatique rend inaccessible dans ce cas la chronologie par le biais de la caractérisation formulaire et discursive.

Le modèle de détection automatique des entités nommées que nous avons développé extrait l'ensemble des entités du cartulaire de Paray en quelques minutes. Le travail qui s'ensuit peut en revanche prendre bien plus de temps. Conformément à nos résultats dans l'évaluation du modèle (partie 3.4), nous pouvons attendre dans les chartes du cartulaire de l'abbaye de Cluny une performance supérieure à 95 % dans la reconnaissance des noms de personnes et de 92 % pour les lieux. L'évaluation faite sur les chartes européennes montre une performance légèrement plus basse pour les XIe et XIIe siècles, période qui nous intéresse particulièrement ici : autour de 90 % pour les personnes, et 86 % pour les lieux. Nous pouvons attendre ici, étant donné que nous n'avons pas annoté un jeu de test sur Paray, une performance entre ces deux chiffres et en tout cas plus proche du résultat observé pour Cluny puisqu'il s'agit d'un cartulaire qui en théorie suit le modèle clunisien et se trouve sous sa sphère d'influence.

Il faut alors planifier le processus de désambiguation des entités en commençant par sa classification canonique en appliquant la distance de Levenshtein. Celle-ci est très utilisée dans le domaine du *data mining* et en général dans la fouille de textes, et permet de calculer la similarité entre deux chaînes de caractères. Pour deux chaînes de caractères, l'algorithme émet une valeur qui exprime le coût que représente la transformation d'une chaîne en une autre chaîne, selon le type de changements nécessaires (suppression, insertion et remplacement). *La mesure de Levenshtein* est très pertinente pour calculer la ressemblance entre petites chaînes de caractères comme c'est le cas des entités nommées avec une portée minimale de 2 mots et maximale de 5 mots (autour de 20-30 caractères). Nous avons calculé qu'une ressemblance de plus de 75 points (la ressemblance s'exprime de 0 à 100) est normalement suffisante pour considérer qu'il s'agit de la même entité nommée. Dans de rares cas, deux entités qui semblent identiques pour l'œil expert peuvent obtenir un résultat plus bas et, *a contrario*, d'autres manifestement différentes peuvent présenter une ressemblance vectorielle très haute, par exemple :

Vldricus Esperon / Heldricus Hesperuns
Guillaume de Centoarbenz / Wilelmus de Centa(r)bent

300. CBMA 7055, 7062, 7105, 7124, etc.
301. CBMA 7236 et CBMA 11403 ; CBMA 7232 et CBMA 11240

Letbaldus de Chopetra / Letbaldus Calvus de Coperia
Ylius de Chavasiget / Ylionis Jhavagist.

L'homonymie et les changements dus à la vernacularisation sous-tendent ces phénomènes qui demeurent numériquement une exception. Hormis cela, une ressemblance supérieure à 75 points nous assure une très bonne classification des entités nommées pour des individus. Cette collection de versions d'une entité sous la forme d'un dictionnaire (voir figure 4.5) contient pour chaque personnage et chaque lieu une clé – forme canonique – et une série de valeurs – formes non canoniques –, de telle sorte qu'on peut accéder à l'entité nommée soit interrogeant la forme la plus commune soit l'ensemble des formes hétérogènes. D'ailleurs, chaque entité est associée à une deuxième collection de cooccurrences selon une liste limitée que nous avons fournie (voir figure 4.4), regroupant des mots qui agissent comme classificateurs sociaux et territoriaux : titres nobiliaires, titres de traitement, qualités ecclésiastiques, offices, pour les noms de personne ; circonscriptions, divisions administratives ou religieuse, lieux-dits pour les noms géographiques. Si l'un de ces mots se trouve associé directement à un nom de personne ou de lieu, on le récupère et on le stocke dans ce deuxième dictionnaire. Il est commun qu'un même personnage cumule plus d'un titre ou qualité (co-occurrences) depuis les différentes catégories : par exemple *dominus comes, miles* ; dans d'autres cas on trouve un personnage dans différents rôles, ce qui peut impliquer des différentes chronologies : par exemple *presbiter, praepositus, sacerdos, camerarius.* Il est bien plus rare de trouver des personnages accumulant des titres prestigieux, tel le plus fameux, Hugues Ier de Chalon qui fut comte et évêque, ou Hugues de Buxolio : *dominus, miles, prior.*

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 religieuses : *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 surnommes : *benedictus, brunus, cantor, grossus, humilis, largus, normandus, paganus, servus, venerabilis*	**Division légale et juridictionnelle :** *areae, castrum, civitas, dioecesis, dominus, ecclesia, provincia, sedes, terra, villa*
Liens de périphrases : *appelatus, cognomen, dictus, nomen, vocatus*	**Micro-espaces :** *altar, atrium, capella, capitulum, castellum, cenobium, domus, ecclesia, hospital, monasterius*
Mots de valeur nominal : *alius, ego, filius, frater, idem, nepos, signum (S), uxor*	**Termes locatives, prépositions, adverbes :** *ad, apud, dicitur, fines, inter, manus, meridies, parte, pro, supra, vocabulum*

FIGURE 4.4 – Liste de cooccurrences pour les entités nommées personnelles et géographiques.

À partir de ce premier ensemble d'informations structurées constitué des listes d'entités nommées et de leurs cooccurrences, nous allons essayer de restituer une

première couche chronologique en les connectant avec d'autres listes d'entités extraites de sources associées au cartulaire de Paray, concrètement :

— La *DBpedia*, base de données qui garde le contenu structuré de la Wikipedia et dans laquelle nous nous attendons à retrouver rapidement les dates des personnages les plus connus qui apparaissent à Paray.
— Le cluster documentaire formé par quatre cartulaires du groupe clunisien contenant des actes datés de la période qui nous intéresse le plus, entre la fin du Xe et la fin du XIIe : les cartulaires de l'abbaye de Cluny, le cartulaire de Marcigny-sur-Loire, le recueil de pancartes de la Ferté-sur-Grosne et le cartulaire de Saint-Vincent-de-Mâcon. Étant donné la proximité géographique et intellectuelle entre les quatre lieux, on constate une forte densité de personnages communs.
— Le cluster documentaire formé par trois documents numérisés indiqués par la bibliographie autour de Paray : la nécrologie de Marcigny-sur-Loire,[302] certaines pages du tome 14 de la *Gallia Christiana*[303] et le dictionnaire topographique de Saône-et-Loire.[304] L'essentiel est ici de profiter d'index et de listes semi-structurées présentant la hiérarchie ecclésiastique locale et les noms anciens récupérés pour les lieux.

wikification

Dans la littérature on appelle wikification le processus qui consiste à connecter une entité trouvée dans un texte avec un article de Wikipédia qui lui est dédié ou, à défaut, qui la mentionne. Ce processus de désambiguïsation se réalise en cherchant le meilleur candidat pour une entité déterminée dans la base de données DBpedia (voir partie 1.6). La DBpedia est une ressource qui doit être interrogée à partir de requêtes écrites en SPARQL à travers d'un point de terminaison[305] pour lequel nous avons écrit un script en Python[306]. Le défi principal dans ce cas n'est pas technique, mais réside plutôt dans la « francisation » correcte des noms médiévaux, les articles sur un personnage étant

302. François Cucherat. *Cluny au onzième siècle : son influence religieuse, intellectuelle et politique*. Dejussieu, 1873.
303. Denis de Sainte-Marthe et Barthélemy Hauréau. *Gallia Christiana in provincias ecclesiasticas distributa*. T. 14. Coignard, 1751.
304. Jean Rigault. *Dictionnaire topographique du département de Saône-et-Loire : comprenant les noms de lieux anciens et modernes*. T. 38. Comité des travaux historiques et scientifiques-CTHS, 2008.
305. SPARQL est un acronyme pour Protocol and RDF Query Language. Il s'agit d'un langage de requête qui permet d'interroger des bases de données qui gardent de contenu sémantique sous le format RDF (Resource Description Framework). RDF est un format de triplètes qui facilite le traitement des descriptions de données les exprimant à partir de triplets, qui correspondent normalement au sujet (le ressource), objet (une autre ressource) et prédicat (une propriété de relation entre les deux ressources).
306. Sur les méthodes de wikification en texte, regarder : (Zhiyuan Cai et al. "Wikification via link co-occurrence". In : *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM. 2013, p. 1087-1096) et (Johannes Hoffart et al. "Robust disambiguation of named entities in text". In : *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2011, p. 782-792)

régis par la version modernisée du nom : *Stephanus*/Étienne, *Gaufredus*/Geoffrey, *Radulfus*/Raoul, *Tetbaldus*/Thibaud, etc. Il n'y a, à notre connaissance, pas de liste complète à laquelle faire appel automatiquement, mais il s'agit d'un travail de recompilation qui ne présente pas de grande difficulté si l'on se sert des travaux déjà classiques sur de l'origine médiévale de l'anthroponomie moderne. De plus, la liste est forcément restreinte puisque les familles reprennent les noms de baptême des ancêtres. La DBpedia nous fournit rapidement un premier groupe de personnages avec leurs chronologies (dates de naissance, mort, événements importants). Il s'agit notamment de membres de la noblesse et de la hiérarchie ecclésiastique. Parmi les plus notables :

- Les rois de France - Les comtes de Chalon - Les comtes de Mâcon - Les ducs de Bourgogne - Évêques d'Autun - Évêques de Mâcon - Évêques d'Auxerre - Évêques de Clermont - Les abbés de Cluny - Les seigneurs de Semur-en-Brionnais. - Les ducs de Lorraine - Les seigneurs de Bourbon-Lancy.

Cela permet de constituer rapidement un groupe de documents, à l'intérieur du cartulaire, bornés chronologiquement par les dates fournies par l'encyclopédie. Cette première couche chronologique va nous servir de base logique pour aborder notre deuxième base de données : les documents de Cluny. Il faut, en tout cas, avant cela, relever un détail important en ce qui concerne la DBpedia. Dans certains cas, nous pouvons être confrontés à un biais, parce que les dates de certains personnages, spécialement pour le bas Moyen Âge, proviennent précisément des documents que nous essayons d'interroger et reprennent des dates incertaines ou douteuses pour l'historiographie. La DBpedia constitue dans ce cas une forme déjà structurée de nos documents. Dans notre cas un seul personnage se trouve dans cette situation : Geoffroi Ier Grisegonelle (940 - 987) dont certains historiens du XIXe siècle prétendent qu'il a été le deuxième époux d'Adélaïs après la mort du comte Lambert. La Wikipedia reproduit cette information alors qu'elle est niée par le cartulaire de Paray et par l'historiographie moderne. Il faut en tout cas prêter attention aux dates présentées comme douteuses ou peu renseignées.

Le cluster clunisien

L'ensemble des documents clunisiens constitue notre principale source de datation. Comme on l'a déjà mentionné, la proximité géographique entre les différents prieurés et l'abbaye mère, la normalisation de l'exercice conjoint de droits sur la terre et en général les liens étendus de parenté et de consanguinité dans la région facilitent l'apparition des membres de certaines familles dans plusieurs cartulaires. L'absence presque totale de datation dans les actes du cartulaire de Paray n'est pas une exception dans les cartulaires de la Bourgogne, nous trouvons aussi des larges séries d'actes non datés dans le cartulaire de Saint-Vincent, qui présente de ce fait de sérieux problèmes de chronologie dans son édition moderne. Le cartulaire de Cluny se trouve à mi-chemin, avec un nombre important de documents portant une date, d'autres portant une fourchette de dates restituée par les éditeurs modernes et une quantité relativement faible de documents sans indication chronologique. Le cartulaire de Marcigny peut, quant à lui, être considéré comme privilégié car il offre un paysage chronologique assez complet avec une grande partie de documents datés et une autre datée dans des

fourchettes serrées, fruit du travail soigné de Jean Richard. [307]

Étant donné que la quasi-totalité du cartulaire de Cluny a été annotée à la main dans le cadre de ce travail, nous n'avons plus besoin d'en extraire les entités. Pour les autres – Paray inclus – nous avons a appliqué le modèle développé qui nous fournit un jeu annoté automatiquement par la machine. Sur les entités extraites, nous appliquons les mêmes protocoles de formalisation que ceux mentionnés précédemment pour Paray, ce qui donne comme résultat des dictionnaires de formes pour les entités que nous pouvons comparer afin de détecter les coïncidences. Les coïncidences détectées peuvent nous fournir des dates de trois manières (le processus est résumé dans la figure 4.5) :

1. nous détectons un personnage dans un document daté précisément et nous associons ce personnage à cette date qui correspond à un événement de sa vie ; il s'agit de la méthode que nous privilégions ;
2. nous détectons un personnage dans un document daté dans une fourchette, ce qui a moins de valeur, spécialement s'il s'agit d'une fourchette de plus de 40 ans, mais qui est très utile pour attester du point de vue documentaire l'existence de certains personnages et pour vérifier des cohérences chronologiques ;
3. nous détectons les personnes mentionnées dans un même acte qu'une personne remarquable ou une personne chronologiquement bien située.

Puisqu'ils sont forcément contemporains, ils peuvent être associés à une fourchette chronologique et qui peut elle-même être transférée aux autres actes non datés ou datés par des intervalles où ces mêmes personnes sont mentionnées. L'objectif est ainsi de définir des séries d'informations chronologiques pour chaque document à partir des personnes mentionnées, informations qui lorsqu'elles sont croisées nous livrent la chronologie la plus précise et la plus cohérente possible pour chaque document. Les différents degrés de valeur et de confiance que nous avons fournis à chacune des trois voies d'information répondent finalement à leur valeur chronologique : date précise, date à fourchette, date transférée.

Cela vaut aussi pour le troisième cluster de documents que nous allons interroger après l'avoir OCRisé, et qui contient de l'information moins structurée parce il s'agit d'éditions érudites parfois lourdement commentées mais dans lesquelles on peut aussi récupérer des listes de noms. En l'occurrence, nous avons surtout récupéré les chronologies associées à la hiérarchie ecclésiastique en cherchant dans les index *abbatum* et *priorum* et les index *praepositorum et decanorum* des diocèses correspondants : Autun (*Augustodunensis*), Mâcon (*Matisconensis*) et Lyon (*Lugdunensis*). Une partie importante de cette information trouve son origine dans l'ensemble des cartulaires que nous avons choisi pour notre étude. Nous profitons de ce travail déjà systématisé – en l'absence duquel il aurait fallu établir un système de récupération similaire à celui que nous proposons ici (extraction d'entités nommées et cooccurrences, désambiguation, restitution de relations) mais en intégrant une quantité substantiellement plus élevée d'informations. Donc, à partir de ces documents, nous avons récupéré les dates des individus suivants :

307. Jean Richard. *Le cartulaire de Marcigny-sur-Loire, 1045-1144 : essai de reconstitution d'un manuscrit disparu*. Sociéte des Analecta Burgundica, 1957.

- Les prieurs de Paray-le-Monial : - Les prieurs de Marcigny sur Loire - Les prieurs de Cluny - Les doyens et archevêques - Les prévôts et chanceliers

Finalement, cette série de comparaisons nous a fourni autour d'une cinquantaine de personnages récupérables depuis les listes et dont les chronologies présentent très peu d'imprécisions. Ils peuvent nous fournir rapidement un arc chronologique très cohérent pour à peu près 60 actes. Dans 18 de ces documents qui présentent au moins deux personnages « datables », nous avons pu restituer une date avec une fourchette très réduite, de moins de 10 ans. En conséquence, c'est à partir de cet ensemble daté que nous allons essayer de fournir une date pour le reste des documents du cartulaire de Paray, pour lesquels nous n'avons trouvé aucune information chronologique précise. La logique de datation demeure la même : à partir de la comparaison des listes de personnes, trouver dans les actes non datés des personnes présentes dans le groupe d'actes déjà datés.

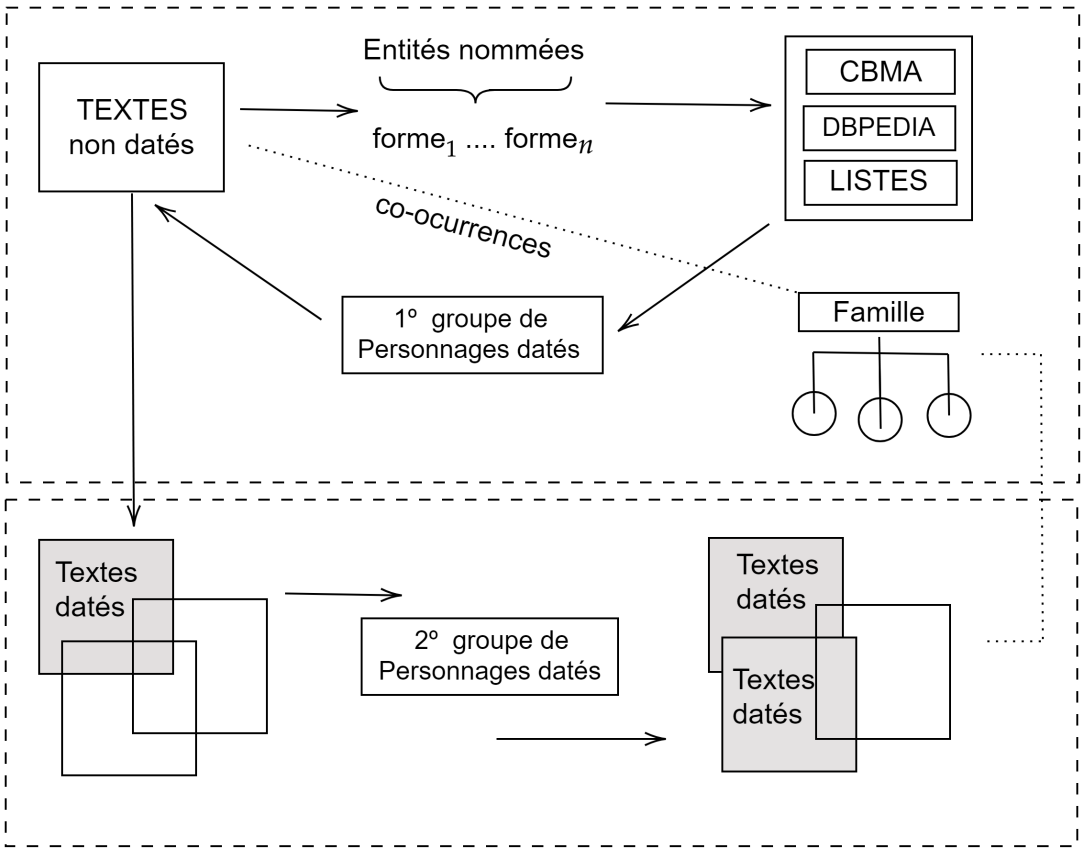

FIGURE 4.5 – Processus de datation assisté par ordinateur en bas au croisement des personnages et des événements.

En résumé, nous avons réalisé une comparaison massive entre des dictionnaires de formes constitués par les entités nommées récupérées dans les textes non datés, principalement les noms de personnes, dont nous avons regroupé toutes les variations d'ordre linguistique et grammatical, et des listes de noms, principalement de personnes, récupérés depuis le contenu déjà structuré – soit par nous, soit dans d'autres travaux – portant une information chronologique. Cette comparaison a été réalisée en deux

étapes : la première entre les chartes de Paray et les ressources externes, et la deuxième entre les chartes de Paray datées et celles qui demeurent non datées (voir figure 4.5).

Afin de bien traiter les textes, nous avons codifié un *pipeline* qui normalise les textes et met le contenu dans un format tabulaire, ce dernier facilitant, à partir des coïncidences trouvées, le transfert automatique des repères chronologiques depuis les listes associées à une date ou série de dates vers celles où elles font défaut. Ce travail permet au chercheur de disposer d'une matrice chronologique qui facilite la validation de la date et qui permet de repérer rapidement les incohérences, étant donné que la matrice peut aussi récupérer des chronologies erronées.

4.6.5 Chronologies utiles

Nous montrons maintenant trois chronologies familiales auxquelles nous avons fait plusieurs fois référence dans ce chapitre. Un cadre familial bien établi facilite l'établissement de chronologies fondées sur la naissance et sur la mort des personnes, ce qui permet d'établir le plus souvent une tranche chronologique plus resserrée qui aide à corriger des dates douteuses. La reconstruction est facilitée par les résultats automatiques autour de la récupération des co-occurrences, mais elle demeure encore un travail lourd à parfaire par le chercheur.

Nous avons considéré, à cet égard, 3 chronologies : celle des prieurs de Paray-le-Monial par rapport aux abbés de Cluny et aux comtes de Chalon et celles de deux familles bien représentées dans le cartulaire et dans d'autres cartulaires : les *Buxol* et les *Digonia*. Dans les graphiques des arbres généalogiques, nous avons inclus sur la droite une barre chronologique qui montre, dans le cas des prieurs, les dates de leur exercice de la fonction, et dans le cas des familles, la décennie estimée de la naissance des membres de chaque génération.

Les prieurs de Paray-le-Monial

La charte 12 du cartulaire[308] nous livre la liste des *prepositi sive procuratores* de Paray dont nous avons estimé la chronologie. Le document indique :

Andraldum *videlicet, virum sapientem et eruditum.... post hunc fuit domnus* ***Gunterius****, vir bonus, castus et rectus, ... huic successit domnus* ***Segualdus****.... non dispar etiam fuit alius domnus* ***Girbertus****....post hos vero quem nominari competit nimium, quia et hoc exigit ipsa operis materies, devenit domnus* ***Hugo*** *hoc tempore moderno.*

1º. *Andraldus* (1000-1030/1035) : *Andraldus* se trouve en poste du temps de l'abbé Odilon. Dans une ancienne notice[309], il apparaît avec le comte-évêque Hugues et dans une autre, que nous avons datée des années 1030-1039, avec Hugues I de Buxol, père du prieur Hugues. Canat de Chizy assure que le premier prieur de Paray fut le comte-évêque Hugues principalement parce que dans la charte de donation de Paray à Cluny ne figure aucun prieur et Hugues pourrait avoir considéré ce titre d'une dignité bien plus basse comme pour l'indiquer ; titre auquel il renonce après cet événement.

308. CBMA 7028
309. CBMA 7161

Le priorat d'*Andraldus* peut être situé entre l'an 1000 – donation de Paray à Cluny – et les années 1030/1035 [310].

2º *Gunterius* (1030/1035-1040-1045), son successeur, apparaît en 1036, lors du voyage du comte-évêque Hugues à Jérusalem qui figure dans l'acte de cession du manse de Belmont à Paray [311]. On en sait davantage sur lui : il apparaît dans deux autres actes de la même époque [312].

3º *Segualdus* (1040/1045 – 1049). La seule référence chronologique qu'on connait à propos de lui est sa démission du poste de prieur de Paray pour prendre le poste de prieur de Cluny abandonné par Hugues de Setmur lors de sa nomination comme abbé de Cluny en 1049.

4º *Girbertus* (1050-1070/1075). *Girbertus* débute comme prieur en même temps que l'abbé Hugues de Cluny. On ne connaît pas la date de sa sortie de charge, mais le cartulaire indique clairement qu'entre lui et le prieur Hugues de Buxol au moins deux prieurs se succèdent brièvement (*quem nominari competit nimium*). On peut alors supposer qu'il demeure en fonction jusqu'à la décennie 1070, ce qui serait cohérent avec les documents où il apparaît [313].

5ª *Aymardus* (?). Dans le cartulaire, figure un prieur dénommé *Aymardus* [314]. Il fait peut-être partie du groupe des prieurs de transition entre *Girbertus* et Hugues. La *Gallia Christiana* le place néanmoins entre *Andraldus* et *Gunterius*.

6º Hugues (1080-1115). On a déjà traité largement la chronologie de Hugues de Buxol, le prieur le plus important du cartulaire. Il est probablement nommé alors que Hugues II comte de Chalon (avant 1079) était en vie, et il est prieur avec certitude à partir de 1080. On peut fixer sa mort en 1115-1117.

7º *Bernardus* (1115-1116/1118). *Bernardus*, qui est ensuite nommé prieur de Cluny passe par Paray très brièvement, probablement pendant une année. Il n'apparaît que dans deux actes souscrits avec l'abbé Pons [315].

8º *Artaldus* (1117-vers 1130). *Artaldus* se trouve dans la lève d'excommunication d'un certain *Karolus* en 1119 [316] et souscrit plusieurs documents du temps des abbés Pons et Pierre. Il demeure en fonction après 1024 [317]. La chronologie de la fin de son priorat est incertaine.

9º *Burchardus* (vers 1130-vers 1145). Nous trouvons *Burchardus* dans deux documents [318] souscrivant avec le comte de Chalon Guillaume I (1113-1166).

10º *Girardus* (vers 1146). *Girardus de Copetra* se trouve dans deux documents datés de l'an 1147 et de l'an 1151 [319]. Il souscrit aussi des actes avec le comte Guillaume Ier (1113-1166) [320].

310. CBMA 7112, 7161
311. CBMA 7127
312. CBMA 7127, 7158
313. CBMA 7059, 7195
314. CBMA 7051
315. CBMA 7206, 7225
316. CBMA 7223
317. CBMA 7192, 7227
318. CBMA 7216, 7217
319. CBMA 7222, 7218
320. CBMA 7111, 7220

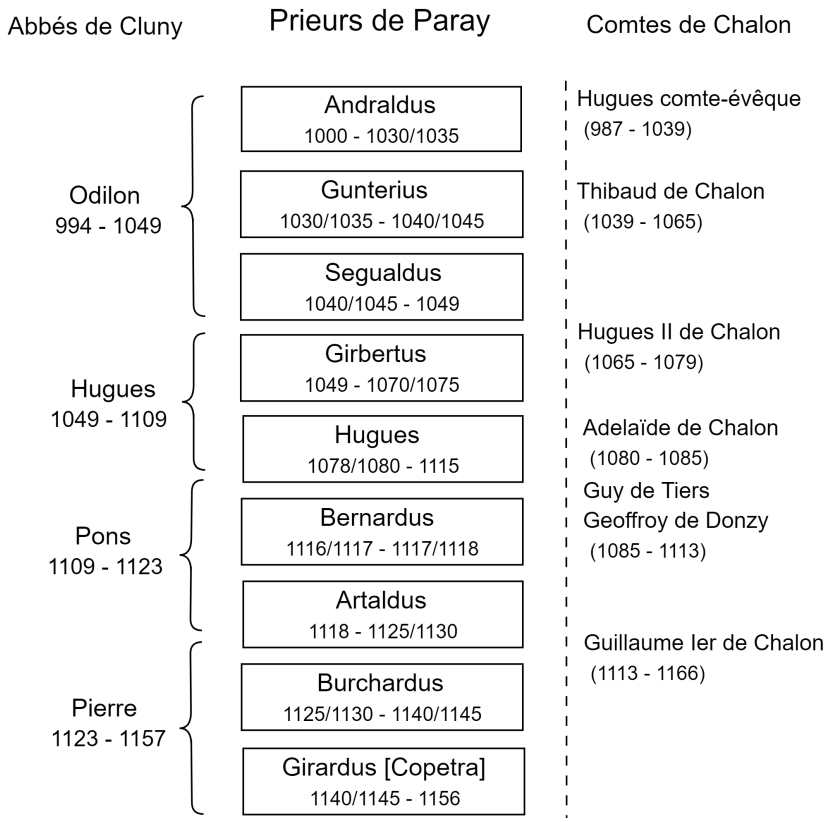

FIGURE 4.6 – Chronologie des prieurs de Paray-le-Monial par rapport aux datés des abbés de Cluny et des comtes de Chalon.

Les Digoin

Le premier sire de Digoin (*Digonia*) est attesté dans la base de données bourguignonne comme étant Josserand de Digonia, repéré dans une donation à Cluny [321] datée par Bernard et Bruel entre l'an 993 et l'an 1048 étant donné la présence de l'abbé Odilon. Il est contemporain du comte *Tetbaldus* [322]. Si on considère cette information et les dates de son fils *Letbaldus Ier*, personnage très actif pendant la première partie du priorat de Hugues (1080-1100), il convient de redater cet acte des années 1030-1048.

Effectivement, Letbaldus Ier, fils de *Jossennus*, est présent dans plusieurs actes à Paray [323] et dans une donation à Cluny de *Bernardus de Cacchiaco* daté de l'an 1105 [324] qu'il souscrit signe avec son fils *Letbaldus II*. *Letbaldus Ier* est déjà vieux au moment de cette donation et sa *donatio pro anima* ne peut pas être postérieure à 1113 [325]. Il avait souscrit un autre acte à Cluny avec son fils [326] confirmant une donation de son

321. CBMA 2022
322. CBMA 7080
323. (CBMA 7067, 7078, 7080, 7082, 7103, 7168, etc.
324. CBMA 5257
325. CBMA 7168
326. CBMA 4488

père *Josseranus*, peut-être à la mort de celui. Cet acte est daté vers 1090.

Letbaldus II n'est connu que pour quatre actes : deux actes souscrits avec son père, un troisième souscrit à Cluny en 1128 avec son fils *Girardus*[327] et un dernier en tant que scripteur (*Hoc autem fecit per manum domni Letbaldi de Digonia*), dans les années 1125-30[328]. Il s'agit de la promesse que le comte de Chalon Guillaume Ier fait à l'abbé de Cluny Pierre (1123-1157) d'en finir avec toutes les *malae consuetudines* dans ses domaines. Il est mort précisément vers 1130, puisque nous trouvons son fils *Letbaldus II* cédant à Marcigny un cens annuel pour son âme à cette époque[329].

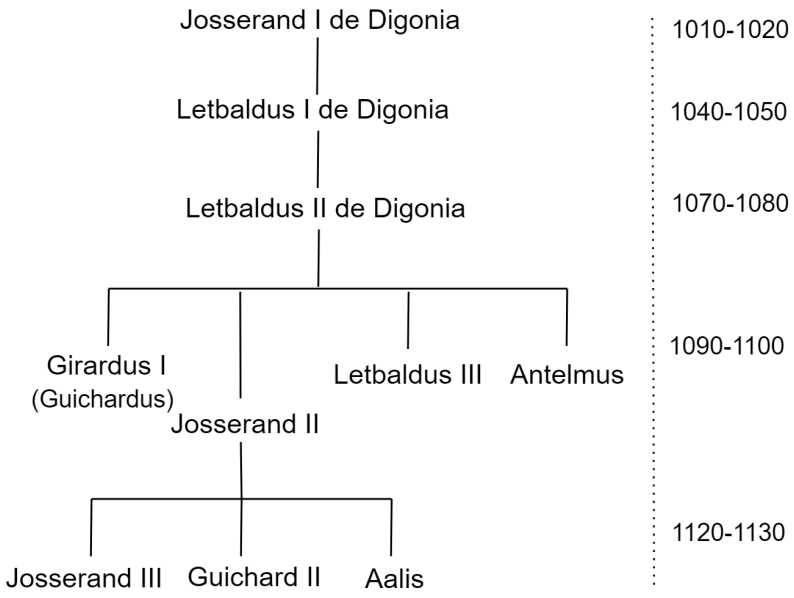

FIGURE 4.7 – Chronologie des sires de Digoin

Avec *Letbaldus II*, qui devrait apparaître plus souvent dans la deuxième partie du prieur Hugues, les mentions aux Digoin en tant que donateurs, disparaissent pour Paray. La famille dirige ses donations vers le monastère de la Ferté-sur-Grosne, fondé en 1113, et dont le recueil des pancartes contient quelques donations de la famille de *Josserannus II*, aussi de *Letbaldus II*. Dans l'une d'elles, datée de 1147, il est attesté avec ses frères (*Jocerannus de Digoni et omnes fratres ejus, Girardus videlicet et Lebaudus et Antelmus*)[330] et dans une autre en 1150 avec ses fils : "*ego Jocerannus de Digonia...et duo filii nostri, Jocerannus videlicet et Guichardus, et filia nostra Aalis*"[331]. Les noms de cette troisième génération des Digoin sont confirmés par deux actes à Paray où ils apparaissent en tant que témoins[332]. *Letbaldus II* et ses fils apparaissent aussi occasionnellement à Cluny[333]. Ce lien établi avec le monastère

327. CBMA 5436
328. CBMA 7225
329. CBMA 11320
330. CBMA 11026
331. CBMA 11042
332. CBMA 7222, 7220
333. CBMA 5436, 1128

de la Ferté-sur-Grosne est sûrement à mettre en relation avec le fait que Josserand II était marié avec la nièce du comte de Chalon Guillaume Ier [334] qui était l'un des promoteurs de la fondation de ce monastère et l'un de ses principaux bienfaiteurs.

Les Busseuil

La chronologie de la famille des seigneurs de Busseuil (*Buxol*) est très complexe à cause de la coexistence dans une même époque de deux ou plus personnages homonymes de différentes générations ; il est donc parfois malaisé d'attribuer avec certitude les documents à l'un ou à l'autre.

Le premier seigneur présent dans le cartulaire est *Hugues Ier de Buxolio* et sa femme *Aya*. Ils apparaissent à partir la décennie 1030 dans une donation faite par le comte-évêque Hugues et *Tetbaldus*, futur comte de Chalon [335]. Ils réapparaissent dans les années 60-70 du XIe siècle [336]. Leurs donations pro anima sont rédigées dans la décennie 1080 [337]. Ce mariage, d'après ce que l'on sait, a donné naissance à 5 enfants : Hugues II, prieur de Paray-le-Monial, *Artaldus, Girardus, Agnès, Gaufredus et Adelaide* [338]. *Artaldus* et *Girardus*, probablement les fils ainés, sont souvent mentionnés dans les donations de Paray à partir la décennie 1070 et surtout pendant la 1ère partie du priorat de leur frère Hugues à Paray [339] *Girardus* est l'un des chevaliers qui accompagnent le comte *Tetbaldus* à Saint-Jacques et qui ramènent son corps à Paray vers 1065 [340]. *Artaldus* assume à la mort de son père (sa mère vit encore quelques années) le rôle de chef de la famille ; il est appelé *avunculus* et *patruus* par ses neveux [341]. Quant à Hugues II, prieur de Paray, nous avons déjà largement détaillé sa vie (voir partie 2.3). *Gaufredus*, chanoine, est aussi mentionné dans quelques actes [342]. Chacune des filles, *Agnès* et *Adelaïde*, est mentionnée dans un acte [343].

Du mariage entre *Gerardus* et *Elisabeth*, nous connaissons bien la descendance parce que les deux fils ainés *Hugues III* et *Bernardus* participent très activement, parfois avec leurs oncles, dans les affaires reflétées dans le cartulaire depuis la décennie de 1090 et pendant le deuxième période de Hugues comme prieur [344]Les donations *pro anima* tant de *Gerardus* que d'*Elizabeth* peuvent être datées de ce moment-là [345]. *Atto*, monachus, et *Artaldus II*, qui sont aussi des fils de *Gerardus* n'apparaissent en revanche que très ponctuellement, jouant le rôle de donateurs et témoins dans des donations importantes de leurs parents ou leurs frères [346] et ils apparaissent tous deux

334. CBMA 11042
335. CBMA 7112
336. CBMA 7232
337. CBMA 7113
338. CBMA 7232, 7130
339. CBMA 7046, 7051, 7070, 7074, 7102, 7154, 7177, etc.
340. CBMA 7026
341. CBMA 7104, 7036
342. CBMA 7126, 7130
343. CBMA 7104, 7233
344. CBMA 7235, 7236, 7036, 7042, 7065, etc.
345. CBMA 7042, 7152
346. CBMA 7104, 7036

avec leur mère dans une donation à Marcigny pour l'âme de leur grand-mère *Aya*[347].

Sur la descendance des autres enfants de *Hugues Ier* et *Aya* nous n'avons très peu d'informations : on sait que *Artaldus* est marié avec une femme nommée *Gertrudis* et qu'ils ont au moins trois enfants : *Hugues IV* et *Artaldus III* qui apparaissent aussi dans les alentours de l'an 1100[348] et *Rotrudis*[349]. La sœur d'*Artaldus*, *Agnès* est moniale (*sanctimoniales*) à Autun[350]. Pour sa part *Adelaïde*, l'autre sœur, est mariée avec *Petrus de Chucy* (c'est le *Petrus de Cacchiaco* dont parle on parle en CBMA 7033) comme cela est rappelé dans une charte de la décennie de 1090 et qu'ils ont au moins deux enfants : *Galterius* et *Guillemus*[351], sur lesquels nous n'avons pas d'autres informations. Dans le cartulaire de Marcigny nous pouvons trouver l'acte de donation de *Rotrudis* de l'an 1105[352] lors de son entrée à la vie religieuse, approuvé par son père *Artaldus* et où participent les trois fils de *Rotrudis* : *Eldinus, Henricus* et *Robertus*.

À la suite à la mort du prieur Hugues et de la désactivation progressive du cartulaire après l'abbatiat de Pons (1109-1123) les informations sur les *Buxol* disparaissent. On trouve encore deux membres : *Guigundus de Buxol* et *Salemon de Bussel* au début du XIIIe siècle[353], mais on ne peut pas rapprocher des liens de parenté avec certitude.

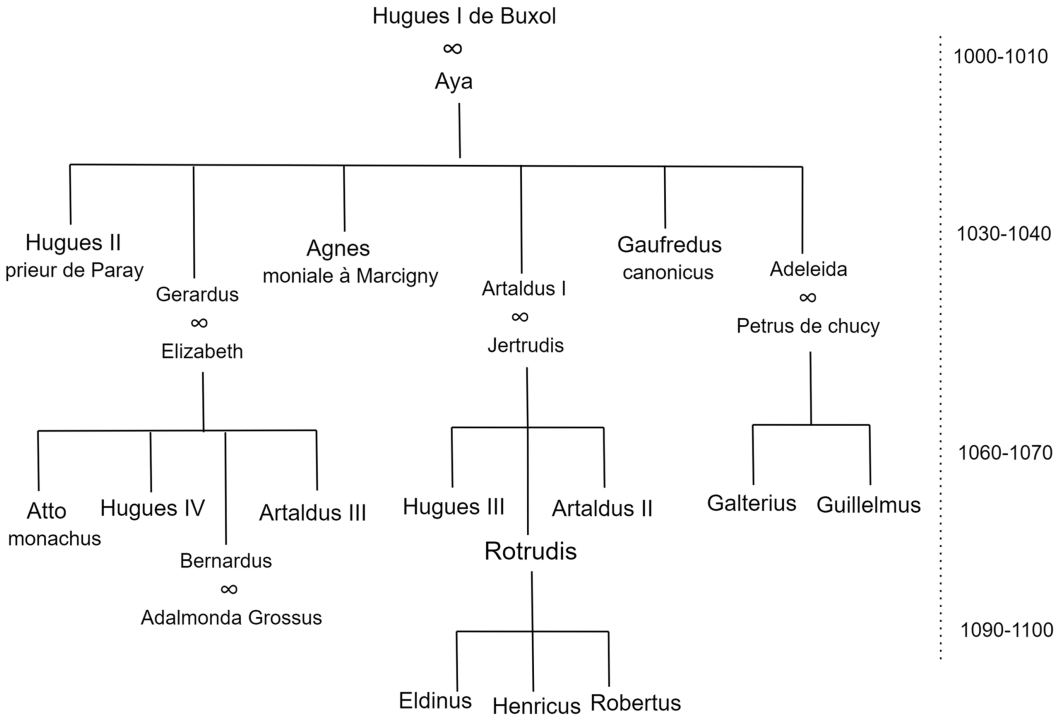

FIGURE 4.8 – Chronologie de la famille de Buxol (*Busseuil*)

347. CBMA 11227
348. CBMA 7042, 7044, 7104
349. CBMA 11304
350. CBMA 7104
351. CBMA 7233
352. CBMA 11304
353. CBMA 7219

Chapitre 5

Détection des parties du discours diplomatique

5.1 Introduction

L'étude présentée ici consiste à proposer une méthode s'appuyant sur les statistiques pour étudier les variations des formules composant le discours diplomatique. Celui-ci se compose de plusieurs parties, chacune correspondant à la séquence d'énoncés ou modules scripturaires utilisés au moment de rédiger un acte, normalement rédigés sur un modèle de formulaire. Nous avons extrait ces différentes parties en nous servant d'un modèle de reconnaissance automatique développé à cet effet, et nous nous sommes concentrés sur le protocole. Une fois les formules compilées, nous avons proposé deux méthodes de travail : l'une pour les organiser et les classer suivant des critères chronologiques, typologiques et juridiques ; et une autre pour étudier le rapport existant entre les changements - ou au contraire les formules conservées - détectés dans une formule (ou groupe de formules) et la nature des actes d'où elles proviennent.

Cette étude est divisée en quatre parties : dans la première nous avons fait une exposition générale des éléments diplomatiques qui nous intéressent particulièrement ; dans la deuxième nous présentons le rapport technique du *pipeline* qui permet d'organiser les résultats du modèle automatique et de les peaufiner à partir de deux mesures de similarité, classiques dans le domaine de la recherche d'information : la *distance Levenshtein*, qui repère des variations mineures au niveau des mots, et le *coefficient de Dice* qui nous aide à repérer les cas plus complexes où les séquences de mots se présentent comme syntaxiquement proches mais avec une distribution différente. Les résultats de l'étude sont présentés dans la troisième partie. La quatrième partie est dédiée aux conclusions qu'on peut tirer de ces résultats.

5.1.1 Les parties du discours

L'étude des éléments du discours diplomatique, c'est à dire, des différentes parties constitutives de l'acte écrit, a une longue tradition dans la diplomatique. On a déjà expliqué (partie 1.2.2) comment les études consacrées à évaluer les différents

degrés d'adoption d'un modèle formulaire lors de la rédaction des actes peuvent nous fournir toute une série d'éléments de premier ordre pour aider à définir les traditions scripturaires dans une certaine région ou à l'intérieur d'un réseau institutionnel. Le formulaire propose lors de la mise en place d'un acte juridique une séquence d'énoncés, autrement dit, des modules discursifs bien définis qui constituent un acte-type pour reproduire des manifestations écrites avec effets de droit. L'usage du formulaire est omniprésent dans les chartes de l'Europe Occidentale depuis le Haut Moyen Âge. Les premiers formulaires royaux proviennent des recueils d'actes d'origine romaine, et les formulaires haut médiévaux mérovingiens cèdent la place à une variété de formulaires au fur et au mesure que les opérations juridiques se multiplient et que les manifestations écrites en accord avec le droit prennent place comme moyen de formalisation des actes de disposition privée, judiciaires, administratives, etc.

Dans les actes émanant des autorités civiles et ecclésiastiques et ceux d'origine privée, qui nous intéressent ici particulièrement, les actes les plus formalisés sont composés de trois parties : protocole, texte et eschatocole, qui acceptent un nombre important de sous-groupes, ce qu'on s'accorde à appeler les parties du discours diplomatique [354]. L'utilisation *in extenso* de ces modules, l'ensemble de formules et en général l'allure du texte peuvent dépendre de divers facteurs, principalement deux : l'adaptation aux modèles en usage pour un même type de document et la nature de l'action juridique puisque les deux étaient déterminés dans les formulaires surtout utilisés dans les *scriptoria* religieux [355]. Il y a d'autres facteurs qui interviennent dans une moindre mesure : le rang des souscripteurs, l'institution rédactrice, le lieu de rédaction, la nature de l'affaire, etc. Tous ces facteurs doivent être exprimés par le scribe en suivant le modèle du formulaire, ce qui souvent l'amène à y introduire des adaptations et variations à différents degrés [356].

Si le *texte*, partie centrale de l'acte, se rapporte directement à l'acte juridique et peut porter diverses clauses destinées à lui conférer sa validité, le *protocole* et l'*eschatocole*, situés respectivement au début et la fin du texte, sont destinés à recueillir, d'un côté les signes protocolaires et juridiques d'authenticité (identification du fait juridique et des circonstances, des auteurs, des destinataires, signatures, sceaux, etc.) et de l'autre côté, les motivations idéologiques de l'acte juridique, sa justification, ses antécédents, etc. en guise de complément au dispositif qui porte usuellement la description de la volonté des parties et de l'action accomplie. C'est précisément cet outillage qui est souvent ignoré dans la notice qui cherche, le plus souvent, tout simplement à attester du succès de l'action.

Le modèle du formulaire est alors un schéma stéréotypé et validé de transmission de l'information et l'acte écrit doit s'attacher aux formes rédactionnelles recommandées comme condition requise de validité pour un document qui vise à devenir preuve de droits. Néanmoins, comme on le sait, le discours diplomatique se trouve lors de la rédaction des actes en constante négociation entre le carcan juridique et idéologique proposé par le formulaire et les manières particulières de mettre par écrit l'action, loin

354. Ortí, *Vocabulaire international de la diplomatique*, p. 120-123

355. Olivier Guyotjeannin et al. *Diplomatique médiévale*. Brepols, 2006, p. 65-70

356. Alice Rio. "Les formulaires et la pratique de l'écrit dans les actes de la vie quotidienne (vie-xe siècle)". In : *Médiévales. Langues, Textes, Histoire* 56 (2009), p. 11-22

d'être univoques, de la part des auteurs des actes. Les protocoles et, dans une certaine mesure, les clauses finales du dispositif destinées à assurer l'exécution de l'acte, se trouvent souvent entravés par l'exigence de légitimité et les scribes suivent strictement le modèle formulaire ; en revanche les dispositifs et les eschatocoles étant plus sujets à la description d'une affaire particulière peuvent montrer un degré plus complexe de variation lié à la capacité du scribe à transmettre les détails de l'action juridique.

Le scribe peut être forcé par des nécessités rédactionnelles très ponctuelles à adapter un modèle formulaire, mais le plus souvent ces altérations ont d'autres causes, en particulier le rapport entre la mémoire et la mise par écrit. Les scribes reproduisent des formules qu'ils connaissent par cœur. À force de répéter ils défigurent, modifient, abrègent et opèrent un agencement parfois très personnel des formules : on voit ici s'opérer le travail de la mémoire, ainsi que celui de l'inventivité du scribe. Les formes mnémotechniques et le goût de l'écrivain sont des vecteurs usuels des variations formulaires, mais aussi son inventivité ou sa négligence, comme le prouvent le nombre très élevé de faux lemmes et imprécisions repérés dans les actes [357]. Cependant, ces changements sont usuellement propres à un scribe ou une institution déterminée. Lorsqu'ils sont soutenus dans le temps et affectent l'assemblage des formules dans l'acte, il faut les identifier comme l'expression de décisions institutionnelles, de pragmatisme scripturaire ou la conséquence de changements dans la réalité historique et sociale.

Ainsi, suivant les évolutions - le plus souvent timides, rarement radicales - des formulaires, la variété de leurs adaptations, influences et emprunts, la position et la disposition des clauses et les variations de vocabulaire, on peut accéder à certains détails concernant des formes personnelles d'agencements des formulaires, mais aussi des mouvements de plus longue durée comme la circulation des idées, la définition des pratiques de l'écrit et les mutations dans la réalité juridique, sociale et spatiale [358].

Il est très important de remarquer ce dernier point parce qu'une étude trop concentrée sur le dénombrement des infinies micro variations dans les formules ne présente que peu d'utilité. Notre objectif principal est le rapport entre des changements permanents dans les formules, ou la mobilisation de certaines d'entre elles, et de certaines spécificités de l'acte : une affaire particulière, un bénéficiaire important, un scribe déterminé ; ou en tout cas, la détection de formulations déplacées ou aberrantes par rapport au modèle. [359] Bien entendu, il faut tout d'abord détecter ce qui était récurrent dans une formulation ; exercice fondamental mais qui ne doit pas constituer le noyau de l'étude pour ne pas amener le chercheur à se noyer dans le fatras de la

357. Michael T Clanchy. *From memory to written record : England 1066-1307*. John Wiley & Sons, 2012, voir aussi les études autour des cultures linguistiques et rhétoriques médiévales en : Benoît Grévin. *Le Parchemin des cieux. Essai sur le Moyen Age du langage : Essai sur le Moyen Age du langage*. Le Seuil, 2013

358. voir à ce sujet : Michel Zimmermann. "Protocoles et Préambules dans les documents Catalans du Xe au XIIe siècle : évolution diplomatique et signification spirituelle I Les protocoles". In : *Mélanges de la Casa de Velázquez* 10.1 (1974), p. 41-76 ; Heinrich Fichtenau. *Arenga. Spätantike und Mittelalter im Spiegel von Urkundenformeln*. Mitteilungen des Institus für Ósterreichische Geschichtsforschung, Ergbd. XVIII, Böhlaus, 1957

359. Robert-Henri Bautier. "Les demandes des historiens à l'informatique [La forme diplomatique et le contenu juridique des actes]". In : *Publications de l'Ecole Française de Rome* 31.1 (1977), p. 179-186.

répétition formulaire.

La diplomatique a déjà bien étudié les mouvements généraux de la formulation dans le cas des actes européens. Nous essayons ici d'ébaucher un portrait minutieux, acte par acte, des recueils de l'abbaye de Cluny, en nous concentrant sur les évolutions dans les parties les plus fixes de la tradition scripturaire. Un tel portrait dépend de diverses sources délivrant de l'information structurée provenant de chaque document, qui constituent un environnement d'analyse privilégié et peuvent en particulier nous permettre de changer rapidement l'échelle depuis l'observation générale vers l'acte particulier. Dans ce sens l'information offerte par les entités nommées se montre précieuse puisqu'elle permet d'associer chaque document à ses auteurs et bénéficiaires et aux lieux de rédaction, ainsi que d'autres données récupérées par les responsables du CBMA : type d'acte juridique, classements par nature de l'acte, classements des auteurs et bénéficiaires, diocèse de rattachement, etc. Dans le cas où l'information est incomplète ou reste douteuse on a dû intervenir, notamment pour mieux préciser la chronologie des abbés, dont la date à fourchette était insuffisamment précise et aurait pu conduire à défigurer les résultats, particulièrement dans les cas des abbés Hugues (1049-1109) et Odilon (994-1049).

Cela étant dit, l'application d'un modèle capable de nous fournir, même imparfaitement, la structure discursive des actes s'avère alors un atout majeur parce qu'il ouvre deux possibilités qui relèvent de capacités surhumaines : d'un côté, celle de pouvoir structurer et classer en peu de temps les documents à partir de ses caractères internes ; et de l'autre celle de pouvoir les comparer massivement au niveau des unités lexicales ayant un sens complet tels que le phrasème, la citation, la formule, la clause. L'importance capitale concédée par la diplomatique à l'étude de la forme de l'acte écrit et aux différentes manifestations de la négociation de cette forme par rapport au modèle du formulaire se voit nécessairement avantagée par l'accès à ces deux manières d'enquêter sur les documents. Celles-ci nous permettront d'ébaucher pour des groupes et séries de documents l'adhésion, l'éloignement, les manières d'assemblage opérés sur les formules, ce qui équivaut presque à esquisser un modèle rédactionnel.

En raison de l'étendue considérable qu'exigerait l'inspection de toutes les parties récupérables, et puisque cette étude est aussi conduite en guise de démonstration de l'utilité de nos modèles d'automatisation, nous allons nous concentrer sur la récupération des protocoles, et notamment des invocations, dans les actes du recueil de l'abbaye de Cluny, tant dans le groupe des originaux et des copies (autour de 2 500 actes) que dans le groupe d'actes provenant des cartulaires A, B et C (autour de 2 700 actes). Nous allons ainsi essayer d'apporter quelques éléments de réponse aux deux questions suivantes :

1. En premier lieu, pouvons-nous proposer une méthode de récupération et de classement qui permette de définir des formules canoniques et de repérer rapidement leurs variations ? Nous nous interrogeons à propos de méthodes validées dans le but d'organiser et de classer les formules des parties du discours à partir des éléments partagés entre elles. Le degré de plasticité observé dans les formules protocolaires complique leur récupération et classement automatique par la voie du remplacement de caractères et nous suggère l'application de méthodes mobilisant l'information sémantique.

2. En deuxième lieu, quels sont les éléments les plus déterminants dans les usages et les variations observées dans la formulation ? Nous cherchons ici à fournir un classement pour les formules récupérées à partir des éléments partagés, permettant ainsi d'associer les formules et leurs variations aux circonstances documentaires. Les réseaux monastiques, spécialement durant les Xe et XIe siècles, se montrent très conservateurs en ce qui concerne les usages formulaires, mais étant donné que Cluny se trouve au carrefour d'un très large réseau de communication, nous pouvons nous attendre un large spectre d'usages ; et il peut être intéressant d'avoir des représentations précises de l'ensemble de phénomènes autour des formules d'invocation, d'adresse et notification qui se montrent en général peu plastiques.

5.2 L'application du modèle.

Nous avons précédemment expliqué les détails de la mise en œuvre du modèle de reconnaissance des parties du discours (voir partie 2.3). Nous allons en rappeler quelques détails : il s'agit d'un modèle formé sur un jeu d'entraînement de plus de 5 000 documents provenant des chartes de Lombardie, le corpus CDML [360], datés entre la deuxième moitié du XIe siècle et la fin du XIIe siècle, issus principalement du réseau de monastères et églises de la région lombarde et de quelques registres de l'autorité civile et des archives de Bergame. Les chartes sont abondantes ; le jeu d'étiquettes d'annotation fait appel à un modèle très formalisé des parties du discours. Les travaux déjà classiques de Giry [361] et de Sickel [362] ainsi que la normalisation proposée par la Commission Internationale de Diplomatique [363] se trouvent à la base du système d'étiquettes. Cependant, les éditeurs utilisent le vocabulaire des parties du discours de tradition allemande, tandis que nous allons nous référer au vocabulaire de tradition française (suscription par intitulation, adresse par inscription, notification par promulgation, etc.)

Pour la validation du modèle, nous avons suivi les mêmes protocoles que pour le modèle d'entités nommées. Les tests de robustesse du modèle sur les chartes bourguignonnes ont montré une performance assez acceptable (voir partie 3.5.1), mais à la différence du modèle d'entités nommées qui évaluait deux étiquettes (personnes et lieux), dans ce cas nous devons évaluer 20 étiquettes (les parties du discours), qui de plus comportent des spécificités rendant la tâche plus complexe :

1) Les parties les plus répandues dans la tradition des actes européens (*invocation, suscription, adresse, notification, dispositif, data de lieu et date de temps*) montrent une plus haute performance que d'autres parties mobilisées dans les chartes les plus formelles (*préambule, exposé, rogation, iussio, recognitio*). Ceci est attribuable à la plus grande plasticité et liberté rédactionnelle attestée dans ces dernières parties, alors que les premières ont une formulation très stable ;

2) La classification des formules et clauses normalement situées à la fin du dispositif

360. http ://www.lombardiabeniculturali.it/cdlm/

361. Giry, *Manuel de diplomatique : Diplomes et chartes.-Chronologie technique.-Éléments critiques et parties constitutives de la teneur des chartes.-Les chancelleries.-Les actes privés.*

362. Theodor Sickel. *Beiträge zur Diplomatik I-VIII.* Georg Olms Verlag, 1975.

363. Ortí, *Vocabulaire international de la diplomatique.*

et au début de l'eschatocole n'a pas pu être établie avec finesse parce que cette information n'est pas disponible dans l'annotation originelle. Même si la modélisation discursive observe un certain ordre, sa disposition est trop souvent altérée. Les parties du discours ne sont pas juxtaposées, et les clauses et formules étant très diverses peuvent en plus se chevaucher et fusionner dans des structures plus grandes. Une plus grande finesse pourrait être mise en œuvre avec des méthodes d'apprentissage automatique, mais ceci n'a pas été abordé dans notre travail ;

3) La sensibilité et la précision du modèle sur les notices et les documents rédigés sous forme de lettres n'est pas complètement satisfaisant. Les notices rédigées dans un style objectif pour porter la notification de l'action juridique se réduisent normalement aux parties essentielles du discours, et les lettres relèvent d'un type diplomatique qui s'éloigne de la charte. Notre modèle entraîné sur des chartes, bien que capable de distinguer les parties essentielles, n'arrive pas à proposer une classification aussi complète pour ces documents que dans les cas des chartes.

Cela étant dit, l'application du modèle a abouti à un très bon résultat, proche de 85 %, pour la détection et récupération des cinq parties intégrant le protocole des actes qui nous intéressent spécialement pour notre analyse :

1. **L'invocation** : formule de dévotion qui met le contenu de l'acte sous la protection de Dieu, d'un saint ou au nom de la Trinité. [364]
2. **La suscription** : formule qui fait connaître le nom de l'auteur de l'acte et nous indique ses titres et qualités. Elle peut porter d'autres formules dites d'humilité ou de dévotion. [365]
3. **L'adresse** : formule qui fait connaître le nom de la personne ou personnes, et éventuellement ses titres et qualités, à qui l'acte est adressé, jouant normalement le rôle de bénéficiaire de l'action. [366]
4. **la notification** : formule très brève qui a pour objet indiquer que ce qui suit est porté à la connaissance universelle ou à celle des intéressés. [367]
5. **La date de temps et la date de lieu** : si elles sont disponibles dans le protocole, ces parties font connaître la chronologie de l'acte et le lieu où il a été rédigé ou commandé. [368]

5.2.1 Les mesures de similarité

Pre-traitement des formules

Une fois les formules récupérées en se servant du modèle automatique de reconnaissance, nous nous confrontons à trois problèmes qui doivent être réglés afin de disposer d'un ensemble complet et bien organisé :

1. La récupération des formules non reconnues ou imparfaitement reconnues par le modèle.

364. Ortí, *Vocabulaire international de la diplomatique*, p. 54.
365. Ibid., p. 54.
366. Ibid., p. 120.
367. Ibid., p. 55.
368. Ibid., p. 132.

Notre modèle de reconnaissance nous fournit 15 % de faux positif ou faux négatifs sur le total des formules étudiées, ce qui peut sembler considérable en regard de notre objectif de proposer un portrait exhaustif des manières dont se présentent les formulations. On pourrait craindre d'ignorer ainsi l'ensemble de grand intérêt constitué par des erreurs concernant des formulations peu représentées dans le corpus d'entraînement ou éloignées de la tradition. Heureusement pour nous, ces formulations sont très attachées au glossaire courant mobilisé par le scribe et il n'existe finalement que peu d'éléments qu'on puisse qualifier d'innovants. Y trouver un *hapax* (en dehors des fautes d'orthographe et des faux lemmes), ou un terme isolé par rapport à la tradition est rare et en général, même dans les formulations aberrantes, nous pouvons attester de l'usage de termes bien connus du vocabulaire diplomatique.

2. Le dénombrement des formules qui ont été récupérées sur des unités plus larges car elles se présentent fusionnées ou imbriquées :

Les parties du discours peuvent être constituées par des multiples formules ou parties de formules. C'est le cas par exemple du préambule, de l'exposé, du dispositif, qui portent souvent des multiples formules, citations et clauses. Dans d'autres cas, surtout dans les parties plus brèves, *mono-formulaires*, les limites entre les formules séquentielles se présentent parfois entravées par d'autres mots donnant l'impression d'être fusionnées, imbriquées ou transposées. La formule se présente alors sous la forme d'une sous-séquence, ce qui complique la récupération automatique et peut exiger un nettoyage par un calcul de similarité au niveau des sous-unités. L'étiquetage originel ne nous a pas fourni un niveau de finesse suffisant pour distinguer les formules à l'intérieur des parties, ce qui exigerait d'ailleurs une annotation bien plus minutieuse.

3. La classement des formules en macro-groupes basés sur les caractères couramment partagés.

Ce que nous cherchons à identifier n'est pas strictement les variations au sein des formules, mais plutôt le rapport existant entre l'utilisation d'une formule et les circonstances de production d'un acte. Pour y arriver, un regroupement des formules selon les types ou les traditions est une étape indispensable. Il doit être possible de rapprocher celles appartenant au même modèle scripturaire au-delà des détails qui les différencient comme la transposition textuelle, les faux lemmes, les changements de caractère syntaxique : abrégement, allongements, suppressions ou au niveau sémantique : synonymes, périphrases, variations dans le style, la personne, etc. Ainsi nous pouvons établir ce qui était de règle, ce qui s'éloigne du modèle, ce qui est innovant et vérifier si les changements observés ont une relation avec l'action juridique notifiée, les souscripteurs présents, le destinataire, l'affaire conclue, le lieu de production, etc.

Ces trois problèmes, bien que relevant d'intérêts fort différents, peuvent être abordés par une même solution, une routine de comparaison massive : transformation en sacs de mots par n-grammes, extraction de la fréquence des termes et bi-grammes (TF-IDF) [369] puis application des mesures de similarité : coefficient de Dice ou Cosinus.

Un exemple est ici plus parlant pour mettre en valeur ce protocole de travail.

369. Mesures statistiques d'évaluation de l'importance d'un terme contenu dans un document relativement à un corpus. TF : Fréquence du terme, IDF : Fréquence inverse de document

Prenons le début de trois chartes dont la troisième présente un problème de détection sur la notification :

> 1ère charte :«*Ego Stephanus, dominus de Neblento, notum facio omnibus tam presentibus quam futuris, quod cum discordia verteretur inter Johannem, fratrem meum, ex una parte, et ecclesiam Cluniacensem, ex altera, pacificata est in hunc modum [...]* » [370]

> 2ème charte :«*Noverint universi presentes pariter et futuri, quod ego Armandus Chabrier, clericus, vendo domui de Grasac et trado et tibi Pagano pro eadem domo recipienti, bona fide et sine dolo, per me et per meos [...]* » [371]

> 3ème charte :«*Quodcumque firmum procliui temporis statu et inmitabile permanere cupimus [...] quapropter omnibus tam presentibus quam absentibus esse notum uolumus, quod placuit atque conuenit inter domnum Oddonem abbtem Cluniesem et Vualfredum [...]* » [372]

Les problèmes de classification sont souvent des problèmes de détection des limites de la formule. Donc, pour mener à bien une comparaison par sous-chaînes il faut proposer une division des parties du discours par n-grammes. Pour rappel, le *n-gramme* fait référence à une sous-séquence de n mots dans une séquence donné de mots. Soit W_f^d une séquence où W^d nous indique le début de la séquence et W_f la fin et soit W_n^k une sous-séquence avec les suivants conditions : $k \geq d$ et $n \leq f$, ayant alors un nombre fixe de $(n - k) + 1$ items pour les sous-séquences on peut les extraire de façon itérative (W_k^d, W_{k+1}^{d+1}....) à partir de la séquence principale.

Ainsi, si on prend le premier exemple comme une séquence et qu'on établit la génération de sous-séquences par fenêtres de 7 *n-grammes* nous avons ce résultat, la sixième sous-séquence étant celle qui nous intéresse le plus car elle contient la formule complète de notification plus adresse :

w_7^1 ['Ego', 'Stephanus', 'dominus', 'de', 'Neblento', 'notum', 'facio'],
w_8^2 ['Stephanus', 'dominus', 'de', 'Neblento', 'notum', 'facio', 'omnibus'],
w_9^3 ['dominus', 'de', 'Neblento', 'notum', 'facio', 'omnibus', 'tam'],
w_{10}^4 ['de', 'Neblento', 'notum', 'facio', 'omnibus', 'tam', 'presentibus'],
w_{11}^5 ['Neblento', 'notum', 'facio', 'omnibus', 'tam', 'presentibus', 'quam'],
w_{12}^6 ['notum', 'facio', 'omnibus', 'tam', 'presentibus', 'quam', 'futuris'],
Etc.

Alors si on opère une division de la phrase par 7 - 9 *n-grammes* à un moment donné on va arriver à isoler les trois sous-séquences de chaque séquence portant la formule d'intérêt :

370. CBMA 6012, daté de 1228. Stephanus, seigneur de Neblanto, fait connaître l'accord conclu entre Johannes son frère et l'abbaye de Cluny

371. CBMA 6036, daté de 1229. Armandus Chabrier clerc vend au prieuré de Grazac le manse de Ceryata ; souscription par l'évêque du Puy-en-Velay, Étienne IV (1220 - 1231)

372. CBMA 1919, daté de 940. Odon, abbé de Cluny, échangent avec Vuarfredus et Vuarrina son épouse quatre champs situés à Lornanto.

w_{12}^{6} ['notum', 'facio', 'omnibus', 'tam', 'presentis', 'quam', 'futuris']

w_{8}^{1} ['noverint', 'universi', 'presentes', 'pariter', 'et', 'futuri', 'quod']

w_{18}^{10} ['quapropter', 'omnibus', 'tam', 'presentibus', 'quam', 'absentibus', 'esse', 'notum', 'uolumus']

Ces trois formules de notification et d'adresse universelle correspondent à des sous-versions d'un même modèle, statistiquement le plus répandu dans notre base de données. Les deux premières sont en fait plus proches puisque l'une est inspirée par l'autre et l'évoque. Toutes les deux débutent par le même verbe *noscere*, mais l'une l'employant dans une périphrase à la première personne du singulier et l'autre le conjuguant à la troisième personne du pluriel, ce qui lui confère un air plus impersonnel ; *omnibus* a été remplacé par un synonyme, *universus*, plus recherché et avec une nuance un peu plus généraliste ; le comparatif d'égalité *tam...quam* a été remplacé par un adverbe plus rarement usité, *pariter*, mais comportant le même sens lexical. La troisième version s'éloigne plus des deux premières parce qu'elle vient d'une notice d'échange et qu'elle dépend d'un formulaire bien plus ancien : la conjonction de coordination *quapropter* ouvre et connecte la formule après un court préambule ; un usage assez commun dans les anciens formulaires. Par contre, le changement portant sur le tandem assez répandu de temporalité : *presentes* - *futuri* remplacé par un autre plus fin, mais appelant plutôt à la spatialité : *presentes* - *absentes* semble une décision personnelle du scribe. L'ordre même de la formule passe de la structure typique « notification-adresse » trouvé massivement depuis le Xe siècle à « adresse-notification » moins usitée ; l'usage du parfait exhortatif *esse notum uolumus* le rapproche des anciens diplômes et des chartes épiscopales.

Malgré les différences entre ces trois formules, expliquées en partie parce que les documents d'où elles proviennent sont séparés de trois siècles, les mots principaux de la formule et le sens général demeurent intacts pour l'œil entraîné ; mais ce n'est pas le cas pour la machine. Les trois genres de différences ici présentes – modifications dans la déclinaison, usage de synonymes et surtout transposition des termes tant dans l'ordre que dans la fonction – ont généré des changements à un niveau sémantique qui ont naturellement affecté leur syntaxe, masquant la ressemblance entre les phrases. Pour y remédier, la mesure de similarité doit se concentrer plus sur la proximité morphologique que sur la détection de l'information partagée entre les formules.

La mesure de *Levenshtein*[373] était très pertinente pour le calcul de la distance entre les entités nommées car on cherchait à surmonter des variations hétérographiques minimales affectant le plus souvent entre trois à cinq caractères. Mais quand la comparaison porte sur des unités lexicales supérieures, avec un sens complet et composées par des chaînes d'une moyenne de 8-10 mots (60-70 caractères) comme dans le cas de l'adresse et de l'invocation, le nombre de caractères à supprimer, insérer et remplacer pour passer d'une chaîne de mots à l'autre est bien plus élevé. De surcroît si nous portons la comparaison au niveau des mots, la mesure de Levenshtein ne gère bien la transposition des termes dont l'édition est coûteuse. En revanche, l'application

373. Li Yujian et Liu Bo. "A normalized Levenshtein distance metric". In : *IEEE transactions on pattern analysis and machine intelligence* 29.6 (2007), p. 1091-1095

	6º	7º	8º	9º	10º	11º	12º	13º	
X_{13}^{6}	nosco	facio	omnis	tam	presens	quam	futurus	quod	cum
	1	0	1	1	1	1	0	0	0
	0	1	1	1	1	0	0	1	0
Y_{18}^{10}	quapropter	omnis	tam	presens	quam	absens	sum	nosco	uolo
	10º	11º	12º	13º	14º	15º	16º	17º	18º

TABLE 5.1 – Formules de notification 1 et 3 comparées au niveau du lemme dans le format de "sac de mots" (*bag-of-words*)

du *Coefficient de Dice*[374] sur la version lemmatisée de chaque sous-séquence semble une méthode de travail très efficace car il calcule les n-grammes partagés en dépit de l'ordre. De fait, étant donné que les formulations utilisent un vocabulaire relativement réduit et avec un niveau élevé de répétition, une version *weighted-Dice*, qui récompense les intersections portant sur des termes et associations-clés peut être empiriquement calculé en se basant sur une mesure aussi simple que la fréquence de termes ou de *bi-grammes*[375].

Pour rappel, l'indice de Dice calcule l'information partagée, rapportée à la somme des cardinalités. Alors, pour calculer $Dice(X, Y)$ étant X et Y deux chaînes à comparer et $\mid X \mid$ le nombre de mots de la chaîne X et $\mid Y \mid$ le nombre de mots de la chaîne Y :

$$Dice(X, Y) = \frac{2 \mid X \cap Y \mid}{\mid X \mid + \mid Y \mid} \tag{5.1}$$

L'exemple ci-dessous montre la comparaison dans la version lemmatisée des formules de notification 1 et 3 sous le format de "sac de mots". L'ordre dans la disposition des termes ainsi que la morphologie flexionnelle cessent d'être un facteur déterminant pour privilégier les intersections dans le vocabulaire diplomatique présent dans chaque typologie formulaire.

Si on fait le calcul : les deux sous-séquences ont une extension de 8 mots et 9 mots et ils ont cinq termes en commun (*nosco, omnis, presens, tam, futurus*). Ainsi l'indice de Dice entre les deux chaînes vaut : $Dice(X, Y) = \frac{2*5}{8+9} = 0.59$

Selon la tâche à évaluer, les exigences métriques peuvent varier. Pour la résolution des problèmes de faux négatifs et de correspondance partielle, étant donné que la comparaison est faite entre le groupe des formules récupérées et les documents suspectés d'avoir une même typologie de formule, un coefficient égal ou supérieur à 0.35 est normalement suffisant pour bien détecter les formules manquantes. En revanche, pour la troisième tâche qui cherche une ressemblance bien plus fine pour définir des sous-groupes dans le groupe de formules, l'exigence doit être portée à 0.60 ou plus.

Cette routine qui mobilise des outils classiques du domaine de la récupération de l'information demeure néanmoins plus empirique que théorique et les paramètres doivent être modifiés selon les caractères particuliers de chaque formule ou partie du

374. Wael H. GOMAA et Aly FAHMY. "A survey of text similarity approaches". In : *International Journal of Computer Applications* 68.13 (2013), p. 13-18

375. William CAVNAR, John TRENKLE et al. "N-gram-based text categorization". In : *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*. T. 161175. Citeseer. 1994

discours. Elle épargne notablement des efforts dans le classement automatiquement, mais on n'évite pas quelques nécessaires corrections à la main [376]

Mesure/seq	(1º , 2º)	(1º , 3º)	(2º , 3º)	frequency	%
Levenshtein	52	67	40	(nosco, sum)	0.58
				(sum, omnis)	0.31
				(nosco, facio)	0.19
Levenshtein lemmes	63	63	52	(tam, presens)	0.18
				(cunctus, fidelis)	0.18
Dice lemmes	0.46	0.63	0.27	(presens, littera)	0.17
				(littera, inspicio)	0.17
				(universus, presens)	0.16
Weighted Dice	0.69	0.87	0.43	(quam, futurus)	0.15
				(sum, cunctus)	0.15

TABLE 5.2 – Comparaison des résultats entre les différentes méthodes de mesure de similarité.

376. Eva PETTERSSON et al. "Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting". In : *Proceedings of the 19th Nordic conference of computational linguistics (Nodalida 2013)*. 2013, p. 163-179

5.3 Analyse des formules

5.3.1 Les invocations

L'invocation introduisait l'élément le plus simple de sacralité religieuse dans l'acte juridique. Il s'agit, comme on le verra, d'un élément du discours très utilisé pendant les Xe et XIe siècles dans les scriptoria clunisiens, bien ancré dans la charte de donation et les diplômes , et qui pourtant, dans notre corpus, est abruptement abandonné au début du XIIe siècle. L'attention portée à cet élément du discours dans les études diplomatiques est souvent insuffisante en partie parce que l'invocation, en tant qu'élément de communication idéologique, n'est pas le plus expressif, ce rôle étant réservé aux préambules et exposés, qui ont attiré l'attention dans la majorité des études des protocoles des chartes [377]. Formule brève et assez répétitive, l'invocation qui apparaît déjà dans les plus anciens formulaires, est un élément parfois ignoré ou abrégé (*in Dei nomine, etc.*) par les scribes eux-mêmes et souvent remplacé par l'invocation monogrammatique (*chrismon*). Mais, comme on verra, les usages, la composition et en général les manières de mobiliser l'invocation par rapport au reste du document étaient très bien réglés dans les habitudes et la conscience des scribes du réseau clunisien. Les observations faites sur le plan général, comme base à une étude statistique qui relève des outils de récupération automatique, nous autorisent à tirer quelques conclusions à propos du mouvement global des usages invocatoires dans les documents et pourraient également nous permettre, une fois défini ce qui était topique, de repérer rapidement des usages plus variés à certaines époques et dans certaines régions ; toutes informations qui peuvent contribuer à mieux comprendre certaines dynamiques internes des pratiques de l'écrit.

Dans le portrait général esquissé ci-dessous (Figure 5.1), l'usage de l'invocation est commun dans la production générale de chartes dans l'abbaye de Cluny jusqu'à l'avènement du XIIe siècle. On la rencontre avant même la fondation de l'abbaye, dans le dernier tiers du IXe siècle dans les diplômes de donation de Charles le Chauve [378] et son usage à cette époque est presque exclusive aux diplômes, préceptes et chartes, qui sont très formalisées jusqu'au début du Xe siècle. À partir de là, son usage se multiplie en même temps que la production des chartes de donation des particuliers et l'invocation, dans toutes ses formes, mais spécialement la plus simple (*in Dei nomine*), apparaît dans le protocole de presque la moitié de la production de l'abbaye vers la fin du Xe siècle.

Pendant un siècle, entre l'an 920 et l'an 1010, l'invocation garde une très forte densité d'usage. Après la décennie 1010, l'observation se complique parce que le recueil clunisien souffre d'une forte baisse de production qui implique aussi des changements assez profonds dans la nature des documents. On perd le style assez régulier et très attaché au formulaire qui avait caractérisé les actes du Xe siècle, au profit d'une forte

377. Voir au sujet des invocations les travaux de Zimmermann, "Protocoles et Préambules dans les documents Catalans du Xe au XIIe siècle : évolution diplomatique et signification spirituelle I Les protocoles" ; Robert-Henri Bautier. "Caractères spécifiques des chartes médiévales". In : *Publications de l'École Française de Rome* 31.1 (1977), p. 81-96 ; Rio, "Les formulaires et la pratique de l'écrit dans les actes de la vie quotidienne (vie-xe siècle)"

378. CBMA 1415, 1416

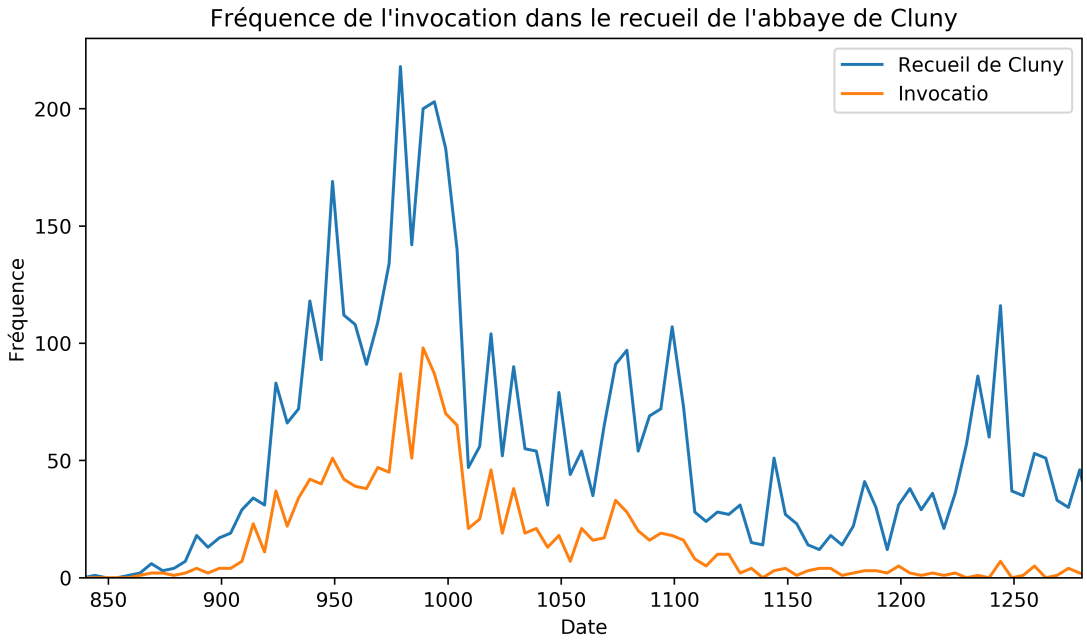

FIGURE 5.1 – Fréquence des actes présentant une invocation dans le recueil de l'abbaye de Cluny.

augmentation de la production des notices, dont le nombre dépasse à certains moments celui des chartes. La production de chartes reprend au cours de l'abbatiat de Hugues (1049-1109), sans toutefois revenir aux niveaux observés pendant l'abbatiat de Maïeul (954-994). Mais l'invocation est déjà entrée en déclin et son utilisation est bannie de l'acte après les premières décennies du XIIe siècle. Après cette époque, elle est pratiquement abandonnée dans le document privé, mais on peut encore la retrouver dans certains diplômes et chartes très formelles pendant le XIIe siècle, par exemple dans deux diplômes de Philippe II Auguste délivrées en 1180 [379], de Louis IX en 1230 [380], dans quelques lettres datées du XIIIe siècle [381] ou des chartes provenant de Castille ou d'Aragon [382]. De fait, après les premières décennies du XIIIe siècle l'abbaye ne l'emploie plus que dans quelques documents d'usage interne ou procès-verbaux : sentences [383], visites [384], litiges [385], etc.

La disparition relativement abrupte de l'invocation entre le XIe et le XIIe siècle, schématisée par le graphique, est expliquée par deux raisons qu'on détaillera dans le chapitre qui suit. Succinctement, on peut dire ici que, d'une part, sa disparition coïncide avec celle de la charte privée de donation qui était l'action juridique la mobilisant le plus ; d'autre part, les cartulaires, et spécialement C et D, témoignent d'une attention particulière à recueillir d'autres types d'actes juridiques, notamment

379. CBMA 5714, 5715
380. CBMA 6047,
381. CBMA 5868, 6008, 6915
382. CBMA 5966, 6007, 6318
383. CBMA 5816, 1198, 6302
384. CBMA 6312
385. CBMA 6881, 6895, 6914, 6945, 6971

des notices, et une documentation provenant des autorités, en particulier la papauté et les évêchés sous la forme de lettres, bulles et privilèges, tous types de documents qui rejettent presque entièrement l'utilisation de l'invocation. La notice, on l'a déjà expliqué, privilégie le dispositif et se débarrasse de la plupart des éléments de la panoplie diplomatique de la charte, y compris les invocations ; pour leur part lettres et bulles utilisent rarement les invocations, la réservant pour quelques cas de litiges [386] et testaments.

Dans le recueil clunisien (env. 4 700 documents entre 842 et 1253) nous avons récupéré 1519 formules d'invocation qui correspondent à un nombre à peu près similaire de documents. Il existe en effet des documents portant des formules doubles d'invocation. C'est le cas de quelques diplômes carolingiens qui portent une invocation finale sous la forme d'appréciation dans l'eschatocole (*in Dei nomine feliciter*) [387] ou de quelques documents qui, ayant utilisé une invocation dans la tête de la charte, la répètent dans la souscription (*Ego, in dei nomine, X, episcopus....*) ou plus rarement dans la date, mais il s'agit de cas atypiques. L'invocation généralement utilisée une seule fois dans le document et la place qu'elle occupe implique presque toujours l'usage d'un style rédactionnel particulier. Elle est normalement placée en tête de la charte, se présentant dans de nombreux cas, chronologiquement bien situés, en combinaison avec le tandem « notification + adresse ». Elle peut se présenter aussi à la suite d'un exposé ou d'une narration, suivant le même modèle, mais cette combinaison est minoritaire. Plus commune est l'invocation dans la souscription (voir point 5.4) qui se présente en tête du dispositif sous une forme assez simple. Très rarement elle se trouve au milieu de la charte ou dans l'eschatocole, exception faite des cas de chartes doubles ou de chartes comportant des extraits d'autres chartes qui, elles, peuvent inclure l'invocation ; dans toutes ces situations, qui ne représentent qu'un faible pourcentage des cas, l'invocation est bien plus difficile à détecter parce qu'elle se présente imbriquée dans d'autres formulations.

Comme on l'a expliqué précédemment, il est de bonne pratique de classer ces formules suivant les sous-types proposés par les études diplomatiques. Ainsi, pour nos 1 519 invocations, nous considérons quatre styles d'invocations, qui correspondent schématiquement aux sous-types présents dans les formulaires :

1. L'invocation trinitaire (264 documents)
2. L'invocation christologique (432 documents)
3. L'invocation divine au nom de Dieu (246 documents)
4. L'invocation divine de souscription (542 documents)

L'invocation trinitaire

L'invocation trinitaire connaît deux versions, l'une abrégée sous le concept de la Trinité et l'autre plus développée et mentionnant chaque personne de la divinité. La première se présente comme :

['In', 'nomine', 'sancte', 'et', 'individue', 'Trinitatis']

386. (CBMA 6955, 6971, 6972
387. CBMA 1434, 1429, 1446, 1491, 1830, 1912

['In', 'nomine', 'summe', 'et', 'individuæ', 'Trinitatis']

La version portant *sanctus* est la plus utilisée, les deux du total se rapportent à ce modèle qui est au début du corpus presque exclusivement employé dans les diplômes. En fait, les diplômes et préceptes se servent très rarement des autres versions trinitaires. Le modèle est rapidement transféré vers les chartes provenant de la noblesse aux environs de la fin du IXe siècle. La version avec *summus* est bien moins utilisée, et exclusive de la charte ; on la retrouve depuis le début du Xe siècle.

La version trinitaire élargie est plus tardive, on l'utilise depuis la deuxième moitié du XIe siècle jusqu'à une bonne partie du XIIIe siècle. On détecte deux sous-versions, mais qui sont parfaitement interchangeables : on les retrouve indistinctement dans les mêmes lieux et aux mêmes dates.

['In', 'nomine', 'sancte', 'et', 'individue', 'Trinitatis', 'Patris', 'et', 'Filii', 'et', 'Spiritus', 'Sancti]
['In', 'nomine', 'Patris', 'et', 'Filii', 'et', 'Spiritu', 'sancti']

Quelques notices, concernant toujours des donations importantes, portent le premier sous-type pendant le XIe siècle. Ensuite on le retrouve plus rarement dans les notices, sous une forme très abrégée :
[In', 'nomine', 'Sancte', 'Trinitatis']

Par ailleurs, la formule trinitaire peut se terminer par la réaffirmation *Amen* à la fin. Il s'agit d'un usage tardif, qui apparaît pour la première fois dans un diplôme de Philippe Ier en 1078 [388] et devient usuel dans les invocations du XIIème siècle, surtout dans les diplômes.

Le modèle trinitaire est le plus riche, mais comme tous les autres modèles d'invocation trouvés dans le recueil il se montre très rigide. On ne dénombre qu'une vingtaine de formulations différentes du modèle cité plus haut, qui s'éloignent de celui-ci par l'emploi de mots de très faible incidence ; ceci est généralement le cas dans des formules très allongées. Ces usages souvent sérialisés apparaissent et disparaissent au cours d'une même décennie, ce qui nous permet d'inférer que leur usage est lié à un scribe ou à *scriptorium* en particulier. Dans d'autres occasions, surtout pendant le XIIe siècle, ils peuvent être introduits dans notre corpus par des chartes des *scriptoria* étrangers :

Quelques exemples :
Avec la déclaration expresse de l'unité divine :

['In', 'nomine', 'sancte', 'et', 'individue', 'Trinitatis', 'Patris', 'et', 'Filii', 'et', 'Spiritus', 'Sancti', 'qui', 'est', 'trinus', 'in', 'nomine', 'et', 'unus', 'in', 'numero'] [389]

388. CBMA 4923

389. CBMA 4941, 4987, 4964, 5007, 5358, 5401. Un bon nombre de versions trinitaires très allongées vient des préceptes et chartes castillanes dont la chronologie est concentrée dans les décennies 1070-1080 et toujours concernant des donations tant privées que royales d'églises et monastères à Cluny

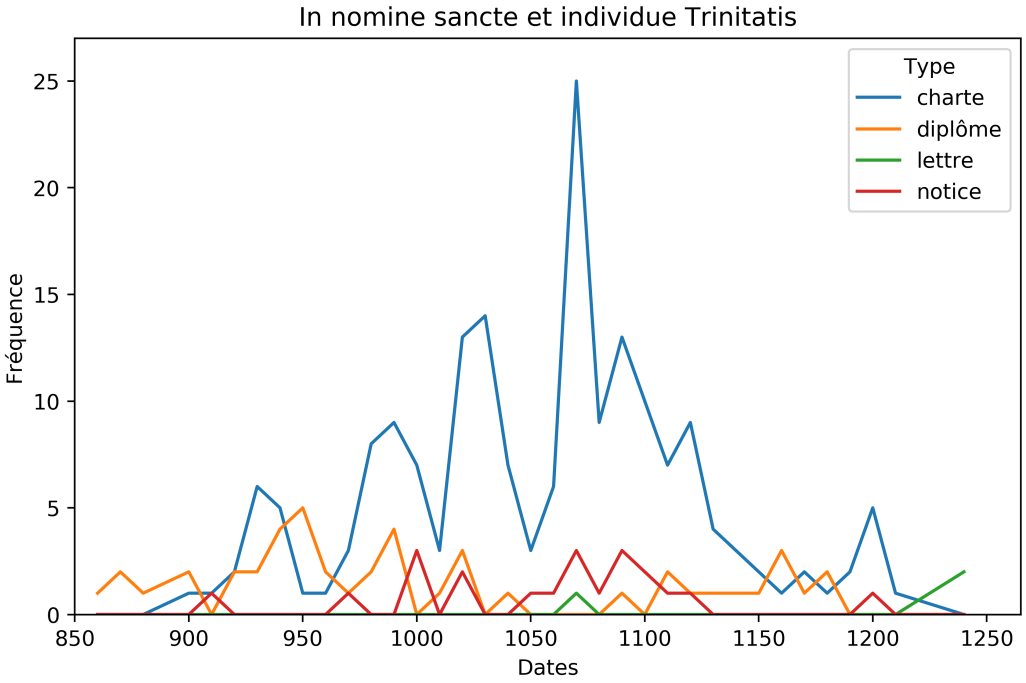

FIGURE 5.2 – Évolution chronologique de l'invocation trinitaire selon le type d'acte juridique.

['In', 'nomine', 'Genitoris', 'et', 'Geniti', 'simul', 'et', 'ex', 'ambobus', 'procedentis', 'Spiritus', 'Sancti,', 'qui', 'est', 'trinus', 'in', 'unitate', 'et', 'unus', 'in', 'deitate'] [390]

Avec clause de perpétuité finale [391] :

['In', 'nomine', 'Patris', 'et', 'Filii', 'videlicet', 'et', 'Spiritu', 'Sancti', 'uni', 'Deo', 'in', 'Trinitate', 'conregnanti', 'per', 'numquam', 'finienda', 'secula', 'seculorum', 'Amen'] [392]

Avec des attributs de la royauté :
['In', 'nomine', 'sanctæ', 'et', 'individuæ', 'Trinitatis', 'potenter', 'cuncta', 'regentis', 'atque', 'disponentis] [393]

Version hellénisée :
['In', 'nomine', 'Dei', 'Patris', 'omnipotentis', 'filiique', 'eius', 'Iesu', 'Christi', 'Domini', 'nostri', 'Salvatoris', 'et', 'Spiritus', 'Sancti', 'paracleti', 'ab', 'utroque', 'procedentis'] [394]

390. CBMA 4876, 4916
391. Une clause de perpétuité exprime le désir de l'auteur de l'acte d'assurer à celui-ci une valeur perpétuelle : ici « secula, seculorum »
392. CBMA 4932, 4964, 5007
393. CBMA 2908, 3336, 3353, 3347. Les deux premiers actes ont été élaborés par le même scribe *Iterius levite et monachi*. Les autres deux viennent des diocèses d'Autun et Dijon respectivement
394. CBMA 2452, 3301, 2568, 3289. Les deux premiers actes suivent le même modèle d'acte formule par formule, donation d'un curtil avec un serf. Les deux derniers ont été écrits par le même scribe : *Data per manus Richardi sacerdotis*

L'invocation christologique

L'invocation christologique mentionne la plupart du temps la seule personne du Fils, mais peut, dans de rares chartes, être intriquée avec une invocation trinitaire ou du Père. Nous considérons trois sous-types pour cette invocation :

Le premier sous-type se présente comme une déclaration de foi du dogme de l'incarnation et représente plus ou moins la moitié des invocations christologiques. On le retrouve exclusivement dans les chartes et trois notices seulement ; aucun diplôme ne le porte. Cette invocation est attestée depuis les premières décennies du Xe siècle, mais elle est très ancrée dans deux périodes assez spécifiques : principalement entre 990 et 1010, et entre 1080 et 1090.

['In', 'nomine', 'Verbi', 'incarnati']

Le deuxième sous-type se présente avec nom et épithète dans un rappel de la majesté du Christ ; une sous-version allongée inclut le titre de Sauveur.

['In', 'nomine', 'Domini', 'nostri', 'Iesu', Christus']
['In', 'nomine', 'domini', 'nostri', Christus']
['In', 'nomine', 'Domini', 'Dei', 'et', 'salvatoris', 'nostri', 'Iesu', 'Christi']

La première sous-version est la plus utilisée et on la retrouve dans une centaine de documents depuis la deuxième moitié du Xe siècle jusqu'aux alentours du XIVe siècle, dans un ensemble de lettres de la papauté. La version avec le terme *sauveur* est, elle, peu usitée. Son usage est attesté dans une vingtaine de chartes depuis la deuxième moitié du XIe siècle, notamment dans un groupe de huit chartes provenant de Pavie.

Finalement, un troisième sous-type est attesté, moins utilisé que les deux précédents. Cette dernière invocation est très abrégée et imite la formule la plus répandue d'invocation : *In Dei nomine* ; elle est de fait employée dans les mêmes circonstances que celle-là :

['In', 'Christi', 'nomine']

Elle est utilisée aux mêmes dates que l'invocation par l'incarnation et s'ancre dans presque les mêmes périodes : 970-1000 et 1080-1100, ce qui la rend parfaitement interchangeable avec celle-là. Normalement située en tête de la charte elle peut aussi se présenter dans le corps des chartes, à la manière de l'invocation divine *in Dei nomine*, mélangée avec la suscription.

['ego/nos', 'quidem', 'in', 'Christi', 'nomine']

Aussi rigide que l'invocation trinitaire, l'invocation christologique présente très peu d'exemples de variation dans ses termes et modèles. De nouveau, les variations sont souvent liées à un scribe ou un scriptorium étranger. Voici quelques-uns parmi les plus intéressants :

Un des très rares exemples d'invocation mariale et apostolique :

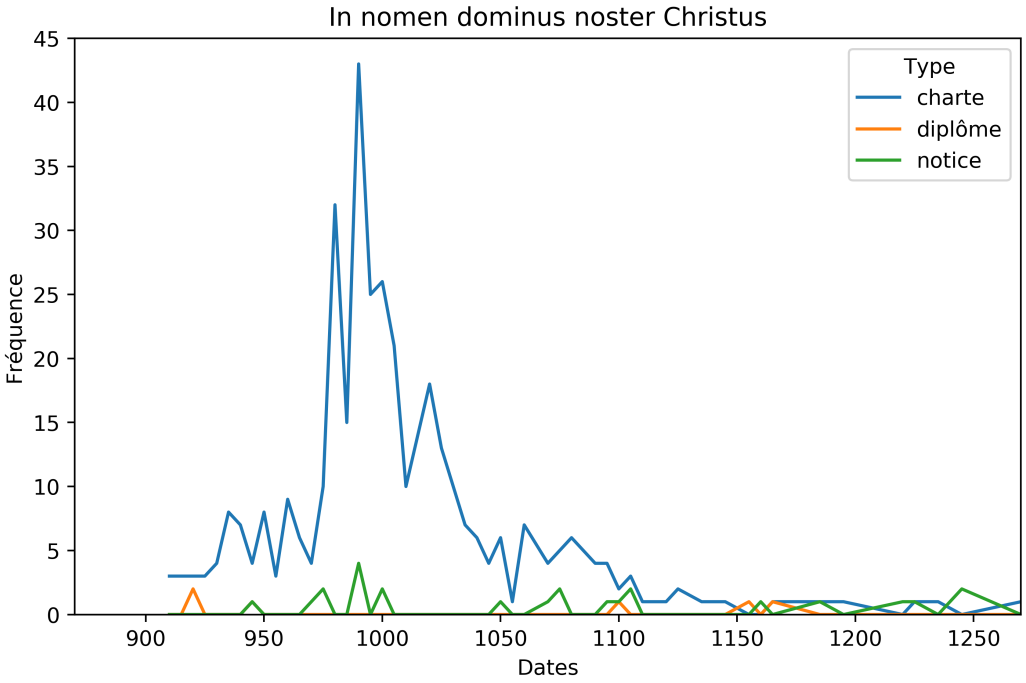

FIGURE 5.3 – Évolution chronologique de l'invocation christologique selon le type d'acte juridique.

['In', 'nomine', 'Domini', 'Jesu', 'Christi', 'et', 'beate', 'Mariæ', 'semper', 'virginis', 'et', 'beatorum', 'apostolorum', 'Petri', 'et', 'Pauli'] [395]

Recours à la formule d'origine biblique (Jean 8 :42) de réaffirmation trinitaire, popularisé par Saint Ambroise [396]
['in', 'nomine', 'Verbi', 'incarnati', 'ex', 'corde', 'Patris', 'eructuatum'] [397]

Avec l'attribut de la *Clementia* dans une charte de donation royale provenant de la chancellerie aragonaise (1145)
['In', 'Christi', 'nomine', 'et', 'eius', 'divina', 'clementia'] [398]

Invocation avec citation à l'occasion d'une charte de donation par fiançailles. [399]

395. CBMA 2027. Il s'agit d'une réutilisation de la formule d'adresse très usitée, pendant le Xe siècle : *Sacrosancto et exorabili loco in honore Dei omnipotentis et beatæ Mariæ virginis ac beatorum apostolorum Petri et Pauli consecrato.*
396. "Quomodo paterno generatus ex utero, quomodo eructuatum ex corde uerbum legitur nisi ut ex intimo et inaestimabili patris intellegatur, ut scriptum est, prodisse secreto ?", Ad Decretum Gratianum, 53
397. 2462, 2482, 2519, 4060, CBMA 2454. Les trois premiers actes rédigés par le même scribe *Evrardus levita indignus scripsit*, en 991-992. Les deux derniers également écrits par la main d'un même scribe : *Ego frater Pontius scripsi, ad vicem cancellarii*, 981-992
398. CBMA 5541. Charte d'origine aragonaise, an 1145. "Divina favente clementia rex" est d'ailleurs une formule de dévotion royale utilisée depuis le Xe siècle. CBMA 1446,1809,2102, 3628, 4873, etc.
399. CBMA 3696. La citation de Matthieu apparaît dans les préambules de quelques chartes de donations entre conjoints de la fin du Xe siècle. CBMA 2118, 2138, 2836, 3454, 3875

['In', 'nomine', 'Domini', 'et', 'Salvatoris', 'nostri', 'Jesu', 'Christi', 'qui', 'vul', 'omnes', 'salvos', 'fieri', 'et', 'acnicionis', 'veritatis', 'venire', [400], 'et' 'quod', 'Deus', 'iunxit', 'homo', 'non', 'separet'] [401]

L'invocation au nom de Dieu

L'invocation la plus usitée est la plus simple et de la plus ancienne tradition. L'invocation à Dieu utilisant l'ablatif instrumental apparaît tôt dans le corpus, depuis les premières décennies du Xe siècle, et son usage est attesté jusqu'aux alentours du XIIIe siècle. Nous avons distingué deux sous-types dans cette invocation qui revêtent en réalité la même terminologie, mais qui se distinguent par leur fonction et leur place dans le document :

Le premier apparaît presque toujours en tête du document sous deux formes simples, la première étant préférée pendant le Xe siècle, mais non abandonnée par la suite, et la deuxième pendant les XIe et XIIe siècles : ['In', 'nomine', 'Dei']

['in', 'nomini', 'domini']

A ces deux formulations quelques attributs divins peuvent être ajoutés, notamment *summus, omnipotens* et *eternus*, spécialement dans les chartes de la deuxième moitié du XIe siècle.

La deuxième version n'est pas plus riche, mais nous amène vers une rhétorique différente. Elle est très caractéristique du corpus clunisien et se présente fusionnée avec la suscription personnelle, comme cela était proposé dans plusieurs formulaires altimédiévaux sous la forme :

[Igitur, in Dei nomine, ego]
[Quapropter, in Dei nomine, ego]
[Idcirco, in Dei nomine, ego]

La liste de conjonctions et adverbes est bien plus longue : *enim, sic, quamobrem, unde, quocirca, quidem,*etc. Et ils peuvent apparaître le plus souvent comme premier terme, ou précédés par le pronom :
['nos', 'enim', 'in', 'deus', 'nomen']

Cette invocation apparaît parfois en tête de la charte, mais est le plus souvent précédée, soit par une adresse individuelle :
['Domino', 'fratribus', 'Leotbert', 'et', 'uxore', 'sua', 'Adalgelt', 'emptores', 'Igitur', 'in', 'Dei', 'nomen']

Soit par une adresse collective :
['Sacrosancto','monasterio','qui','est','constructus','in','honore'beatorum', 'apostolorum','Petri', 'et', 'Pauli']

400. 1 Timothée 2,4 : "Qui omnes homines vult salvos fieri et ad agnitionem veritatis venire"
401. Matthieu 19, 6 : "Itaque iam non sunt duo sed una caro quod ergo Deus coniunxit homo non separet"

Soit par un préambule, qui déplace l'invocation à un deuxième rang. Cette formulation est très utilisée dans les chartes de donation de la deuxième moitié du Xe siècle :

['Inspirante', 'omnium', 'rerum', 'Creatore', 'divinaque', 'benignitate', 'favente', 'cunctis', 'bona', 'temporalia', 'possidentibus', 'concessum', 'atque', 'attributum', 'constat', [...], 'Igitur', 'in', 'Dei', 'nomine', 'ego'.....]

Dans les deux cas l'usage de la conjonction consécutive *igitur* et dans une moindre mesure *quapropter* pourrait sembler superflu ou déplacé si elle n'était pas une particule obligatoire dans les formulaires anciens. Ce dernier adverbe est particulièrement rencontré dans les documents de ventes ou de donations-ventes de la deuxième moitié du Xe siècle.

La majorité des invocations rédigées sur ce modèle intégré dans la suscription se fait à la première personne, mais il n'est pas rare de trouver le pluriel *nos* quand il s'agit d'une donation collective, notamment des couples, même si ensuite le récit revient au singulier.

Par ailleurs, la réaffirmation *Amen* qui est utilisée dans les diplômes portant l'invocation trinitaire pendant les XIe et XIIe siècles et qui apparaît souvent associée à l'invocation, n'est presque jamais retrouvée ici. Il n'existe que trois documents la portant, provenant de notaires apostoliques en 1246 [402].

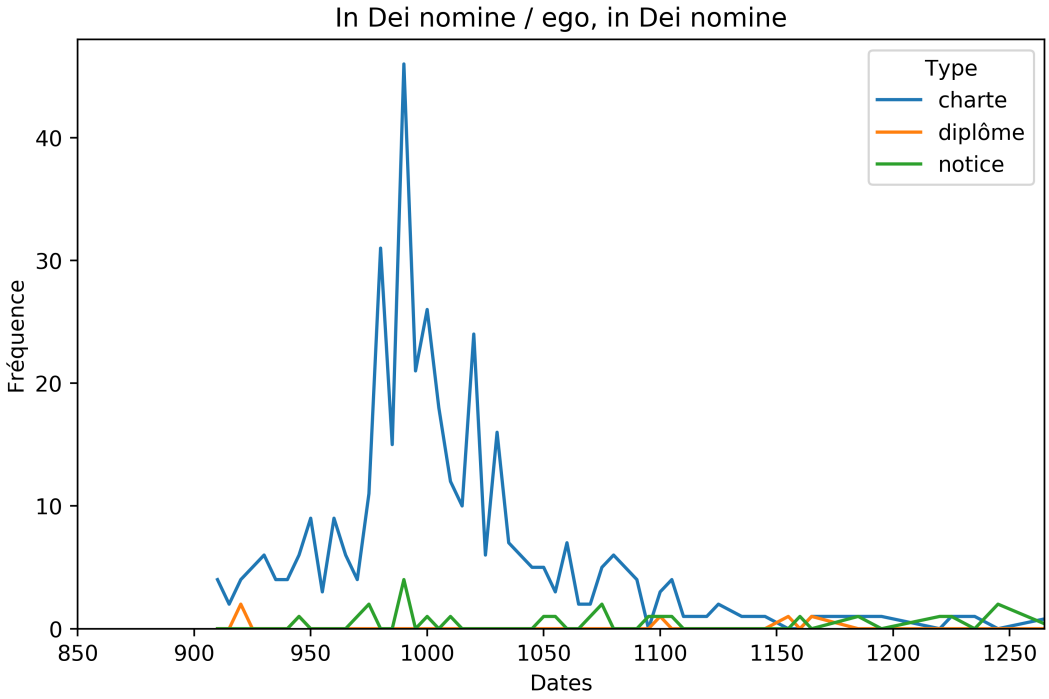

FIGURE 5.4 – Évolution chronologique de l'invocation divine selon le type d'acte juridique.

Enfin, si les invocations trinitaires et christologiques étaient déjà peu sujettes à l'inventivité et à l'altération de la part des scribes, dans le cas de l'invocation divine

402. CBMA 6361, 6362, 6955

ces variations sont réduites au minimum. Les changements sont concentrés sur la mobilisation de quelques attributs divins bien connus et de quelques conjonctions de conséquence. Les exceptions sont très rares, de même que l'utilisation d'autres attributs que ceux que nous avons listés. On peut néanmoins exposer quelques exemples :

Combinaison des attributs séculaires et divins :
['In', 'nomine', 'Dei', 'omnipotentis', 'pii', 'et', 'misericordis'] [403]
['In', 'nomine', 'regis', 'eterni'] [404],

Invocation divine et hagiographique :
['In', 'dei', 'nomine', 'et', 'sanctorum', 'omnium'] [405]

Invocation sous le nom du *Pater*. Il s'agit du seul exemple identifié dans tout le corpus d'invocation du *Pater* sans la trinité.
['in', 'nomen', 'pater', 'omnipotens'] [406]

5.4 Facteurs déterminant l'usage de l'invocation dans les actes

5.4.1 Les auteurs et bénéficiaires des actes

L'usage de l'invocation dans les actes est loin d'être hasardeux. Dans le portrait général que nous venons d'esquisser, on peut bien percevoir les limites typologiques et chronologiques qui leur sont imposées. Néanmoins, ce panorama, s'il offre des éléments de réponse sur les facteurs qui déclenchent l'usage d'une invocation dans un acte, ne nous révèle pas s'il existe des distinctions entre les actes avec invocation et celles qui n'en portent pas. Est-ce que la présence de certains individus dans l'acte ou un acte commandé par un personnage de haut rang peuvent être la raison de la mention d'éléments plus solennels ou formels, dont l'invocation semble faire partie ? Est-ce que certaines institutions sont plus attachées à l'usage des invocations ? Ou encore l'invocation se présente-t-elle plutôt comme un élément obligatoire d'un certain style rédactionnel et optionnel ou banni dans d'autres ?

Un des plus grands avantages de disposer d'une base de données avec une information structurée concernant les différents caractères des chartes, est que nous pouvons pousser l'analyse sur ce genre de questions assez spécifiques dont la réponse implique la consultation ou le dépouillement de milliers d'actes.

Pour répondre à la première des questions nous avons comparé le groupe d'actes portant une invocation à celui que n'en porte pas, classés selon la catégorie du commanditaire de l'acte, comme on peut les voir ci-dessous (Tables 5.3 et 5.4). Nous

403. CBMA 5637, "ut pius et misericors Deus [Dominus] eripere dignetur" est une formule d'usage dans les exposés de plusieurs actes depuis la deuxième moitié du Xe siècle : CBMA 1616, 1876, 2058, 2085, 2213, 2356, etc.
404. CBMA 3335, 3413, 5107
405. CBMA 3758
406. CBMA 4830

avons limité les catégories de commanditaires à douze, même si la liste est plus longue et qu'il est possible d'affiner ; cela nous fait perdre à peu près 5 % des actes répartis dans des catégories très minoritaires de personnages.

La catégorie majoritaire, QU (*Quidam*), correspond aux actes commandés par un particulier, couramment un membre de l'aristocratie locale ou quelqu'un dont on ne connaît pas de titres et dignités. Le rapport 40 % - 60 % entre les actes, essentiellement des chartes, portant une invocation et ceux n'en portant pas, détectés pendant les Xe et XIe siècles est naturellement répliqué puisque la majeure partie des donateurs appartient à cette catégorie.

Auteur/ Date	AB	AE	CL	CO	DO	DU	EP	IM	MI	QU	RE
850	0	0	0	0	0	0	0	1	0	9	3
900	3	0	15	7	0	1	6	7	1	178	14
950	13	3	77	5	0	0	4	3	0	462	13
1000	8	2	10	3	0	2	3	1	8	283	5
1050	1	4	5	22	1	3	10	0	12	112	12
1100	0	2	0	2	4	1	8	1	2	32	9
1150	2	2	0	2	0	0	2	0	0	0	10
1200	5	1	0	1	1	0	1	0	0	0	1
1250	0	0	0	0	1	0	0	0	1	0	0
Total	32	14	107	42	7	7	34	13	24	1076	67

TABLE 5.3 – Évolution chronologique des actes portant une invocation selon le type de commanditaire. Légende : AB (*abbas*), AE (*archiepiscopus*), CL (*clericus*), CO (*comes*), DO (*dominus*), DU (*dux*), EP (*episcopus*), IM (*imperator*), MI (*miles*), QU (*quidam*), RE (*rex, regina*).

Auteur/ Date	AB	AE	CL	CO	DO	DU	EP	MI	PA	QU	RE
850	1	0	3	2	0	0	2	0	0	34	0
900	18	4	14	10	0	1	7	0	8	288	0
950	60	1	81	12	1	0	6	8	4	720	1
1000	25	1	13	17	0	3	14	8	6	370	2
1050	19	1	19	24	3	6	20	54	45	246	5
1100	15	14	6	13	7	1	26	12	84	81	7
1150	9	6	0	9	1	4	15	1	71	0	7
1200	161	11	9	25	43	10	50	5	104	5	6
1250	4	2	1	2	0	2	33	1	148	1	12
Total	312	40	146	114	55	27	173	89	470	1745	40

TABLE 5.4 – Évolution chronologique des actes ne portant pas d'invocation selon le type de commanditaire. Légende : AB (*abbas*), AE (*archiepiscopus*), CL (*clericus*), CO (*comes*), DO (*dominus*), DU (*dux*), EP (*episcopus*), IM (*imperator*), MI (*miles*), QU (*quidam*), RE (*rex, regina*).

Ainsi, la disparition assez drastique de l'invocation au début du XIIe siècle se produit en partie parce que, comme on voit, l'acte privé de donation commence à

disparaître du corpus au cours de la deuxième moitié du XIe siècle. Donc, dans l'acte privé provenant d'un *quidam*, ce qui correspond à presque le 60 % du total des actes dans le corpus (77 % si on se limite aux IX-XIe siècles) il est habituel de ne pas porter d'invocation : 6 sur 10 n'en portent pas. Mais est-ce que cela a plutôt à voir avec le type d'affaire ici conclue ? On le verra par la suite.

Les actes provenant des rois (RE) et des empereurs (IM) carolingiens, on l'a déjà vu dans la description de l'invocation trinitaire, sont très attachés à l'invocation qui est quasiment de règle jusqu'au XIIe siècle. Il y a 79 diplômes et préceptes dans le corpus, un nombre important pour ce type d'actes, mais assez réduit par rapport aux autres catégories. Or, malgré ce petit nombre d'actes on observe très bien la disparition presque absolue de l'invocation au XIIIe siècle.

Les actes provenant de la haute noblesse sont très peu nombreux. Les documents comtaux (CO) sont les mieux représentés. On voit leurs actes s'attacher rapidement à l'invocation à l'instar des documents royaux, sur le modèle de l'invocation trinitaire depuis les premières décennies du Xe siècle, mais c'est un usage qui reste peu répandu : on observe que hormis dans la période 1050-1100, l'invocation apparaît dans moins d'un tiers des actes. La situation est similaire pour les actes provenant des ducs (DU), des seigneurs (DO) et des chevaliers (MI) même s'ils sont plus difficile à étudier, car peu nombreux.

En ce qui concerne les actes provenant du pouvoir public ecclésiastique, évêques (EP) et archevêques (AE), ils suivent un usage partagé de l'invocation jusqu'à la fin du XIe siècle. À partir de là, les actes des évêques se multiplient sous la forme des lettres, bulles, privilèges et confirmations, tous documents qui, comme on l'a vu, négligent systématiquement l'usage de l'invocation et la relèguent à quelques litiges et procès-verbaux.

Pour sa part, la deuxième catégorie la plus nombreuse, AB, correspond aux actes commandités par des membres des communautés monastiques : abbés, prieurs, et autres officiers, ou simples moines (personnages portant les titres : *abbas, prior, decanus, monachus*). Il s'agit pour l'essentiel de documents traitant d'échanges, achats et de concessions - la vente demeure très rare - de la part de l'abbaye et de quelques litiges jusqu'au début du XIIIe siècle. À partir de ce moment, il y a une forte augmentation des actes commandités par les membres de l'abbaye de Cluny sous la forme de lettres et échanges entre l'abbaye, l'évêché et la papauté qui sont conservés dans le corpus à partir du cartulaire C, D et E. Ces actes correspondent ainsi, dans ces deux périodes, aux affaires avec les laïcs et avec la hiérarchie ecclésiastique. Hormis la période de 950-1000, l'invocation y est très peu mobilisée par rapport à la production générale. Elle figure dans seulement huit documents entre 1050-1250.

Finalement la catégorie CL correspond aux actes commandés par des clercs, prêtres et diacres (personnages portant les titres *presbyter, levita, sacerdos. clericus, diaconus*). Il s'agit, pour la plupart, de donations *pro animae* à l'abbaye de la part de divers membres de la basse hiérarchie ecclésiastique de la région. Il y a un nombre important de chartes portant l'invocation avant le XIe siècle, mais cette proportion diminue par la suite, et les donations provenant de ces personnages ont presque disparu au siècle suivant.

Donc, pour récapituler, hormis dans les cas des diplômes royaux, la dignité du

personnage qui commande l'acte n'est pas, à strictement parler, un facteur qui détermine l'usage de l'invocation dans un acte. La plus forte concentration des invocations se trouvant dans les actes émanant d'un *quidam*, qui est la seule catégorie (avec les diplômes royaux), où on trouve un relatif équilibre entre les actes avec ou sans invocation, et la seule qui la conserve chronologiquement jusqu'à sa disparition, l'invocation étant adoptée pendant le Xe siècle par les documents émanant des autres producteurs, mais presque abandonnée dans la deuxième moitié du XIe siècle.

Il serait vain de croiser les données à propos des bénéficiaires des actes, puisqu'on va retrouver, naturellement, un univers réduit à deux catégories : AB (abbaye) et QU (quidam), ce qui correspond aux affaires conclues avec l'abbaye ou entre particuliers. En effet, si on isole les documents qui émanent d'AB et QU dans les deux tableaux afin de connaître les chiffres à propos des bénéficiaires, le rapport reste équilibré :

Actes avec inv : 84 % dirigés à AB ; 14 % dirigés à QU ; 2 % autre
Actes sans inv : 87 % dirigés à AB ; 12 % dirigés à QU ; 2 % autre

Si commanditaires et récepteurs de l'acte semblent avoir une influence limitée et équivoque dans l'emploi de l'invocation comme élément formel du discours, la faible affinité voire le rejet de l'usage invocatoire par certains types documentaires tels que notices, lettres, privilèges, actes de vente, etc. observés tout au long de la chronologie nous amènent à penser que la typologie de l'acte écrit, et par extension l'affaire ici conclue, pourraient avoir un impact plus profond dans la composition de l'acte ou en tout cas dans la mobilisation d'un certain modèle rédactionnel d'acte. Nous allons, dans le tableau ci-dessous (Table 5.5) inspecter cette question, en la limitant aux actes antérieurs à 1200 - car ensuite l'occurrence de l'invocation est infime. Nous baserons la comparaison sur les typologies majeures qui contiennent une invocation : chartes, notices et diplômes.

Inv | Action jur. / Date | Typologie		+ Inv	- Inv	+ Inv	- Inv	+ Inv	- Inv
		Donation		Confirmation		Vente/échange	
800	Charte	3	10		1	6	24
	Notice	1	2		1		3
	Diplôme	2	1	2			
900	Charte	611	731	1	3	120	246
	Notice	3	59	2	5		124
	Diplôme	16		15	1		
1000	Charte	412	573	10	9	28	77
	Notice	24	102	1	3		35
	Diplôme	2	2	4			
1100	Charte	39	71	2	4	1	2
	Notice	4	38	3	3		1
	Diplôme	8	6	6	4		
Total		1125	1595	46	34	155	512

TABLE 5.5 – Comparaison entre les ensembles avec (+Inv) et sans (-Inv) invocation selon la typologie de l'acte et le type d'action juridique accomplie.

Dans le tableau, on observe que les diplômes et préceptes, naturellement, ne portent jamais sur des ventes ou échanges et sont répartis de manière équivalente entre donations et confirmations, qui sont souvent associées à une même action juridique. L'invocation est très usitée dans les diplômes jusqu'au XIIe siècle, et pratiquement systématique entre le IXe et le Xe siècle. Puis le diplôme, et donc l'invocation dans le diplôme, commence à se raréfier dans notre corpus et l'invocation est finalement abandonnée sous Philippe le Bel (1285-1314). Par ailleurs, comme on le voit, les typologies décrivant ventes et échanges emploient l'invocation, mais bien moins fréquemment que les donations : les notices sur des ventes et échanges ne portent quasiment jamais d'invocation (il existe quelques rares exemples non présentés dans ce tableau), et les chartes sans invocation sont deux fois plus rares que celles qui en comportent, tout au long de la chronologie.

Dans la donation ce rapport est plus équilibré, hormis dans les notices, où l'invocation reste présente dans quelques groupes de chaque siècle étudié. La vraie relation d'équilibre se trouve alors dans la charte de donation où les invocations se trouvent réparties dans un ensemble numériquement proche (rapport 4 : 5) de celui sans invocation, les deux ayant une progression similaire jusqu'à la fin du XIIe siècle.

Ainsi, l'identité de l'auteur et du bénéficiaire ont peu d'influence quant à l'usage de l'invocation, si ce n'est dans les actes datant d'avant la première moitié du Xe siècle, où la noblesse et les abbés adoptent pour leurs documents les plus solennels l'invocation divine et trinitaire empruntée souvent à l'acte souverain. Ultérieurement, l'utilisation de l'invocation est déterminée presque exclusivement par d'autres facteurs, le plus important étant la typologie de l'acte et le modèle auquel il se rattache.

Notices, lettres et bulles négligent l'invocation. La notice ne se débarrasse pas seulement d'elle mais également de la plupart du protocole et des formules du dispositif, bien que cela n'empêche pas de retrouver quelques ensembles la portant - on verra plus loin lesquels. Pour leur part, les bulles, privilèges et lettres, coulées dans un moule diffèrent de celui de la charte, ne l'incluent que rarement dans leur tradition dès leur origine, la faisant figurer dans des actions juridiques d'exception (quelques litiges et testaments).

5.4.2 Les invocations dans la charte de donation

L'invocation est alors un élément intimement lié à la charte, mais celle-ci étant si répandue, les situations d'usage sont multiples. Les chartes de ventes et échanges peuvent utiliser l'invocation mais de façon minoritaire. La charte de donation, elle, fait un usage très prolifique de l'invocation, qui est présente dans 45 % des actes écrits, et particulièrement pendant les Xe et XIe siècles. Ensuite, lorsque la charte commence à disparaître et à adopter des nouvelles formulations durant le XIIe siècle, l'invocation tombe en désuétude.

Il existe, bien évidemment, d'autres facteurs impliqués dans l'usage de l'invocation, mais il faudra explorer plus avant la charte de donation, qui constitue le noyau et l'ensemble le plus large de notre corpus (environ 2 400 documents) et dont les changements affectent sérieusement le graphique général du corpus.

Nous avons donc concentré l'analyse sur les chartes provenant d'un *quidam* (QU)

ayant comme bénéficiaire l'abbaye de Cluny (AB) ou un autre *quidam* (QU). Cela limite l'ensemble à plus ou moins deux mille documents (∼ 38 % du corpus), et nous offre un tableau qui permettra de bien mettre en évidence les variations du fait de son homogénéité. Il est en effet composé d'actes ayant les mêmes chronologie, typologie, action juridique, commanditaires et bénéficiaires.

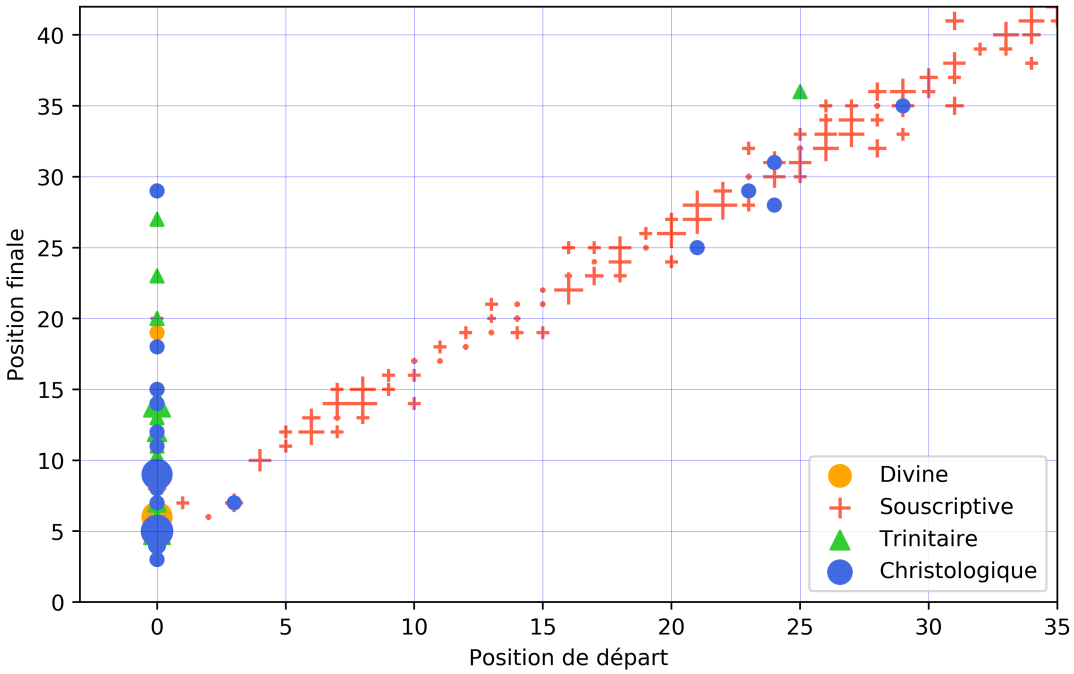

FIGURE 5.5 – Place des invocations dans les chartes privées de donation. Légende : les invocations sont ici représentées suivant leur place dans l'acte, en fonction du nombre de mots qu'elles occupent entre leur début (nommé « position de départ ») et leur fin (nommé « position finale ») du document.

Une première analyse consiste à relever les invocations employées dans ces chartes et les classer selon les quatre modèles définis plus haut. La répartition est comparable à celle observée au niveau du corpus :

Invocation christologique : 274 documents, 31.5 %)
Invocation trinitaire : 87 documents. 10 %)
Invocation au nom de Dieu : 75 documents, 8.5 %)
Invocation dans la suscription : 434 documents, 50 %)

Le seul point discordant est l'invocation trinitaire qui représente 18 % du total d'invocations dans le corpus, mais seulement 10 % dans la charte de donation ; cette différence est bien expliquée par le fait que cette invocation est surtout utilisée dans les diplômes, comme on l'a vu plus tôt, et que ceux-ci ne sont pas représentés dans ce graphique.

Un point intéressant est révélé par la place occupée par chaque invocation dans la charte, en s'intéressant tout d'abord à la longueur de celle-ci : nous nous concentrons

donc sur la représentation verticale en position de départ 0. Les invocations au nom de Dieu, trinitaires et christologiques n'apparaissent, à quelques exceptions près, qu'en tête de charte, leur position de départ étant donc 0. Il existe une très forte concentration de ces trois invocations ayant une position finale comprise entre 0 et 8, ce qui correspond au développement normal des types les plus simples d'invocation en tête de charte, comprenant un nombre réduit de mots, tels que (*In nomine Verbi incarnati, In nomine Sanctae et individue Trinitatis*). Le groupe d'invocations ayant une position finale entre 10 et 15 est moins dense et correspond aux sous-versions plus longues qu'on a renseigné auparavant telles que : (par ex. *In nomine Dei Summi Patris, et Filii et Spiritus Sancti.*), (voir 6.2). Finalement, les invocations finissant entre les positions 16 et 28 correspondent à des exemples particulièrement longs et des invocations uniques, par exemple fusionnées avec d'autres formules.

L'invocation dite dans la suscription peut aussi apparaître en tête de charte comme ici : "*Ego (Nos), in Dei nomine,*". Il y a 79 documents qui commencent ainsi et dans dix cas la conjonction (*quapropter*), usage de formulaire, est présenté comme premier mot. L'invocation de suscription ne devrait même pas être considérée comme faisant partie du protocole puisqu'elle constitue en réalité le début du dispositif ; elle conduit immédiatement à l'exposition du fait de l'action juridique et il n'y a presque jamais d'autres formules au milieu, exception faite de la formule de motivation (*pro anima mea, pro redemptionis anima*, etc.). Ce rapport entre le dispositif et l'invocation est très normé dans les chartes de donation. Lorsque l'invocation non-suscriptrice apparaît en tête de charte, on ne voit pas souvent mobilisées d'autres parties protocolaires, exception faite de la notification, normalement dans sa version universelle (*In Dei nomine, notum sit omnibus quod*) ; et évidemment l'utilisation de l'invocation de suscription est liée à l'apparition, auparavant, d'autres parties du discours à différents degrés de complexité.

Le graphique (figure 5.5) illustre ainsi la place occupée par l'invocation par suscription dans le protocole, en s'intéressant aux différentes positions de départ qu'elle peut occuper, et à sa longueur (figurée par la coordonnée entre position de départ et position finale). La plupart de ces invocations démarrent après la position 5 de la charte (soit après 5 mots). Quand elle débute entre les positions 5 et 15 elle est normalement précédée soit par une adresse du type : *Dilecta uxore mea, nomine*, soit par une notification individuelle du type : *notum sit omnibus quod*. La majorité des invocations par suscription débute entre les positions 15 et 35 : ici l'invocation est précédée par différentes parties du discours, normalement la notification à l'abbaye du type : "*Sacrosancto et exorabili loco Cluniensi cenobio, in honore beatorum apostolorum Petri et Pauli consecrato...*" dans les deux modèles très formulaires : "notification + adresse" et parfois de courts préambules. Plus l'invocation est précédée par d'autres formulations, plus les conjonctions et adverbes de coordination *igitur* et *quapropter* prennent place comme premier mot de l'invocation. Les invocations de suscription débutant après la position 36 n'apparaissent pas dans le graphique. Il s'agit d'une cinquantaine d'invocations : certaines mobilisent l'outillage complet des parties protocolaires, et d'autres consistent en des exposés et récits riches en citations bibliques.

La version de suscription est d'ailleurs complètement inflexible dans sa formulation.

Dans quelques rares cas le scribe ajoute un attribut au nom de Dieu (*summus, omnipotens*), mais en général *Ego, in Dei nomine* est la seule version observée. En fait, lorsque l'invocation christologique et trinitaire sont utilisées en tant qu'invocations de suscription, comme on le voit dans le graphique, elles le font adoptant ce même modèle : *Ego, in Christo nomine et Ego, in Trinitatis nomine.*

5.4.3 Les chartes de donations sans invocation

Le portrait du protocole diplomatique auparavant esquissé correspond à l'ensemble de 870 chartes de donation portant une invocation. Est-ce que ces situations sont répliquées dans l'ensemble de 1 079 chartes de donation qui ne portent pas d'invocation ? Quelques mesures statistiques peuvent être ici très éloquentes pour une analyse de corrélation. Dans le tableau ci-dessous (table 6.6) nous étudions la longueur des textes dans 3 groupes de chartes de donation : chartes avec invocation régulière (415 items), chartes avec invocation dans la suscription (455 items) et chartes sans invocation (1079 items).

groupes/ statistics	(1)	(2)	(3)
moyenne	159.8	204.2	203.2
var. std	76.9	113.1	138.2
%ile 25	110.3	134.7	138.0
%ile 50 (médiane)	146.0	185.0	179.0
%ile 75	190.75	244.0	237.0

TABLE 5.6 – Mesures statistiques selon la taille des textes. (1) : actes avec invocation régulière ; (2) : actes avec invocation dans la suscription ; (3) : actes sans invocation.

La longueur de la charte peut sembler un indicateur très relatif de sa composition, mais étant donné que nous avons porté la comparaison sur des ensembles par ailleurs très homogènes, cette différence de longueur peut nous permettre dégager rapidement une constatation : les documents sans invocation (3) et ceux portant l'invocation dans la suscription (2) coïncident dans toutes les mesures, alors que les chartes portant une invocation (1) sont beaucoup plus éloignées. De fait, si nous portons la comparaison au niveau des parties du discours sur les ensembles proches de la médiane, souvent la seule différence entre les documents du groupe 2 et ceux du groupe 3 est justement l'invocation :

Voici quelques-uns des exemples, parmi les plus significatifs, de chartes des groupes 2 et 3, respectivement, datées de la même décennie :

> ['*Sacrosancto et exhorabili loco in honore Dei et beatorum apostolorum Petri et Pauli consecrato, in pago Matisconense, cui preest domnus Hemardus venerandus abba. Igitur ego Teotbertus,* ***in Dei nomine****, dono predicto loco...*']407

407. CBMA 2183, an 950

['*Sacrosancto et exorabili loco Cluniaco, in honore beatorum apostolorum Petri et Pauli, ubi preest domnus Emardus abbas, ego Guilerada dono aliquid ex rebus meis....*'] [408]

['*Divina largitate sanccitum est ut de rebus transitoriis æterna valeat merces promereri. Quapropter ego,* ***in Dei nomine****, Bodo, venturi judicii examen precavens....*] [409]

['*Divina largitate sanccitum est ut de rebus transitoriis æterna valeat merces promereri. Quapropter ego Erveus, pro remedio animæ meæ ....*] [410]

Les 40 mots qui séparent la médiane et la moyenne entre le groupe 1 et les groupes 2 et 3, correspondent en grande partie, à des éléments de formalisation protocolaire, notamment exposés et préambules, dont les chartes de groupe 1 sont normalement dépourvues. Le formulaire employé dans le groupe 1 nous propose un style de charte relativement simple et objectif portant un protocole constitué de façon récurrente par l'invocation ou par la formule invocation + notification ; d'un dispositif occupant la moitié ou plus de la charte, munie uniquement des clauses indispensables de sanction et d'opération, et un eschatocole assez plat qui ne recueille dans la plupart des cas que les signa des témoins et, le cas échéant, la souscription. Les structures des groupes 2 et 3 partagent ce schéma, mais en nous proposant un développement plus riche des parties essentielles et des protocoles plus complexes, apportent plus d'informations.

Voici quelques exemples qui synthétisent les manières de développer le protocole dans le groupe 1 :

['*In nomine Verbi incarnati. Ego igitur Heldinus, recogitans eterne remunerationis gloriam, dono Deo et sanctis apostolis ejus Petro et Paulo, et ad locum Cluniacum...*'] [411]

['*In nomine Dei summi. Noverint cuncti tam presentes quam etiam futuri, quod ego Petrus, pro remedio animæ meæ dono Deo et sanctis apostolis ejus Petro et Paulo, et ad locum Cluniacum...*'] [412]

['*In nomine Verbi incarnati. Notum sit cunctis christianis fidelibus, quod nos Rodbertus videlicet et Wichardus donamus Deo et ejus apostolis Petro et Paulo...*'] [413]

['*In nomine sancte et individuæ Trinitatis. Pateat omnibus fidelibus qualiter ego Teutbertus animarum Salvatori obediens dicenti : «Facite vobis amicos de Mammona iniquitatis,» et cetera, dono Deo et sanctis apostolis eius Petro et Paulo...*'] [414]

408. CBMA 2259, an 953
409. CBMA 2196, an 950
410. CBMA 1983, an 942-954
411. CBMA 2377, an 990
412. CBMA 2549, an 993-996
413. CBMA 2589, an 1015
414. CBMA 2599, an 1005

5.5 Les modèles rédactionnels

L'observation antérieure nous renvoie à plusieurs éléments de réflexion au sujet de ces différents types de chartes et du choix du formulaire employé par le scribe. Tout d'abord l'invocation dans la suscription est l'invocation la plus ancienne, et celle qui apparait comme de règle dans les formulaires anciens comme celui de Marculf et les formulaires angevin et de Tours. Ainsi, ce type d'invocation est typique des formulaires employés couramment pour les chartes de donation à l'abbaye tels que les modèles sous les rubriques *Cessio ad loco sancto, Cessio a diae presentae ad ecclesiam, Donatione de parva rem ad acclesia*, etc. ainsi que des modèles correspondant aux donations et échanges entre particuliers [415]. L'invocation en tête de charte n'y est proposée que dans très peu de modèles et toujours dans ceux en relation avec les autorités publiques.

En deuxième lieu, les chartes de donation ne portant pas d'invocation et celles qui portent l'invocation de suscription utilisent un style de charte très proches, qui nous autorise à les considérer comme un ensemble homogène. Elles portent le même schéma formulaire et développent une structure discursive, qui sauf en ce qui concerne la présence de l'invocation, est souvent indistinguable. A contrario, les chartes débutant par l'invocation utilisent un schéma moins riche et le développent en suivant un ordre diffèrent.

En troisième lieu, le groupe des chartes de donation, entre le Xe et le XIe siècle, qui débutent par l'invocation nous renvoie directement au style de la notice. Cela n'est pas seulement dû à leur apparente simplicité mais aussi à l'utilisation récurrente de la notification associée directement à l'invocation, marque habituelle de la notice. Ces chartes sont rédigées dans un style subjectif mais suivant un cadre proche de la notice, qui, elle, propose un style objectif, tandis que les deux autres groupes empruntent leurs structures aux modèles traditionnels et stéréotypés de la charte de donation utilisés bien avant et ailleurs qu'à Cluny.

Ces distinctions observées au sein des chartes de donation à partir de l'usage et du style de l'invocation peuvent se montrer encore plus profondes et riches si nous élargissons la portée de la comparaison incluant les autres parties du protocole. Dans le graphique qui suit (Figure 5.6), et en guise de résumé statistique, on peut rapidement visualiser les grandes solutions rédactionnelles dans l'agencement initial privilégiées par chaque groupe.

On observe en particulier que les chartes avec invocation à leur tête présentent, comme on l'a déjà signalé, une association étroite avec la notification, et lorsque celle-ci est absente, l'invocation est connectée directement avec la suscription. Il s'agit alors d'un style très direct et objectif de charte. Ces deux compositions représentent 90 % du total. Il y a une vingtaine de cas de double invocation, c'est à dire de chartes portant une invocation en tête et une dans la suscription, cette dernière étant alors comme accessoire. Enfin, il y a quelques documents où une invocation différente de celle de la suscription (*ego, in Dei nomine*) est placée dans la charte après un préambule ou une adresse collective. Lorsque l'invocation est en tête, ces parties ne sont pas

415. Karl Zeumer. *Monumenta Germaniae historica : Formulae Merowingici et Karolini aevi accedunt ordines iudiciorum Dei*. Impensis Bibliopolii Hahniani, 1886, p. 53, 75Marculfi formularum, liber I, 12 ; Liber II, 4,6

mobilisées, mais en quelques occasions le scribe utilise entre l'invocation et le dispositif à proprement parler quelques formulations qui rappellent une adresse ou une narration, mais très courtes et rédigées dans un style subjectif.

En ce qui concerne les chartes portant l'invocation de suscription, elles sont plus riches en solutions rédactionnelles. Les deux tiers (65 %) d'entre elles s'ouvrent soit par un préambule, soit par l'adresse collective à l'abbaye, parties qui apparaissent très peu dans le cas des chartes avec l'invocation en tête. De ce fait le style de ces chartes est perçu comme bien plus élaboré et formalisé. Ces chartes monopolisent aussi l'adresse personnelle des donations entre individus privés - normalement de la même famille. Comme dans les formulaires anciens, l'invocation de suscription peut ouvrir la charte et connecter directement avec le dispositif, mais c'est une solution peu employée (18 %). D'ailleurs, la relation entre cette invocation et la notification n'est pas aussi étroite que dans les autres invocations. La notification, qui porte souvent une adresse universelle, n'est plus nécessaire puisque ce type de chartes a déjà une adresse personnelle ou collective. Enfin, vers la fin du XIIe lorsque les chartes introduisent la date de temps et de lieu dans le protocole, l'invocation est déplacée au deuxième rang car la date occupe la tête de charte.

Finalement, si les chartes avec une invocation en tête se montrent en général dépourvues des préambules et des parfois très longs adresses et exposés présents dans les autres groupes de chartes, est-il possible d'envisager que dans certains cas il s'agisse du résultat de l'intervention des copistes ultérieurs du cartulaire ? Il n'est en effet pas rare de trouver des marques d'abrègement dans les têtes de charte lorsque le copiste, qui connaît par cœur les formules, se fatigue de les copier et les abrège avec un "etc." ou ajoute un "ut supra". La statistique semble étayer cet argument, montrant une différence mais modeste entre le cartulaire et sa copie :

Chartes avec invocation en tête
Originaux : 33 (7.5 %), Copies : 177 (40.5 %), Cartulaires : 226 (52 %)

Chartes avec invocation suscriptrice
Originaux : 31 (7 %), Copies : 215 (49.5 %), Cartulaires : 189 (43.5 %)

Chartes sans invocation
Originaux : 101 (9.5 %), Copies : 463 (43.5 %), Cartulaires : 501 (47 %)

L'origine des chartes avec invocation en tête est en effet plus tributaire des cartulaires que les autres deux groupes, mais sans qu'on puisse en être certain. Chronologiquement les trois groupes sont distribués de manière similaire, comprenant des chartes écrites avant et après le période des cartulaires. Par ailleurs, il ne faut pas oublier que dans certains cas, l'invocation en tête, notamment dans le cas de l'invocation divine, peut-être une restitution du cartulariste qui interprète ainsi l'invocation monogrammatique dessinée dans les originaux, mais l'édition de Bruel néglige malheureusement presque toujours l'information relative à cet élément. Et les sondages que nous avons effectués n'appuient pas l'argument. Ainsi, s'il est possible que pour certains actes le style notificatif remarqué dans les chartes du groupe 1 ait son origine dans la copie du cartulaire, ceci semble bien moins décisif que les choix stylistiques du modèle adopté par le scribe.

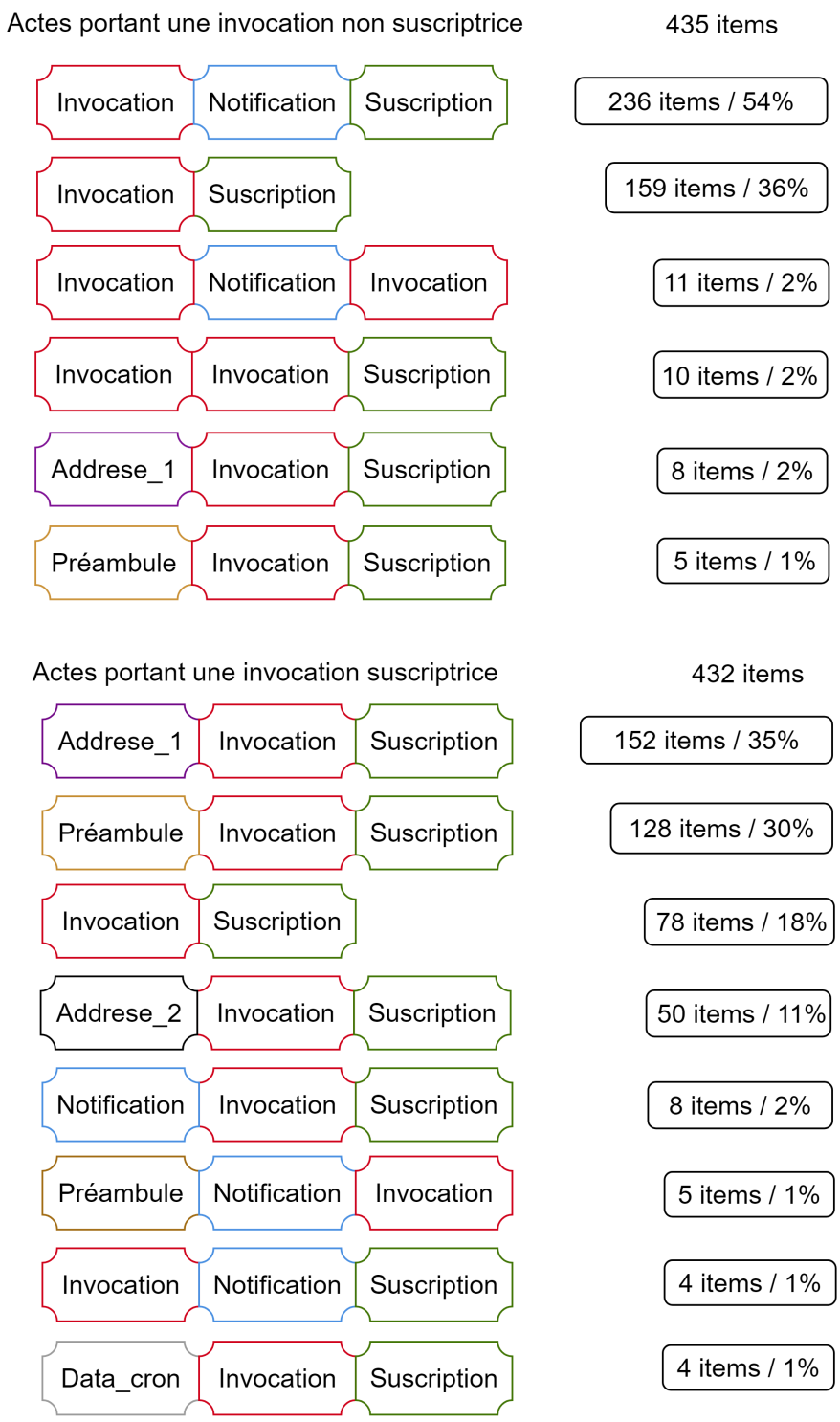

FIGURE 5.6 – Solutions rédactionnelles dans les protocoles des chartes de donation. Adresse_1 : adresse collective ; Adresse_2 : adresse personnelle ; Data_cron : Date chronique. Le préambule inclut préambules et narrations.

Enfin, devant ce panorama nous proposant deux styles pour transcrire le même genre de document, le plus logique - et ce qui semble plus en accord avec la tradition des actes de Cluny - est de les rapporter à des évolutions naturelles dans les pratiques de l'écriture, notamment en ce que concerne la rédaction des actes les plus fréquents à partir du Xe siècle. Il ne s'agit pas d'une confrontation entre deux styles de rédaction, mais de l'adoption progressive d'une solution rédactionnelle développant la matière centrale de l'affaire conclue et ne contenant que les détails indispensables. L'invocation, qui pourrait sembler plus en accord avec un style d'acte très formalisé, devient ainsi un marqueur assez expressif du réductionnisme des parties protocolaires qui rapproche beaucoup la charte du modèle utilisé par la notice. Sans doute le rang des personnes ou les traditions opérant dans les régions jouent un rôle important dans le développement du modèle utilisé, mais moins influent, comme nous l'avons constaté, que ce que l'on peut estimer. C'est en réalité le croisement entre la fonction juridique de l'acte et l'affaire conclue ce qui en semble le plus décisif.

5.6 Conclusion

Le travail que nous avons effectué, consistant à récupérer et regrouper les formules invocatoires, puis à étudier la corrélation existant entre l'usage de l'invocation et les caractères et typologies des actes nous amènent à deux conclusions :

Tout d'abord, le classement automatique des invocations par modèles nous permet de distinguer rapidement les formulations qui ne sont pas régulières ou qui en tout cas s'éloignent des formes invocatoires normalement mobilisées. L'invocation est une formule brève qui répond à une composition très inflexible. Les modifications sur les formules et les formules *sui generis* occupent une place très réduite. Dans la plupart des cas, ces modifications répondent à des situations assez spécifiques dans la production de l'acte : une origine étrangère (notamment espagnole), une affaire autre que le transfert foncier, ou la présence d'éléments n'appartenant pas au formulaire, telles que les citations bibliques. Derrière ces opérations, il existe souvent un scribe ou une institution qui reprend ces formulations au cours de la même décennie dans un *scriptorium* déterminé.

En deuxième lieu, la chronologie de l'invocation nous indique un usage assez répandu, depuis le IXe siècle jusqu'aux alentours de l'an mille. Notre vision est relativement biaisée par la production d'actes à Cluny dans la première moitié du XIe siècle, mais cela n'empêche pas de constater un fort déclin des usages invocatoires qui disparaissent au début du XIIe siècle avec la charte de donation privée. Les scribes qui mobilisent l'invocation se montrent très attachés aux quatre formes invocatoires topiques et les altèrent à peine dans les actes. Cependant, les usages invocatoires, sauf dans les cas d'invocation trinitaire pour les diplômes, ne sont en général pas proposés dans les modèles de charte des formulaires altimédievaux. Cette question épineuse nous a amené à faire une enquête plus en profondeur distinguant les chartes de donation portant une invocation en tête, la plus commune, de celles portant une invocation dite dans la suscription, usage qui est occasionnellement proposé dans les formulaires. Les résultats montrent que les premières chartes, émises entre les Xe et XIe siècles, s'éloignent des protocoles formulaires et se rattachent à un modèle proche

de la notice, alors que les secondes, dans la même chronologie, possèdent des protocoles plus denses en accord avec les formulaires anciens et développent les mêmes solutions rédactionnelles que celles attestées dans les chartes sans invocation.

Ainsi, l'étude de la partie du discours la plus inflexible et brève nous apporte déjà plusieurs éléments qui nous renvoient à des différences bien marquées entre deux modèles de développement d'un protocole, à l'intérieur d'une même typologie, auxquels les scribes s'attachent très fortement. Les sondages que nous avons effectués sur les autres parties du protocole, notamment notifications et adresses, semblent étayer ce raisonnement, mais sans doute une étude plus en profondeur sur ces parties apportera d'autres certitudes pour nuancer cette conclusion. L'information portant sur le type d'action juridique et les solutions rédactionnelles que notre modèle automatique peut détecter dans les protocoles semble suffire pour proposer des classements assez précis des actes.

Chapitre 6

Le vocabulaire de l'espace

6.1 Introduction

Dans l'étude qui suit, nous allons analyser le vocabulaire spatial présent dans les actes des recueils de l'abbaye de Cluny et de Saint-Vincent de Mâcon. Cette étude se concentre sur les dix termes les plus courants accompagnant les entités nommées géographiques, en tant que cooccurrences. À l'aide de différentes matrices statistiques et de techniques de distribution sémantique, on essayera d'apporter une définition précise du rôle joué par chaque terme à l'intérieur des formules de spatialisation. Puis, nous décrirons les différentes mutations des cadres territoriaux de localisation auxquels se rapportent les scribes ainsi que les évolutions topo-spatiales reflétées dans les actes. Ensuite, à l'aide de la version numérique du dictionnaire topographique de Saône-et-Loire, nous proposerons quelques recompositions cartographiques au niveau de l'*ager*, unité intermédiaire du tissu spatial, pour illustrer les différents états du processus de mutation du paysage.

L'étude est divisée en trois parties. Dans la première partie, nous avons fait une présentation sommaire de la structure notariale utilisée par les actes au moment de réaliser les descriptions foncières. Dans la deuxième partie nous présentons les différentes matrices d'exploitation des champs sémantiques autour des termes spatiaux ainsi que la routine de pré-traitement sur les entités nommées. La troisième partie quant à elle présente une étude chrono-spatiale des cadres territoriaux en s'appuyant sur les résultats obtenus auparavant et quelques recompositions géographiques.

6.2 La description foncière dans le dispositif des actes

La plupart des milliers d'actes conservés dans le recueil de l'abbaye de Cluny, et en général dans toute la base CBMA, datés de la période antérieure au XIIe siècle, correspondent à des enregistrements de mutations foncières. Les collections et les cartulaires recueillent les actes qui attestent la réception ou l'acquisition de divers biens, droits et biens-fonds par l'abbaye de la part de particuliers ou des agents du pouvoir public, sous la forme d'une donation, le plus souvent, ou d'une vente, d'un échange, précaire ou d'un déguerpissement. Des titres antérieurs à la fondation de l'abbaye ou ne la concernant pas directement sont aussi conservés dans les recueils

puisque l'abbaye considère parfois utile de reproduire l'histoire juridique des biens entrés dans son patrimoine. Ces actes de transfert de la terre suivaient de façon assez stricte les modèles formulaires qui depuis le Haut Moyen Âge prescrivaient les manières de traduire une action juridique dans un acte écrit en conformité avec le droit et la tradition. Ils peuvent présenter des développements très divers pour un large spectre de raisons, mais en général ils devaient répondre à deux exigences légales de premier ordre : la localisation des biens aliénés, et la pratique descriptive et énumérative de la nature et de la composition de ces biens. Pour bien accomplir cette tâche, le scribe mobilise un vocabulaire et une série de coordonnées spatiales qui, même s'ils ne sont pas d'actualité pour lui, lui servent à communiquer avec plus ou moins de précision la localisation et l'inventaire du bien-fonds donné, vendu ou racheté.

Occupant la place centrale du dispositif des actes, ces pratiques nous transmettent une information spécifique et évènementielle précieuse qui devrait représenter un ou plusieurs points géolocalisables dans le tissu historique d'un anthroposystème perdu. La haute densité tant des termes du vocabulaire spatial que des entités géographiques dans les dispositifs nous l'indiquent et peuvent nous convaincre de tenter une recomposition topo-spatiale en tirant profit des nouvelles technologies. Une même ambition animait déjà des historiens comme M. Chaume et A. Déleage qui ont dédié des vastes volumes au classement et à la désambiguation des toponymes trouvés dans les actes du Recueil tout comme à la reconstruction des systèmes de référencement géographique [416].

Mais comme ils l'ont montré, ainsi que l'ont également fait plusieurs études depuis la thèse de G. Duby [417], il existe diverses raisons, inhérentes au mode de rédaction des actes, pour lesquelles on ne peut pas toujours prendre ces indications de localisation spatiale pour des points géographiquement bien déterminés, ni ces inventaires pour des descriptions bien conformes avec la réalité, ce qui représente un obstacle majeur dans cette démarche visant à des reconstitutions au niveau du cadastre.

Tout d'abord, le système de référencement utilisé par les scribes se trouve à la croisée de deux réalités : l'une disparue - ou en passe de s'estomper - qui se rapporte à l'encadrement territorial de l'Antiquité tardive et dont la définition et les frontières sont difficiles à cerner car elles ne sont plus soutenues dans la réalité ; et l'autre, encore d'actualité, vérifiable sur le terrain, mais qui est décrite en employant un vocabulaire souvent imprécis qui définit les petites unités agricoles et les bâtiments d'exploitation ou de culte qu'elles abritent.

En deuxième lieu, tant que les scribes gardent ce système à deux réalités, soutenu en plus par le formulaire, d'autres formes de distribution de l'espace et d'autres processus d'ordonnancement social du territoire prennent place au fil des trois siècles

416. Maurice Chaume. *Les Origines du duché de Bourgogne : 2ème partie, Géographie historique.* E. Rebourseau, 1927, étude topographique du *pagus Matisconense* aux pages 1014-1170 ; André Déléage. *La Vie économique et sociale de la Bourgogne dans le haut moyen âge : thèse pour le doctorat ès lettres présentée à la Faculté des lettres de l'Université de Paris, par André Déléage.* Protat frères, 1941

417. Voir à ce sujet : Duby, *La Societe aux Xle et Xlle siecles dans la region maconnaise* ; Rosenwein, *To be the neighbor of Saint Peter : the social meaning of Cluny's property, 909-1049* ; Bange, "L'ager et la villa : structures du paysage et du peuplement dans la région mâconnaise à la fin du Haut Moyen Age (IX e-XI e siècles)"

que couvrent les actes du recueil de l'abbaye de Cluny. L'espace dépourvu d'un bornage clair et même de frontières administratives, caractéristique de la période altimédiévale, souffre de tentatives successives de réarrangement en rapport avec l'essor des diverses juridictions imposées par les anciens puis les nouveaux pouvoirs locaux, certains durables, d'autres plus éphémères. Cette nouvelle territorialisation de l'espace ne s'impose pas seulement dans la réalité mais aussi dans les actes, qui reflètent les bouleversements dans la conscience et la perception spatiale des hommes [418]. Cela affecte surtout les unités supérieures à la *villa* dont les limites réelles ou supposées se chevauchent dans le temps et l'espace, générant de nombreuses erreurs dans la détermination des biens fonciers de la part des scribes. La majorité des imprécisions, doublons, et superpositions qui avaient rendus très incohérents, aux yeux des historiens des deux derniers siècles, les formes de localisation employées dans les chartes, ont ici leur origine.

Si les frontières supérieures se montrent globales et imprécises, les frontières dans les structures inférieures, vers lesquelles se dirige la nouvelle réorganisation de l'espace, se montrent denses et dépourvues de bornage. Les cellules fondamentales du terroir, la *villa* et le *locus*, ne sont ni homogènes ni extensives. C'est un caractère qui définit le domaine foncier bourguignon : le sol se trouve réparti entre plusieurs maîtres de la terre qui ont de surcroît un héritage dispersé dans plusieurs *villae*. Cette profonde fragmentation de la domination de la terre nous empêche, sauf dans de rares cas, de mieux définir des reconstructions géographiques plus précises. Ceci était effectivement un problème pour les rédacteurs. L'acte de donation est souvent précédé d'une visite sur le terrain afin de mesurer le bien-fonds à donner et d'établir les frontières locales par rapport au voisinage, et d'enquêter sur sa composition en termes de richesse (*inquisitio*). En l'absence de cartes géographiques et de cadastres officiels auxquels se rapporter, la détermination spatiale est improvisée, fortement dépendante de la connaissance directe et de l'auto-gestion des hommes [419].

Finalement, bien que les trois points précédents relèvent des situations d'actualité pour les donateurs et les scribes, il existe d'autres altérations dans l'information transmise par les actes. Dans le cas de Cluny, une bonne partie des actes a été transmise dans le contexte d'un cartulaire et, en conséquence, nous allons parfois croiser des modifications a posteriori de la main des copistes. Parfois ces modifications sont intéressées : l'abbaye cherche à uniformiser son patrimoine, à réarranger quelques

418. Sur l'organisation du territoire et de l'habitat haut-médiéval et les structures d'origine romaine voir : Benoît Cursente. "Autour de Lézat : emboîtements, cospatialités, territoires (milieu X-milieu XIII siècle)". In : *Les territoires du médiéviste. Rennes, Presses universitaires de Rennes* (2005), p. 151-167 ; François Favory et al. "Le territoire : un modèle de l'organisation de l'espace en archéologie rurale ; étude de cas dans la cité antique de Nîmes". In : *Habitat et société, actes des XIXe rencontres internationales d'archéologie et d'histoire d'Antibes* (1998), p. 499-518 ; Édith Peytremann. "Les structures d'habitat rural du haut Moyen Age en France (Ve-Xe s.). Un etat de la recherche"'. In : *Lorren and Perin (eds), l-28* (1995)

419. À propos des cadastres et parcellaires voir : Gérard Chouquer. "Aux origines antiques et médiévales des parcellaires". In : *Histoire & sociétés rurales* 4 (1995), p. 11-46 ; Monique Bourin. "Délimitation des parcelles et perception de l'espace en Bas-Languedoc aux 10è et 11è siècles". In : *Campagnes médiévales : l'homme et son espace. Etudes offertes à Robert Fossier* (), p. 73-85 ; Jean-Loup Abbé. "Permanences et mutations des parcellaires médiévaux". In : *G. Chouquer, sd, Les formes du paysage* 2 (1996), p. 223

structures, à donner la prééminence à quelques centres et en oublier des autres. Mais le plus souvent les altérations sont dues à la méconnaissance des scribes d'une réalité dont l'empreinte est déjà très faible ou disparue au moment de la rédaction des cartulaires. Ainsi, le copiste, confronté à un récit spatial presque oblitéré de sa conscience, hésite, corrige ou même abrège des indications et des descriptions pour lesquelles quelques originaux peuvent se montrer plus riches, dans l'intérêt de transmettre uniquement la partie dispositive de l'acte, celle qui lui donnait sa validité comme preuve de droit [420].

Or, ce qui représente un problème pour situer géographiquement les points spatiaux des actes, représente également une riche carrière pour l'historien soucieux d'embrasser toutes les dimensions cohabitant dans une même réalité historique. Si les sources nous empêchent de reconstituer le maillage foncier avec précision car elles sont lacunaires et partielles, elles ouvrent par contre la possibilité de cerner les réalités entremêlées dans un riche période de transformation du territoire et d'appréhender sa conception dans la mentalité des hommes. Pour y arriver, néanmoins, un dépouillement minutieux de tous les points spatiaux disponibles dans les actes est nécessaire. Cela ne consiste pas seulement en l'élaboration des *index locorum*, dont l'historien regrette l'absence dans l'édition du Recueil de Bernard et Bruel [421], mais aussi la récupération de toute l'expression textuelle concernant la localisation géographique dont le toponyme est le centre. Ceci n'est en rien nouveau. Les historiens ont traditionnellement pris tous les points spatialement déterminés et les ont disposés dans une série d'observations au fil des siècles dans le but de découvrir les grandes lignes des mouvements fonciers. Mais ceci était un travail qui leur a coûté beaucoup de temps pour développer un nombre toujours très limité d'observations.

À présent, les tâches qui concernent le repérage et le classement des points spatiaux tout comme leur contexte textuel sont automatisables et le nombre d'observations qu'on peut tirer de cette information s'est multiplié sur la base des contenus structurés. Nous allons ainsi profiter de l'identification des entités nommées pour interroger à grande échelle le vocabulaire qui leur est associé (co-occurrences) et les développements formulaires dans lesquels elles s'insèrent dans des pratiques de l'écrit. Ceci correspond à une exploitation séquentielle des actes qui coïncide avec le parcours technique suivant :

1. Extraction des descriptions foncières dans les dispositifs ;
2. Détermination des entités géographiques, du vocabulaire et des systèmes de référencement spatial utilises par les scribes ;

420. Nous pouvons invoquer d'autres raisons de moindre impact qui contribuent à ce même résultat, par exemple, les nombreux cas d'homonymie ; le dépeuplement et la disparition de nombreux lieux, ce qui implique souvent la disparition du micro-toponyme ; l'absence des informations concernant l'alleu (techniquement opposé à la mutation), etc.

421. Un tome final incluant les index personarum et locorum dans l'édition de Bernard et Bruel avait été projeté mais il n'a jamais vu le jour (Atsma et Vezin, "Autour des actes privés du chartrier de Cluny (Xe-XIe siècles)"). A. Bernard a néanmoins publié partie de ses travaux d'indentification dans un annexe de l'édition du cartulaire de Savigny (Auguste Bernard. *Cartulaire de l'abbaye de Savigny : suivi du petit cartulaire de l'abbaye d'Ainay*. T. 2. Impr. impériale, 1853). Le projet avait été repris pour J. Richard dans les années 60 et postérieurement par le groupe de travail de Münster mais sans que nous sachants de l'existence d'un produit terminé, exception faite des travaux publiés dans les volumes de l'Orbis Latinus (Johann Georg Theodor Graesse et Friedrich Benedict. *Orbis latinus : Lexikon lateinischer geographischer Namen des Mittelalters und der Neuzeit*. T. 1. Klinkhardt & Biermann, 1972

3. Représentation des relations sémantiques entre les termes et leurs évolutions ;
4. Représentation chronologique et cartographique des unités d'encadrement spatial.

6.3 Localisations et inventaires des biens-fonds dans les actes

La localisation et l'inventaire des biens-fonds constituent le centre du dispositif des actes et sont assez récurrents entre les IXe et XIe siècles. L'indication géographique plus ou moins précise se rapportant au système de découpage du territoire est naturellement complétée par la définition des limites de la terre cédée au bénéfice de l'abbaye, de la place géographique qu'elle occupait dans le parcellaire local, de son orientation, de leurs mesures (en perches ou *perticae*) et enfin de sa valeur en monnaie locale. Ces deux pratiques répondent à des usages de formulaire, tributaires des pratiques notariales, qui suivent une structure assez répétitive facilitant, d'un côté, la détermination spatiale du bien dans un *territorium*, et de l'autre, assurant la démarcation précise de ses composants pour répondre à un besoin légal, puisque la terre se trouve très morcelée dans un entourage, la région bourguignonne, où son exploitation est assurée par plusieurs maîtres.

La description et l'inventaire présentent différents degrés d'exhaustivité selon le type d'acte, l'action juridique accomplie ou la nature des biens associés à la terre. Elles sont systématiques et souvent détaillées dans le cas des actes de donation privée. Les diplômes, concessions royales ou confirmations, portant sur des nombreux biens, abondent de riches descriptions, de même que les privilèges obtenus de la papauté ou de l'évêché concernant la concession ou l'usufruit des droits sur la terre ou sur des bâtiments. Dans les notices de la même période, la description peut se contenter de donner la localisation du bien selon le système d'organisation territoriale opérant à l'époque, ne mentionnant qu'occasionnellement les détails. De même, échanges, déguerpissements, loyers et mises en gages entre privés mobilisent souvent un style de charte très concis et direct. En outre, selon l'époque ou la tradition, l'inventaire peut être raccourci ou, ce qui est plus rare, allongé ; nous allons trouver des descriptions assez riches dans les chartes de donation du Xe siècle, richesse qu'on ne retrouve plus après la première moitié du XIe siècle ; inversement dans les cartulaires on voit quelques actes portant de descriptions plus génériques, puisque dans ceux-ci plusieurs détails trouvés dans les originaux ont été supprimés par le cartulariste, ce qui n'est pas très étonnant étant donné que les actions juridiques copiées sont déjà passées et que l'encadrement territorial de tradition ancienne se trouve sérieusement affecté.

Pour mieux visualiser les différents développements formulaires dans les localisations et inventaires, voici quelques exemples de descriptions foncières dans le dispositif dont nous nous servirons :

(i) "*Carta scammiaria quam fecerunt duo fratres [...] contra sanctorum apostolorum Petri et Pauli Cluniensis monasterii, de terra qui adjacet in pago Matisconense, in* villa *que nominatur Binzono : hoc est una petia de vinea, de longo habet dextros XXIII, de latus dextros*

II et tres pedes, et terminat ipsa vinea a mane via publica, a cercio Sancti Stephani, a medio die terra Sancti Letgerii, de alia parte ipsius Sancti Petri ; et faciant ipsi monachi quicquid facere voluerint". [422]

(ii) "*Ego Maiolus, dono Deo et sanctis ejus apostolis Petro et Paulo [...] duos mansos in territorio Matisconensi, in* villa *quæ vocatur Varengo ; et in uno manso habetur unus servus, nomine Maiolus, quem etiam dono cum uxore sua et infantibus suis, cum silvis et pratis et omnibus ad ipsum mansum pertinentibus. Alius mansus habet duas mansiones, et vineam quæ terminatur a mane terra Sancti Petri, et a meridie et a sero via puplica, a cirtio terra Sancti Petri*". [423]

(iii) "*Diletto Domino Deo et Sancti Petri et Sancti Pauli de Cluniaco, et donni Maioli, abatis, entores. Igitur ego, in Dei nomine, Ermennardus et uxor sua Leotgar vendimus vobis aliquit de res nostras que sunt citas in pago Matisconens, in agro Matornens, in* villa *Escozolas, in Ramonda vocat ; oc es curtilus qui terminet a mano via pullica, et de alias partes terra Sancti Petri. Infra istos terminios, la una medietate vobis donamus. Et dono vobis vinea qui terminet a mano terra Sancti Petri, a medium die terra Leotaldi, a sero terra Dodoni, a sercio de ipsa ereditate. Infra istos terminios, parcionem nostram vobis vendimus, et acepimus precium invalentem solidos V [...]*" [424]

(iv) "*Domino fratribus Girbalt sacerdote, entores. Ego Ansaldus et uxor sua Rotrudis vendimus nos tibi campo in pago Matisconensi, in agro Maciacense, in* villa *Vitriaco, ubi a Vias Becias vocat ; termined de uno latus terra Evardi, de alio latus et uno fronte terra Sancti Petri, ad alio front conturnos ; et abet in longo perticas XXXX, in lato ad uno front perticas V et pedes VI, et in alio front perticas III et dimidia. Infra istas terminationes et perticacio, totum ad integrum tibi vendimus, et accepimus nos de te precium invalentem solidos IIIIor, et pro ipsa precium manibus recepimus, et de juro nostro in tuo tradimus [...]*" [425]

(v) "*Ego Be[r]trannus sacerdos Domino Deo et sanctis ejus apostolis Petro et Paulo, ad locum Cluniensis monasterii, ubi dunnus Odilo abba preesse videtur, villam quæ dicitur Falgerias, totum ad integrum, cum campis, pratis, boscum cum aquis aquarumque decursibus, et omne quod ad ipsam villam respicere videtur, dono ad supradictum locum [...]*" [426]

Ainsi, voici cinq actes appartenant au dernier quart du Xe siècle, un échange (i), deux donations (ii et v), une vente (iv) et une donation-vente (iii) mobilisant des éléments descriptifs communs et des éléments exclusifs selon les différents besoins rédactionnels et juridiques. Parmi les éléments en commun figure la localisation et l'inventaire des biens :

1) La localisation du bien par rapport aux structures d'organisation du terroir [427] :

422. CBMA 2903. Constantius et Teudoinus son frère échangent avec les moines de Cluny une mesure de vigne située à Bonzon (com. de Saint-Gengoux-de-Scissé), au pagus de Mâcon. Vers l'an 1000

423. CBMA 2772. Maiolus donne à l'abbaye de Cluny des manses et leurs dépendances et d'autres biens situés dans les villae Flagiaco et Varengo, au pagus de Mâcon. An 994-1007

424. CBMA 3376. Ermennardus et Leotgar son épouse, Maaliodus et Jodceldis son épouse, vendent à l'abbaye de Cluny des biens situés dans la villa Escozolas et alii, au pagus de Mâcon. An. 966-970

425. CBMA 2395. Ansaldus et Rotrudis, son épouse, vendent à Girbalt, prêtre, un champ à Vitry-lès-Cluny, au pagus de Mâcon, moyennant 4 sous. An. 990

426. CBMA 2749. Bertrannus, prêtre, donne à l'abbaye de Cluny la villa Falgerias. An 993-1010

427. Il s'agit des usages notariaux trouvés dans les formulaires, ainsi par ex. : Zeumer, *Monumenta Germaniae historica : Formulae Merowingici et Karolini aevi accedunt ordines iudiciorum Dei.* Cartae Senonicae : 41, p.203 : "hoc sunt res proprietatis meae in ipso pago in agro illo portione mea..." ; Formulae Arvenenses, 6, p. 31 : "...manso nostro in pago Arvenico, in vico illo, in villa illa.." ;

— In pago Matisconense, in *villa* nominatur Binzono (i)
— In territorio Matisconense, in *villa* quae vocatur Varengo (ii)
— In pago Matisconens, in agro Matornens, in *villa* Escozolas, in Ramonda vocat (iii)
— In pago Matisconensi, in agro Maciacense, in *villa* Vitriaco ubi a Vias Becias vocat (iv)
— Villam quae dicitur Falgerias (iv)

La détermination spatiale est la formule prééminente qui initie toujours la description du bien. Elle fait référence au modèle hiérarchique territorial hérité de l'Antiquité tardive et opérant en Bourgogne depuis au moins le VIIIe siècle, modèle qui présente la *villa* comme une cellule territoriale fondamentale. L'unité immédiate supérieure, l'*ager*, se présente comme un ensemble de *villae* dont l'unité spatiale est unifiée le plus souvent par la cohérence géographique. Pour sa part, le *pagus*, en tant que macro-unité géographique et historique [428], renferme dans ses mailles des ensembles d'*agri*. Néanmoins, ni l'*ager* ni le *pagus* n'ont une définition territoriale précise et font plutôt référence au territoire ou au ressort d'un chef-lieu dont ils prennent le nom. Au-dessous du niveau de la *villa*, et entre elle et les différents biens individualisés qu'elle contient, le *locus* désigne les lieux-dits, secteurs clairement séparés dans la conscience des hommes mais souvent pauvrement localisés dans les actes qui s'y réfèrent sous les périphrases : *qui vulgo dicitur, qui vocatur, qui nominatur*, un usage régulièrement transféré aussi aux *villae*.

C'est à ce système quadripartite - *pagus*, *ager*, *villa* et *locus* - que les hommes se rapportent lorsqu'ils veulent indiquer la position d'un bien dans l'espace [429]. Il fonctionne sur un modèle des cercles concentriques prenant une direction centripète, allant de la zone extérieure vers l'intérieur mais sans toujours mobiliser l'ensemble des éléments (ou les mêmes éléments) de hiérarchique topo-spatiale établie. Par exemple, les localisations utilisant l'association *pagus* - *villa* comme dans l'exemple (i) vont se présenter bien plus nombreuses que celles utilisant l'association *ager* - *villa* qui serait la plus naturelle et informative. On examinera plus tard cette question qui à première vue semble paradoxale étant donné que la plupart des donations sont circonscrites au *pagus Matisconense*, dont la mention devrait être sur entendue. La mention du lieu-dit comme dans l'exemple (iii et iv) n'apparaît qu'occasionnellement. Les donations portant sur

Formulae Turonensis, 1, p. 159 : "hoc est locum, rem proprietatis meae illum, situm in pago illo, in condita illa, cum omni integritate vel adiecentiis suis"

428. Sur l'origine du *pagus* médiéval voir les études de : Hans Hubert Anton. *Pagus und Comitatus in Niederlothringen : Untersuchungen zur politischen Raumgliederung im früheren Mittelalter (Bonner Historische Forschungen, Bd. 49)*. 1986 ; Andrea Castagnetti. *L'organizzazione del territorio rurale nel Medioevo : circoscrizioni ecclesiastiche e civili nella" Langobardia" e nella" Romania"*. T. 3. Pàtron, 1982 ; Laurent Schneider. "Du pagus aux finages castraux, les mots des territoires dans l'espace oriental de l'ancienne Septimanie (IXe-XIIe siècle)". In : *Les territoires du médiéviste* (2005), p. 109

429. D'après différents calculs des mesures des terres, Bange, *op.cit.*, p.538 ; Deleage et Gérard Chouquer. "La forme juridique et cadastrale des actes "notariés" de Cluny en 870-935". In : *www.formesdufoncier.org* (), p. 6 coïncident en indiquer une surface d'entre 250 – 300 ha (2,5 – 3 km^2) pour une *villa* et 2500 – 3500 ha pour un *ager*. À titre indicatif le *pagus* de Mâcon avait une surface d'environ 2500 km^2.

une *villa* entière, comme dans l'exemple (v), sont rares et normalement il s'agit d'un bien ou d'un ensemble de biens dont la localisation se trouve dans un ou deux lieux-dits proches ; mais il semble que l'indication précise au niveau du lieu-dit n'est pas toujours disponible pour le scribe ou qu'il considère que la localisation indiquée par l'indication de la *villa* est suffisante.

Ainsi, la localisation par le trinôme *pagus-ager-villa* est assez commun dans les documents du IXe siècle et de la première moitié du Xe siècle, mais devient de moins en moins utilisée au fur et au mesure que on avance dans le temps. Et la localisation mobilisant les quatre niveaux spatiaux, *pagus-ager-villa-locus*, apparaît faiblement dans la documentation. Malgré cela, la direction centripète est privilégiée par les scribes et toutes les altérations s'adaptent à ce sens qui va depuis le maillage extérieur vers la *villa* (ou plus tard vers le *locus*) et reprend depuis là la description des biens emboités.

D'autre part, Il faut faire attention, à partir de la fin du Xe siècle, à l'usage intermittent dans les actes d'autres types de circonscriptions utilisés pour remplacer les niveaux les plus fréquents : dans l'exemple (ii) on trouve *territorium* à la place de *pagus*, il peut arriver aussi de trouver *centena* ou *iacis*, mais c'est rare ; plus communément on trouvera *comitatus* et, surtout au XIe siècle, *episcopatus* ; *fines* et plusieurs exemples de *vicaria* comme remplacement de l'*ager* et à l'occasion *locus* ou *ecclesia* à la place de *villa* (métonymie) ou encore *cultura* à côté du *locus* dans les documents de l'époque de la fondation de l'abbaye et dans ceux du XIe siècle. Cela répond à de nombreuses situations que nous allons détailler plus loin (voir 6.4). Ce qu'il faut signaler maintenant c'est que dans la deuxième moitié du Xe siècle le système régi par le *pagus* entre dans sa phase finale de décomposition, ce qui oblige les scribes à se rapporter plus souvent à d'autres structures d'encadrement en place dans la région. Dans les réseaux clunisiens, d'autres termes spatiaux circulent afin de permettre une définition territoriale plus précise, mais aussi de rendre compte de la superposition de différents ressorts juridictionnels. La mutation principale de l'ancien système concerne la destruction de l'*ager* et la modification de l'encellulement humain qui se polarise prenant les chefs-lieux des anciens *agri* comme centres de l'essor des villages. Les différents degrés d'adaptation au nouveau système selon la localité et la variabilité des structures intermédiaires d'encadrement génèrent un sérieux bouleversement, bien reflété dans les actes tout au long du XIe siècle, des points de référence spatiale dans la région. Il s'agit d'un phénomène que l'on va examiner plus minutieusement dans le chapitre final. On ne doit pas, en tout cas, confondre le vocabulaire appartenant à l'ancien système, même si les scribes le font parfois dans les documents des XIe-XIIe siècles, avec celui provenant des structures de remplacement depuis la deuxième moitié du XIe siècle : *conventus*, *parochia*, *diocesis*.

2. Le deuxième point en commun de la description dans les actes est constitué par l'énumération sommaire des unités patrimoniales objets de la donation :

1. Une *petia* (mesure) d'une vigne (i) ;
2. Deux manses, l'un avec toutes ses dépendances *cum silvis et pratis* (c'est à dire les parties non labourables) et un serf avec sa femme et leurs enfants ; l'autre avec deux maisons et une vigne (ii) ;
3. Un *curtile* (jardin clos près de la maison), une vigne et une portion d'une autre

(iii)

4. Un champ vendu par 4 sous (*solidos*) (iv)
5. Une *villa* dans son intégrité (v).

L'inventaire des biens est souvent sommaire surtout dans les affaires entre personnes privées, mais il n'est pas rare, dans des donations des biens plus vastes, de trouver des descriptions détaillant toute la richesse contenue dans les différentes dépendances foncières allant du paysage humanisé au paysage naturel. Dans l'acte, une fois les biens localisés, on mobilise le vocabulaire socio-économique décrivant le milieu et le paysage appliqué à l'espace d'une *villa* ou d'un *locus* (lieu-dit) [430]. On peut reconnaître trois types d'unités : d'un côté les unités agricoles associées à des bâtiments d'exploitation : *curtilis, mansus, jardin, predium, clos*, etc. (exemples ii et iii). Il s'agit des unités normalement fermées dans une enceinte, habités dans les cas du manse, ou près des bâtiments d'habitation continue ou saisonnière. D'un autre côté ceux précisant la spécialisation de la culture : *vinea, campus, pratum, grangia, pascua, prairie*, etc. (exemples i, ii, iii et iv). Ces sont les unités de production qui définissent la valeur de la donation en termes de richesse. Et finalement, les unités décrivant les espaces non cultivables dans les périphéries : *aquas, nemus, pascuis, mons, silvas*, etc. (exemple iv).

L'espace naturel n'est pas toujours enserré entre les mailles de la *villa*, mais il peut être l'objet de donations lorsqu'il a été intégré comme partie des *terras incultas* ou représente une ressource, comme dans le cas de la forêt ou des rivières et sources d'eau. La précision de la description peut continuer à descendre jusqu'au contenu des biens : maisons, moulins, potagers, forges, canaux, serfs, etc., mais ce genre de description spécialisée est trouvée surtout lorsque la donation porte spécifiquement sur ce bien. Ce vocabulaire descriptif est limité en nombre (il y a autour d'une trentaine de termes [431]) mais très complexe dans leurs définitions et très variable dans le sens attribué par chaque scribe. Les termes, surtout dans les unités d'habitat humain, peuvent renvoyer à des réalités bien différentes selon l'époque, le scribe, l'intérêt du donateur ou le formulaire utilisé.

Ainsi, la localisation et l'inventaire forment un tandem qui ouvre le dispositif et que on trouvera régulièrement dans la quasi-totalité des actes de mutation foncière, spécialement dans le type le plus utilisé : l'acte de donation privée. Dans les cas les plus formels, le scribe nous amène depuis le *pagus*, passant par la *villa* et allant jusqu'aux éléments indivisibles de richesse dans le cadre d'un lieu-dit.

430. On peut consulter au sujet des descriptions et les inventaires foncières dans les actes médiévaux : Michel Zimmermann. "Glose, tautologie ou inventaire ? L'énumération descriptive dans la documentation catalane du Xe au XIIe siècle". In : *Cahiers d'Études Hispaniques Médiévales* 14.1 (1989), p. 309-338 ; Anne Mailloux. "Perception de l'espace chez les notaires de Lucques (VIIIe-IXe siècles)". In : (1997) ; Chouquer, "La forme juridique et cadastrale des actes "notariés" de Cluny en 870-935"

431. Plus de 1000 observations : *vinea, campus, mansus, pratum, silva, aqua, domus*, plus de 300 : *mons, molina, fructus, casa, vallis, pascua*, plus de 100 : *alodium, boscus, nemus, mansio, fons, curtis, arbor, cultura, fluvius, grangia*, plus de 50 : *mansura, flumen, area, fundus, curtiferum, predium*, moins de 50 : *iter, vercheria, finagium*

De plus, chacun des actes fait appel à des structures rédactionnelles particulières qui permettent de préciser la description du bien : définition des limites, orientation dans le parcellaire, voisinage, surfaces, prix.

- Dans l'exemple (i) est attesté l'échange de *res propriae* familiales entre deux particuliers. Cet acte de permutation d'une mesure de terre, c'est-à-dire d'une portion mineure d'un bien partitionné, doit répondre à deux obligations légales : d'un côté, la description de la surface et de son extension et de l'autre l'indication de ses limites et de la place qu'occupe la section donnée par rapport à la structure immédiate supérieure, ce que les actes appellent, respectivement, *perticatio* et *terminatio*. Dans cet exemple les mesures en pieds (*pedes*) ne sont pas des mesures de surface mais des indications géométriques donnant parfois le périmètre, parfois la longueur des bords latéraux. La description du voisinage avec le terme générique *terra* ajouté au nom du propriétaire connu est aussi utile pour combler l'absence d'un cadastre et de plans de terroir.

- Dans l'exemple (ii), seule la vigne du deuxième manse mérite l'indication du voisinage, orientée selon les points cardinaux. Les vignes et les champs comme biens de production dans une structure de culture intensive sont normalement exploités par différents tenanciers, donc ils se trouvent souvent partitionnés ou en tout cas peuvent être facilement dénombrés. Dans cet exemple l'abbaye avait tout l'intérêt à recevoir cette donation pour arrondir ses biens, puisque la vigne en question se trouve, naturellement, entourée par d'autres parcelles de vignes appartenant déjà à Cluny. Les manses, en revanche, comme structure domestique habitée au centre d'une exploitation ne se trouvent pas normalement partagés par différents propriétaires et ils sont donnés ou vendus souvent en entier avec les serfs qu'y habitent ; dans d'autres cas, est procédé à un dénombrement faisant la distinction entre les maisons (*domus*) et les bâtiments d'exploitation (courtils) et les terres, cultivées ou non, associés au manse.

- L'exemple (iii) reproduit ces mêmes usages à l'occasion d'une donation-vente ou peut-être d'une vente complétée par une donation dont l'église à tout l'intérêt puisque le courtil acheté contribue à attendre ses possessions périphériques et la vigne donnée contribue à augmenter sa présence dans les terres cultivées dépendantes des bâtiments d'exploitation du courtil. Ce genre d'affaires motivées par l'abbaye est régulièrement trouvé dans les actes à partir de la fin du Xe siècle. Dans le cas de la vente, le prix de 5 sous (*solidos*), qui paraît élevé par rapport à des autres prix qu'on a pu constater dans des biens similaires à l'époque (ce qui peut indiquer qui s'agit en réalité d'un dédommagement), est normalement indiqué.

- Dans l'exemple (iv) deux particuliers, un prêtre et un couple, accomplissent la vente d'un champ moyennant 4 sous. Des cinq exemples copiés ici celui-ci est l'exemple le plus rigide en terme de séquence formulaire. La localisation est suivie par la *terminatio*, puis par la *perticatio*. Le dispositif est terminé par deux clauses très répandues qui attestent la partie et la contrepartie du contrat : le transfert du bien dans son intégralité (*totum ad integrum tibi vendimus*) et l'accomplissement de l'échange pécuniaire en main (*precium manibus recepimus*).

- Finalement l'exemple (iv) présente un style assez sommaire d'acte qui s'éloigne en plusieurs points de ces autres exemples, même s'il relève, en comparaison, du transfert d'un bien beaucoup plus riche. Son style rédactionnel nous indique une probable réduction dans une étape de sa tradition. À la place de la description, que

on imagine détaillée, des biens donnés, le cartulariste mobilise la formule classique *totum ad integrum* completée par une petite liste de substantifs à l'ablatif. Même la localisation du bien est réduite à la seule mention de la *villa*, sousentendant le reste. Le travail de réduction, s'il a eu lieu, est néanmoins peu hasardeux et on respecte le sens universellement employé des descriptions de l'univers géographique que constitue une *villa* : les bâtiments, les espaces cultivés, les espaces non cultivés, l'espace naturel, la technologie d'irrigation, c'est-à-dire depuis l'habitat humain vers l'habitat naturel mais humanisé.

Cela dit, s'il est possible de constater différences entre les originaux et les copies des cartulaires, elles se rapportent presque toujours à des éléments assez neutrales, concernant la suppression de préambules [432] ou des clauses prolixes (ou son allongement [433]), des réorganisations de la matière de l'acte ou de la phrase [434] et surtout à des actualisations de la grammaire - suppression des incorrections flexionnels du latin - et des noms de personnes et de lieux [435]. Ce dernier cas étant le plus important pour notre travail car affecte la canonisation automatique des toponymes que nous avons entamé. On peut considérer que l'ordonnancement spatial fondé sur le *pagus* et l'*ager* se trouve désactivé à l'époque des cartulaires, ce qui pourrait expliquer quelques cas d'omission de hiérarchie spatial dans ces copies, mais en général les cartularistes (et le copiste moderne) n'altèrent que dans de très peu cas les indications concernant la localisation ou les inventaires des biens fonds. Si dans certains cas on trouve des actes qui semblent fortement abrégés ou ayant adopté un style très direct, cela se rapporte le plus souvent à l'état des actes du chartrier originel. Plusieurs chercheurs se sont intéressés á l'état de la tradition textuelle des actes de Cluny et aux modes d'opérer des copistes, nous renverrons le lecteur vers ces études [436].

Donc, les descriptions foncières dans les actes montrent diverses adaptations par rapport aux modèles formulaires. Comme on l'a montré dans l'étude des invocations, les actes de Cluny sont très proches des formulaires anciens mais ils présentent des modifications selon des besoins rédactionnels précis, des exigences juridiques dûes à la nature du bien transferé et enfin des mutations dans la tradition des actes. Un certain niveau de variabilité est aussi observé dans le vocabulaire descriptif des biens et dans sa localisation. Dans les inventaires il peut exister un décalage entre le terme utilisé pour définir un bien, peut-être proposé par le formulaire, et la réalité actuelle que veut transmettre le rédacteur.

De son côté, la localisation des biens dans l'anthroposystème, instance partagée par tous les actes, peut présenter différents degrés de précision. La localisation se

432. Voir quelques exemples : CBMA 1534, 1542, 1924

433. Alexandre Bruel. "Note sur la transcription des actes privés dans les cartulaires antérieurement au XIIe siècle". In : *Bibliothèque de l'École des chartes* 36 (1875), p. 445-456

434. Voir aussi dans l'édition de Sébastien Barret et al. *Les plus anciens documents originaux de l'abbaye de Cluny, t. II : Documents nos 31 à 60.* 2000 les actes nº 21 et nº24 du tome 1 et l'acte nº32 du tome 2

435. Voir quelques exemples : CBMA 1533, 1538, 1711, 2263

436. Sébastien Barret. *La mémoire et l'écrit : l'abbaye de Cluny et ses archives (Xe-XVIIIe siècle).* T. 19. LIT Verlag Münster, 2004, p. 115-123 ; Iogna-Prat, "La confection des cartulaires et l'historiographie à Cluny (XIe–XIIe siecles)"' ; Harmut Atsma et Jean Vezin. *Les responsables de la transcription des actes juridiques et les services de l'écriture au Xe siècle : l'exemple de Cluny*, p. 10-20

référant au système hiérarchique du *pagus* et de ses divisions est le seul système de positionnement géographique mais dont l'utilisation paraît variable. Jusqu'à la deuxième moitié du XIe siècle, seule la *villa*, cellule élémentaire du système territorial, peut agir comme référent indispensable de localisation foncière ; tous les autres niveaux sont contingents. Il est vrai que les actes n'indiquant que le nom de la *villa* sont peu nombreux et elle est normalement accompagné d'au moins un des autres éléments hiérarchiques de l'espace, mais pour les scribes la *villa* ne peut pas être omise, d'où son omniprésence dans les actes, et la quasi-inexistence de chartes se rapportant directement à un *locus* (Dans quelques cas un lieu-dit a été pris par une *villa*).

	Localisation				Inventaire	
	Unités superieures	Unités intermédiaires	Unités de base	Unités inférieures	Unités locales	Détermination
1	pagus Matisconense		villa Binzono		Partie d'une vigne	Perticatio Terminatio
2	territorio Matisconense		villa Varengo		Deux manses	Descriptio Terminatio
3	pagus Matisconense	ager Matornens	villa Escozolas	locus Ramonda	Un courtil, deux parties de vigne	Terminatio Pretium
4	pagus Matisconensi	ager Maciacense	villa Vitriaco	locus Vias Becias	Un champ	Perticatio Terminatio Pretium
5			villa Falgerias		Une villa	Descriptio

FIGURE 6.1 – Localisations et inventaires des actes selon leurs unités d'encadrement et la description de leurs biens.

Ainsi, c'est à partir de la récupération des noms des *villae* en tant qu'entités nommées que nous allons essayer de restituer les mouvements des structures de l'espace reflétés dans les actes du recueil clunisien et du cartulaire de Saint Vincent. Bien entendu, la récupération des entités se référant aux autres instances d'encadrement sont indispensables dans ce portrait, mais leurs apparitions se trouvent presque toujours subordonnées à la *villa*. C'est à travers de la stabilité que nous procure la *villa* que nous allons vérifier les changements survenus dans les structures qui la renferment : d'abord, l'*ager*, la structure immédiatement supérieure qui se rapporte à un ensemble de *villae* ; et en deuxième place le *pagus* en tant que macro-unité englobant le réseau des *agri*. En ce qui concerne le lieu-dit (*locus*), sous-unité de la *villa*, même si toute analyse au niveau de la commune devrait impérativement s'y référer, le nombre de leurs mentions, bien que non négligeables, ne permettent pas d'essayer une reconstruction vraiment cohérente. Les lieux-dits les plus connus, apparaissent dans un maximum de six actes — qui sont une exception — et le plus souvent dans un seul. Même les *villae* (on verra ensuite la statistique) ne nous sont connues le plus souvent que par quelques chartes. Il nous faut alors rester sur le niveau le plus sûr, celui de la *villa*, dans notre exercice de reconstruction.

6.4 Le pré-traitement des entités nommées et des co-occurrences

6.4.1 Co-occurrences et vocabulaire de l'espace.

Afin d'avoir une vision générale des vocabulaires et formes de détermination spatiale nous allons extraire d'abord la liste de co-occurrences accompagnant les toponymes puis construire une matrice classique de co-occurrences autour de la *villa* qui nous permette d'exploiter le champ sémantique de la dénomination spatiale : la liste est constituée de tous les bi-grammes - ou collocations - ayant une entité nommée de lieu (par ex. ***loco*** *Cluniaco,* ***villa*** *Rufiaco, Matisconense* ***pago***, etc.) [437], la matrice, quant à elle, est constituée des fenêtres contextuelles ayant entre deux et quatre unités co-occurrentes (par ex. « *donamus duas colonias quæ sunt sitæ in* ***pago*** *Matisconensi et in* ***villa*** *Lotchiaco que Letbaldus tenebat...* » ; « *donamus vobis aliquid de redibus nostris, que sunt sitas in* ***pago*** *Ostudonens, in* ***agro*** *Moncioscosens, in* ***villa*** *Corjoan resedunt : hoc est...*») .Ceci pour deux raisons : la liste de co-occurrences va nous servir à révéler le vocabulaire spatial le plus communément mobilisé pour présenter une entité nommée, autrement dit, il va nous informer sur les structures de l'espace depuis la cellule fondamentale jusqu'à la maille extérieure. La matrice va nous permettre d'interroger les formules de localisation qui présentent le plus souvent les coordonnées spatiales de façon séquentielle ; elle va alors nous informer des relations de dépendance et la hiérarchie que gardent entre elles les différentes unités du vocabulaire spatial.

Afin de que cela soit clair, lorsque nous parlons de co-occurrence nous faisons référence aux différentes associations habituelles de deux termes dans une phrase. Ce rapprochement n'est pas arbitraire, spécialement dans des textes suivant un formulaire, et on peut dire, si leur fréquence est très élevée, qu'il s'agit d'associations fixées par le langage diplomatique. Dans ce sens, on pourrait aussi parler de collocations avec deux composants : la base, constituée par le terme du vocabulaire spatial et le *collocatif*, constitué par l'entité nommée [438]. Donc, nos deux matrices montrent les relations statistiques que les termes co-occurrents entretiennent à deux niveaux : en tant qu'unités et en tant qu'unités contextuelles.

La liste de co-occurrences spatiales peut être rapidement obtenue et formalisée.

Dans le graphique (figure 6.2) nous avons repéré les co-occurrences ayant plus de 120 apparitions. L'extraction s'est appliquée sur trois parties des actes : le dispositif, les dates de lieu et les dates temporelles. Les termes peuvent être regroupés en trois ensembles : vocabulaire de définition de l'espace ; verbes d'action juridique ; et usages de formulaires d'actes de donation. Les 5 premiers termes (*villa, pagus, ecclesia, terra, locus*) et cinq des files suivantes (*ager, finis, episcopatus, comitatus, civitas*) appartiennent à la première catégorie, qui définit les structures du paysage humanisé. La *villa* et le *pagus*, utilisés intensément, permettent de définir les coordonnées spatiales

437. Dans les cas de villa et locus il a fallu intervenir sur la périphrase car ces deux termes peuvent se présenter utilisant les verbes *vocere, appellatur, nominatur*, etc. Par ex. « villa quae vocatur Varengo »

438. Stefan Evert. "The statistics of word cooccurrences". Thèse de doct. Dissertation, Stuttgart University, 2005, p. 15-31

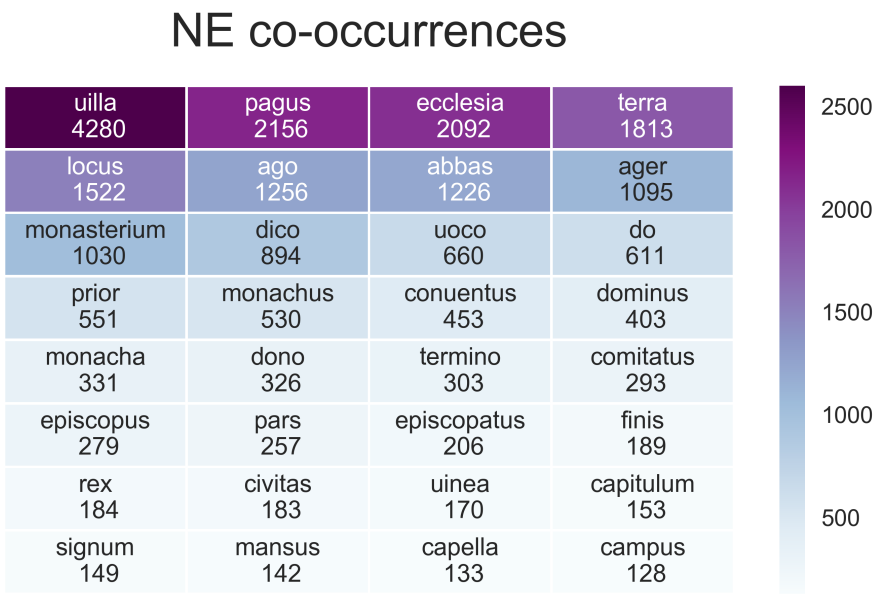

FIGURE 6.2 – Liste de co-occurrences des entités nommées de lieu mentionnées plus de 120 fois.

des biens. *Ager, finis, episcopatus et comitatus* se présentent quant à eux comme des structures intermédiaires entre la *villa* et le *pagus* ou équivalents à ce dernier. *Terra, locus*, sont termes de sens local et général, définissent le découpage à l'intérieur des unités fondamentales du territoire, dans notre cas, la *villa*. Pour leur part, *ecclesia* ainsi que *monasterium*, font effectivement référence aux bâtiments ecclésiastiques qui font partie du paysage d'un lieu. Mais dans les actes ils ont aussi acquis un sens spatial plus complet, comme on le verra plus loin, au fur et au mesure que les lieux contenant une église ou un monastère (abbaye, monastère, couvent, etc.) se sont transformés en pôles territoriaux peuplés.

Les termes *ager* et son synonyme *finis* (utilisé par quelques scribes) qui appartiennent aux catégories majeures de l'espace, juste en dessous du *pagus*, méritent une attention particulière. L'*ager* est en réalité une structure pré-carolingienne qui fait référence non à une vraie division administrative, mais plutôt au territoire - au sens large du terme - d'une *villa* qui joue le rôle de chef-lieu territorial et de pôle organisatrice des autres *villae*. Le mot *finis* transmet à peu près le même concept. Mais les deux sont sous-utilisés par rapport aux autres termes de la même catégorie ; nous en développerons plus loin les raisons.

Deux autres termes, de plus faible incidence dans le corpus, *comitatus* et *episcopatus*, appartiennent aux catégories se rapportant au maillage supérieur, auquel nous pouvons ajouter *conventus* et *civitas*. Chronologiquement ces quatre termes apparaissent vers la fin du IXe siècle mais ils sont bien moins mobilisés si on les compare avec ceux qui apparaissent dans les premiers résultats de notre tableau. Ils relèvent de juridictions importantes de l'espace mais comme ils sont peu utilisés on ne saisit pas bien les limites de leurs définitions en termes géographiques. En tout

cas, l'intensité dans leur usage coïncide avec la disparition d'abord de l'*ager* puis du *pagus*, avec qui ils cohabitent depuis le Xe puis remplacent partiellement, comme une ressource de la part des scribes pour se rapporter à des juridictions opérationnelles dans l'espace contre l'imprécision des termes territoriaux hérités de l'Antiquité tardive.

Ces unités majeures et intermédiaires sont complétées par *vinea, mansus, capella* et *campus* qui apparaissent à la fin du tableau car beaucoup moins utilisées comme co-occurrentes d'entités nommées. Celles-ci ne portent souvent pas de toponyme qui les individualise puisqu'il s'agit des structures mineures de spécialisation de la culture ou des ensembles de bâtiments d'exploitation normalement non habités.

Le deuxième groupe comprend cinq verbes : *dico, voco, do, dono* et *ago*. La présence de *dico* et *voco* est circonstancielle car ils répondent aux formules de dénomination alternative ou vernaculaire des noms géographiques, se référant généralement à un lieu-dit (*villa quae vulgo* ***dicitur*** *Belma* ; *terra ubi* ***vocant*** *Vallis*, etc.). Dans ce cas la co-occurrence est périphrastique. Pour leur part *do* et *dono* (donner et faire un don) sont, naturellement, deux des verbes les plus mobilisées dans un acte de donation qui précèdent souvent une entité nommée, spécialement dans les notices et les actes de cartulaire (*terra quae Rotbertus* ***dedit*** *Sancto Petro* ; *in villa Bainas* ***donamus*** *unam vineam*). *Dono* est bien plus utilisé que *do* comme verbe central de la donation, employé à la première personne du singulier dans les chartes et à la troisième dans les notices. *Do*, par contre, même s'il est parfois utilisé comme synonyme de *dono*, apparaît dans la plupart des cas comme co-occurrent du fait de son usage intensif dans la date de lieu des donations émanés de pouvoirs publics (***Datum*** *Rome* ; ***Datum*** *Laterani*). Cet usage est complété par le verbe *ago* qui apparaît comme co-occurrence dans les dates topiques des actes locaux d'une certaine importance (***Actum*** *Lordono* ; ***Actum*** *Civiniono*). À ceux-ci on peut ajouter *capitulum* qui agit de manière similaire lorsque la signature ou la validation du document a été réalisée dans le chapitre ou la salle qui remplissait ses fonctions (*fecimus autem hanc communicationem in* ***capitulo*** *Santi Stephani* ; *sigillo* ***capituli*** *Sancti Dyonisii*)

Finalement le troisième groupe, moins hétérogène, accueille des termes co-occurrents définis par leur usage comme adjectifs locatifs. Cela correspond partiellement au cas de *monasterium*, mais surtout de *abbas* dans les actes de Cluny car ils répondent aux formules d'adresse très utilisées dans les actes de donation (*donamus ad* ***monasterium*** *Cluniaco*) et de suscription (*Odoni venerabili* ***abbati*** *Cluniacensis monasterii*). Pour leur part *rex, episcopus, dominus, monachus* et *prior*, désignent les titres personnels des participants des actes en tant que donateurs ou récepteurs, utilisant le génitif d'appartenance au lieu de leur juridiction ou provenance (***rex*** *Anglorum,* ***prioris*** *Marcinicensis,* ***episcopi*** *Aruenensis,* ***monachi*** *Cluniacenses,* ***dominus*** *Sinemuri*).

Donc, cette liste nous révèle rapidement une trentaine de termes fonctionnant en tant que co-occurrents des entités géographiques. Une relative connaissance du corpus et du vocabulaire spatial utilisé pendant le Moyen Âge central permet de reconnaître des catégories de rassemblement. Quinze parmi eux sont directement classés dans le vocabulaire de l'espace en trois rangs d'unités : unités majeures (*villa, pagus, ager, finis, comitatus, episcopatus, civitas*), unités intermédiaires (*locus, terra, ecclesia, conventus, monasterium*), unités de base (*campus, capella, vinea, mansus*). Cette

quinzaine de termes sont la base du vocabulaire spatial que les scribes mobilisent dans la localisation des biens-fonds dans les actes. D'autres termes, on le verra ensuite, peuvent aussi servir à décrire l'espace mais leur usage est minoritaire ou il n'accompagne pas les entités nommées car ils décrivent des unités très élémentaires sans mention de toponyme. Pour le reste des termes, il s'agit, d'un côté, des usages promus par les formulaires, surtout dans les protocoles au début de l'acte afin de déterminer donateurs et récepteurs et à la fin pour signaler le lieu de rédaction et de validation du document. De l'autre côté, nous avons des déterminants topographiques des titres. Le latin, langue flexionnelle, les mobilise en tant que co-occurrences directes mettant normalement l'entité nommée au génitif.

Le vocabulaire de l'espace qui rassemble les termes les plus utilisées est, comme on le voit, relativement restreint. Dans tout le corpus, nous avons repéré 41 termes utilisés régulièrement dans la détermination spatiale, dont les 15 les plus utilisés, dont la fréquence est bien plus élevée que dans les autres 25, sont indiqués par la liste [439]. Pour mieux contextualiser ceci, il faut se rappeler que dans la plupart des cas chaque terme n'apparaît qu'une seule fois dans un document - excepté dans le cas de *terra* qui apparaît normalement entre deux et trois fois - et que le cœur de notre analyse correspond à l'acte privé de donation (autour de 3 550 items), car les déterminations spatiales étant très peu présentes dans les lettres, bulles, jugements, etc. et moins riches dans les notices et actes d'affaires entre individus privés.

6.4.2 Représentation vectorielle du vocabulaire

En revanche, la liste ne nous dit pas grand-chose des relations qu'entretiennent entre elles les différentes structures et unités de description du paysage dont nous venons d'identifier les termes. Une manière de dresser un portrait général les décrivant consiste à reprendre la liste antérieure afin d'entraîner un modèle de "plongements de mots" (*word2vec*) [440] et ensuite générer une représentation des composants principaux (*PCA*) associés aux termes en question dans notre corpus. Le principe sous tendant le "plongement de mots" vient de la théorie distributionaliste qui prévoit des régularités dans le langage comme base à l'analyse du contexte de mots. L'algorithme capture cette information et la transforme en vecteurs de nombre réels. Si deux mots apparaissent dans des contextes similaires, ils possèdent des vecteurs qui se rapprochent dans

439. Plus de 5 mille observations : *terra, ecclesia, villa, locus* ; plus de 2 mille : *monasterium, vinea, pagus, campus*, plus de 1 mille : *mansus, pratum, conventus, curtile, monacha* ; plus de 200 : *finis, mons, capella, cenobium, feodum, comitatus, episcopatus, alodium, castrum, parochia*, plus de 100 : *civitas, conventum, diocesis, vicaria, burgus, cultura, territorium, provincia*, plus de 50 : *colonia, urbs, vicus, patria, suburbium*, moins de 20 : *iacis, centena*

440. Le plongement de mots (*words embedding* en anglais) cherche à projeter un ensemble de mots d'un vocabulaire dans un espace vectoriel continu où les vecteurs de ces mots ont une taille de quelques centaines de dimensions. De plus, on réalise une représentation vectorielle de chaque mot du vocabulaire de façon à ce que les mots aux contextes similaires apparaissent proches dans la représentation et les mots sémantiquement distants s'en éloignent. La méthode est aussi appelée plongement lexical car elle tient compte des mots entourant chaque séquence de l'analyse. Word2vec est un groupe de modèles utilisé pour le plongement de mots. Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In : *Advances in neural information processing systems*. 2013, p. 3111-3119

l'espace et ils peuvent être mutuellement commutés dans la phrase ou le phrasème. Une représentation distributionnelle sémantique est très intéressante afin de vérifier l'uniformité des rôles descriptifs qu'assument les termes de notre vocabulaire dans un formulaire.

Un pré-traitement du corpus est nécessaire : nous avons transformé en *uni-grammes* les *bi-grammes* formés par les entités nommées et son colocataire. Il s'agit tout simplement d'ajouter un tiret bas sur les termes à gauche (le plus souvent) ou à droite des entités nommées géographiques (par ex. villa_Colonias, ager_Ibgiacensis, etc.). Ceci est une méthode suggérée pour modéliser information mutuelle.

Par ailleurs nous avons uniformisé les entités nommées en les remplaçant dans le corpus par un seul mot (PERS et GEO respectivement). La raison en est simple : le principe du formulaire est de recueillir une information variable - par exemple les entités nommées - sur une structure fixe. Si on n'intervient pas, chaque entité sera prise comme une unité de vocabulaire et chaque bi-gramme formé avec les co-occurrences comme un mot différent, altérant ainsi la robustesse de l'analyse qui nous intéresse. Les données statistiques autour de cette question peuvent être observée dans le graphique ci-dessous.

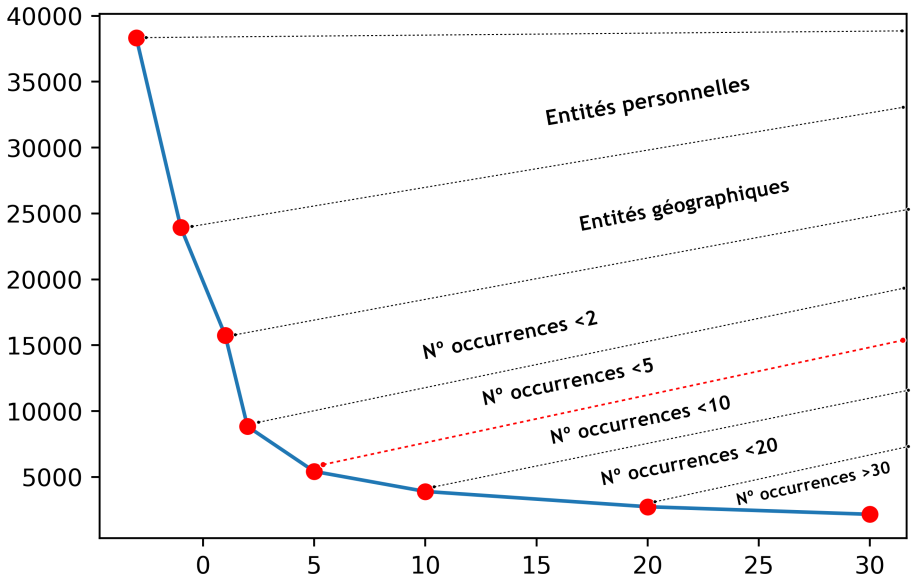

FIGURE 6.3 – Vocabulaire du Recueil de l'abbaye de Cluny classé par le nombre d'occurrences (en abscisse, le nombre d'occurrence d'un mot, en ordonnée, le nombre de mots ayant ce nombre d'occurrences).

Le vocabulaire de notre corpus est constitué d'environ 38 700 mots uniques. Le nombre exact est indéterminé puisqu'il existe une quantité non négligeable de *faux lemmes*, phénomène d'ailleurs commun dans cet état de la langue, mais pas toujours bien maîtrisé par notre outil de lemmatisation. À peu près la moitié de ce vocabulaire correspond aux entités nommées – environ quatorze mille aux personnes, huit mille aux lieux - ce qui montre le poids de ces composants dans le discours diplomatique. Les mots comptant entre 1 et 4 observations se comptent autour de dix mille. La stabilité

se trouve à partir des mots avec 10 ou plus apparitions, ce qui constitue un vocabulaire d'autour de cinq mille cinq cents (5 500) mots, que nous avons pris comme vocabulaire d'entrainement. Ces chiffres sont donnés toutes chronologies confondues ; sans doute, les textes du XIe et du XIIe siècle, plus éloignés du formulaire, constituent la plupart des observations de termes de faible incidence, plus rares dans les époques antérieures.

Par rapport à la taille des corpus normalement utilisés dans ce type d'entrainement, le nôtre est un corpus de petite taille (autour de 750 000 mots) qui pourrait difficilement être utilisé pour des tâches plus complexes de plongements de mots (par ex. prédiction de contextes ou modélisation des sujets) ; néanmoins ce qui nous intéresse le plus sont les fonctions de base : les représentations lexicales distribuées fondées sur les relations de co-occurrence.

Nous n'avons pas effectué d'autres adaptations sur tous les autres termes de notre vocabulaire de corpus.

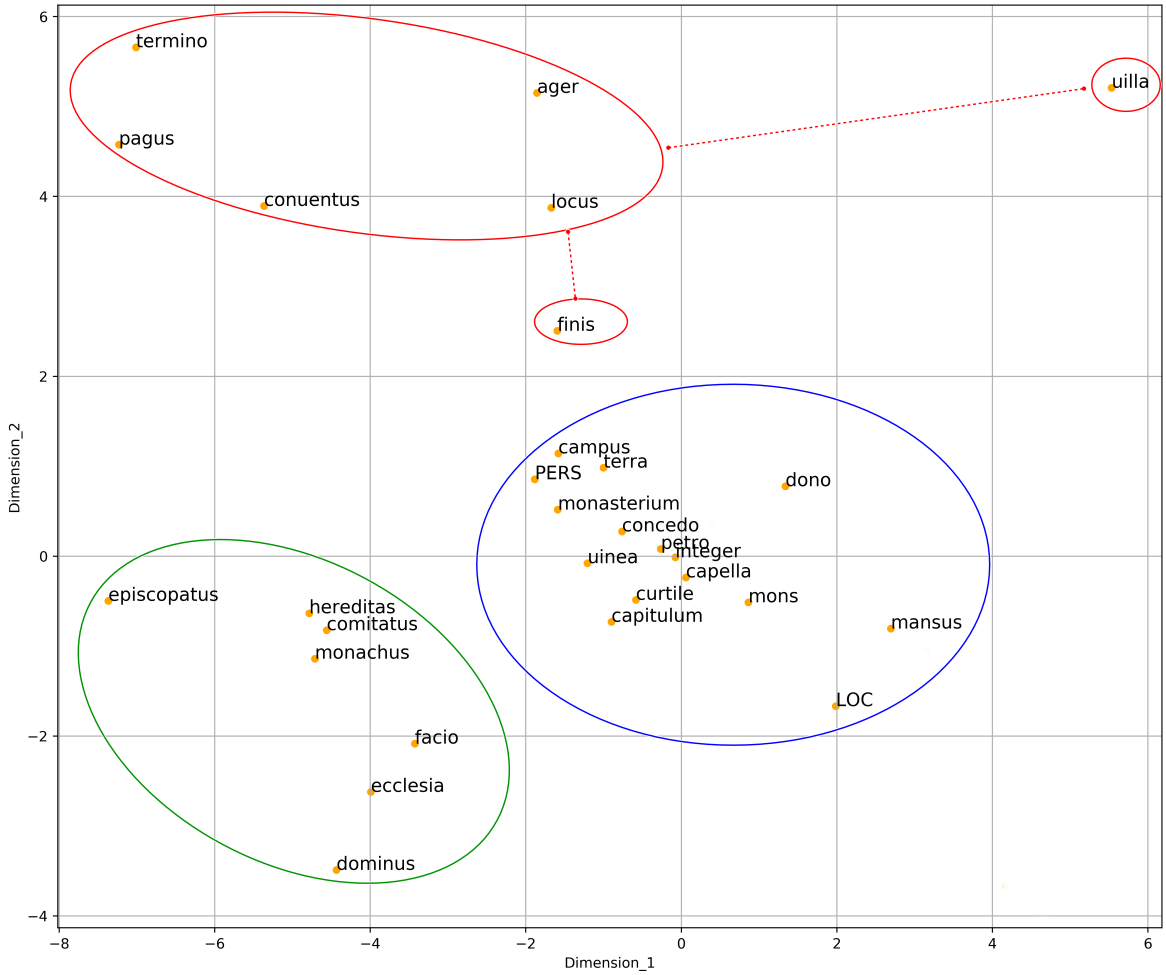

FIGURE 6.4 – Représentation en deux dimensions du vocabulaire principal de l'espace d'après les données du modèle word2vec

L'interprétation du graphique de la figure 6.4 relève d'une certaine complexité. La disposition des mots dans le plan bidimensionnel forme effectivement des secteurs

de regroupement par similarité sémantique qui sont assez significatifs. On peut distinguer trois secteurs qui correspondent au vocabulaire des structures foncières de l'espace (cercle rouge), au vocabulaire des divisions juridictionnelles (cercle vert) et au vocabulaire des structures élémentaires, bien ancré autour du point central (cercle bleu). Les mots de chaque groupe non seulement partagent beaucoup de co-occurrences contextuelles mais peuvent aussi être remplacés les uns par les autres dans la phrase en conservant du sens.

Les termes du premier groupe *pagus, ager, locus* et *conventus* apparaissent ainsi associés dans des contextes similaires mais sur une distribution plus éparse que celle des autres. On peut reconnaître une certaine séquentialité à partir de la formule qui présente souvent les structures énonçant d'abord le *pagus*, ensuite l'*ager* et après la *villa* et le *locus*. Donc, *pagus* qui apparaît deux fois plus que l'*ager* et cinq fois plus que *locus*, a un voisinage contextuel plus large que les deux autres termes. Le terme *finis*, classé avec les structures majeures et assimilable à l'*ager* est proche de celui-ci mais en dehors du cercle puisque son incidence bien plus faible ne permet pas d'obtenir un plongement de qualité suffisante.

Ceci se distingue de la *villa* qui apparaît au même niveau de l'ordonnée que les autres termes de la même catégorie mais s'en éloigne notablement sur l'abscisse car elle présente une variabilité bien plus élevée. Le concept de variabilité est très important. Plus le contexte dans lequel sont employés les termes est variable, plus le graphique sera dispersé. La variabilité dans notre corpus est, comme on le voit, discrète et prend la direction d'une diagonale ; la *villa* se montre comme le terme de vocabulaire le plus variable, les termes du centre du graphique étant les moins variables. Les raisons sont facilement compréhensibles : la *villa* accompagne à peu près 4200 entités nommées mais elle apparaît plus de six mille fois dans le corpus car elle est aussi utilisée en tant que concept, institution, coréférence, ce qui multiplie sensiblement ses contextes sémantiques. Lorsqu'elle apparaît avec une entité nommée, elle est entourée des autres termes du cadre territorial, mais, dans les autres cas le contexte est très contingent. On n'observe pas le même phénomène dans les autres structures majeures de l'espace qui sont très rarement utilisées dans d'autres fonctions que celle de co-occurrences d'une entité nommée.

Une situation similaire à celle de la *villa* est constatée pour celui d'*ecclesia* (cercle vert) dont les apparitions n'accompagnent pas toujours une entité nommée, car il s'agit d'un terme de sens plus général qui peut désigner une église en tant que bâtiment, un centre peuplé ou se référer à l'institution. Les églises accompagnées par les noms de saints car elles leur sont dédiées, formant ainsi une entité géographique, sont parfois l'objet de donations, à part entière ou en partie, mais elles sont le plus souvent mobilisées en tant que déterminants spatiaux des limites locales des biens-fonds. Lorsqu'elle joue ce rôle, l'église a une fonction similaire à un *locus* ou une *curtis*.

En ce qui concerne les termes accompagnant *ecclesia* dans ce groupe : *monachus, episcopatus, comitatus*, il s'agit d'institutions ecclésiastiques qui ont développé une juridiction territoriale. Elles se présentent comme proches parce qu'elles appartiennent en partie au même contexte topique et se retrouvent souvent entourées par des mots de contenu religieux. *Monachus/monacha* est plus un terme institutionnel que spatial puisqu'il fait habituellement référence au monastère en tant que personne juridique

incluant l'ensemble de personnes qui y habitent et qui jouent normalement la fonction de récepteurs de la donation. Dans ce sens il est proche de la fonction remplie par d'autres termes comme *monasterium, cenobium, ecclesia*. Par ailleurs, *episcopatus* et *comitatus* qui relèvent aussi d'un sens institutionnel organisé autour des figures de l'évêque et des cités épiscopales (ou du comte dans le Haut Moyen Âge) se régularisent tard dans le corpus et se trouvent un peu plus éloignés des formules classiques de référencement spatial des IXe et Xe siècle. Les deux peuvent reprendre l'encadrement territorial fourni par le *pagus*, mais de deux manières différentes : le *comitatus* le remplace dans la formule tripartite [441] tandis que l'*episcopatus*, plus tardif, forme un tandem avec la *villa* lorsqu'elle est encore opérationnelle (première moitié du XIe), et après avec le *locus* et surtout *ecclesia* lorsqu'ils prennent le rôle d'unités fondamentales auparavant remplies par la *villa* [442].

Finalement, autour de l'origine (0,0) on retrouve regroupés la plupart des termes qu'on a observé à la fin du tableau de cooccurrences. Leur variabilité est très limitée, ce qui pourrait expliquer leur place au centre du graphique, en partie aussi parce que le nombre d'exemples est limité. Ces termes correspondent aux structures élémentaires de l'espace qui parfois sont distinguées dans les actes par un odonyme ou un micro-toponyme mais qui le plus souvent en apparaissent dépourvus [443]. Leurs vecteurs sont très proches dans le graphique et presque chevauchés, ce qui indique que sont parfaitement commutables entre eux et qu'ils participent à des formules dont l'ordre et la composition sont très similaires.

Si on affine l'analyse, on peut distinguer deux sous-groupes formés autour des termes PERS et GEO que nous avons introduit lors de l'entrainement comme remplacement des entités nommées. Ceux situés autour de PERS (personnes), au nord du point zéro, *terra, dono, campus, monasterium*, apparaissent bien plus souvent liés en tant que co-occurrences aux entités nommées personnelles tant que ceux autour de GEO (lieux), au sud du point zéro, qui correspondent aux espaces nettement géographiques (*vinea, capella, mons, curtile, mansus, dono*). Par exemple *dono* est un verbe dépendant d'une entité personnelle mais qui fonctionne comme co-occurrent spatial [444]. *Monasterium* peut aussi jouer ce même rôle [445]. Dans le cas de *terra*, il s'agit d'un terme amplement mobilisé dans la détermination des voisinages des lopins de terre, ce qui explique qu'il apparaisse entouré par des noms de personnes [446].

Les termes du deuxième groupe s'éloignent du contexte des entités nommées personnelles. Par exemple, *campus* correspond à une unité de culture [447], du même

441. CBMA 1529. « . . . site in comitatu Uiennense, in agro Cominaco, in villa que dicitur Craponi»
442. CBMA 4585 : «. . . .ecclesiam unam Sancti Iohannis, in episcopatu Pictauensi, in villa Alosio. . . » ; CBMA 3352 : « . . . hereditate que sita coniacet in episcopatu Eduensi, et in comitati Avalense, in villa que nuncupatur Martiniacus. . . »
443. En plusieurs occasions ces termes se présentent comme co-occurrentes d'entités nommées que ont été déjà présentées par d'autres termes (par ex. « dono in villa Nugerold vineam unam quam.. » ; « donant etiam in uilla Donziaco mansum indominicatum quem »), mais il s'agit de cas facilement distinguibles car ils sont saisis dans le contexte droit de l'entité (les mots postérieurs).
444. CBMA 2348 ". . . ego, pro anima senioris mei Girardi, dono ad Cluniacum monasterium. . . "
445. CBMA 1544 ". . . ego Rannulfus dono Sancto Petro Cluniensis monasterii de res meas..."
446. CBMA 4055 : "a mane Vuandelfredi terra, a medio die Ermengeri, a sero Sancti Petri"
447. CBMA 2255 : "habet campus sancti Martini in longum perticas novem" ; CBMA 3190 : "...et dono in Colonias, campum unum, terminat a mane increpito"

que *uinea*[448] ,*mons*[449] ,*curtile*[450] et *mansus*[451]. Comme on le voit les unités de culture sont très proches du contexte de la *villa* puisqu'elles sont présentées à leur suite dans l'inventaire des biens-fonds donnés. Elles portent aussi dans plusieurs cas des noms de saints, spécialement dans les cas de *terra* et *campus* car il s'agit de possessions ecclésiastiques appartenant (*res sancti*) ou sous le patronage d'un saint et dont l'ampleur leur fait régulièrement jouer le rôle de frontière ou de point de bornage. *Capella* est un autre terme régulièrement accompagné par des noms de saints, mais qui à la différence d'*ecclesia* fait référence exclusivement au bâtiment de culte ou à une partie d'une église dédiée à un saint[452]. Finalement on a déjà vu les cas du verbe *ago* et de *capitilum* qui présentent dans la date de lieu les lieux de signature d'un acte[453].

6.4.3 Matrice croisée de co-occurrences dans le contexte de la villa

Les catégories de co-occurrences qui dépendent d'une entité nommée deviennent assez claires, ainsi que les termes que nous devons considérer comme le vocabulaire de l'espace le plus significatif pour une exploitation automatique. Le graphique formé à partir des vecteurs sémantiques nous a de plus montré la proximité qu'elles entretiennent dans le discours diplomatique, ce qui continue à être cohérent avec les résultats de la première matrice. La deuxième matrice, que nous allons voir à présent, prend comme base ces résultats et analyse la question de la présentation formulaire des structures de l'espace. Elle nous servira comme complément informatif avant de passer à une reconstruction à partir des points géographiques décrits dans les actes.

Dans le discours formulaire, comme on l'a vu dans plusieurs exemples, la liberté d'expression est normalement subordonnée à un modèle et à un vocabulaire limité de description de la réalité. En conséquence, les entités présentes dans chaque formule de localisation, qui est aussi une phrase dont le sens est complet, gardent forcément de multiples relations de dépendance et de coordination. Suivant cette idée et ayant montré que la *villa* se trouve au cœur de ces descriptions et qu'elle est statistiquement le terme le plus significatif, au moins jusqu'à la première moitié du XIe siècle, nous nous proposons de représenter les relations qu'elle a établies à l'intérieur de la formule avec les autres entités mobilisées par le scribe afin de préciser la localisation d'un bien.

Sur un plan technique cette matrice ne relève pas d'une grande difficulté. Il s'agit d'une matrice classique de co-occurrences croisées. Nous sélectionnons toutes

448. CBMA 2630 : "...dono iam dictum loco in uilla Bieria uineam unam que habet viginti duas perticas..." ; CBMA 3501 : "Et donamus un villa Arelia uineam unam cum curtilo insimul tenente..."
449. CBMA 2517 : "Sunt autem heedem ecclesie : capella in Algoio monte et ecclesia de Tresda" ; CBMA 3919 : "...sitas in pago Uiuarensi, in agro Albanense, in monte qui uocatur Rompone"
450. CBMA 2110. "Donamus in uilla Casanicas curtilum cum sibi pertinentibus..." ; CBMA 2541 : "Ego Girardus dono ad locum Cluniacum curtilum unum in uilla quam dicunt Casal Uiral"
451. CBMA 2206 : "dono...in uilla Boyaco mansos duos et quicquid in ipsa uilla uisus sum habere..." ; CBMA 3982 : "donant etiam in uilla Donziaco mansum indominiatum quem inibi habent"
452. CBMA 3385 : "Vendo itaque uobis capellam Sancti Genesii, cum rebus sibi circa adiacentibus..." ; CBMA 2517. "...Sunt autem heedem ecclesie : capella in Algoio monte et ecclesia de Tresda..."
453. CBMA 1620. "Actum capella Sancta Maria" ; CBMA 4173. "Actum Lordono castello" ; CBMA 4488. "...in capitulo Cluniacensi laudauit et fecit."

les fenêtres contextuelles contenant un minimum de deux et un maximum de quatre des bi-grammes formés auparavant dont au moins un correspond à la *villa*. Ensuite, nous présentons le nombre de coïncidences des termes dans le même contexte, c'est à dire, nous calculons la fréquence de distribution du vocabulaire de l'espace ayant la *villa* comme terme spécifique central.

	A	B	C	D	E	F	G	H	I	J	K	L
A	3	789	32	945	111	5	1	0	32	13	1	6
B	789	2	77	1771	196	42	11	0	1	105	2	36
C	32	77	22	373	57	6	6	3	14	7	4	10
D	945	1771	373	431	558	425	92	19	180	161	60	103
E	111	196	57	558	58	20	4	0	24	18	4	5
F	5	42	6	425	20	142	8	3	17	1	13	2
G	1	11	6	92	4	8	4	1	5	1	6	1
H	0	0	3	19	0	3	1	2	0	0	2	0
I	32	1	14	180	24	17	5	0	4	3	4	11
J	13	105	7	161	18	1	1	0	3	4	0	5
K	1	2	4	60	4	13	6	2	4	0	0	0
L	6	36	10	103	5	2	1	0	11	5	0	2

FIGURE 6.5 – Matrice de co-occurrences autour du terme *villa*. A : *ager*, B : *pagus*, C : *locus*, D : *villa*, E : *terra*, F : *ecclesia*, G : *monasterium*, H : *conventus*, I : *comitatus*, J : *finis*, K : *episcopatus*, L : *vicaria*. Le but de la matrice est de présenter le nombre de fois où chaque entité des lignes apparaît dans le même contexte que chaque entité des colonnes.

Dans le premier quadrant de la matrice s'affichent les résultats les plus importants concernant les six termes les plus utilisés du système de découpage du paysage, de A à E : *ager, pagus, locus, villa, terra et ecclesia*. Les six termes suivants, de G à L, correspondent aussi à des structures d'ordonnancement géographique, mais moins utilisés que les autres ou juridiquement différents. Pour mieux comprendre les résultats nous avons repris depuis les résultats du modèle *word2vec* la liste des termes sémantiquement les plus proches de ceux qui nous intéressent ici (en police *courrier*). La valeur de la proximité sémantique y est exprimée de 0 à 1.

De la matrice, on peut facilement dégager quelques remarques :

Les termes principaux

En premier lieu, le tandem *villa-pagus* est effectivement bien plus utilisé que celui d'*ager-villa*, le rapport est de 1 : 1,87 ; et les trois termes – *villa, ager, pagus* – apparaissent dans la même fenêtre contextuelle dans au moins 790 documents. *Pagus* et *ager*, qui correspondent aux unités majeures et intermédiaires de l'espace rural, sont utilisés presque exclusivement associés à la mention de la *villa*, formant tous trois un trinôme à différentes échelles ; mais il est évident que souvent pour les scribes l'indication *pagus-villa* est suffisante (ou moins problématique car l'*ager* est en processus de dislocation) pour localiser un bien par rapport à l'indication tripartite ou celle du binôme *ager-villa*.

uilla [454]

(ager : 0.837), (uinea : 0.822), (curtile : 0.805), (uoco : 0.801), (sino : 0.793), (coniaceo : 0.754), (uicarius1 : 0.724), (comitatus : 0.719), (pagus : 0.715), (uocabulum : 0.711), (campus : 0.705),(territorium : 0.700), (alodium : 0.687), (mansus : 0.680),(pratum : 0.652), (condamina : 0.650), (peciola : 0.637)

Le contexte sémantique de la *villa* est partagé avec d'autres termes d'encadrement spatial : *ager, vicaria, pagus, comitatus, territorium* avec lesquels elle apparaît souvent associée dans les formules de localisation, mais aussi avec les termes précisant les structures foncières de base comme *vinea, curtile, campus, alodium, mansus, pratum* qui apparaissent normalement dans les formules précisant l'inventaire des biens-fonds. Le terme *villa* fonctionne normalement comme terme charnière entre localisations et inventaires. En outre, la *villa* se montre sémantiquement proche de deux périphrases très récurrentes formées avec le verbe *sino* qui renforce l'idée de situation [455] et *uoco / uocabulum* qui introduit un toponyme vernaculaire dans le but de bien identifier l'endroit [456].

D'autre part, il n'est pas surprenant de détecter dans la matrice un nombre important de tandem *villa-villa*, 431 cas, ce qui atteste la situation de deux *villae* juxtaposées dans la phrase soit parce que s'agit d'une énumération de *villae* [457] soit parce que la formule de donation concerne deux *villae* localisées dans le même *ager* ou *pagus* [458]. Cette situation se présente à peine dans le cas des autres termes selon la matrice : deux *pagi* et deux *agri* n'apparaissent presque jamais ensemble, et deux *loci* exceptionnellement.

En ce qui concerne le *pagus*, il fonctionne comme terme englobant des *villae*, d'où son rapprochement sémantique à d'autres termes faisant la même fonction : *ager, vicaria, comitatus, episcopatus* mais comme il apparaît souvent dans le contexte ordinaire des localisations quelques termes de base se rapprochent de lui comme *curtile* et *vinea*. Dans la matrice, la relation très étroite entre le *pagus* et la *villa* se produit en partie du fait de la décadence, à partir de la deuxième moitié du XIe siècle, de l'*ager* comme circonscription. Le *pagus* n'est pas mieux défini que l'*ager*, mais en tant qu'unité globale il offre un niveau de certitude plus affirmé ; les scribes hésitent rarement à propos de la localisation d'un bien dans un *pagus* ou dans un autre et la plupart des donations ne concernent que le *pagus Matisconense*. D'ailleurs, parfois, les scribes utilisent le terme *pagus* pour désigner d'autres circonscriptions mineures, et ce, en construisant un nouveau toponyme à partir du nom d'un chef-lieu. Le *pagus* devient alors un terme de sens général pour se rapporter au maillage supérieur englobant les *villae*.

454. Tous ces termes et chiffres correspondent aux mots plus proches de chaque terme expliqué dans l'espace vectoriel, dont la similarité est exprimée de 0 à 1.

455. CBMA 3417. " dono etiam una uineam in uilla Wissandone sitam »

456. CBMA 2277. "...reas meas que sunt site in comitatu Matisconense, in uilla que Seya uocatur..." ; CBMA 2352. "...rebus meis que sunt site in pago Matisconense, un uilla uocabulo Galmiriaco..."

457. CBMA 2231. « ... et quicquid in uilla Petronna et in uilla Berneto et in Filnarias uisi sumus habere... »

458. CBMA 1921. « ...site in pago Cabilonense, in uicaria uel agro Talmariaco, in uilla Cuulesia et in uilla Brogoldio et in Ragia... »

pagus

(comitatus : 0.762), (ager : 0.725), (uilla : 0.649), (uicarius1 : 0.636), (uocabulum : 0.625), (curtile : 0.623), (coniaceo : 0.565), (sino : 0.553), (uinea :0.551), (episcopatus :0.548),(uoco : 0.530), (indominicatus1 : 0.521), (territorium : 0.511)

Enfin, l'*ager* présente des similarités sémantiques presque calquées sur le *pagus*. Il fonctionne effectivement comme unité intermédiaire englobant des *villae*, mais à différence du *pagus*, l'*ager* fait plutôt référence au territoire d'un chef-lieu, normalement une *villa* d'une certaine importance, donc, sa définition géographique manque de certitude pour les scribes qui le mobilisent moins que le *pagus* devenu un terme "passe-partout". Ainsi, l'*ager* est concurrencé dans la même fonction par d'autres termes qui proposent leurs ressorts administratifs ou domaniaux comme unités intermédiaires. Cela produit une répartition de ces unités entre les 4 termes spatiaux ayant des relations sémantiques similaires (*ager, vicaria, finis* et *comitatus* à quelques occasions).

ager

(uilla : 0.837), (sino : 0.817), (curtile : 0.729), (uinea : 0.728), (*pagus* : 0.725), (comitatus : 0.712), (uicarius1 : 0.703), (uocabulum : 0.689), (finis : 0.665), (indominicatus1 : 0.658), (coniaceo : 0.653), (campus : 0.646), (uulgo1 : 0.636), (territorium : 0.629), (termino : 0.612), (manus1 : 0.608)

Pour en finir avec les termes principaux, même si cette information n'est pas reflétée dans la matrice, l'ordre dans la séquence est très important. L'*ager* et le *pagus* apparaissent dans 9 observations sur 10 (88 %) dans le contexte de gauche (l'avant) de la *villa* ; le *locus*, par contre, dans 8 observations sur 10 (77 %) dans le contexte de droite (l'après) ; ce qui nous indique un respect rigoureux de la part des scribes de la formule hiérarchique de présentation des coordonnées spatiales : *pagus – ager – villa – locus*. Cet ordre formulaire reste canonique même lorsque *comitatus* et *episcopatus* prennent le rôle de cadres territoriaux.

Les termes concurrents

Comme nous l'a indiqué la matrice, les termes concurrents d'*ager* ne semblent pas se présenter emboîtés avec lui ; il y a seulement 13 localisations mobilisant ensemble un *ager* et une *finis* et 6 mobilisant un *ager* avec une *vicaria*. On peut dire la même chose concernant les termes concurrents du *pagus* : on détecte seulement 2 cas de *pagus-episcopatus* et 1 cas de *comitatus-episcopatus*[459]. Autrement dit, les termes peuvent dans certains cas coïncider mais que le plus souvent l'un remplace à l'autre dans la formule hiérarchique. Si on regarde d'un peu plus près ces 4 termes concurrents (*finis, vicaria, comitatus* et *episcopatus*) on verra qu'effectivement ils montrent des similarités sémantiques avec leurs termes « phare », mais aussi des rapprochements avec d'autres termes moins associés aux formules de localisation du fait d'une variabilité d'usage un peu plus large.

459. Il nous conste qu'ils existent quelques cas de plus d'unités concurrents se présentant ensembles, mais étant donné que notre matrice se limite aux cas où la *villa* est le centre, ils n'ont été pas récupérés. (Dans le tableau de la partie 3.2 formé sous de paramètres différents ces cas sont en revanche récupérés)

Dans le cas de *finis* on ne distingue pas clairement ce qui motive l'utilisation de ce terme ou de l'*ager* dans la formule. A. Déleage et M. Chaume avaient proposé que *finis* était une subdivision du finage, c'est-à-dire des terres exploitées par une même communauté rurale, mais le finage est une structure qui n'est pas bien attestée par les sources dans la Bourgogne altimédiévale car les actes ne parlent que rarement des populations. En tout cas, il semble exister un certain rappel à une structure ancienne dans le cas de la *finis* qui coïncide dans les grandes lignes avec l'*ager* et qui prend le même réseau de chefs-lieux. Si l'on compare les entités nommées accompagnant chaque terme, on voit qu'il n'existe quasiment aucune *finis* qui ne soit également dénommée comme *ager*.

finis

(*pagus* : 0.639), (aiacis : 0.636), (ortus1 : 0.635), (superpositus : 0.621), (ager : 0.596), (dexter1 : 0.582), ,(circuitus : 0.574), (portio : 0.568), (diuisio : 0.566), (territorius : 0.550), (consorto : 0.549), (ultra1 : 0.549), (terminum : 0.547), (flumen : 0.545), (molina : 0.544)

Les termes proches de *finis* ne coïncident pas forcement avec ceux proches d'*ager* qui apparaît 6 fois plus utilisé. On repère naturellement les termes de la formule de localisation (*pagus, ager, territorium*) et un autre terme synonyme de l'*ager*, mais qui n'apparaît qu'une vingtaine de fois dans les actes : *iacis*. Les autres termes qu'on voit apparaître sont mobilisés dans les indications de découpage territorial (*superpositus, dexter, circuitus, portio, diuisio, ultra, terminum*), etc. sont proches de l'homonyme de *finis* dans le sens de limite, confins des terres [460].

Dans le cas de la *vicaria*, traduit comme viguerie, le terme se rapporte comme dans le cas de *finis* et *comitatus*, aux chefs-lieux qui se trouvent dans le centre des *agri* et s'insère dans les rubriques spatiales classiques de localisation (voir 3.2). Son apparition dans le corpus est attestée depuis les premiers documents de l'abbaye au début du Xe siècle, mais son usage est très intermittent. La matrice ne nous en montre qu'une centaine d'exemples. Vers la moitié du XIe siècle les actes attestent un changement notable dans son champ lexical, *vicaria* perd le sens de sous-unité spatiale pour désigner un aspect fiscal et judiciaire. Le mot apparaît associé aux termes de l'exercice de la justice publique [461], dans les formules qui traduisent l'exigence de l'abbaye de libérer les biens donnés de tous les charges imposées par la seigneurie banale et de leur épargner tout droit et prérogative de domination.

vicaria

(coniaceo : 0.900), (adiaceo : 0.894), (territorium : 0.871), (castrum : 0.860), (uoco : 0.852), (curtilis : 0.848), (uilla : 0.847), (comitatus : 0.838), (nuncupo : 0.833), (colonia : 0.815), (curtis : 0.810), (alodium : 0.805), (burgus : 0.782), (*episcopatus* : 0.774), (uinea : 0.767), (uocabulum : 0.757), (capella : 0.755), (parochia : 0.753), (curtile : 0.752)

460. CBMA 1591. "... dono...una peciola de terra que est in pago Arueniaco, in uilla Linaria ; sunt autem fines eius, de una parte terra Ebrei, de alia gutta mortua...." ; CBMA 1642. "Infra as fines et terminaciones una cum arboribus, cum omnem superpositum, exsiuis, totum et sub integro tercia porcione tibi dono..."

461. CBMA 4736. « consuetudinibus et uicariis omnibus » ; CBMA 5673. « iusticiariis, vicariis et omnibus. »

Pour ce qui concerne *comitatus*, il semble un synonyme de *pagus* puisque dans la formule séquentielle (*pagus-ager-villa*), qui sert de repère pour la localisation, *comitatus* apparaît à la place de *pagus*. En fait, selon la matrice, et toujours dans le contexte de la *villa*, ils ne coïncident que deux fois. Mais *comitatus* fait plutôt référence à la juridiction locale établie autour des chefs-lieux dont quelques-uns sont pris comme centres de pouvoir des *pagi* et de quelques *agri*. Ainsi parfois *comitatus* et *pagus* accompagnent les mêmes entités nommées (*Uienense, Matisconense, Cabilonense, Aruenico, Lugdunense*), l'une faisant référence au domaine immédiat, plus restreint, du chef-lieu et l'autre à un cadre territorial théoriquement supérieur qui prend le nom du chef-lieu. La juridiction locale évoquée par le terme *comitatus* peut se référer, comme son nom l'indique, à celle d'un comte, mais dans nos actes ce sens apparaît à de rares occasions. Il est plus commun de le trouver faisant référence au réseau de juridictions ecclésiastiques établies au XIe siècle prenant comme nœuds les cités épiscopales. Si on fait une recherche par similarité, de nombreuses entités présentées par le terme *comitatus* sont aussi accompagnées par *civitas* (*Uienna, Ticinum, Niuerni, Bresie, Pictauense, Diensis*, etc.), dont le sens est également transformé car le concept antique de *civitas* n'est déjà à cette époque qu'un simple rappel de formulaire sans soutien dans la réalité (voir partie 6.6.1).

comitatus

(uicarius1 : 0.862), (episcopatus : 0.858), (uocabulum : 0.831), (nuncupo : 0.824), (territorium : 0.823), (coniaceo : 0.798), (*pagus* : 0.762), (curtis : 0.761), (adiaceo : 0.755), (castrum : 0.744),(capella : 0.735), (uilla : 0.719), (uulgo1 : 0.717), (ager : 0.712), (sino : 0.706), (territorius : 0.687), (abbatia : 0.686)

Naturellement *comitatus* se présente sémantiquement proche des autres entités d'encadrement territorial avec lesquels il cohabite (*vicaria, territorium*) mais surtout remplace (*pagus, episcopatus, ager*).

Finalement l'usage du terme *episcopatus*, à peu près au niveau du *pagus* est attesté vers la fin du Xe siècle dans notre corpus. *Episcopatus* et *comitatus* connectent vers les mêmes entités : les cites épiscopales. La matrice nous indique que même si à quelques exceptions près *comitatus*, *episcopatus* et *pagus* peuvent coexister, le plus souvent il prend le rôle d'unité majeure de l'espace dans la formule. *Episcopatus* est en fait la moins utilisée des structures alternatives précisément parce que *comitatus* exerce déjà ce rôle. Mais *episcopatus* ne sert pas que comme terme d'encadrement, il garde aussi son sens originel pour se référer à la juridiction d'un siège épiscopal et de leurs possessions. De ce fait, une partie importante des mentions à l'*episcopatus* sont faites dans le cadre de quelques bulles et privilèges papaux émis dans la deuxième du XIe siècle [462] qui confirment les monastères, biens et droits possédées par Cluny dans la région et où on les énumère, évêché par évêché. *Episcopatus* en fait, comme on le verra plus tard, continue d'être opérationnel lorsque le système du *pagus* est abandonné pendant la deuxième moitié du XIe siècle. Les termes tardifs d'encadrement spatial comme *parochia, provincia, burgus* se montrent comme sémantiquement proches à lui.

462. Voir par example, CBMA 4775, 4922, 5241, 5282, 5439

episcopatus

(capella : 0.899), (cella : 0.871), (castrum : 0.858), (comitatus : 0.849), (uocabulum : 0.839), (suburbium : 0.824), (construo : 0.824), (fundatus : 0.814), (uulgo1 : 0.813), (sanctae : 0.806), (insula : 0.805), (parochia : 0.797), (territorium : 0.794), (presbyteratus : 0.791), (prouincia : 0.787), (abbatia : 0.779), (curtis : 0.778), (consecro : 0.777), (ciuitas : 0.776), (uicarius1 : 0.774), (urbs : 0.763), (burgus : 0.750)

Les termes des unités de base

Par ailleurs, la matrice nous montre la relation entre la *villa* et les trois unités principales qu'elle abrite : *locus, ecclesia* et *terra*. La relation entre chacune des trois unités principales, *pagus-ager-villa*, et le *locus* (qui se présente souvent comme une sous-unité de la *villa* ou lieu-dit) est beaucoup moins intense : 373 mentions du tandem *villa-locus*, 77 pour *pagus-locus* et seulement 32 pour *ager-locus*. Le *locus* n'est donc associé régulièrement qu'à la *villa*. En particulier, dans 32 cas, car cela correspond à l'*ager-locus*, on observe que le scribe mobilise séquentiellement les 4 unités pour la détermination spatiale. Il n'est pas étonnant que les termes sémantiquement proches de *locus* soient *monasterium* et *cenobium* car *loco Cluniaco* est une collocation très répétée et exclusive de nos actes pour se référer à l'abbaye [463]. Dans ce sens *locus* joue un rôle similaire à ces deux termes en tant que synonymes pour se référer à l'abbaye.

locus

(monasterium : 0.668), (cenobium : 0.640), (alodio2 : 0.622), (cœnobio : 0.556), (res : 0.540), (uicarius1 : 0.525), (offero : 0.521), (mansus : 0.518), (seruiens : 0.514), (comitatus : 0.508), (ius : 0.502), (domus : 0.497), (possessio : 0.481), (edifico : 0.481), (curtile : 0.481)

Il faut dire que dans les formules de spatialisation, un lieu-dit n'est pas toujours présent avec le terme *locus*. C'est une pratique commune de présenter un lieu-dit en se rapportant au micro-toponyme vernaculaire. Dans ce sens on trouvera des nombreux exemples de lieux-dits décrits par des périphrases utilisant les verbes *voco* ou *dico* [464] ou par des prépositions de localisation (*apud, ad, in, supra*). La référence au nom vernaculaire est une ressource très répandue chez les scribes afin de préciser dans la mesure du possible la localisation d'un bien. C'est un usage que l'on trouvera aussi pour les unités de culture (*vinea, campus, curtile*, etc.). Les mentions de *loci* en tout cas, lorsqu'ils commencent à se faire de plus en plus nombreux, comportent un sens similaire à celui de la *villa* en tant qu'unité de base du découpage spatial ; mais malheureusement nos actes ne nous laissent pas percevoir son évolution au XIIe siècle.

463. CBMA 2104. "Ego Maluinus et uxor mea Guntrudis damus aliquid de rebus nostris ad locum Cluniacum" ; CBMA 2531. « Sacrosancto et exorabili loco Cluniaco, in honore beatorum apostolorum Petri et Pauli dicato… »

464. CBMA 1647 : "vendedisimus nos vobis campo qui es in agro Laliacense, in villa Belplano, ubi vocat Carnedo" ; CBMA 1673 : "Dedit Erembertus et Rainulfus in cambio Sancto Petro vel rectoribus Cluniacensis monasterii campum in pago Matisconense, in agro Rufiacense, in loco ubi dicitur Atfagia, per illum campum qui vocatur In illo Monte" ; CBMA 1611. "… vendedi ego tibi aliquid de res meas que sunt sitas in pago Matisconense, in agro Marciacense, in fine Cavaniacens : oc sont prael II, ubi vocat Asamuniago…"

La situation antérieure est aussi attestée dans le cas des deux autres termes l'accompagnant dans le groupe : *terra* (E) et *ecclesia* (F) qui apparaissent presque exclusivement après la *villa* (ou le lieu-dit) parce que dans la plupart des cas, ces deux termes font partie des formules de description des biens et non de celles de la localisation. Effectivement *terra* et *ecclesia* suivent les termes de localisation puisqu'ils sont mobilisés majoritairement dans deux cas : pour la définition de la situation spatiale des biens lorsqu'il s'agit de biens dénombrés (partie d'une vigne, de terre cultivée, etc.) ou pour signaler l'existence d'un bâtiment religieux (le cas de *ecclesia*) dans l'inventaire des biens donnés. Les formules métonymiques "*in terra* + nom du propriétaire" ou "*a mane ecclesia* + nom du saint" sont très utilisées par les scribes dans le but d'établir les confronts d'un bien donné que se trouve inséré dans un milieu où l'exploitation de la terre est assurée par plusieurs maîtres. Ceci explique que les termes sémantiquement proches de *terra* soient les points d'orientation (*oriens, serum, meridies, septentrion, condamina*, etc.) , des termes de segmentation territoriale (*partio, circius, pecia*) et des aires naturelles qui agissent aussi comme frontières (*boscus, campus, rivus, pratum*). Enfin, comme le terme *terra* est très utilise dans le but d'établir la situation spatiale des biens-fonds, il peut acquérir une nuance ethnographique et sociale : *in terra francorum, in terra hebreorum, in terra vicinorum* [465].

terra

(boscus : 0.781), (similiter : 0.773), (oriens : 0.754), (serum : 0.743), (pecia : 0.743), (partio1 : 0.733), (meridies : 0.724), (arabilis : 0.721), (circius : 0.719), (condamina : 0.707), (meridianus : 0.703), (septentrion : 0.702), (campus : 0.700), (occidens2 : 0.698), (riuus : 0.691), (uoluo : 0.683), (circuitus : 0.675), (pratum : 0.674), (curro : 0.672)

Pour sa part, le terme *ecclesia* présente une certaine polysémie. Il peut se référer au bâtiment de culte et aux propriétés associées (i), dont la totalité ou une partie est l'objet d'une donation [466], à une localité peuplée (ii), normalement une *villa* ou une paroisse, formées autour d'un lieu de culte [467] ; il peut prendre le sens de personne juridique réceptrice de donations (iii) [468], désigner la *sedes episcopalis* (iv) [469] et enfin, il peut former une collocation comme *locus* et *monasterium* pour se référer à l'abbaye de Cluny (v) [470]. Les termes sémantiquement proches agissent dans les mêmes fonctions : *cella, altar, cappella, predium, decima* pour (i) ; *parochia, provincia, diocesis* pour (ii) ; *presbiteratus, episcopatus* pour (iii et iv) et *abbatia, monasterium, cenobio* pour (v).

465. CBMA 2317 : "In primis donamus curtilum, qui terminat a mane terra Sichelmi, a medio die Sancti Petri, heredibusque meis, a sero Sancti Stephani et Argadi, a cercio terra Hebreorum." ; CBMA 3417 : "...videlicet unum mansum cum omnibus quæ ad ipsum aspiciunt ; qui terminatur a mane via publica, de alia parte terra Francorum, a sero terra Constantii..."

466. CBMA 3609. « ego dono...aliquid est rebus meis...hoc est medietatem ecclesie Sancti Germani, et capellam que est in uilla Domaniaco ».

467. CBMA 1519. "dono....in comitatu Secus Tyronense, Alairacum cum ecclesia in honore Sancti Petri... »

468. CBMA 3612. "pro remedio animarum nostrarum...donauimus ecclesie Sancte Marie Paterniensis nostrum alodum per cartas... »

469. CBMA 3283. "Dum ego Gausfredus, sancte Cabilonensis ecclesie presul..."

470. CBMA 2818. "ego Girbaldus senioribus de monasterio Cluniaco uendo duos campos in pago Matisconense, in uilla Besorniaco..."

ecclesia

(parochia : 0.606), (cella : 0.599), (altar : 0.595), (capella : 0.594), (diocesis : 0.590), (prioratus : 0.577), (prouincia : 0.574), (presbyteratus : 0.567), (predium : 0.563), (decima : 0.556), (episcopatus : 0.533), (abbatia : 0.532), (cœnobio : 0.531), (monasterium : 0.530), (martyr : 0.528), (consecro : 0.527), (patronatus : 0.524)

Les bâtiments

Finalement, la matrice nous présente deux termes, *monasterium* (G) et *conventus* (H), qui même s'ils ne font pas partie du vocabulaire d'encadrement apparaissent souvent dans le contexte de la *villa*. *Monasterium*, comme on l'a déjà vu, est surtout mobilisé par la formule d'adresse des modèles de l'acte de donation faisant normalement référence à l'abbaye de Cluny en formant une collocation. En conséquence les termes qui lui sont proches sont des synonymes désignant des bâtiments de la congrégation (*cenobio, abbatia, cella, prioratus*), mais aussi les verbes qui alternent dans la formule d'adresse des actes : *ad monasterium Cluniensem quod in honore beatorum apostolorum Petri et Pauli* ***constructum/consecratum/fundatum/sacratum/*** *est*

monasterium

(cœnobio : 0.781), (abbatia : 0.736), (prioratus : 0.730), (cœnobium : 0.680), (locus : 0.668), (consecro : 0.664), (ordo : 0.664), (fundo1 : 0.635), (instituo : 0.633), (regimen : 0.629), (specto : 0.621), (cella : 0.621), (martyr : 0.615), (insula : 0.611), (destitutio : 0.609), (construo : 0.602)

Conventus, pour sa part, est un terme qui apparaît tardivement dans le corpus (vers 1030) et qui cohabite avec *monasterium*, dont il est un synonyme, mais avec un sens moins institutionnel que celui adopté par *monasterium*. En fait les deux termes accompagnent de manière récurrente trois congrégations (*Cluniacense, Balmense, Trenorchiensis*) à partir de la deuxième moitié du XIe siècle. *Conventus* n'est pas un mot du contexte courant de la *villa* (seulement 19 coïncidences), mais davantage un mot référentiel mobilisé par les formulaires.

conventus

(canonica : 0.661), (religiosus2 : 0.645), (conuentum : 0.615), (prior : 0.572), (frater : 0.563), (capitulum : 0.557), (obedientia : 0.543), (nominatus2 : 0.542), (abbas : 0.525), (congregatio : 0.524), (uenerabilis : 0.519), (pater : 0.515), (eligo : 0.510), (cenobium : 0.506), (camerarius2 : 0.497)

En résume, nous avons observé la fréquence de la distribution des douze termes de classification spatiale les plus mobilisés dans les actes du recueil de l'abbaye de Cluny. Ils peuvent répondre au moins à deux regroupements, selon un critère géographique ou sémantique. Au niveau géographique (figure 6.6 ci-dessous), ils s'intègrent dans une hiérarchie spatiale centripète ayant le *pagus* comme circonscription globale et la *villa* comme le nœud élémentaire du paysage. Entre ces deux extrêmes, d'autres unités foncières cohabitent, la plus importante étant l'*ager*, qui agit comme unité intermédiaire, suivie de *comitatus, finis, episcopatus* et *vicaria*. Toutes ces unités font

référence à un territoire autour d'un chef-lieu. Ce chef-lieu n'est pas forcément dans le centre géographique mais peut agir comme centre organisateur d'une circonscription judiciaire, domaniale, ecclésiastique, etc. Sous le maillage de la *villa* se développe tout un système de découpage du paysage dont le *locus* agit comme une sous-unité portant souvent un micro-toponyme. D'autres termes comme *terra*, *ecclesia* et dans une moindre mesure *monasterium*, avec une myriade d'autres vocables mineurs, servent, d'un côté, à déterminer les frontières des biens distribués dans la *villa* qui sont objet de la donation, et de l'autre côté, à décrire les bâtiments de culte et exploitations qui forment le tissu fondamental de l'habitat.

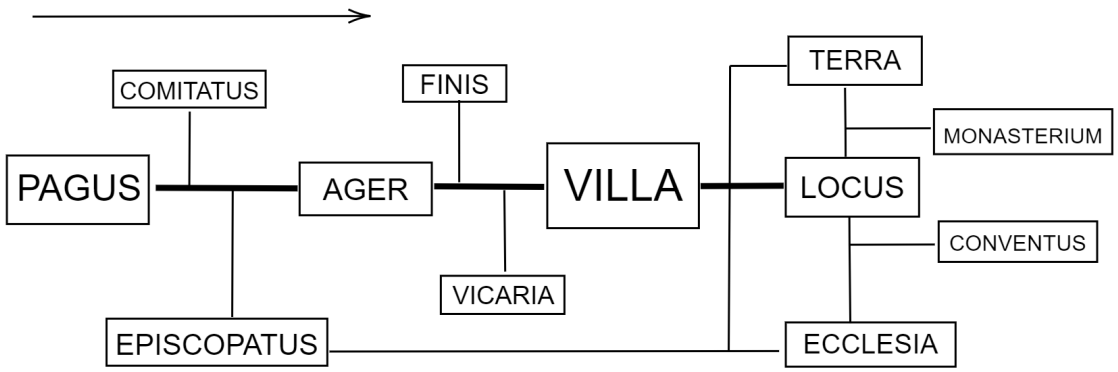

FIGURE 6.6 – Représentation des relations sémantiques entre les termes de classement spatial

Au niveau sémantique (figure 6.7) le panorama n'est pas très différent. Il existe un attachement très fort de la part des scribes au modèle hiérarchique et centripète sous le trinôme *pagus-ager-villa*. D'autres modèles non hiérarchiques et centrifuges existent, mais leur usage est bien plus limité. Les 8 unités d'encadrement de l'espace que nous avons analysées s'adaptent à cette rubrique spatiale. La présence de la *villa* y est frappante, au moins jusqu'à la première moitié du XIe siècle. Elle monopolise tout le système de détermination spatiale, au point qu'il existe très peu d'indications géographiques de biens-fonds qui ne la mentionnent. Mais s'il est vrai que la *villa* domine, il n'est pas moins vrai que le *pagus* l'accompagne dans 7 observations sur 10 (68 %). La robustesse du lien *pagus-villa* contraste avec la décadence progressive du lien entre la *villa* et l'*ager*, malgré le recours aux autres unités intermédiaires. En fait, ce que nous dit la matrice c'est que pour les scribes la mobilisation de l'*ager* était problématique et que le rapport aux autres unités n'avait pas réussi à complètement contrebalancer cette difficulté. En ce que concerne les unités abritées par la *villa*, leurs présentations suivent aussi les patrons des formulaires. La mention de la *villa* ou optionnellement du *locus*, qui lentement prend le rôle d'unité fondamentale, ouvre la voie aux formules de description et d'inventaire, qui déclenchent la multiplication des mentions aux *terrae, ecclesiae* et *monasteria* dans le but de préciser les limites des biens donnés ainsi que leur nature et leur valeur.

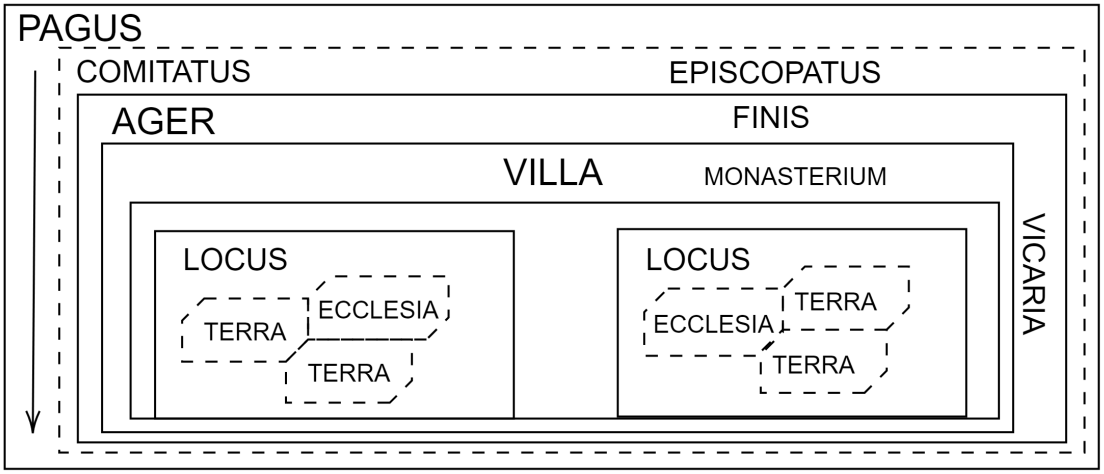

FIGURE 6.7 – Emboitement géographique du système régi par le *pagus*

6.5 Vision spatio-temporelle des cadres territoriaux.

L'analyse des champs sémantiques du vocabulaire de l'espace nous a montré quelques caractéristiques assez spécifiques de nos actes concernant les pratiques de localisation et de descriptions des biens-fonds :

1. Le système de coordonnées spatiales utilise un quadrinôme hiérarchique ayant la *villa* comme le nœud central. La grande majorité des localisations s'y rapportent et elle joue le rôle dans le formulaire d'élément charnière entre localisation et inventaire des biens.
2. L'unité majeure d'encadrement territorial, le *pagus*, et surtout l'unité intermédiaire, l'*ager*, se trouvent concurrencés par d'autres unités de bornes plus définies dont les ressorts se chevauchent sur un même territoire.
3. Les unités inférieures, notamment *locus* et *ecclesia* voient augmenter leurs occurrences et leurs relations sémantiques contrairement à ce que survient dans les unités majeures d'encadrement. Les matrices les présentent comme troisième et quatrième termes les plus utilisés.

Nous avons déjà introduit quelques remarques temporelles, repérées a posteriori, dans notre analyse sur les relations sémantiques, sans lesquelles les résultats demeuraient trop statiques. À présent nous allons vérifier cette coévolution du quadrinôme, en nous servant d'un histogramme chronologique. Dans le graphique qui suit (figure 6.8) nous représentons le classique polygone de fréquences formé par l'union des différents points médians des fréquences de l'usage de chacun de ces termes cumulés à chaque décennie. La distribution est destinée à refléter les cycles et tendances des cinq variables (*villa, locus, ecclesia*, unités intermédiaires et unités supérieures) sur des séries temporelles.

En observant le graphique un premier constat est clair : notre vision sur les rythmes de transformation dans les structures du paysage est conditionnée par les cycles de conservation et production des actes. On voit que l'utilisation des unités de classement

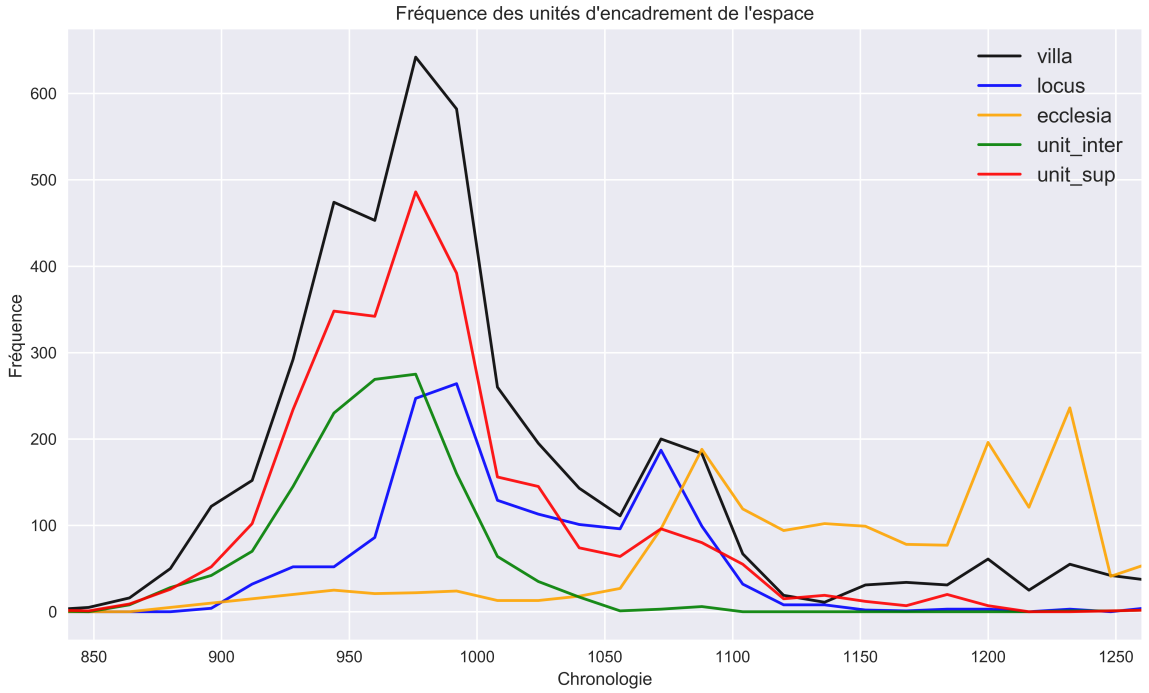

FIGURE 6.8 – Fréquence d'usage des unités d'encadrement de l'espace. Les unités intermédiaires (unit_inter) incluent *ager, finis* et *vicaria* ; les unités supérieures (unit_sup), *pagus, comitatus et episcopatus*.

spatial augmente de façon exponentielle depuis l'époque de l'abbé Odon (926-942), connaît un point culminant entre les décennies de 960-990, période de l'abbatiat de Maïeul (954-994) et décline après l'an mille. Cela coïncide avec les grandes lignes de la production générale des actes. Donc, est-ce que les transformations attestées sur le plan sémantique correspondent plutôt à des fortes variations dans la production et conservation qui biaiseraient les statistiques ?

Il n'est pas étonnant que ces périodes en question coïncident avec les longs abbatiats. Un nombre important d'actes dépourvus de datation sont datés dans une fourchette à partir des périodes des abbatiats qui y sont mentionnés. Nous avons resserré ces dates en mobilisant des matrices de datation comme celles construites dans l'étude autour de Paray-le-Monial. Mais comme on l'a vu dans cette étude, une meilleure précision requiert un travail bien plus minutieux pour un nombre très élevé d'actes. En tout cas, on sera toujours dépendants des dates des abbés dans le cas de manque de datation.

Cela dit, nous devons nous concentrer sur les deux fluctuations cycliques présentées par la distribution fréquentielle de nos données : celle des décennies 960 à 1000 (période l'abbé Maïeul) et 1050 à 1090 (période de l'abbé Hugues).

Dans le graphique, si on regarde l'ascension (point i), le sommet (point ii) et le flanc (point iii) du mont formé entre les décennies 920 à 1010 au cours de laquelle la conservation et production montre une relative stabilité, on peut dégager quelques observations :

1. Le rythme d'évolution est symétrique entre les trois unités de la hiérarchie entre

les points i et ii, à ce moment chacun prend sa propre vitesse de progression, mais garde une similarité dans les grandes lignes ;

2. Les unités intermédiaires se dégagent du rythme de progression deux décennies avant les autres et subissent un déclin plus précoce et plus prononcé, au point de quasiment disparaître au cours des point ii et iii ;
3. Les unités supérieures et la *villa* suivent, en proportion, une évolution en miroir depuis le point ii et gardent la même proportionnalité jusqu'à la récupération de la production dans la décennie 1050 ;
4. Les unités inferieures se joignent tardivement aux rythmes de transformation. Le *locus* remonte dans la décennie 980 et *ecclesia* dans la décennie 1050. Tous deux souffrent du déclin de la production des actes, mais résistent bien mieux que les unités d'encadrement.

En regardant plus en détail, l'abandon de l'*ager* avait commencé quelques décennies avant les autres termes comme le montre le graphique ci-dessous qui atteste que les unités concurrentes, *finis* et *vicaria*, étaient déjà en usage dans les documents antérieurs à la fondation de l'abbaye de Cluny (910). Tout l'appareil formé par les unités intermédiaires décroît après la décennie de 960 et est en net déclin jusqu'à la fin du siècle. Dans la première décennie du XIe siècle l'*ager* est déjà peu usité et il disparaît peu après. *Finis* et *vicaria* le suivent. Quand le niveau de conservation et production reprend à l'époque de l'abbé Hugues (1049-1109), l'*ager* a disparu alors que la *villa* et les unités supérieures demeurent opérantes.

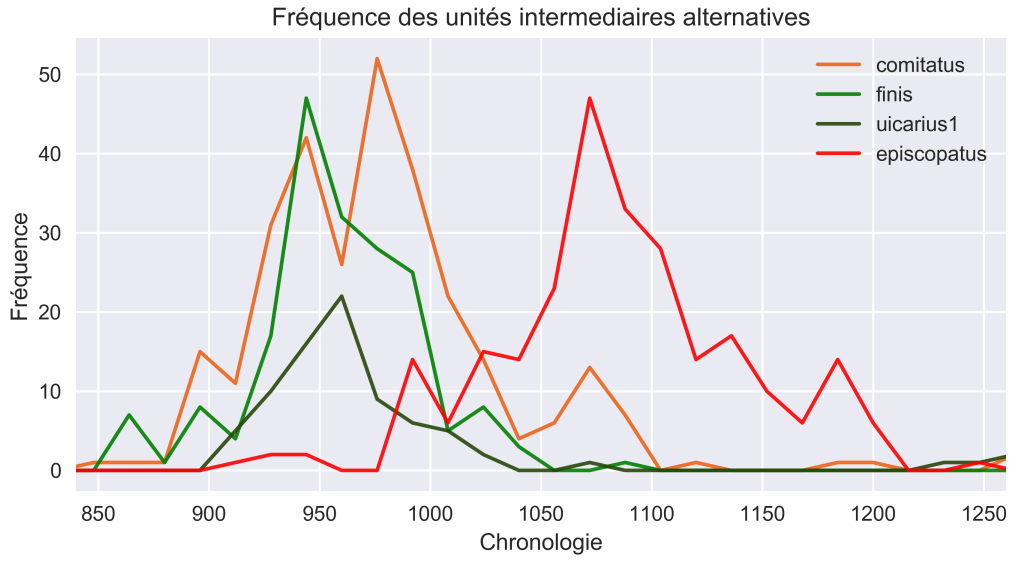

FIGURE 6.9 – Fréquence des unités alternatives intermédiaires et supérieures

Dans nos actes le *pagus* commence son déclin deux décennies après l'*ager*, mais sa progression est marquée par deux spécificités. Les termes concurrents, *comitatus* et *episcopatus* émergent à deux moments décalés. Le *comitatus*, qui est parfois pris dans le sens de l'*ager*, le rejoint très tôt, depuis la fin du VIIIe siècle, et suit une

évolution similaire à celle de l'*ager*, présentant un point culminant à partir de la décennie 950 et disparaissant après les premières décennies du XIe siècle. Par contre, l'*episcopatus* commence à être mobilisé autour de l'an mille et connaît un point culminant précisément lorsque la production des actes reprend. À ce moment (la décennie de 1060) on utilise plus l'*episcopatus* que le *pagus* pour se rapporter au maillage supérieur et en fait, l'*episcopatus* perdure en tant que terme spatial jusqu'à la fin du XIIe siècle. Le recours à ces deux structures adoucit le déclin des unités supérieures et en conséquence on les trouve encore actifs après la décennie de 1090, moment où notre corpus change brusquement car l'acte privé de donation disparaît pratiquement des copies et cartulaires.

D'ailleurs, bien que les unités alternatives ne dépassent pas les 50 occurrences par décennie à leur point culminant, on voit bien qu'elles suivent les mêmes rythmes de progression que les unités "phare" et qu'elles sont abandonnés au même moment. L'intention des scribes n'était pas en effet de remplacer l'*ager* ou le *pagus* par d'autres structures plus d'actualité. Mais, comme on le constate dans le graphique, les six termes co-évoluent ensemble vers l'abandon qui se produit d'abord au sein de la hiérarchie intermédiaire et plus lentement dans celui des unités supérieures, au fur et à mesure que les localisations se circonscrivent à l'intérieur du réseau territorial.

Dans ce sens, la coévolution la plus intéressante est celle instaurée entre la *villa* et le *locus*. Dans les formules classiques de localisation le *locus* est la quatrième instance après la *villa* et cela persiste effectivement jusqu'à la première moitié de XIe siècle. Ensuite, la *villa* perd sa prééminence comme unité fondamentale précisément par l'action du *locus* dont les observations prennent une direction différente de celle des unités d'encadrement. Nous avons vu dans les matrices que la relation entre le *locus* et la *villa* qui devrait être par nature plus proche ne l'est pas et que le quadrinôme de l'espace incluant le *locus* est en réalité peu mobilisé. Cela se produit parce que le *locus* ne se rapporte pas vraiment à une unité de délimitation, mais agit comme terme de sens général qui désignait sûrement des situations très différentes, ce qui l'amenait parfois à être pris comme une *villa*, un *mansus*, un *curtile*, un *castrum*, etc. Le *locus* se présentait alors comme un terme de précision spatiale optionnel pour les scribes qui continuent à se rapporter essentiellement à la *villa*. En tout cas, *locus* comme *villa* souffrent du fort déclin de la production d'actes en 990, mais dans les décennies qui suivent le *locus* progresse jusqu'à arriver au même niveau que la *villa* vers 1050. Le *locus* passe alors d'un terme optionnel dépendant de la connaissance directe à un terme quasiment obligatoire concurrent de la *villa*.

Ainsi, les quatre décennies de conservation et de production maximale entre 960 et 990 nous montrent un dégagement séquentiel des trois éléments de la hiérarchie classique de détermination territoriale – *pagus*, *ager*, *villa* - qui évoluent différemment pendant la période de césure de l'abbatiat d'Odilon et cèdent la place à une nouvelle structuration formée pendant cette période et pleinement opérationnelle à l'époque de la reprise, pendant l'abbatiat de Hugues entre les décennies 1060 et 1090. Il est vrai que cette reprise de production ne la fait pas revenir aux niveaux observés auparavant, mais elle affiche un portrait bien différent à celui de 990. En 1050, le *locus* et la *villa* montrent des niveaux assez similaires d'usage, alors que les unités supérieures, dominées par l'*episcopatus*, sont d'un usage déjà restreint. Avant cette période, les

relations entre les unités se sont déjà transformées, mais la pénurie des actes pendant l'abbatiat d'Odilon (994-1049) ne nous permet pas d'observer précisément le processus. En tout cas, dans la décennie de 1060, le trinôme classique *pagus-ager-villa* a été remplacé par un binôme alternant constitué par *villa*/*locus* – *pagus*/*episcopatus* et leurs différentes combinaisons.

Ce portrait qui débouche sur les unités locales est complété par la progression frappante, dans le deuxième cycle du graphique (après 1060), d'*ecclesia* qui se montre le terme le plus polysémique de tout le vocabulaire de l'espace (voir 2.3). L'*ecclesia* est à peine mobilisé durant les transformations subies par le reste du vocabulaire et son niveau d'usage semble immuable jusqu'à la deuxième reprise des actes vers la décennie 1060. Mais à ce moment-là son usage se multiplie exponentiellement jusqu'à dépasser celui de tous les autres termes du vocabulaire spatial. Nous sommes en fait dans le moment de la réforme grégorienne dont la restauration des structures ecclésiastiques propose nouveaux points d'ancrage, les bâtiments ecclésiastiques, pour la structuration de l'espace [471].

Ecclessia est un terme qui s'intègre en partie dans les hiérarchies formulaires au niveau du *locus* ou d'un sous-*locus* ou, ce qui est plus rare, agit comme référent au territoire d'un siège ecclésiastique. Mais, il garde (et multiplie) aussi son sens de bâtiment de culte et d'un centre peuplé autour de ce bâtiment. Cette polysémie est une des raisons pour lesquelles, lorsque l'acte de donation, disparaît le terme semble statistiquement bien moins affecté que les autres car cela n'affecte qu'un ou deux de ses multiples *semas*.

Cette croissance de l'usage d'*ecclesia*, alors que l'ancien système d'encadrement territorial est encore opérant, est aussi le symptôme de l'apparition d'un deuxième système de localisation bien plus simple car il se rapporte directement au bien donné. Afin de regarder la cohabitation de ces deux systèmes, dans le graphique ci-dessous (figure 6.10) nous avons affiché une nouvelle matrice avec les relations contextuelles (fenêtre de 20 mots) entretenues par nos 5 termes de détermination spatiale entre les décennies 1050 et 1090.

D'après le graphique il existe encore, pendant le deuxième cycle, des actes incluant une localisation foncière suivant l'ancien système, ce qui explique que le *pagus* soit encore mobilisé presque toujours en association avec la *villa* et le *locus*. L'*episcopatus* par contre, bien qu'associé partiellement à la *villa* et mobilisé dans le trinôme spatial, est surtout utilisé en présence de l'*ecclesia*. Cela correspond en partie aux premiers privilèges et confirmations des biens de Cluny émis par papes et évêques dont la localisation des biens ne se trouve plus dans le cadre d'un acte de donation, mais d'une liste énumérative d'*ecclesiae* et de *res ecclesiae* en possession de l'abbaye dont la localisation se rapporte au binôme *episcopatus-ecclesia*. Ce qui semble plus important est la différence entre le nombre d'observations et le nombre de connections : les termes se trouvent très déconnectés les uns des autres. La *villa* continue à être le terme le mieux connecté et le plus utilisé mais d'une manière déjà très dispersée (571

471. Voir au sujet des transformations des repères spatiaux dans le cadre de la réforme grégorienne : Florian Mazel. *La réforme" grégorienne" dans le Midi, milieu XIe-début XIIIe siècle*. 2013 ; Didier Panfili. *L'évolution des repères spatiaux en Bas-Quercy et Haut-Toulousain de 930 à 1130 : une approche des transformations sociales et des paysages agraires*. 2004

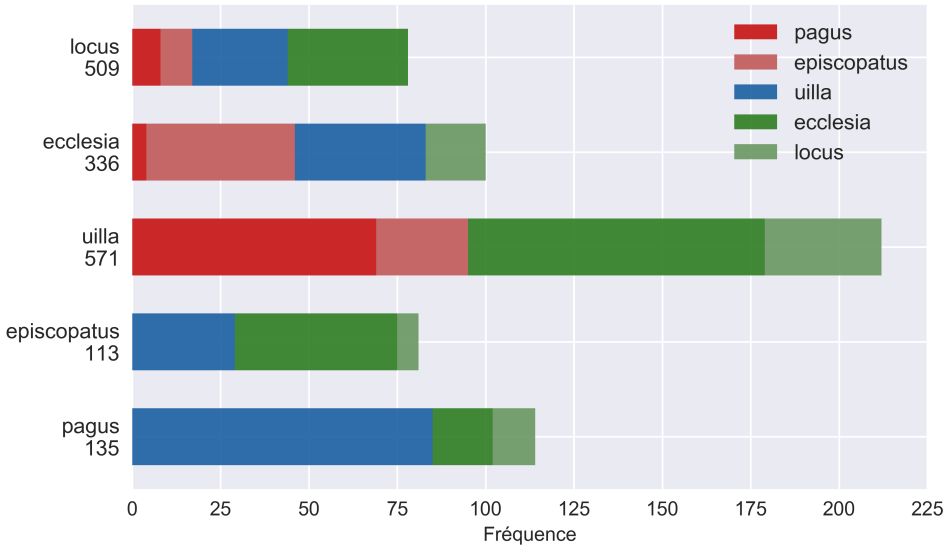

FIGURE 6.10 – Fréquence des relations contextuelles entre les 5 termes d'encadrement spatial pour la période 1050-1090. Les chiffres sous les termes indiquent le nombre d'observations totales, les bars indiquent le nombre d'observations connectées à d'autres termes.

observations dont 212 connections). À l'autre bout se trouve le *locus*, le terme le moins connecté mais utilisé par les scribes à une fréquence similaire à celle de la *villa*. Il ne s'agit pas seulement du fait que le *locus* et l'*ecclesia* se rapportent peu aux hiérarchies de l'espace, jouant plutôt le rôle du point solitaire de localisation, mais que *loci* et *ecclesiae* se sont multipliés massivement à l'intérieur des cadres territoriaux naguère et encore définies par la *villa*. C'est la raison qui explique pourquoi par exemple la connexion *villa-ecclesia* est bien plus forte que la relation *ecclesia-villa*.

En résumé, la formule hiérarchique classique devenue un binôme persiste pratiquement jusqu'à la fin du XIe siècle, mais alors la polarisation dans les localisations s'est complétée. Les unités intermédiaires ont disparu et la mobilisation des unités supérieures se trouve très altérée. Le couple principal de localisation formé par le *locus* et la *villa* est statistiquement bas. A ceux-ci s'ajoute abruptement *ecclesia* qui parmi ses multiples significations permet aussi de désigner le centre peuplé formé autour d'un lieu de culte. Donc, en gros, nous passons d'un système qui vers 950 se rapportait encore strictement à l'*ager*, au *pagus* et à la *villa* à un autre, un siècle plus tard, qui se rapporte de façon moins stricte et plus éparse à l'*episcopatus* comme maillage supérieur et au *locus* et l'*ecclesia* comme topos, cellules de l'habitat humain [472].

472. Voir au sujet de la territorialisation du sacré et de la polarisation de l'espace sur les lieux-dits les travaux de : Michel FIXOT et Élisabeth ZADORA-RIO. "L'environnement des églises et la topographie religieuse des campagnes médiévales". In : *Documents d'archéologie française* 46 (1994) ; LAUWERS et RIPART, "Représentation et gestion de l'espace dans l'Occident médiéval" ; Étienne HUBERT. "Quelques considérations sur l'organisation de l'espace, la propriété foncière et la géographie du peuplement dans la vallée du Turano (IXe-XIIIe siècle)". In : *Quelques considérations*

Une dernière observation est nécessaire. Même si d'après la figure 6.8 dans la deuxième partie du XIIe siècle on voit une légère récupération de l'usage du terme *villa*, une augmentation de l'utilisation de *ecclesia*, et la survivance déjà mentionnée de l'*episcopatus*, ils n'appartiennent plus aux usages formulaires de localisation spatiale. Ils sont effectivement mobilisés en tant que co-occurrences d'entités nommées mais exclusivement dans le cadre, déjà mentionné, de recensement de biens et propriétés de l'abbaye cédés ou confirmés par les bulles, privilèges ou lettres. Ceci doit être compris comme une conséquence de la rédaction des cartulaires D et E de l'abbaye de Cluny qui rassemblent plutôt les documents concernant ses relations avec le pouvoir public, notamment la papauté et l'évêché à partir du XIIe siècle.

6.6 La reconstruction cartographique des unités intermédiaires.

Un dernier exercice peut être très intéressant dans le but de montrer plus en détail un des aspects les plus remarquables de nos actes : la concurrence de termes dans le système topo-spatial, notamment entre les unités intermédiaires. Cet exercice est facilité par deux solutions techniques déjà mobilisées dans les chapitres précédents de cette thèse :

a. La récupération et le regroupement, par la voie de la similarité, des entités nommées accompagnées par les termes de l'espace en tant que co-occurrences ;
b. L'interrogation, à partir de la collection d'entités nommées, des volets de toponymie ancienne contenus dans le dictionnaire géographique de Saône-et-Loire [473] dont l'édition numérique en format XML [474] peut être interrogée par des requêtes XQUERY.

Dans cet exercice nous allons présenter un tableau renseignant les cas où sur une même entité nommée convergent trois ou plus termes de l'espace, ce qui peut ici être très éloquent pour vérifier différents phénomènes autour des décisions de localisation géographique prises par les scribes et de la nature des chefs-lieux présentés comme centres locaux du système de découpage territorial. Une connexion avec le cadastre actuel, moyennant les coordonnés fournies par le dictionnaire topographique, est exigée, afin de tirer une corrélation géographique à partir des descriptions textuelles.

Deux considérations générales au sujet de notre corpus sont ici nécessaires :
Nous avons identifié dans les recueils de l'abbaye de Cluny environ de 582 *villae*, 151 *agri* et 46 *pagi* [475] différents. Le nombre exact est impossible à fournir du fait des

sur l'organisation de l'espace, la propriété foncière et la géographie du peuplement dans la vallée du Turano (IXe-XIIIe siècle) (2000), p. 1000-1025

473. Rigault, *Dictionnaire topographique du département de Saône-et-Loire : comprenant les noms de lieux anciens et modernes*

474. http ://cths.fr/dico-topo/dictionnaires/cartes.php ?cdep=71

475. Il faudra plutôt dire 580, 156 et 47 toponymes nommés comme étant *villa*, *ager* et *pagus* respectivement, sans que cela signifie qu'effectivement ce nombre d'unités a eu une existence véritable, comme on le verra ensuite.

	Toponyme	uilla	*ager*	finis	uicaria	*pagus*	episcop	comitatus	ciuitas
1	Ambianensi	1	0	0	0	1	6	0	0
2	Amniaco	2	1	1	0	0	0	0	0
3	Antisiodorensi	0	0	0	0	1	6	0	1
4	Aruernensi	0	0	0	0	7	8	5	1
5	Atensi	0	0	0	0	1	1	3	0
6	Augustudunense	0	0	0	0	81	10	13	1
7	Aurelianensi	0	0	0	0	1	2	0	1
8	Beluacensi	1	0	0	0	1	5	0	0
9	Bisumtinensi	0	0	0	0	1	5	0	1
10	Buferiacensi	28	2	0	4	0	0	0	0
11	Burgundia	13	0	0	1	1	0	4	0
12	Cabilonense	0	0	0	0	123	5	22	11
13	Cadiacense	0	4	2	2	0	0	0	0
14	Canauiacense	1	4	0	1	0	0	0	0
15	Caturcensi	0	0	0	0	6	2	1	0
16	Cauaniacense	4	6	2	0	0	0	0	0
17	Casniaco	0	2	1	1	0	0	0	0
18	Ciciacense	6	11	14	0	0	0	0	0
19	Claromontense	0	0	0	0	1	1	1	0
20	Diense	0	0	0	1	3	4	1	1
21	Domciacensi	18	9	4	0	0	0	0	0
22	Euoriacense	1	11	4	0	0	0	0	0
23	Fabriacensi	3	14	14	0	0	0	0	0
24	Flagiacensi	24	1	1	0	0	0	0	0
25	Forense	0	1	0	0	4	0	5	0
26	Fusciacens	0	40	1	1	0	0	0	0
27	Galunniacense	3	64	1	0	0	0	0	0
28	Ibgiaco	34	11	1	1	0	0	0	0
29	Lanciaco	5	1	1	0	0	0	0	0
30	Lugdunense	0	0	0	0	172	13	19	2
31	Maciacense	15	74	6	1	0	0	0	0
32	Marziago	26	102	2	0	0	0	0	0
33	Matisconense	2	0	0	1	1504	32	143	6
34	Maxiriacense	0	16	2	0	0	0	0	0
35	Mariaco	1	5	1	0	0	0	0	0
36	Mediolanense	0	12	2	0	0	0	0	3
37	Meldensi	1	0	0	0	3	3	0	1
38	Miliacensi	2	1	1	0	0	0	0	0
39	Parisiensi	0	0	0	0	1	1	0	1
40	Pictauensi	1	0	0	0	3	3	2	3
41	Prisciacens	19	17	1	1	0	0	0	0
42	Prissiaco	3	1	0	1	0	0	0	0
43	Regense	0	0	0	0	2	1	3	0
44	Rofiacense	61	62	18	1	0	0	0	0
45	Salorniacense	7	21	1	0	0	0	0	0
46	Sancti Germani	2	0	1	1	0	0	0	0
47	Sancti Mauricii	1	2	0	0	0	0	0	0
48	Tolosano	1	0	0	0	2	6	3	1
49	Trecassino	0	0	0	0	0	1	3	1
50	Ualentiaco	3	1	0	0	5	5	2	1
51	Uersiacense	23	3	1	1	0	0	0	0
52	Uienense	0	0	0	0	59	6	8	12
53	Uiriaco	24	1	0	1	0	0	0	0
54	Uualdense	1	0	0	0	1	0	1	0
55	Uzetico	0	0	0	0	3	0	1	1
Total obs.	-	337	500	83	20	1987	126	240	49
% Total	-	9 %	46 %	44 %	25 %	90 %	51 %	82 %	38 %

TABLE 6.1 – Toponymes concentrant trois ou plus structures de l'espace. Total obs : total d'observations récupérées par le tableau ; % total : pourcentage selon le total d'observations existantes dans le corpus

nombreux problèmes d'identification mentionnés auparavant (homonymie, ambiguïté des frontières, doublons, imprécisions, erreurs du copiste, etc.). Cela correspond à peu près à 4280 observations de *villa*, 2156 de *pagus* et 1095 d'*ager*, comme il est affiché dans la liste analysée de cooccurrences. Le tableau ci-dessus (tableau 6.1), qui cherche à afficher les cas de multiples coïncidences sur un même toponyme, relève surtout des chefs-lieux et cités, car sont les toponymes prises comme centres des unités intermédiaires et supérieures respectivement et par conséquent vont accumuler deux ou plus termes de l'espace. Pour former les collections de toponymes nous avons de plus appliqué la ressemblance sur la racine communément partagée car la transformation en unité d'encadrement se fait normalement en ajoutant des suffixes : *-ico, -aco, -ense, -ensi*[476].

Par ailleurs, bien que le tableau nous montre un nombre restreint de toponymes (54), cet ensemble accumule un pourcentage important de toutes les observations d'usage des unités intermédiaires et supérieures, notamment dans le cas du *pagus* (90 %) et du *comitatus* (82 %) en grande partie grâce à la présence du *comitatus Matisconense* et du *pagus Matisconense* auquel se rapportent les deux tiers des donations. Dans les cas de l'*ager*, la *finis*, la *civitas* et de l'*episcopatus*, les toponymes affichés dans le tableau cumulent environ la moitié de tous les usages observés dans le corpus.

De ce tableau on peut dégager différentes observations :

Premier constat : il existe une division très claire entre le groupe des unités intermédiaires et celui des unités supérieures. Dans deux cas seulement (*Matisconense* et *Forense*), un toponyme est partagé entre les deux groupes. Donc, les chefs-lieux qui se trouvent au centre de la structure de l'*ager* et de la *finis* et auxquels est aussi affectée la *vicaria* sont rarement pris comme le centre d'unités supérieures ; et à l'inverse, les centres peuplés et cités dont la circonscription est prise comme une unité supérieure ne sont quasiment jamais pris comme centres de référence pour les unités intermédiaires. La différence se rapporte à celle observée entre d'un côté les espaces auto-organisés qui peuvent souffrir des multiples réarrangements (unités intermédiaires) et d'un autre côté les territoires organisés et produits par un pouvoir se montrant moins sujets aux dynamiques temporelles et aux changements (unités supérieures) [477].

Deuxième constat : les chefs-lieux sont normalement dénommés comme *villae*. En fait, la *villa* apparaît dans le tableau parce que dans plusieurs cas elle partage le toponyme avec l'*ager* et la *finis* et à quelques occasions avec la *vicaria*. Dans le tableau 4 cas seulement attestés (*Cadiacense, Casniacence, Maxiriacense, Forense*) où un toponyme d'*ager* n'apparaît pas comme étant une *villa*. Dans le cas de la *finis*, aucun toponyme n'est indépendant se montre indépendant de l'*ager*, s'agissant alors d'un synonyme statistiquement parfait.

Troisième constat : il existe parmi les unités supérieures une association relativement forte entre *pagus* et *episcopatus*, deux termes très liés à la *civitas* car 15 des 16 *civitates* sont mobilisées comme centre d'un *episcopatus* et *pagus*. Les anciennes

476. Par ex. Ruffiaco —> Ruffiacense, Cluny —> Cluniaco.

477. Voir à ce sujet les études de : Cursente, "Autour de Lézat : emboîtements, cospatialités, territoires (milieu X-milieu XIII siècle)" ; Gérard Chouquer. "Une année d'exception pour l'archéogéographie". In : *Études rurales* 173-174 (2005), p. 297-324

cités prennent en effet rapidement place comme évêchés et dans plusieurs cas les scribes s'y rapportent en tant qu'unité d'encadrement mais aussi parfois comme *pagus*, même si cela peut ne pas correspondre à une autre réalité que rédactionnelle.

Quatrième constat : on a vu dans les matrices sémantiques que le *comitatus* est dans certaines occasions mobilisé comme unité intermédiaire, mais dans le tableau les coïncidences *comitatus* - *ager* sur un même toponyme sont inexistantes. Le *comitatus* est en réalité peu mobilisé et lorsqu'on le mobilise il est surtout englobé par les mêmes macro-toponymes des *pagi* les plus connus (*Augustodonense, Matisconense, Cabilonense, Lucdunense, Valentianense*, etc.).

Finalement, le tableau n'affiche que 25 % d'observations de la *vicaria*. Il s'agit du terme le moins mobilisé (102 observations dans tout le corpus) et plus épars. Dans le tableau, il se rapporte surtout au chef-lieu de l'*ager* et par extension à la *villa*. Ici, dans seulement deux cas on voit une *vicaria* se rapportant aussi à un macro-toponyme (*Matisconense* et *Diense*). Dans le reste des observations, la *vicaria* est effectivement prise comme unité intermédiaire.

Regardons de plus près quelques-uns des toponymes les plus connus afin de mieux observer les différents phénomènes que le tableau nous indique :

6.6.1 Unités supérieures autres que celle de Mâcon

1. *Augustudonense* (l'Autunois), *Cabilonense* (le Chalonnais), *Lugdunense* (le Lyonnais), *Meldensi* (Meldes, Meldorum, Meaux), *Tolosano* (Toulouse), *Pictavense* (Poitiers), *Valentiaco* (Valence), *Arvenense* (Auvergne), *Bisumtinensi* (Besançon), *Viennense* (Vienna-le-Delphinat), *Trecassino* (Tricassium), etc.

Les *pagi* confrontant celui de Mâcon sont mentionnés dans notre corpus avant même la fondation de l'abbaye en 910. L'ensemble des possessions des puissants de la région étant dispersées entre plusieurs *villae*, les donations localisées dans des *villae* de l'Autonois, le Lyonnais et le Chalonnais étaient très communes pendant le Xe et la première moitié du XIe siècle. Environ 95 % des donations enregistrées dans le recueil avant la décennie de 1030 se rapportent à ces quatre *pagi*. Les mentions des *pagi, comitatus* ou *episcopatus* plus proches, comme ceux de l'Auvergne, l'Yonne, Valence, Forez, Besançon et même d'autres, plus éloignés, comme ceux de Beauvais, Paris, Meaux ou du Poitou commencent a augmenter a l'occasion de la grande entreprise de confirmation et de récapitulation des biens-fonds (principalement des églises) à laquelle se livre l'abbaye à partir de la deuxième moitié du XIe siècle, comme cela est attesté dans des nombreux actes provenant des évêchés et de la papauté copiés dans les cartulaires D et E.

Les anciennes cités gallo-romaines devenues sièges épiscopaux sont avant même la décennie de 1050 prises comme le centre de la dénomination de l'*episcopatus* dans les actes. Les privilèges et confirmations de propriétés, nombreuses après de cette décennie, multiplient les localisations sous forme d'inventaires mobilisant principalement le tandem *episcopatus-ecclesia*[478]. Ici on peut aussi trouver l'origine de plusieurs usages

478. CBMA 5518 : « ... item in episcopatu Meldensi, ecclesiam de Firmitate Ansculfi, ecclesiam de Chailli, de Chamini, de Bussei, de Bellovidere, ecclesiam Sancti Christophori in suburbio... » ;

de *civitas* (plus rarement *urbs*[479]) pour chacun de ces toponymes faisant référence tant au centre peuplé qu'à sa juridiction mais introduits tardivement, à partir de la fin du XIe siècle[480].

Sur ces toponymes relativement nouveaux quelques scribes mobilisent la dénomination *pagus* qui peut être repérée dans une chronologie quelques décennies plus tôt (1030-1060) en utilisant surtout le binôme *pagus-villa* qui à l'époque est encore opérationnel. Les exemples sont cependant peu nombreux[481]. La dénomination *comitatus* est utilisée à la même époque que le terme *pagus* mais en présentant deux légères différences : elle est plus adoptée par les chartes émanant du pouvoir civil[482] et lorsqu'elle est adoptée par les papes et les évêques elle semble répondre à une équivalence parfaite avec l'*episcopatus*[483] au sens géographique.

Comme on l'observe dans le tableau, toutes les unités supérieures partagent largement les mêmes centres de référence qui coïncident en grande partie avec les cités, les sièges épiscopaux et les grands centres peuplés. Il peut ainsi arriver que dans certaines localisations il existe une superposition des termes qui se rapportent aux unités supérieures. Il s'agit d'un emboîtement dans la hiérarchie spatiale, le *pagus* étant la circonscription englobant un *episcopatus* ou un *comitatus*, mais le plus souvent il s'agit d'un chevauchement des ressorts[484]. Les scribes sont normalement conscients

CBMA 5241 : « ... In Episcopatu Pictaviensi Ecclesiam Dei dilectricis sanctæ Mariæ Magdalenæ juxta Mirebellum castrum... » ; CBMA 3037 : «... in episcopatu Valentinensi : ecclesia de Monte Ison, ecclesia de Aleso... »

479. Par exemple : CBMA 5060 : «... monasterium Sancte Marie infra urbem Bisunticam quod vocatur Jusanum » ; CBMA 4840 : « Ex quibus in suburbio Nevernis urbis unum extitit in honore beate Mariæ semper virginis»

480. CBMA 6218 : « ... In civitate Tolosana monasterium Sancte Marie Deaurate, prioratum Sancti Petri de Quoquinis, hospitale quod dicitur Bernardi Mainaderii ; in diocesi Tolosanensi, abbatiam Sancti Petri Lesatensis, monasterium et villam de Conquitis... » ; CBMA 5364 : «In Episcopatu Augustodunensi, Abbatiam Vizeliacensem. In civitate Antissiodorensi, Abbatiam sancti Germani... » ; CBMA 5008 : « Actum civitate Mediolanensium feliciter. »

481. CBMA 5902 : « ... tradidi Domino Deo et beate Marie .... quamdam Sanctae Columbae aecclesiam in Tolosanensi pago sitam, in territorio Chercorbensi, juxta fluvium Erz, cum omnibus ecclesiis sibi pertinentibus... » ; CBMA 1540 : « dono Deo et sanctis ejus apostolis Petro et Paulo aliquid de mea hereditate, que sita est in pago Augustodunensi, in villa Monte : hoc est unum mansum cum omnibus appenditiis suis... » ; CBMA 2943 : « ego, in Dei nomine, Raimfredus et uxor mea Aldegaldis, venditores, vendidimus tibi curtilum cum uinea....quæ est sita in pago Cabilonense, in fine Rofiacense, in Vetus Molinum vo[cant] ; et accepimus a te precium de casa beati Petri Cluniensis cenobii denariorum solidos VIIIº »

482. CBMA 4904 : « Ego Rogerius comes et uxor mea Sicardis...ipsas res quas offerimus ipsi...in comitatus Tolosano, in valle Savartensi, quas etiam nominatim sicut sunt per loca singula necessario in hac donationis nostræ carta exprimendas censuimus. » ; CBMA 4141 : « Ego Ramnulfus...dono eidem sanctisque apostolis ejus Petro et Paulo...sunt autem ipse res site in commitatu Cabilonensi, in villa Noglas : hoc est campum cum vinea sibi conjuncta» ; CBMA 2359 : « Notitia vuerpitionis quæ facta fuit apud Cabilonensem civitatem, ante presentiam incliti comitis Hugonis ejusque genetricis Adeleidis, de quibusdam terris quæ conjacent in comitatu Cabilonensi seu Matisconensi, hoc est de æcclesia Sancti Jangulfi atque Masnile vocabulo... ».

483. CBMA 4340 : « Ego Ademarus comes et uxor mea, nomine Roteldis, una cum filiis nostris, Pontio videlicet episcopo, Ugone, Lamberto, Gontardo, Geraldo, quendam locum Salciacum nomine, situm in episcopatu et comitatu Valentinæ civitatis, consecratum in honore sancti Marcelli martiris »

484. Par ex. CBMA 2285 : « Sunt autem ipsæ res sitæ in pago Arvernensi, in comitatu Brivatensi,

de cette réalité et l'expriment ainsi dans les actes [485].

Par ailleurs, si on observe des cas où un terme d'unité intermédiaire accompagne l'une des unités supérieures il s'agit, dans tous les cas, de toponymes homonymes, par exemple la *villa Meldensi* [486] dans le comté *Arvernensi* et *ager Valentiaco* dans le *pagus* d'Auvergne [487].

6.6.2 Unités intermédiaires à l'intérieur du *pagus* de Mâcon.

2. *Medionalense* (Meulin-le-Bourg), *Donziacense* (Donzy-le-National), *Rufiacense* (Ruffey), *Bufferiacense* (Buffières), *Burgundia* (com. de Saint-Point), *Ibgiaco* (Igé), *Prisciacens* (Prissé), *Galoniaco* (Jalogny), *Salorniaco* (Salornay), *Verziaco* (Verzé), etc.

Par sa nature le tableau nous laisse bien voir la plupart des chefs-lieux opérant à l'intérieur du *pagus Matisconense* . La relation *villa-ager* se montre alors très étroite car souvent les scribes vont signaler les chefs-lieux comme étant des *villae*. Lorsque le scribe doit rédiger une donation qui se trouve dans le chef-lieu, centre de l'*ager*, qui est également une *villa*, nous avons observé différentes stratégies par ordre de fréquence :

— Mobiliser la formule tripartite répétant le toponyme à la place de l'*ager* et de la *villa* [488] ;
— Faire omission de l'*ager* [489] (voici un des causes de la sous-utilisation de l'*ager*) ou plus rarement de la *villa* [490] ;
— Se rapporter à un *ager* limitrophe ;
— Ou créer un nouvel *ager*.

in vicaria Sancti Germani, in villa quæ vocatur Rogiacus... » ; CBMA 3691 : « aliquid ex rebus meis que sunt site in pago Lugdunensi, atque in comitatu Forensi posite : hoc est curtem meam que vocatur Poliacus... »

485. CBMA 3757 : « aliquit ex rebus nostris, quæ sunt site in pago atque in comitatu Lugdunensi : hoc est æcclesiam in honore sancti Martini in villa quam Oratorias vocant » ; CBMA 4339 : « aliquid de rebus nostris que sunt site in pago Dyensi seu in episcopatu» ; CBMA 4340 : « quendam locum Salciacum nomine, situm in episcopatu et comitatu Valentinæ civitatis, consecratum in honore sancti Marcelli martiris. »

486. CBMA 3400 : « Quapropter, in Dei nomine, ego Dacbertus, presbiter, dono aliquit ex rebus meis...in comitatu Arvernensi, in villa Pedronensi Montis...dono etiam in Meldensi villa mansos duos»

487. CBMA 3251 : « ego Leodegarius presbiter dono... rebus meis que sunt site in pago Arvernico, in agro Valentiaco, in villa Saligniaco, in loco qui dicitur ad Novem Fontibus, scilicet appenderiam unam cum vinea et prato. »

488. CBMA 2160 : "...Dono quoque in ipso pago [Matisconense], in agro Donziaco, in ipsa villa Donziaco, curtilium ubi Martinus mansit, cum omnibus apendiciis et pertinentiis suis... » ; CBMA 1672 : « ... In primis donat atque commutat Fredoenus et Rotardus partibus Sancti Petri aliquid ex rebus eorum que sunt site in pago Matiscense, in agro Rofiacense, in villa Rofiaco... » ; CBMA 2128 : « Domno fratribus Ainart sacerdote, ego Gertrudis femina venditor, ego tibi res meas in pago Matisconense, in agro Salorniacense, in villa Salorniago, curtilo cum superposito..."

489. CBMA 1708 : "Sunt vero ipse res site in pago Matiscense, in villa Prisciaco. Totum et ad integrum cum omnibus suprapositis a die presenti dono... » ; CBMA 3676 : "Igitur ego Randefredus, hæc cogitans, dono ... aliquid ex rebus meis quæ sunt sitæ in pago Matisconensi, in villa Galoniacensi, quicquid in ipsa villa visus sum habere vel possidere, excepto unum campum quem vocant ad Granda."

490. CBMA 2190 : "ego Rotfredus, uxor sua Erelt et Arfredus, vendimus vobis aliquid de res nostras, qui sunt sitas in pago Matisconense, in agro Prisciaco, in ipsa villa : oc est curtilus cum vinea et manso..."

Le tableau nous montre bien l'état des *agri* du quadrant nord du *pagus* de Mâcon, au sud de l'abbaye et entre la Grosne et Mâcon, puisque le réseau des sièges vicariaux coïncide en grande partie avec le réseau des sièges agraires de cette partie du *pagus*. L'abondance des actes correspondant à quelques-uns de ces *agri* comme l'*Ibgiacense, Salornicense, Prisciacense* et *Verziacense* a permi à F. Bange pour proposer l'existence d'un processus de dislocation des grands *agri* dans la deuxième moitié du Xe siècle [491]. Effectivement, comme il le signale, l'*ager Ibgiacense* se fragmente vers 980 et son territoire est réparti entre le territoire de deux de ses *villae* Salorniaco [492] et Verziaco [493] qui dans les actes après l'an 1000 sont présentées comme étant des *agri*. Ces nouveaux *agri*, formés sur des *villae* relativement importantes il les appelle « *agri* fictifs » car leur existence n'est plus soutenue par la réalité originelle de l'*ager*. Mais l'exemple de l'*ager Ibgiacense* est presque le seul dans tout le corpus - avec l'*ager Ruffiacense* - dont on peut tirer avec certitude et en détail une telle observation. La composition du reste des grands *agri* comme le *Galoniacense, Fusciacense, Marciacense, Donziacense*, etc. est très cohérent jusqu'à l'extinction totale des unités intermédiaires vers la décennie 1030. On peut estimer que plusieurs chefs-lieux perdent la prééminence ou sont abandonnés et d'autres *villae* prennent place en tant que chef-lieu de nouveaux *agri*, transitoires ou fictifs, qui participent à la genèse des paroisses. Mais il s'agit d'un phénomène qui a dû avoir lieu précisément dans les décennies de baisse de production entre 990 et 1040, et la faible quantité d'actes nous empêche de bien le cerner.

La question de la dislocation des *agri*, bien que compliquée à observer, car les actes des petits *agri* ne sont pas nombreux, est cependant très pertinente et nous pouvons la représenter statistiquement en interrogant toutes les mentions aux *agri* du *pagus* de Mâcon, ce qui offre une vision globale qui, dans une certaine mesure, semble étayer le raisonnement de F. Bange qui l'avait émis à partir d'observation de l'*Ibgiacense* et du *Galoniacense*.

Dans le figure 6.11 se trouvent ainsi représentées les lignes de vie de chaque *ager*, établies entre deux points [494] : les dates de leur première et de leur dernière mention. Nous avons établi trois groupes selon la durée chronologique : une vingtaine d'*agri* (en vert) jouent le rôle de concentrateurs et ont une longue vie dans notre corpus (de 60 à 150 ans) ; parmi eux on peut trouver les *agri* classiques centrés sur les *villae* de *Galoniaco, Rufiaco, Dunziaco, Maziaco, Verziaco, Tissiaco, Fusciaco, Ibgiaco*, etc. sur le sol desquels la plupart des transferts fonciers ont lieu. Une deuxième vingtaine (en rouge) ont une vie documentaire bien plus courte (moins de 60 ans), là on peut trouver des *agri* plus excentrés par rapport à l'abbaye mais dont les chefs-lieux sont relativement bien attestés comme ceux de *Bufferias, Chessiaco, Miseriaco, Matorn,*

491. Bange, "L'ager et la villa : structures du paysage et du peuplement dans la région mâconnaise à la fin du Haut Moyen Age (IX e-XI e siècles)", p. 551-559

492. CBMA 2545 : "Ego Aalber vendo vobis prato in pago Matisconense, in agro Salorniacense, in villa Ibiaco, Alabadesi vocat. . . ."

493. CBMA 2993 : "Constantinus et usore sua donet Sancti Petri uno curtilo que est in Matisconense, in fine Verziacense, in villa Ipgiaco sedi. . . "

494. Le nombre d'agri que nous avons repéré dans le pagus de Mâcon s'élève à entre 65 et 70. On ne peut pas être sur du nombre, notamment dans les cas d'agri attestés qu'une seule fois, à cause de l'homonymie et proximité phonétique entre quelques-uns d'entre eux. (par ex. Maceacense/Mazeriaco ; Laxiaco / Laisiaco, etc.)

Auriaco, etc. Finalement une autre vingtaine d'*agri* attestés dans un seul acte (points en bleu), la plupart prenant comme centre des *villae* relativement connues appartenant traditionnellement a d'autres *agri*, par exemple : *Cluniacense* (Cluny), *Flagiacense* (Flagiaco), *Veschennias* (Vetis Canivas), *Miliacense* (Miliaco), *Saliacense* (Saliaco), etc.

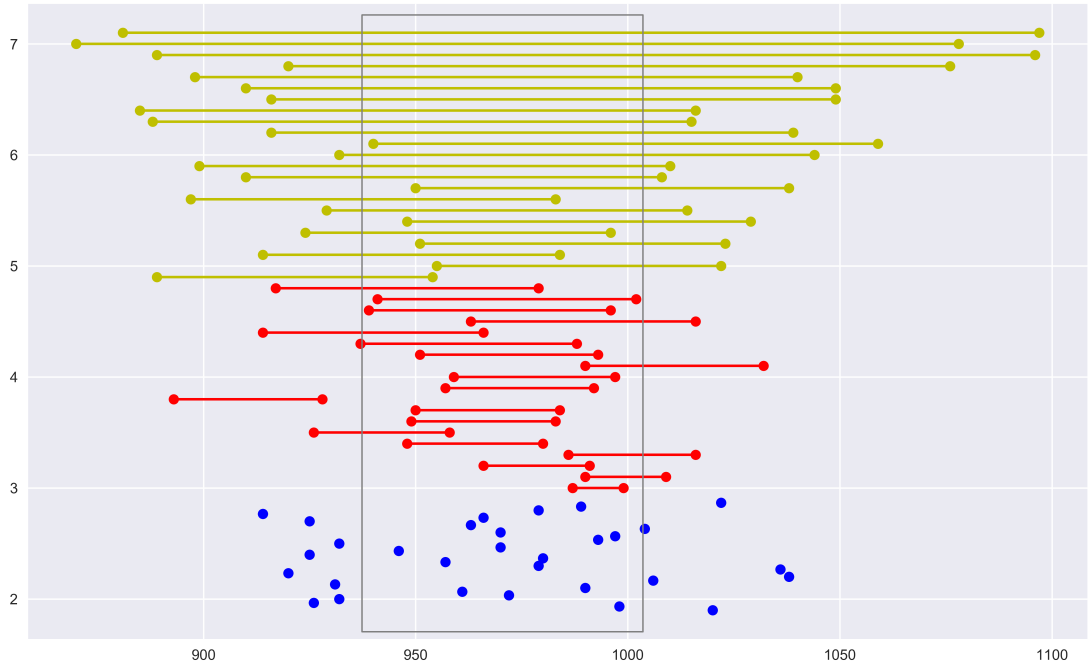

FIGURE 6.11 – Chronologie des *agri* dans le *pagus* de Mâcon. Les lignes en vert représentent les *agri* observés dans une période de plus de 60 ans ; en rouge les *agri* mentionnés dans une période de moins de 60 ans ; en bleu les *agri* attestés dans un seul acte.

Ce que nous indique le graphique est l'existence d'une période de grand bouleversement (cadre gris) pour l'*ager* entre les décennies de 940 et 990, ce qui coïncide avec le point culminant de la production d'actes pendant l'abbatiat de Maïeul (954-994). Dans cette période est attestée l'apparition de la plupart des *agri* "de courte durée", beaucoup d'entre eux correspondent assez bien à la définition de l'*ager* transitoire ou fictif suggérée par F. Bange pour désigner les nouveaux *agri* fondés à partir du morcèlement des grands *agri*. Ils sont fondés sur des *villae* appartenant traditionnellement à d'autres *agri* de plus longue tradition (voir la liste ci-dessous).

En ce qui concerne les *agri* de plus longue durée : la plupart entre eux sont mentionnés depuis la fin du IXe siècle et le début du Xe siècle mais on en voit quelques-uns commencer leur vie documentaire précisément aux alentours de 950 et durer jusqu'à la décennie de 1030 lorsque l'*ager* s'éteint. Cela peut correspondre à de nouveaux *agri* qui ont une vie documentaire plus longue ou à des *agri* déjà existants mais dont on ne voit pas de mentions jusqu'à la décennie de 940 car ils sont plus excentrés par rapport à l'abbaye. À l'époque de Maïeul, sont aussi attestés la plupart des *agri* qui ne sont mentionnés que dans un seul acte, dont les observations se compliquent car ils n'ont été pas réutilisés comme unités de rapport spatial : ces

derniers *agri* peuvent correspondre effectivement à quelques cas d'*agri* fictifs fondés sur des *villae* prééminentes qui ont grandi au détriment des celles de leur entourage, mais dans d'autres cas la mention semble correspondre à d'autres causes : une manque d'information, une actualisation postérieure du cartulariste, et enfin, à une confusion de la part de l'auteur de l'acte [495].

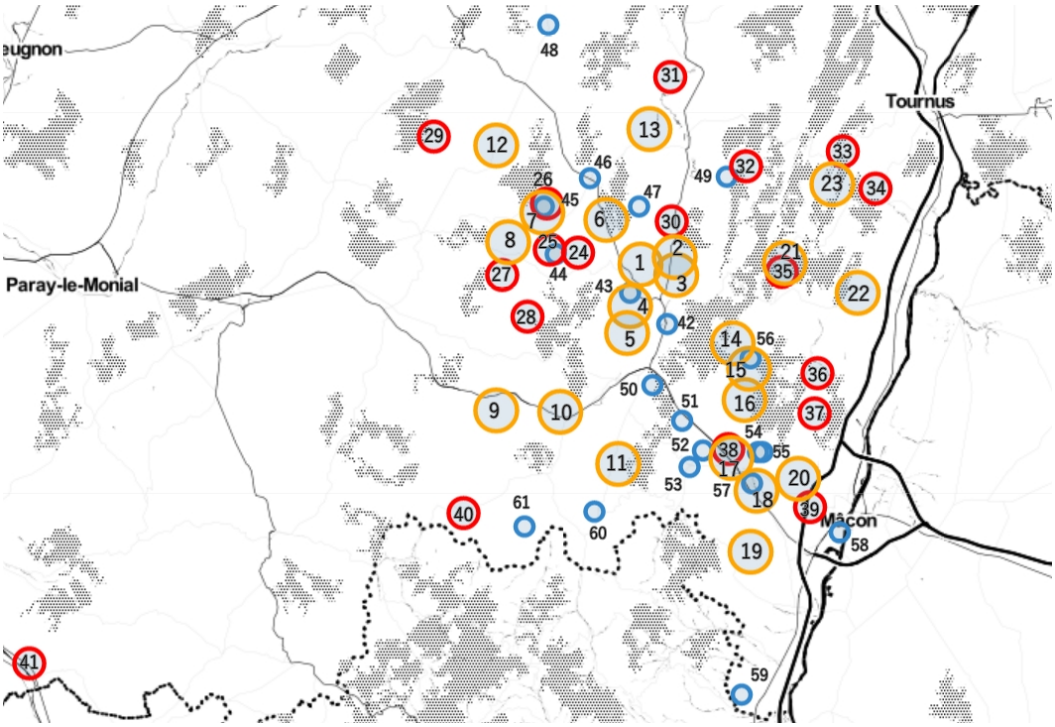

FIGURE 6.12 – Localisation des chefs-lieux des *agri* attestés dans le *pagus* de Mâcon d'après les coordonnées fournies par le dictionnaire de Saône-et-Loire. Les *agri* avec plus de 60 ans de vie sont représentés en jaune, ceux de moins de 60 ans en rouge, et ceux attestés dans un seul acte en bleu.

1. Mattiaco, 2. Marciacense, 3. Dariacense, 4. Rufiacense, 5. Galoniacense, 6. Maciacense, 7. Aigiaco, 8. Donziacense, 9. Mediolanense, 10. Briendonense, 11. Tassiaco, 12. Cavaniensis, 13. Ainacensis, 14. Euriacense, 15. Ibgiacense, 16. Verziacense, 17. Potziacense, 18. Pressiacense, 19. Fusciacense, 20. Salorniacense, 21. Circiaco, 22. Voriacense, 23. Grivilliacense, 24. Fenestiacense, 25. Arniacense, 26. Cadiacense, 27. Fabriacense, 28. Buferiacense, 29. Bociaco, 30. Masciliaco, 31. Saviniacense, 32. Chessiaco, 33. Miseriaco, 34. Cardiniacensi, 35. Auricense, 36. Laxiaco, 37. Arpagiaco, 38. Gigniacensi, 39. Meleniacense, 40. Matornensi, 41. Evuirando, 42. Cluniaco, 43. Veschanivas, 44. Cortboso, 45. Craiacense, 46. Athanacense, 47. Flagiacense, 48. Jovenciaco, 49. Disiacense, 50. Miliacense, 51. Coliniacense, 52. Maloniacense, 53. Cluvacense, 54. Arelariacense, 55. Saliacense, 56. Nuiriaco, 57. Noviliacense, 58. Masconense, 59. Torrense, 60. Tremaias, 61. Cuvleciacens. ***agri* non trouvés** : 62. Vallis Caponiensis, 63. Nirneio, 64. Cocenacens. 65. Morgonico. ***agri* erronés :** 66. Valiciacense (Galoniacense ?), 67. Martiacensium (Matisconense), 68. Vueriæ (Evoriacense ?), 69. Jaciacense (Cadiacense ?)

495. Il y a clairement 3 agri en cette situation : *Valiciacense* (CBMA 1595) où le scribe range 3 *villae* du *Galoniacense*, donc, probable mauvaise écriture ; *Martiacensium* (CBMA 1666) confusion avec *Matisconense*, car le scribe range dans cet *ager* l'abbaye de Cluny et *Jaciacense* (CBMA 1846), confusion évidente avec un *ager* de nouvelle création, le *Cadiacense*.

Ainsi, le dénommé *ager* fictif répond effectivement à plusieurs situations où des agglomérations importantes à l'intérieur d'un ancien *ager* se sont concentrée ou ont grandi suffisamment pour servir de référence spatiale aux scribes, pour mieux préciser la localisation d'un bien en se rapportant à son ressort géographique, c'est par exemple le cas des *agri Verziacense, Potziacense, Pressiacense, Salorniacense, Cadiacense*. On ne le sait pas si ce dénombrement du territoire de l'*ager* a eu des conséquences à un niveau juridictionnel ou administratif. Comme on peut le voir dans la carte ci-dessous (figure 6.12) le phénomène de l'*ager* fictif et des petits *agri* se concentrent dans deux zones : à l'est du site de Cluny qui touche les actuels communes de Donzy-le-National, Saint-André-le-Désert et Pressy-sous-Dondin et surtout la zone entre la Petite Grosne et Mâcon, de plus grande hétérogénéité géographique (communes actuelles de Verzé, Prissé et Hurigny.)

Dans d'autres cas, les nouveaux *agri* ne semblent être que une réalité rédactionnelle. C'est par exemple le cas des *agri Masconense*, formé sur Mâcon, *Cluniense* formé sur Cluny, *Veschanivas*, formé sur Vetus Canivas ou *Athanacense* sans existence véritable et dont leur apparition est due à l'intervention de l'auteur ou du copiste probablement afin de revendiquer le prestige d'une *villa* qui avait pris de l'importance au fil du temps. Et enfin, dans d'autres cas, les nouveaux *agri* répondent tout simplement à une mauvaise lecture ou écriture d'autres *agri*, c'est le cas par exemple des *agri Valiciacense* (*Galoniacense*), *Jaciacense* (*Cadiacense*) et *Martiacense* (*Matisconense*), car lorsque ils sont mentionnés le scribe range de *villae* appartenant à ces autres *agri*.

Pour finir avec les unités intermédiaires, les actes ne nous apportent pas de solides éléments de distinction entre *finis* et *vicaria* par rapport à l'*ager*. La *vicaria*, comme on l'a dit, apparaît dans nos sources en même temps que l'*ager*, et est à l'origine une sous-division du comté, c'est-à-dire une unité théoriquement supérieure à l'*ager*. Mais dans très peu d'actes, on observe une *vicaria* contenant un *ager*[496]. Si les chefs-lieux qui cumulent une dénomination d'*ager* et de *vicaria* accueillent quelques compétences judiciaires, comme on peut l'observer dans les vigueries un siècle plus tard, nous ne pouvons que le spéculer parce que, dans les actes des Xe et XIe siècles, elles ont un sens exclusivement territorial.

Pour la *finis* la différence est vraiment imperceptible. Les scribes mobilisent *finis* d'une manière un peu erratique sans que son usage n'apporte de précision territoriale. Hormis dans le cas de l'*ager Ruffiacense* (analysé ensuite) qui cumule plusieurs actes l'appelant *finis* (et de l'*ager Fabriacense* dans le *pagus* de Chalon), le réseau des chefs-lieux du finage coïncide pleinement avec celui de l'*ager* affiché dans le tableau. Éventuellement le seul détail qu'on peut apporter est la constatation que, lorsque la *finis* est utilisée, l'acte est souvent rédigé *in loco Cluniaco* ou *in villa puplica*[497]. C'est un argument qui reste à explorer mais qui pour l'instant est faible car on peut trouver plusieurs exemples d'autres actes qui y sont rédigés mais emploient l'*ager* sur le même toponyme.

496. 6 actes (CBMA 1904, 3060, 3563, 3600, 3641, 3935) dont 4 concernent l'ager Thyernensis et la vicaria Dorotensis.

497. Voir par exemple : CBMA 2997, 2279, 2254, 1533, 2213, 4075, 4668, 2136, etc.

6.6.3 L'*ager* Rufiacense et la *villa* Rufiaco (Rufey).

D'après le tableau, l'*ager Rufiacense*, fondé sur la *villa Rufiaco*, nous présente la relation *ager-villa* la plus densement représentée dans notre corpus et mérite quelques commentaires car de surcroît le site de Cluny est affecté à cet *ager*. Tout d'abord, il faut dire qu'il existe deux chefs-lieux appelés Rufiaco (ou Ruggiaco), le plus connu est celui qui abrite l'abbaye, et un autre dans l'actuelle commune de Sennecey-le-Grand dans l'ancien *pagus* de Chalon sur le site d'un *castrum* gallo-romain, l'actuel Château de Ruffey. Dans nos actes il est presque exclusivement reconnu comme étant une *finis* et une série d'actes s'y référant nous est parvenu car l'abbaye avait acquis l'église de Sancta Maria Belmontense avec toutes ses propriétés dans une *villa* appelée *Vetus-Molinum*[498].

Dans l'*ager* de Ruffey de l'actuelle commune de Cluny, ce que les nombreux actes nous laissent observer est un bouleversement dans son arrangement territorial. Dans cet *ager* nos actes localisent au moins 12 *villae* (Cluniaco[499], Vetis Canivas[500], Ruffiaco[501], Lordonum[502], Bainas[503], Colonias[504] , Maliaco[505], Lornant[506], Castello[507], Cariniaco[508], Belusia[509] (?), Turro[510]) dont on sait que sept n'appartiennent pas à cet *ager* mais à l'un des autres *agri* limitrophes, le *Galoniacense* (Castello, Colonias, Belusia) et le *Maciacense* (Bainas, Maliaco, Cariniaco, Lornant). L'abbaye de Cluny était à l'origine fondée dans la *villa* de Cluny dans l'*ager Rufiacense*[511] comme cela est indiqué dans une adresse que les scribes mobilisent dans quelques actes du début du Xe siècle. Cette adresse disparaît rapidement ainsi que les mentions de transferts à la *villa* de Cluny attestées dans seulement cinq actes[512]. Les actes du transfert foncier se concentrent principalement sur les trois *villae* de son entourage immédiat : *Vetus Canivas*, *Ruffiaco* et *Lornant* que dans le parcellaire actuel on peut identifier avec Montaudon, Ruffey et le château de Lourdon, respectivement. Bientôt Cluny reçoit et promeut la donation et l'achat des terres dans ces *villae* et au fur et au mesure que ses possessions avancent, notamment à *Vetus Canivas*, les actes reflètent un changement de statut de ces trois endroits. Dans les actes du début du Xe siècle *Vetus Canivas* est progressivement considérée non plus comme un

498. Voir CBMA 1533, 2943, 3264, 3169, 3173, 3195, 3215, 3216, 4147, 4196
499. CBMA 2213, 1799
500. CBMA 2552, 3372, 2864, 3881, 2120, 2135, 3381, etc.
501. CBMA 1663, 1672, 2080, 2895, 3052, 4061, 4595, etc.
502. CBMA 3687
503. CBMA 1779, 1851, 2228, 3481,
504. CBMA 1500
505. CBMA 3271
506. CBMA 3102, 3844, 1890
507. CBMA 2867
508. CBMA 1700
509. CBMA 1700
510. CBMA 3770
511. CBMA 1772 : "Domino sacrosancte ecclesie Sancti Petri Cluniensi, qui est fundata in pago Matisconense, in agro Rofiacense, ubi domnus ac venerabilis Otdo abba ad regimen tenet."
512. CBMA 2084, 2516, 2186, 4829, 4973, 5236. Après « Actum villa Cluniaco » apparaît régulièrement comme date topique, par ex. CBMA 1926, 1969, 2019, 2034, etc.

simple *locus* associé à *Rufiaco*[513] mais comme une *villa* et plus tard même comme un chef-lieu d'*ager*[514]. La *villa* Lornant, dont l'abbaye de Cluny avait reçu le château comme donation à sa fondation est rapidement acquise par l'abbaye qui ici même vient de recourir à un évènement rare : la location de terrains. Pour sa part la *villa* Rufiaco semble perdre son entité en tant que chef-lieu d'*ager*. Avant la décennie 950 les localisations dans les actes se rapportent systématiquement à l'*ager* Rufiacense, alors qu'ensuite il devient une mention optionnelle, privilégiant plutôt le tandem *pagus-villa*. Mais on ne peut pas en tirer une observation complète.

Ainsi de ces 12 *villae* initialement affectées à cet *ager*, 7 sont empruntés aux autres *agri*, 3 ne sont mentionnées que dans un seul acte chacune, concentrant les transferts fonciers sur les trois restantes. Le parcellaire théorique de l'*ager* Rufiacense nous suggère l'existence d'un nombre plus élevé de *villae* que les trois qui reviennent constamment, mais les actes ne nous les transmettent probablement par pour deux raisons : soit parce que s'agissant de son territoire immédiat (ces actes sont le plus souvent datés comme *actum Matiscono publice* ou *actum Rufiaco publice*) les scribes sous-entendent le nom de la *villa* en évoquant seulement le nom d'un *locus*[515] ; soit parce que les autres *villae* mises à part Ruffiaco, Lornant et Vetus Canivas, sont en passe de s'estomper ayant été absorbées rapidement dans le ban de l'abbaye de Cluny[516].

D'après les actes, il demeure évident que le site de Cluny et les villes de son entourage forment rapidement un territoire plus unifié que celui trouvé ailleurs car l'abbaye devienne rapidement dans la seule propriétaire. Mais cela n'implique pas que Cluny se transforme formellement, même s'il est de facto, en un chef-lieu. En tout cas, la croissance écrasante de l'abbaye a dû avoir des fortes conséquences sur son modeste chef-lieu et sur les chefs-lieux de son entourage et la croissance de son pouvoir et de son prestige[517] peut être l'explication de la localisation d'autres *villae* appartenant aux *agri* limitrophes dans l'*ager* Rufiacense, voire, dans le voisinage du *locus* Cluniaco ; une imprécision dont la faute n'est pas toujours par négligence car plusieurs de ces actes sont rédigés dans la *villa* concernée.

Un point intéressant concernant cela est apporté par un acte du cartulaire A de l'an 926[518] qui mentionne dans le *comitatus Matisconense* (un comte est témoin dans l'acte) un *ager Cluniense* dans lequel il range la *villa* Rusciaco (il faut lire Rufiaco) et Bieria et Castello, qui comme on l'a dit appartiennent aux *agri* limitrophes. Le style de cet semble avoir été corrigé par le cartulariste , comme le pense M. Chaume, qui a changé la référence à l'*ager* Rufiacense, à l'époque déjà disparu, par une nouveauté : un *ager* Cluniense, mais plus probablement il s'agit d'une imprécision du scribe. En

513. CBMA 1607
514. CBMA 2365
515. Par exemple : CBMA 1700 et 3009 (locus Cardonari / Cardonarium), CBMA 1673 (locus In illo Monte), CBMA 4025 (locus Lonbesco), CBMA 3101 (locus Vaurelia)
516. L'évolution du ban clunisien a été bien montrée dans la thèse de Didier Mehu : Méhu, "Paix et communautés autour de l'abbaye de Cluny (Xe-XVe siècle)"
517. Voir Rosenwein, *To be the neighbor of Saint Peter : the social meaning of Cluny's property, 909-1049*, p. 145-155
518. CBMA 1688

tout cas l'*ager Cluniense* n'avait jamais eu, selon les actes, une existence véritable [519].

Donc, pour résumer, le tableau nous suggère que la réalité autour des unités concurrentes correspond à un chevauchement et non à un emboitement. Les *vicariae* et *fines* coïncident dans les mêmes chefs-lieux que l'*ager* de même que *episcopatus* et *comitatus* sont centrés par les mêmes sièges épiscopaux, et les *civitae* que le *pagus*. Les matrices chronologiques nous ont montré de plus que, tant les principaux que les concurrents, tous les termes suivent un même rythme d'évolution chronologique ; nous nous trouvons alors devant d'une substitution motivée par des changements sociales (notamment l'essor des cadres ecclésiaux) à laquelle les scribes se rapportent aussi par souci de précision faisant référence à ces réseaux spatiaux plus territorialisés. Naturellement, les unités supérieures englobent les inférieures, mais ne se mêlent pas entre elles. Ou pour parler plus précisément, un chef-lieu d'*ager* n'est quasiment jamais pris comme le centre d'un *pagus* ni à l'inverse. Á l'intérieur de chaque unité quelques emboîtements peuvent avoir lieu entre termes concurrents, mais ce n'est pas le plus courant. Ce qui arrive le plus souvent sont les doublons et les imprécisions au moment de rapporter une *villa* à un certaine unité intermédiaire car ces unités souffrent différents phénomènes : dénombrement territorial, indéfinition dans leurs frontières, croissance d'un chef-lieu au détriment d'un autre, etc.

6.7 Conclusion

Les actes étudiés permettent d'observer que la mutation principale de l'écriture de l'espace concerne les *villae* et les unités que les abritent, les *agri*, unités communautaires qui sont par ailleurs affectées par une forte simplification dans leurs réseaux. La *villa*, unité fondamentale de ce topo-système, demeure valide jusqu'à la disparition de l'acte de donation à la fin du XIe siècle, mais le réseau dispersé, formé par environ 600 *villae* observées dans notre corpus se rétracte fortement à partir de la deuxième moitié du Xe siècle. Certaines *villae* définies en tant que chefs-lieux qui voient leur puissance s'accroître se distinguent des autres qui, elles, sont abandonnées ou réorganisées. Il s'agit là d'un phénomène qui a un impact profond bien qu'inégal sur les unités intermédiaires, car elles sont définies comme le domaine de ces chefs-lieux. Cette concentration des centres de référence, et probablement du peuplement, provoque le morcèlement des grands *agri* dans la région, suivi par la création de nouveaux *agri* éphémères basés sur les nouveaux centres. Selon plusieurs chercheurs, c'est sur ce nouveau réseau que s'établissent les villages et paroisses, unités dominantes de structuration de l'espace à partir du XIIe siècle.

Avec la disparition de l'*ager*, au cours de la décennie de 1030, le système semble plus lâche, mais polarisé avec des nœuds plus densement peuplés. Cette mutation a beaucoup à voir avec la multiplication dans les actes des références de localisation fondées sur les unités de base, d'abord le *locus* et ensuite l'*ecclesia*, précisément après 1030. Si la localisation des lieux de culte transformés rapidement en pôles sacrés se

519. Chaume, *Les Origines du duché de Bourgogne : 2ème partie, Géographie historique*, p. 1054. F. Bange discute l'origine de quelques agri « fictifs » dont celui de Cluny en : Bange, "L'ager et la villa : structures du paysage et du peuplement dans la région mâconnaise à la fin du Haut Moyen Age (IX e-XI e siècles)", p. 546-551

trouve à l'origine du remplacement de certaines *villae* par autres, ce phénomène reste à examiner avec bien plus de détail. Par contre, il demeure assez clair que les *villae* ou chefs-lieux organisés dans le nouveau réseau formé dans la deuxième moitié du XIe siècle abritent un nombre sensiblement plus élevé de lieux de culte et que leur domaine recouvre un grand nombre de petits espaces d'habitat et de production définies encore sous la dénomination de *loci*. Dans les dernières séries de donations après la décennie 1050, on peut voir la manière dont ces lieux commencent se transformer en points de repère spatiaux car, dans la localisation des biens-fonds, les scribes abandonnent les références aux cadres territoriaux. et commencent à se rapporter directement au réseau de lieux, depuis le maillage extérieur, sans passer obligatoirement par l'*ager* ou le *comitatus* : autrement dit, en allant du *pagus* au *locus* et de l'*episcopatus* à l'église. Ce renforcement des unités locales comme centres de localisation marque l'abandon formel du système de référencement spatial ancien et l'adoption d'un autre système plus direct, lié au processus de renforcement spatial de l'Église correspondant à la période de la Réforme grégorienne.

En parallèle, les actes nous montrent que les unités supérieures résistent mieux aux changements du XIe siècle, car à l'abandon et la désaffection documentaire des structures intermédiaires, floues et mal distribuées, les scribes répondent avec un renforcement du lien *villa-pagus* qui est concurrencée par celui de *villa-episcopatus* et remplacée par *ecclesia-episcopatus*. Le *pagus* ne souffre pas des mêmes vicissitudes que l'*ager*, mais il se voit transformé en une structure généraliste définie par la conscience spatiale des hommes et soutenue dans sa dernière étape par la définition plus claire de l'*episcopatus* avec lequel il entretient, en certaines occasions, un rapport mimétique.

La concurrence des termes de *vicaria, fines, episcopatus* et *comitatus* n'implique pas nécessairement l'emboîtement des territoires, mais leur simple superposition parce que la plupart de ces unités s'articulent autour des mêmes centres : les supérieures se rapportent principalement au réseau des *civitates*, sièges épiscopaux et grands centres peuplés de la région, et les intermédiaires à l'ample réseau de chefs-lieux, normalement des *villae* d'importance modeste. En tout cas, ces réseaux géographiques et ces espaces organisés par le pouvoir, dont l'existence répond aussi à une logique d'évolution sociale, n'arrivent pas à s'imposer sur le vocabulaire emprunté au monde ancien, ni sur les formes stéréotypées de localisation des biens proposées dans les formulaires haut-médiévaux. À l'exception de l'*episcopatus*, soutenu par la configuration spatiale de l'église, ces unités disparaissent toutes ensemble. Mais le corpus nous empêche de bien cerner avec précision la fin de ce processus ainsi comme l'établissement des nouvelles structures de remplacement car l'acte de donation, qui avait été notre fil conducteur, disparaît du corpus précisément à l'avènement du XIIe siècle.

Conclusion

Ce travail de thèse a permis d'illustrer les nombreux avantages apportés par le développement et l'application des outils et méthodes de traitement automatique pour l'étude de grands corpus médiévaux numérisés, sans masquer les défis à relever pour dépasser les difficultés rencontrées. Le lecteur a pu trouver dans le développement d'un modèle de reconnaissance d'entités nommées un exemple des multiples solutions possibles afin de produire un outil qui demeure efficient même s'il n'est pas appliqué à son corpus d'origine, ce qui représente une des premières difficultés.

Le modèle que nous proposons, en nous appuyant sur un corpus annoté à la main, corrigé et augmenté, a fait la preuve de sa robustesse sur des multiples sous-corpus différant par la taille, la typologie, la chronologie et l'origine. Ce modèle peut ainsi être maintenant mobilisé pour l'étude d'une grande variété de sources diplomatiques. On peut en attendre une performance élevée, autrement dit qu'il fournisse une information structurée d'un niveau aussi complexe que celui des entités nommées. Il en est de même pour le modèle développé pour les parties du discours diplomatique, sur les bases du modèle précédent, dont la performance, bien que moindre, reste encore élevée et tout à fait acceptable comme moyen d'indexation et d'automatisation de données. Nos deux modèles permettent ainsi d'économiser d'énormes efforts humains, autrefois nécessaires, dans le but de structurer les éléments exigés pour l'exploitation massive de documents.

La rencontre entre l'informatique et le document médiéval ne s'est pas montrée aussi brutale qu'on aurait pu le redouter. L'algorithmique peut s'adapter de manière idéale aux documents relevant d'une structure très formalisée comme les chartes qui procèdent d'une rédaction formulaire ou une documentation sérielle comme les censiers et terriers, les enquêtes et les recensements. L'appréhension par la machine, à travers une itération massive, surhumaine, des structures sous-tendant la composition de ces documents est un point d'ancrage solide pour libérer le texte de son carcan et disposer ses données au service nécessaire pour répondre à des multiples questions. Le chercheur en histoire peut ici trouver un outil d'une extraordinaire puissance qui s'adapte à ses besoins scientifiques spécifiques et peut mettre à sa disposition en un temps record une masse de données qui l'aidera à étayer ou invalider ses hypothèses.

Nous avons aussi cherché à préfigurer un modèle standard d'annotation et de modélisation pour les corpus historiques. La recherche actuelle visant à appliquer des outils informatiques aux études des grands corpus s'est beaucoup appuyée sur des outils qui viennent d'autres domaines, et qui ont été générés en utilisant d'autres

textes que les textes historiques, d'autres langues que les langues anciennes et d'autres domaines que celui de la recherche en humanités. Nous avons montré qu'ici se trouve le verrou le plus important qu'il faut lever avant d'entamer des exploitations massives de la matière textuelle. De ce fait, le processus de prétraitement des sources, avant de faire appel au traitement algorithmique, acquiert un statut fondamental et doit être bien défini. D'une part, les questions qui intéressent l'historien doivent être abstraites et transformées en tâches contrôlées ; de l'autre, les outils de traitement automatique de la langue doivent se confronter avec efficacité aux hauts niveaux de variabilité orthographique et discursive que l'on peut trouver dans les artefacts textuels anciens, qui sont fortement dépendants d'un contexte historique et social.

Par ailleurs, l'informatique n'est pas une panacée et force est de constater que ses limites sont rapidement attentes pour le chercheur. Il n'est pas envisageable que l'ordinateur produise de façon autonome des éditions documentaires ou de connaissance pertinente pour l'historien. Ce travail a bien confirmé que la production et l'application des outils d'automatisation sont des processus nécessitant une communication constante entre le regard critique, qui demeure l'apanage du chercheur, et les vastes capacités des outils pour récupérer et indexer l'information. Le sens des résultats offerts par l'algorithme doit être construit si l'on veut pouvoir les employer comme coadjuvants dans la recherche de la science historique.

Les études historiques de la deuxième partie de cette thèse essaient précisément de montrer trois cas où la structuration et l'indexation des noms de personnes, des noms de lieux et des parties du discours permettent de disposer d'une information extraite à partir de formats textuels assez arides. Dans le domaine de la datation nous avons montré que l'indexation des noms de personnes est un moyen très efficace pour resserrer la datation des documents qui en sont dépourvus. Par ailleurs, la récupération des noms de lieux a permis d'étudier les co-occurrences les accompagnant, ce qui équivaut à indexer chronologiquement le vocabulaire de l'espace de manière exhaustive. Nous avons montré qu'il est possible d'offrir à l'historien des indices sur le découpage de l'espace et les nouvelles formes d'organisation spatiale, en reprenant des propositions historiques que nous avons essayé de modéliser. Enfin, nous nous sommes aussi occupés de montrer que l'extraction de formules permet de structurer et de classer en peu de temps les documents à partir de leurs caractères internes. Ce travail a permis de les comparer massivement en se fondant sur deux unités lexicales fondamentales pour la recherche diplomatique : la formule et la clause. Ainsi nous avons enquêté sur différentes questions qui intéressent historiens et diplomatistes, touchant les domaines de la datation, de la localisation et de la formulation de manière à proposer un cadre de recherche renforcé par le pouvoir de l'ordinateur en ce qui concerne l'obtention de indices pour étayer ou écarter certaines hypothèses.

Finalement, si nos méthodes de travail peuvent être considérées comme novatrices, elles se trouvent néanmoins fortement attachées aux méthodes d'exploitation déjà utilisées par les historiens. Dans ce sens, l'enjeu est d'adapter ces méthodes à l'automatisation par l'informatique et de surmonter leurs limitations pour les rendre plus puissants et extensibles. En amont, on évite la répétition de la tâche qui limite les dépouillements systématiques dans les éditions de textes, ce dernier se trouvant transformés en matrices de données. En aval, on renforce le lien avec les

méthodes consacrées de l'érudition et la critique scientifique. Les cadres interprétatifs de nos résultats demeurent tributaires aux méthodes historiques traditionnels. L'interprétation de matrices, de lignes d'évolution, de chiffres et calculs contextuels reposent sur de nombreuses études consacrées à l'histoire sociale, aux formes de territorialisation, de polarisation et de sacralisation de l'espace qui transforment l'Europe à partir du Moyen Âge central. Du même, l'interprétation des données structurées reposent sur les apports d'études traditionnelles ou plus récentes dans le domaine de la diplomatique de l'acte privé et des pratiques de l'écrit dans le cas de l'étude des formulations.

De plus, une attention particulière doit être portée à l'introduction des outils de traitement automatique de la langue par les praticiens des "humanités numériques". Il s'agit d'un mouvement naturel compte tenu que les intérêts des humanités reposent aussi sur le texte. Or, si le nombre de corpus numérisés et d'outils produits pour les étudier est grandissant, ils sont rarement l'objet d'études véritables. Il est frappant de constater l'absence quasi totale d'une réflexion méthodologique portant sur la construction d'une connaissance à partir des outils développés, tout comme la rareté d'une réflexion tournée vers l'efficience heuristique des outils mises au point. Cette thèse essaie de constituer un premier pas dans cette voie, en fournissant des modèles et un cadre démonstratif pour le traitement massif de sources historiques médiévales.

Tria digita scribunt, totum corpus laborat.

Annexe A

Patterns pour les modèles

```
\# current token (non-case-sensitive)

U01:%x[0,0]

\# token features (fenêtre contextuelle +/-4)

U02:%x[0,1]
U03:%x[-1,0]
U04:%x[-2,0]
U05:%x[-3,0]
U06:%x[-4,0]
U07:%x[1,0]
U08:%x[2,0]
U09:%x[3,0]
U10:%x[4,0]
U11:%x[-1,1]
U12:%x[-2,1]
U13:%x[-3,1]
U14:%x[-4,1]
U15:%x[1,1]
U16:%x[2,1]
U17:%x[3,1]
U18:%x[4,1]

\# token features (fenêtre contextuelle B +/-4)

U19:%x[-4,0]/%x[-3,0]
U20:%x[-3,0]/%x[-2,0]
U21:%x[-2,0]/%x[-1,0]
U22:%x[-1,0]/%x[0,0]
U23:%x[0,0]/%x[1,0]
```

```
U24:%x[1,0]/%x[2,0]
U25:%x[2,0]/%x[3,0]
U26:%x[3,0]/%x[4,0]
U27:%x[-4,1]/%x[-3,1]
U28:%x[-3,1]/%x[-2,1]
U29:%x[-2,1]/%x[-1,1]
U30:%x[-1,1]/%x[0,1]
U31:%x[0,1]/%x[1,1]
U32:%x[1,1]/%x[2,1]
U33:%x[2,1]/%x[3,1]
U34:%x[3,1]/%x[4,1]

\#Bigrams

B
```

Dépôts Git des modèles

- Modèle pour la reconnaissance des entités nommées

    https://github.com/magistermilitum/latin_NER_model

- Modèle pour les parties du discours diplomatique

    https://github.com/magistermilitum/latin_diplo_model

Annexe B

Outils et bibliothèques logicielles

Outils :
— Wapiti (segmentation et étiquetage de séquences)
https ://wapiti.limsi.fr/
— TreeTagger (marquage morpho-syntaxique)
https ://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
— Omnia project (lemmatiseur du latin médiéval)
https ://glossaria.eu/outils/lemmatisation/
— BRAT (outil d'annotation)
https ://brat.nlplab.org/
— BratEval (évaluation des annotations)
https ://bitbucket.org/nicta_biomed/brateval/src
— Qgis (logiciel SIG)
https ://www.qgis.org/
— Gephi (analyse et visualisation de réseaux)
https ://gephi.org/

Python modules :
— nltk (outils de TAL) : https ://www.nltk.org/
— scikit-learn (machine learning) https ://pypi.org/project/scikit-learn/
— fuzzy wuzzy (Levenshein distance) : https ://pypi.org/project/fuzzywuzzy/
— Gensim (plongement de mots) https ://pypi.org/project/gensim/
— spaCy (répresentation de mots) : https ://pypi.org/project/spacy/
— folium (visualisation spatiale) : https ://pypi.org/project/folium/
— sparql-client (rêquetes avec sparql) : https ://pypi.org/project/sparql-client/

- R modules :
— ggplot2 (graphiques et tableaux) : https ://ggplot2.tidyverse.org/
— threeJS (3D graphes et réseaux) :
https ://cran.r-project.org/web/packages/threejs/index.html

Annexe C

Chronologie des actes du cartulaire de Paray-le-Monial

Le numéro CBMA correspond au numéro fourni dans la base de données des *Chartae Burgundiae Medii Aevi*. Entre parenthèses nous avons inclus la référence bibliographique de l'édition du cartulaire par Ulysse Chevalier. De cette édition vient aussi le contenu inclus dans le « Titre », ce qui inclut le plus souvent une description sommaire du texte et parfois des informations non présentes dans le texte mais que l'éditeur avait en main au moment d'éditer les actes.

Ensuite nous proposons une date selon les critères chronologiques adoptés et dont nous avons expliqué la méthode d'identification dans le chapitre 4 de cette thèse. Lorsqu'il s'agit d'un document qui notifie différentes actions juridiques séparées dans le temps, nous allons inclure une date pour chaque action. Quand il s'agit d'un acte qui notifie une action très décalée dans le temps nous avons inclus la date estimée de rédaction et entre parenthèses la date estimée dans laquelle a eu lieu cette action. Lorsque l'acte est daté nous l'indiquons comme « daté de ».

À continuation, dans la partie du commentaire nous avons signalé brièvement la typologie documentaire et les critères que nous avons utilisés pour proposer une date au document. Dans les documents problématiques ou qui relèvent de conflits chronologiques, le commentaire développe plus en détail l'idée.

Finalement nous avons inclus deux titres : personnages primaires et personnages secondaires. Comme personnages primaires nous avons inclus tout ceux personnages qui portent un titre, fonction, dignité, poste, bref tous ceux dénommés avec des co-occurrences ; les personnages secondaires correspondent aux personnages non connus dans les bases de données externes, notamment provenant des autres cartulaires et recueils bourguignons, mais qui apparaissent deux ou plus fois dans notre base de données. La référence alphanumérique que nous utilisons pour indiquer les personnages provient de l'*index personarum* que se trouve à la fin.

Numéro CBMA : 7017 (Cartulaire de Paray-le-Monial, p. 1-2, n° 1.)
Titre : Incipit præfatiuncula.
Dates proposées : 1080 (977)
Commentaire : Acte mémoriel / de fondation. Texte incomplet. Le compilateur expose les motifs de la récupération des anciennes chartes écrites « *in veteribus pittaciis ac membranis* » dans un seul volume « *in unius codecelli* » entre autres, pour éviter la cupidité et les querelles sur la terre : « *hominumque cupidorum sœpius emergentium* ».

Numéro CBMA : 7018 (Cartulaire de Paray-le-Monial, p. 2-3, n° 2.)
Titre : Incipit textus. — caput i'.
Dates proposées : 1080 (daté de 977)
Commentaire : Acte mémoriel / de fondation. Texte incomplet. Acte de consécration du coenobium de Paray-le-Monial.
Personnages principaux : le comte de Lambertus ; Ingeltrudis ; Maieul (abbé de Cluny).

Numéro CBMA : 7019 (Cartulaire de Paray-le-Monial, p. 4-5, n° 3.)
Titre : Caput ii. — Quæ et quanta contulit in sacratione hujus ecclesiæ.
Dates proposées : 1080 (daté de 977)
Commentaire : Acte mémoriel / de fondation. Texte incomplet. Acte de consécration du coenobium de Paray-le-Monial.
Personnage principal : Lambertus comte de Chalon

Numéro CBMA : 7020 (Cartulaire de Paray-le-Monial, p. 5, n° 4.)
Titre : Caput iii. — Quod longius a propriis obiit suumque corpus huc deferri jussit.
Dates proposées : 1080 (daté de 988)
Commentaire : Acte mémoriel / dogmatique. Cet acte est en connexion avec le précédent, qui décrit la concession de Lambertus, comte de Chalon, à Paray lors de la fondation du monastère. La référence « decessit e mundo isdem egregius comes » signale la mort de Lambertus en 988 et non en 979 comme l'affirment quelques chronologies. Cette confusion a nourri l'hypothèse selon laquelle après la mort de Lambertus, Hugues I étant encore enfant, Geoffroi Ier Grisegonelle, présumé deuxième mari de la comtesse Adelaïde, prend les fonctions de comte de Chalon (979-987).
Personnage principal : Lambertus comte de Chalon

Numéro CBMA : 7021 (Cartulaire de Paray-le-Monial, p. 6, n° 5.)
Titre : Caput iiii. — Quod post ejus finem in ejus loco surrexit filius ejus Hugo.
Dates proposées : 1080 (999 – 1002)
Commentaire : Acte mémoriel. Première période du comte-évêque Hugues. Texte incomplet
Personnages principaux : Hugues I,comte-évêque ; Adelaïde, comtesse ; Henri IV, duc de Bourgogne

Numéro CBMA : 7022 (Cartulaire de Paray-le-Monial, p. 6-8, n° 6.)
Titre : Caput v. — quam largus ..... in hunc locum....
Dates proposées : 988 - 998
Commentaire : Acte mémoriel. Acte de concession foncière du comte-évêque Hugues à Paray. Texte incomplet. Daté très probablement du début de la première période d'Hugues et en tout cas avant la donation de Paray à Cluny (999).
Personnage principal : Hugues I, comte-évêque

Numéro CBMA : 7023 (Cartulaire de Paray-le-Monial, p. 8-9, n° 7.)
Titre : Capitulum vi'. — quod post ejus decessum exsurrexit in loco ejus donnus Theobaldus, nepos ejus, comes cabilonensis.
Dates proposées : 1080 (1039 - 1065)
Commentaire : Acte mémoriel. Suite de la chronologie des comtes de Chalon. Thibaud de Chalon († 1065), comte de Chalon de 1039 à 1065, continue avec la tradition familière des concessions de propriétés à Paray. Texte incomplet.
Personnage principal : Thibaud, comte de Chalon

Numéro CBMA : 7024 (Cartulaire de Paray-le-Monial, p. 9, n° 8.)
Titre : Carta Rodberti vicecomitis.
Dates proposées : 990 - 999
Commentaire : Acte mémoriel. Robertus, vicomte de Chalon, frère de Lambertus de Chalon apparait dans deux autres chartes de donation du CBMA avec sa famille (vers 994/999). Étant donné le manque d'autres dates on peut supposer un arc chronologique identique pour cette charte.
Personnage principal : Rotbertus,vicomte de Chalon

Numéro CBMA : 7025 (Cartulaire de Paray-le-Monial, p. 9-10, n° 9.)
Titre : Cap. vii'. — De mala consuetudine in vineis de Rosers.
Dates proposées : 1066 - 1080
Commentaire : Première charte qui correspond à l'époque de confection du cartulaire. Elle se situe après l'époque du comte Thibaud (1039-1065). Le groupe de personnages qui y est trouvé est très connu pour ses donations avant l'époque du prieur Hugues et pendant la première période de son prieuré. Il est peu probable qu'ils soient vivants après 1100, mais cette charte est probablement antérieure à 1080, étant donné le ton de proximité avec lequel est évoquée l'affaire et sa place dans le cartulaire. Texte incomplet.
Personnage principal : Thibaud, comte de Chalon
Personnages secondaires : ['W4', 'G8']

Numéro CBMA : 7026 (Cartulaire de Paray-le-Monial, p. 10, n° 10.)
Titre : Cap. viii. — Quod Tolosæ obiit.
Dates proposées : 1080 -1085 (1065)
Commentaire : Acte mémoriel. Poursuite de la chronologie des comtes de Chalon. Instructions testamentaires du comte Thibaud, mourant à Tolosa, demandant d'être inhumé dans le panthéon familial à Paray. Il y mentionne son fils Hugues, encore enfant et dont on ne connait pas la date de naissance (vers 1055).
Personnages principaux : Thibaud, comte de Chalon ; Hugues II, comte de Chalon
Personnages secondaires : ['D4', G13', 'Y1']

Numéro CBMA : 7027 (Cartulaire de Paray-le-Monial, p. 10-11, n° 11.)
Titre : Cap. viiii. — Quod in ejus locum infans filius ejus Hugo successit.
Dates proposées : 1080 - 1085
Commentaire : Acte mémoriel. Suite du précédent. Des évènements postérieurs à la mort de Tetbaldus comte de Chalon (1065). La mort prématurée de son fils Hugues II lors de son pèlerinage à Saint-Jacques de Compostelle.
Personnages principaux : Thibaud, comte de Chalon ; Hugues II, comte de Chalon

Numéro CBMA : 7028 (Cartulaire de Paray-le-Monial, p. 11, n° 12.)
Titre : Cap. x. — Nomina et utilitas quorumdam præpositorum hujus loci partim notata.
Dates proposées : 1080 - 1085
Commentaire : Enumération des prieurs de Paray-le-Monial : *Andraldus, Gunterius, Segualdus, Girbertus* et *Hugues*. Nous avons déjà présenté la chronologie du prieur Hugues (1080-1115). Étant donné que Hugues est présenté comme : « *Hugo hoc tempore moderno* » le document date de son époque et on peut le considérer comme le premier de la série contemporaine du cartulaire.
Personnage principal : Hugues, prieur de Paray

Numéro CBMA : 7029 (Cartulaire de Paray-le-Monial, p. 11, n° 12.)
Commentaire : Texte manquant

Numéro CBMA : 7030 (Cartulaire de Paray-le-Monial, p. 12, n° 14.)
Titre : Cap. xii. — .....
Dates proposées : 1080 - 1085
Commentaire : Supplément du CBMA 7028, cette fois-ci on présente la liste des abbés de Cluny depuis Maieul. Le dernier à être mentionné est effectivement l'abbé Hugues. Texte incomplet.

Personnage principal : Hugues, abbé de Cluny

Numéro CBMA : 7031 (Cartulaire de Paray-le-Monial, p. 12-13, n° 15.)
Titre : Cap. xiii. — De præsulibus æduensibus qui hunc locum adcreverunt et de Bertranno vicecomite arvernensi.
Dates proposées : (979 - 989) et (1025 – 1030)
Commentaire : Document à date doublée. Les premiers évènements datent du temps de Gautier, évêque d'Autun (977 - 1024) et de Bertrannus, fils de Robert II de Clermont (après 962), vicomte d'Auvergne ; il est détecté dans une donation similaire à Saint-Julien de Brioude (980/985). La donation date probablement d'avant l'époque du comte Hugues (989). Les deuxièmes évènements datent du temps du successeur de Gautier, Hermuinus, évêque d'Autun (1025 - † 1055). Présence du comte-évêque Hugues I et de Tetbaldus, qui ne porte alors pas encore le titre de comte mais nepos (voir CBMA 7117), donc, vers 1030. Il existe une autre copie de cet acte dans le cartulaire (CBMA 7156).
Personnages principaux : Gautier, évêque d'Autun ; Bertrannus, vicomte d'Auvergne ; Hermuinus, évêque d'Autun ; Hugues, comte de Chalon ; Tetbaldus, comte de Chalon.
Personnages secondaires : ['D2', 'A6']

Numéro CBMA : 7032 (Cartulaire de Paray-le-Monial, p. 13-14, n° 16.)
Titre : Cap. xiiii. — Privilegium domni Aganonis episcopi huic loco, canonicorumque ejus.
Dates proposées : 1080 - 1098
Commentaire : Privilège d'Aganon de Mont Saint Jean, évêque d'Autun (1055 - † 25 juin 1098) ; le prieur Hugues apparait parmi les signataires.
Personnages principaux : Aganon de Mont Saint Jean ; Hugues, prieur de Paray

Numéro CBMA : 7033 (Cartulaire de Paray-le-Monial, p. 14-15, n° 17.)
Titre : Cap. xv. — De æcclesiis conquisitis recapitulatio. i. de ecclesia ragi. conquisita noticia.
Dates proposées : 1080 - 1100
Commentaire : Donation *pro anima* d'Adelais après le meurtre (*"gladio interfectus"*) de son mari Petrus de Cacchiaco. Adelais est la sœur du prieur Hugues. Signataires associés à la première période du prieur Hugues.
Personnages secondaires : ['I9', 'A23', 'G1']

Numéro CBMA : 7034 (Cartulaire de Paray-le-Monial, p. 15, n° 18.)
Titre : Cap. xvi. — Notitia de æcclesia vitriacensi.
Dates proposées : 1080 - 1100
Commentaire : Donation de Hugues miles et sa femme Stephana. Nous avons identifié ce couple de donneurs mentionnés dans d'autres chartes de la même période (CBMA 7037) comme Hugo Rubeus de Castello (Châtelperron) et sa femme Stephana. Signataires associés à la première période du prieur Hugues.
ersonnages secondaires : ['W4', 'H10', 'H18']

Numéro CBMA : 7035 (Cartulaire de Paray-le-Monial, p. 15-16, n° 19.)
Titre : Cap. xvii. — Carta ex ecclesia curdin.
Dates proposées : 1080 - 1100
Commentaire : Document supplémentaire au précédent, mêmes dates.
Personnage principal : Hugues II, comte de Chalon.
Personnages secondaires : ['A13']

Numéro CBMA : 7036 (Cartulaire de Paray-le-Monial, p. 16, n° 20.)
Titre : Cap. xviii. — Carta Attonis Buxol monachi de æcclesia Poisson aliisque rebus.
Dates proposées : 1080 – 1100
Commentaire : Donation d'Atto de Buxol, neveux du prieur Hugues. Parmi les signataires apparaissent son oncle Artaldus (appelé *avunculus*), sa mère Helizabeth et ses frères Hugues et Bernardus. Les signataires sont associés à la première période du prieur Hugues.

ersonnages secondaires : ['A26', 'A7', 'I9', 'H23']

Numéro CBMA : 7037 (Cartulaire de Paray-le-Monial, p. 16-17, n° 21.)
Titre : Cap. xviiii. — Carta domni Antelmi de æcclesia Sancti Leodegarii et aliis rebus.
Dates proposées : 1080 - 1100
Commentaire : Présence du prieur Hugues. Donation à Paray de Antelmus. Signataires associés à la première période du prieur Hugues.
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['H18', 'A23', 'W4', 'L3']

Numéro CBMA : 7038 (Cartulaire de Paray-le-Monial, p. 17-18, n° 22.)
Titre : (cap. xx.) — Carta domni Ylionis de ecclesia Novas Casas.
Dates proposées : 1090 - 1115
Commentaire : Donation *pro anima sua* d'Ylius de Chavasiget. La plupart des signataires sont attestés depuis la décennie de 1090 et pendant la deuxième période du prieur Hugues.
Personnages secondaires : ['A13', 'I5', 'H20', 'L3', 'A1', 'Y1']

Numéro CBMA : 7039 (Cartulaire de Paray-le-Monial, p. 18-19, n° 24.)
Commentaire : Texte manquant

Numéro CBMA : 7040 (Cartulaire de Paray-le-Monial, p. 18-19, n° 24.)
Titre : (cap.) xx.ii. — Carta domni Artaldi Blanchi, de ecclesia Sanctæ Mariæ de Bosco.
Dates proposées : 1080 - 1100
Commentaire : Donation d'Artaldus Blanchus à Paray d'une partie de l'église de Sancta Marie de Bosco. Témoins très connus de première période du prieur Hugues.
ersonnage principal : Hugues, prieur de Paray
Personnages secondaires : ['A23', 'G9', 'G16']

Numéro CBMA : 7041 (Cartulaire de Paray-le-Monial, p. 19-20, n° 25.)
Titre : (cap.) xxiii. — Carta donni Gaufredi, ex ecclesia Curbiniaco et aliis rebus.
Dates proposées : 1090 - 1098
Commentaire : Notice. Donation de Gaufredus de Cassagnias. Présence du prieur Hugues et de Aganon, évêque d'Autun (1055-1098). Les signataires sont associés à la deuxième période de Hugues ou en tout cas aux dernières années du XIe siècle.
Personnages principaux : Aganon, évêque d'Autun ; Hugues, prieur de Paray
Personnages secondaires : ['G9', 'W13', 'D6', 'T1', 'G16', 'H32']

Numéro CBMA : 7042 (Cartulaire de Paray-le-Monial, p. 20, n° 26.)
Titre : Cap. xxiiii. — Carta Girardi militis, de ecclesia Prisiaco et aliis rebus.
Dates proposées : 1100 - 1110
Commentaire : Notice. Donation *pro anima sua* de Girardus [de Buxol], frère du prieur Hugues. On peut supposer que ce Girardus sans nom de famille est Girardus Buxol à cause de la présence de son frère Artaldus et du fils de ce dernier Artaldus iunior. Un des témoins, Hugues Gaufredus est attesté dans deux autres chartes datés en 1102, 1105 et cette charte doit sûrement être datée de la deuxième période du prieur Hugues étant donné la chronologie des Buxol.
Personnages principals :
Personnages secondaires : ['H35', 'A23', 'G13', 'H9']

Numéro CBMA : 7043 (Cartulaire de Paray-le-Monial, p. 20-21, n° 27.)
Titre : Cap. xxv. — Adquisitio domni Hugonis prioris de quibusdam ecclesiis.
Dates proposées : 1080 - 1088
Commentaire : Notice. Donation de quelques églises à Paray de la part de Girardus et Rotbertus. Présence du prieur Hugues et de l'abbé Hugues. Affaire privée de la famille de Semur. Geoffroy III de Semur (-1123) futur prieur de Marcigny (1110-1123) et son frère Hugues-Dalmas de Semur (-après 1118), neveux de l'abbé Hugues ; les deux font une donation à Marcigny (CBMA 11186) lors

de l'entrée de Geoffroy à la vie monastique en 1088. L'absence de titre monastique pour désigner Geoffrey pourrait nous indiquer que la donation s'est produite avant 1088. Un groupe de signataires est détecté dans la première période du prieur Hugues.
Personnages principaux : Geoffroy III de Semur ; Hugues, prieur de Paray ; Hugues, abbé de Cluny.
Personnages secondaires : ['W17', 'G23', 'G17', 'P14', 'I3', 'T1', 'H8', 'H13', 'H20', 'H33', 'W14', 'H29']

Numéro CBMA : 7044 (Cartulaire de Paray-le-Monial, p. 21-22, n° 28.)
Titre : ? ?
Dates proposées : 1090 - 1098
Commentaire : Notice. Donation privée de la famille de Buxol, dans ce cas Artaldus, sa femme Gertrudis et leurs fils. Hugues de Buxol, frère d'Artaldus et prieur de Paray, apparait aussi mentionné comme laudator et certificateur. Aganon, évêque d'Autun (1055-1098) se trouve parmi les signataires. Il s'agit de la génération des neveux du prieur Hugues pour lesquels il faut considérer la décennie de 1090. Personnages principaux : Aganon, évêque d'Autun ; Hugues, prieur de Paray
Personnages secondaires : ['H21', 'A23', 'A26', 'D6', 'G12', 'G1']

Numéro CBMA : 7045 (Cartulaire de Paray-le-Monial, p. 22, n° 29.)
Titre : Cap. xxvi. — Carta domnæ Stephanæ.
Dates proposées : 1080 - 1100
Commentaire : Notice. Stephana, femme de Tatardi Renensis fait donation à Paray. Son mari est repéré dans 3 autres documents du cartulaire depuis la décennie de 1070. Probablement de la première période du prieur Hugues.
Personnages secondaires : ['B1', 'T1']

Numéro CBMA : 7046 (Cartulaire de Paray-le-Monial, p. 22, n° 30.)
Titre : Cap. xxvii. — Item ejusdem.
Dates proposées : 1080 - 1100
Commentaire : Notice. Supplément de CBMA 7045 avec les mêmes donneurs. Parmi les signataires apparait Artaldus de Buxol, signataire régulier pendant la première période de son frère, le prieur Hugues.
Personnages secondaires : ['A23']

Numéro CBMA : 7047 (Cartulaire de Paray-le-Monial, p. 23, n° 31.)
Titre CAP. XXVIII. — CARTA ARTALDI.
Dates proposées : 1080 - 1115
Commentaire : Notice. Donation d'Artaldus Grossus. Le nom « Grossus » n'est pas vraiment identifiable avec une famille parce qu'il a une origine similaire à celui du cognomen romain ; donc, même si « Grossus » est un nom porté par d'autres personnages dans le cartulaire on ne peut pas assurer qu'il y ait un lien familial.
Personnages secondaires : ['A18']

Numéro CBMA : 7048 (Cartulaire de Paray-le-Monial, p. 23, n° 32.)
Titre : (cap.) xxviiii. — Carta Evæ.
Dates proposées : 1080 - 1115
Commentaire : Notice. Eva de Veura fait *donation pro anima* de son mari à Paray, en présentant ses frères comme témoins. La seule connexion existant se trouve en CBMA 4673 où on trouve à Wilelmus de Veura, probablement le frère d'Eva, faisant donation à Cluny sous l'abbatiat de Hugues (1049 -1109).

Numéro CBMA : 7049 (Cartulaire de Paray-le-Monial, p. 23-24, n° 33.)
Titre : (cap.) xxxi. — Carta Girbergi.
Dates proposées : 1000 - 1049
Commentaire : Notice. Girberga donne deux pièces de terre à Cluny, qui seront transférées, peu de temps après à Paray par l'abbé Odilon. Il est impossible de déterminer la date parce que les

personnages portent seulement un prénom. Le plus probable est que ce transfert ait eu lieu après la donation de Paray à Cluny par le comte Hugues de Chalon (999). Curieusement, il y a une Girberga qui fait la donation pro remedio anima à Cluny de « duos curtilos » dans le site de Jaligny précisément dans la période de l'abbé Odilon, ici mentionné (CBMA 4718).
Personnage principal : Odilon, abbé de Cluny.

Numéro CBMA : 7050 (Cartulaire de Paray-le-Monial, p. 24, n° 34.)
Titre : (cap.) xxxiii. — Carta Rainerii de Villon.
Dates proposées : 1080 - 1100
Commentaire : Notice. Rainerius de Villon et sa femme Atala offrent un de leur fils au service du monastère. Deux signataires connus de la première période d'Hugues. La dénomination de Paray comme « *locum Aureœ Vallis* » est ancienne, rarement trouvée après 1100.
Personnages principals : Hugues , prieur de Paray
Personnages secondaires : ['H21', 'B2', 'D6']

Numéro CBMA : 7051 (Cartulaire de Paray-le-Monial, p. 24-25, n° 35.)
Titre : (cap.) xxxvi. — Carta Golferii.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation *pro debito* de Gulferius à Paray. Parmi les signataires se trouve Artaldus de Buxol (1080-1100) à qui le service était dû. Personnages secondaires : ['A23']

Numéro CBMA : 7052 (Cartulaire de Paray-le-Monial, p. 25, n° 36.)
Titre : (cap.) xxxvii. — Carta Bertranni de Chasuit.
Dates proposées : 1090 - 1100
Commentaire : Notice. Double acte de donation des frères Bertrannus et Girardus de Chasuit. Groupe de signataires très connus dans cette charte de donation double. Quelques-uns sont réguliers de la première période du prieur Hugues, mais y figurent deux signataires qui continuent à signer pendant la deuxième période d'Hugues.
Personnages secondaires : ['H21', 'A23', 'A10', 'G22', 'I3', 'B8', 'G19']

Numéro CBMA : 7053 (Cartulaire de Paray-le-Monial, p. 25, n° 37.)
Titre : Cap. xxxviii.
Dates proposées : 1080 - 1115
Commentaire : Acte assez laconique, impossible de dater avec plus de précision : « *Carta Roberti Mala Testa, qui dedit curtilum unum in Parriniaco.* »

Numéro CBMA : 7054 (Cartulaire de Paray-le-Monial, p. 25-26, n° 38.)
Titre : Cap. xlii. — Carta Hugonis de Parriniaco.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation pro anima sua de Hugues de Parriniaco. Laudatores et témoins assez connus de la première période du prieur Hugues.
Personnages secondaires : ['H21', 'A23', 'L4', 'A13']

Numéro CBMA : 7055 (Cartulaire de Paray-le-Monial, p. 26, n° 39.)
Titre : (cap.) xlv. — Donum Humberti de Scotia.
Dates proposées : 1080 - 1115
Commentaire : Notice. Donation d'Humbertus. Il est impossible de bien déterminer la date étant donné que « Humbertus » est la seule référence nominale.

Numéro CBMA : 7056 (Cartulaire de Paray-le-Monial, p. 26, n° 40.)
Titre : (cap.) xlvii. — Carta Lamberti militis.
Dates proposées : 1100 - 1115
Commentaire : Notice. Donation de Lambertus à Paray. La succession de noms de baptême « Ansedei, Lamberti, Rotberti » parmi les témoins correspond à la succession des frères Angleduris dont deux se trouvent dans quelques chartes de la deuxième période du prieur Hugues.

Personnages secondaires : ['B3', 'B4', 'R7']

Numéro CBMA : 7057 (Cartulaire de Paray-le-Monial, p. 26-27, n° 41.)
Titre : (cap.) xlviii. — Carta Walterii de Mardialgo.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation *pro susceptione* de Walterius de Mardialgo. Cet acte est associé au suivant. Parmi les témoins le couple Walterius et Rotrudis avec leurs fils et les frères d'Angleduris, tous détectés dans la première période du prieur Hugues.
Personnages secondaires : ['B3', 'B4', 'R7']

Numéro CBMA : 7058 (Cartulaire de Paray-le-Monial, p. 27, n° 42.)
Titre : (cap). l. — Item carta Rotrudis uxoris ejusdem Walterii et filiorum ejus.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation relative à la précédente, mais cette fois-ci faite par un de fils du couple mentionné dans la précédente, Enricus avec sa mère Rotrudis de Mardialgo. Ce personnage se trouve comme signataire, de nouveau avec la famille Angleduris dans deux autres chartes (CBMA 7174, CBMA 7121) datées de la même période.
Personnages secondaires : ['W20', 'G5']

Numéro CBMA : 7059 (Cartulaire de Paray-le-Monial, p. 27, n° 43)
Titre : Cap. lxii.
Dates proposées : 1050 - 1075
Commentaire : Notice. Document assez laconique. Il est indiqué « *acta tempore domni Girberti prioris* ». Entre l'époque du prieur Guntherius, détecté à la fin de la période d'Odilon de Cluny, et celle d'Hugues existent selon CBMA 7028 les prieurs Segualdus et Girbertus, repérés dans deux chartes à Paray. Segualdus prend la poste de prieur à Cluny en 1049, nous pouvons supposer alors un terminus a quo pour Girbertus des débuts de l'abbé Hugues jusqu'aux alentours de l'arrivée du prieur Hugues (voir la chronologie des prieurs de Paray, supra).
Personnage principal : Girbertus, prieur de Paray.

Numéro CBMA : 7060 (Cartulaire de Paray-le-Monial, p. 28, n° 44.)
Titre : (cap.) lxv. — Carta Ansedei præpositi Quadrilensis.
Dates proposées : 1065 - 1079
Commentaire : Notice. Donation familier d'Ansedeus « *miles Quadrilensis* » (Charolais). La chronologie de cet acte est déterminée par le suivant qui est complémentaire.
Personnage secondaire : ['W18']

Numéro CBMA : 7061 (Cartulaire de Paray-le-Monial, p. 28-29, n° 45.)
Titre : (cap.) lxvi. — Carta Wichardi filii Ansedei præpositi.
Dates proposées : 1065 - 1079
Commentaire : Notice. Donation à Paray de Wichardus, fils d'Ansedeus, avec ses frères lors de son entrée dans la vie religieuse. Il y a quatre témoins connus de la première période du prieur Hugues qui sont détectes après la décennie 1060, ce qui laisse supposer que le « domnus Hugo comes » mentionné dans l'acte est Hugues II comte de Chalon (1065-1079).
Personnage principal : Hugues II, comte de Chalon.
Personnages secondaires : ['G13', 'H38', 'W18', 'L3', 'G1']

Numéro CBMA : 7062 (Cartulaire de Paray-le-Monial, p. 29, n° 46.)
Titre : (cap.) lxvii. — Carta Aganonis fratris ejus.
Dates proposées : 1065 - 1079
Commentaire : Notice. Troisième donation de la série de la part de Aganon, fils du sus-mentionné Ansedeus. Ses frères apparaissent comme signataires. Un des témoins, Hugues de Chialoet, est détecté en CBMA 7155 (vers 1080) en donation avec les frères Buxol.
Personnages secondaires : ['H34']

Numéro CBMA : 7063 (Cartulaire de Paray-le-Monial, p. 29, n° 47.)
Titre : Cap. lxviii.
Dates proposées : 1080 - 1100
Commentaire : Notice laconique. Donation d'un moulin de la part du prieur Hugues à Wichardus, frère du monastère.
Personnage principal : Hugues, prieur de Paray

Numéro CBMA : 7064 (Cartulaire de Paray-le-Monial, p. 29-31, n° 48.)
Titre : (Parrochia Sancti Leodegarii.)
Dates proposées : 1080 - 1100
Commentaire : Notice complémentaire à la précédente qui recueille d'autres donations à Paray et notifie surtout différentes affaires privées liées à la donation du moulin à Wichardus qui avait suscité quelques problèmes. La famille Angleduris et la famille du Quadrilensis (CBMA 7060) apparaissent parmi les témoins.
Personnages principaux : Hugues, abbé de Cluny ; Hugues, prieur de Paray
Personnages secondaires : ['H10', 'L5', 'A8', 'G21', 'H23']

Numéro CBMA : 7065 (Cartulaire de Paray-le-Monial, p. 31, n° 49.)
Titre : (cap.) lxxviiii. — Carta Hugonis et Bernardi de Buxol.
Dates proposées : 1090 - 1110
Commentaire : Notice. Affaire privée impliquant trois des frères Buxol : Hugues, Bernardus et Atto, neveux du prieur Hugues, assez connus entre 1090-1110.
Personnages secondaires : ['A23', 'A26', 'B10', 'H9']

Numéro CBMA : 7066 (Cartulaire de Paray-le-Monial, p. 31-32, n° 50.)
Titre : (cap.) xci. — Carta Wilelmi de Maringis.
Dates proposées : 1049 - 1079
Commentaire : Notice. Donation à Paray de Willelmus de Maringis. Parmi les signataires se trouve Artaldus de Castel, frère de Hugo Rubeus de Castello, attesté depuis la décennie 1050. Willemus et Artaldus de Castel signent un acte à Cluny (CBMA 4293) datée en 1047-1049. Il s'agit de personnages d'avant l'arrivée du prieur Hugues.
Personnages secondaires : ['A16', 'W15']

Numéro CBMA : 7067 (Cartulaire de Paray-le-Monial, p. 32, n° 51.)
Titre : (cap. xc...) — Carta Letbaldi mulieris.
Dates proposées : 1080 - 1100
Commentaire : Notice très laconique, l'éditeur identifie à la donneuse Virberga comme la femme de Letbaldus (de Digonia), personnage du premier période de Hugues.
Personnages secondaires : [L3 ?]

Numéro CBMA : 7068 (Cartulaire de Paray-le-Monial, p. 32, n° 52.)
Titre : Cap. xcvii.
Dates proposées : 1080 - 1115
Commentaire : Notice incomplète. Aucune référence personnelle.

Numéro CBMA : 7069 (Cartulaire de Paray-le-Monial, p. 32, n° 53.)
Titre : (cap.) xcviiii. — Carta Ansedei de Parriniaco.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation d'Ansedeus de Parriniaco. Présence du prieur Hugues et de Deodatus, praepositus, qui apparait régulièrement dans les chartes de la première période du prieur.
Personnage principal : Hugues , prieur de Paray
Personnages secondaires : ['W21', 'H40']

Numéro CBMA : 7070 (Cartulaire de Paray-le-Monial, p. 32-33, n° 54.)
Titre : Cap. ci. — Carta Dalmacii de Centarbent.

Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Dalmatius de Centarberent à Paray. Dalmacius est nommé comme étant le grand-père (*avunculus*) de Robertus Dalmacius dans une charte de Cluny datée en 1128 (CBMA 5436), et il est aussi le père de Jocerannus de Centarbens (oncle de Robertus) et grand-père de Hugues de Centarbens qui apparaissent dans cette notice et dans plusieurs chartes à Cluny et Marcigny jusqu'à la première décennie du XIIe siècle (Cluny 3874, Marcigny 98,83, Cluny 3636, 3827). Parmi les signataires on trouve Girardus Buxol.
Personnages secondaires : ['G13', 'D4']

Numéro CBMA : 7071 (Cartulaire de Paray-le-Monial, p. 33, n° 55.)
Titre : Cap. c.ii. — carta Iodceranni de Varennis.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Jotcerannus de Varennas. Son frère Petrus, apparait comme signataire à Cluny (CBMA 5002) en 1080. Letbaldus de Copetra parmi les signataires, personnage de la première période d'Hugues.
Personnages secondaires : ['L4']

Numéro CBMA : 7072 (Cartulaire de Paray-le-Monial, p. 33-34, n° 56.)
Titre : Cap. ciiii. — carta girberti.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Girbertus et sa femme Raingardis. Présence de Deodatus, praepositus, très associé à la première période d'Hugues.
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['B5']

Numéro CBMA : 7073 (Cartulaire de Paray-le-Monial, p. 34, n° 57.)
Titre : Cap. cv. — Carta Hugonis de Giverzi.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Hugues de Giverzi. Signataires trouvés dans la première période d'Hugues.
Personnages secondaires : ['L4', 'H32']

Numéro CBMA : 7074 (Cartulaire de Paray-le-Monial, p. 34, n° 58.)
Titre : Cap. cvi. — cCrta Raimodis.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Raimodis de Vernol. Wido de la Rochi, son mari, est trouvé dans le cartulaire de Cluny (CBMA 4560) dans une charte datée entre 1049 et 1109. Les frères Buxol, Girardus et Artaldus, souscripteurs, offrent comme référence la première période du prieur Hugues.
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['W12', 'A23', 'G13']

Numéro CBMA : 7075 (Cartulaire de Paray-le-Monial, p. 35, n° 59.)
Titre : (cap.) cvii. — Carta Rainaldi de Copetra.
Dates proposées : 1065 - 1100
Commentaire : Notice. Donation de Girardus [de Moncellis]. Wichardus de Cavazola qui apparait parmi les signataires est détecté dans plusieurs chartes depuis la décennie de 1060 (CBMA 11293). L'éditeur mentionne à Rainaldus de Copetra dans le titre, mais il n'est pas mentionné dans l'acte.
Personnages secondaires : ['W4']

Numéro CBMA : 7076 (Cartulaire de Paray-le-Monial, p. 35-36, n° 60.)
Titre : (cap.) cviii. — Carta Dalmatii de Centarbent.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Dalmatius de Centarbent et de sa femme Ada. Deuxième charte de donation de Dalmatius de Centarbent (CBMA 7070). Ici sont présents une partie des personnages trouvés dans cette charte-là. Parmi les signataires Girardus de Buxol et le praepositus Deodatus.

Personnages secondaires : ['D4', 'H13', 'G13']

Numéro CBMA : 7077 (Cartulaire de Paray-le-Monial, p. 36, n° 61.)
Titre : (cap.) cxiiii. — Carta Widonis de Fracto Puteo.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Wido et de ses frères Girardus et Bonet. Trouvés aussi en CBMA 7069. Stephanus de Parriciacus, témoin, est repéré entre 1085-1100.
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['S7']

Numéro CBMA : 7078 (Cartulaire de Paray-le-Monial, p. 36, n° 62.)
Titre : (cap.) cxv. — Charte de Liebaud de Digoine, confirmée par sa [femme] et ses fils, (sans date).
Dates proposées : 1090 - 1105
Commentaire : Texte manquant. Letbaldus [de Digonia] et sa femme signent un autre acte ensuite (CBMA 7080) daté entre 1090 et 1105.
Personnages secondaires : ['L3']

Numéro CBMA : 7079 (Cartulaire de Paray-le-Monial, p. 37-37, n° 63.)
Titre : (cap.) cxvi. — Carta Widonis de Pinet.
Dates proposées : 1080 - 1115
Commentaire : Nous n'avons pas trouvé de références pour dater cette charte.

Numéro CBMA : 7080 (Cartulaire de Paray-le-Monial, p. 37, n° 64.)
Titre : (cap.) cxvii. — Carta Letbaldi Digonia.
Dates proposées : 1105 - 1115
Commentaire : Notice. Donation de Letbaldus I de Digonia avec sa femme et son fils (Letbaldus II) comme *laudatores*. On peut retracer ce personnage depuis la fin de la décennie de 1060 et on le trouve dans une charte de dispute à Cluny (CBMA 5257) appelé comme *pater* à côté de son fils. Cette charte de dispute avec l'abbaye après une donation de Bernardus de Cacchiaco est datée en 1105. Letbaldus I pourrait avoir disparu peu après étant donné sa longévité. En tout cas l'acte qui nous occupe doit avoir été émis à la fin de sa vie, alors 1105 peut marquer un *terminus a quo* pour ce document-ci, avec un *terminus ante quem* proche étant donné la présence de Jotcerannus de Copetra et de Hugues de Olsola repérés depuis la décennie de 1090.
Personnage principal : Tetbaldus, comte de Chalon ; Richardus prepositus.
Personnages secondaires : ['H21', 'G1', 'L3']

Numéro CBMA : 7081 (Cartulaire de Paray-le-Monial, p. 37-38, n° 65.)
Titre : Carta pro sepultura Rodulfi.
Dates proposées : 1080 - 1115
Commentaire : Nous n'avons pas trouvé de références pour dater cette charte.

Numéro CBMA : 7082 (Cartulaire de Paray-le-Monial, p. 38, n° 66.)
Titre : Carta Letbaldi Dilon.
Dates proposées : 1100 - 1115
Commentaire : Notice. Arrangement, avec l'intervention du prieur Hugues, entre Letbaldus de Digonia et Letabaldus de Bilon. Letbaldus II de Digonia et Petrus de Civignon, témoins, sont des personnages actifs depuis la dernière décennie du XIe. Nous sommes probablement dans la deuxième période du prieur Hugues comme en CBMA 7080.
Personnage principal : Hugues, prieur de Paray ; Richardus prepositus.
Personnages secondaires : ['P8', 'H27', 'L3']

Numéro CBMA : 7083 (Cartulaire de Paray-le-Monial, p. 38, n° 67.)
Titre : Cap. lxxix
Dates proposées : 1100 - 1115
Commentaire : Charte très laconique : « Carta Wlberti de Fracto Puteo, Puisrompu. »

Numéro CBMA : 7084 (Cartulaire de Paray-le-Monial, p. 38-39, n° 68.)
Titre : (cap.) lxxxii. — carta Gelini Meschins.
Dates proposées : 1070 - 1090
Commentaire : Notice. Déguerpissement de Gelinus Meschins en faveur de Paray. Presque tous les témoins sont détectés dans le cartulaire de Marcigny (CBMA 11174, 11181) dans la décennie de 1070. Le « Wilelmus de Centa(r)bent », témoin, est l'ancêtre direct de la famille Centarbent de CBMA 7070 et 7076.
Personnages secondaires : ['G30']

Numéro CBMA : 7085 (Cartulaire de Paray-le-Monial, p. 39, n° 69.)
Titre : (cap.) lxxxiii. — Carta Artaldi de Simirie.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Artaldus de Simirie. Le seul témoin connu de cette donation est Letbaldus de Copetra, repéré régulièrement dans la première période du prieur Hugues.
Personnages secondaires : ['L4']

Numéro CBMA : 7086 (Cartulaire de Paray-le-Monial, p. 39-40, n° 70.)
Titre : (cap.) lxxxv. — Carta Lamberti, fratris Duranni de Gurbiniaco.
Dates proposées : 1090 - 1115
Commentaire : Notice. Donation de Lambertus. Présence de deux témoins très connus de la deuxième période du prieur Hugues, mais présents depuis la dernière décennie du XIe.
Personnages secondaires : ['G9', 'G16', 'A25']

Numéro CBMA : 7087 (Cartulaire de Paray-le-Monial, p. 40, n° 71.)
Titre : (cap.) lxxxviii. — Carta Ansedei de Avingo.
Dates proposées : 1080 - 1115
Commentaire : Impossible de préciser la chronologie de cet acte. Tous les personnages portent seulement le nom de baptême.

Numéro CBMA : 7088 (Cartulaire de Paray-le-Monial, p. 40, n° 72.)
Titre : (cap.) lxxxviiii. — Carta Dodanæ.
Dates proposées : 1080 - 1115
Commentaire : Impossible de préciser la chronologie de cet acte. Tous les personnages portent seulement le nom de baptême.

Numéro CBMA : 7089 (Cartulaire de Paray-le-Monial, p. 41, n° 73.)
Titre : (cap.) xciiii. — Carta Girbaldi et Uncbergiæ.
Dates proposées : 1080 - 1115
Commentaire : Impossible de préciser la chronologie de cet acte. Tous les personnages portent seulement le nom de baptême.

Numéro CBMA : 7090 (Cartulaire de Paray-le-Monial, p. 41-42, n° 74.)
Titre : (cap.) xcv. — carta Landrici.
Dates proposées : 1080 - 1115
Commentaire : Impossible de préciser la chronologie de cet acte. Tous les personnages portent seulement le nom de baptême.

Numéro CBMA : 7091 (Cartulaire de Paray-le-Monial, p. 42, n° 75.)
Titre : Cap. xcvi.
Dates proposées : 1090 - 1115
Commentaire : Notice assez laconique : « Quidam homo nomine Dominicus vendidit monachis degentibus... »

Numéro CBMA : 7092 (Cartulaire de Paray-le-Monial, p. 42, n° 76.)
Titre : (cap). xcvii. — Carta Detcendæ filiorumque ejus.

Dates proposées : 1090 - 1115
Commentaire : Donation et placitum entre Letbaldus de Digonia (I) et Paray.
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['I4', 'G28', 'L3']

Numéro CBMA : 7093 (Cartulaire de Paray-le-Monial, p. 42-43, n° 77.)
Titre : (cap. xcix). — Xarta Ioceranni de Digonio.
Dates proposées : 1100 - 1115
Commentaire : Notice. Acte potentiellement complémentaire de 7092, étant donné qu'il s'agit d'une donation de Iotcerannus de Digonia qui apparait ici avec son fils Girardus, il s'agit donc de la deuxième période du prieur Hugues.
Personnages secondaires : ['I10', 'G28', 'L3']

Numéro CBMA : 7094 (Cartulaire de Paray-le-Monial, p. 43, n° 78.)
Titre : (cap.) c.i.
Dates proposées : 1090 - 1115
Commentaire : Impossible de préciser la chronologie de cet acte. Tous les personnages portent seulement le nom de baptême.

Numéro CBMA : 7095 (Cartulaire de Paray-le-Monial, p. 43-44, n° 79.)
Titre : (cap.) c.ii. — Carta Hugonis monachi, fratris Humberti Blanchi.
Dates proposées : 1090 - 1110
Commentaire : Notice. Donation de Hugues de Blanchi. Hugues de Saleniaco, témoin, détecté à Cluny et Paray depuis 1087 et pendant la deuxième période du prieur Hugues, comme Seguinus de Culmines qui apparait dans trois chartes datées à Cluny et Marcigny entre 1106-1108. Artaldus Ruilus, témoin, apparait à Cluny dans une charte datée par Bernard et Bruel de l'an 1050 (CBMA 4751), néanmoins une nouvelle date doit être proposé pour cette charte-là parce que Bernardus Ruil est un personnage 30 ans postérieur. Bernard et Bruel pensaient avoir corrigé Lambert de Barive qui avait daté cette charte de Cluny de l'an 1100 environ, ce qui selon nous était une date effectivement correcte.
Personnages secondaires : ['A15', 'S1', 'H23']

Numéro CBMA : 7096 (Cartulaire de Paray-le-Monial, p. 44, n° 80.)
Titre : (cap.) c.iii. — Carta Humberti de Domziaco villa.
Dates proposées : 1080 - 1115
Commentaire : Impossible de préciser la chronologie de cet acte. Tous les personnages portent un nom de baptême.

Numéro CBMA : 7097 (Cartulaire de Paray-le-Monial, p. 44, n° 81.)
Titre : (cap.) c.iiii. — Beraldi carta.
Dates proposées : 1080 - 1115
Commentaire : Impossible de préciser la chronologie de cet acte. Tous les personnages portent seulement le nom de baptême.

Numéro CBMA : 7098 (Cartulaire de Paray-le-Monial, p. 44-45, n° 82.)
Titre : (cap.) c.vi. — Bernardi carta Uriul.
Dates proposées : 1080 - 1115
Commentaire : Impossible de préciser la chronologie de cet acte. Tous les personnages portent seulement le nom de baptême.
Personnage principal : Adraldus decanus

Numéro CBMA : 7099 (Cartulaire de Paray-le-Monial, p. 45, n° 83.)
Titre : (cap.) c.vii. — Hugonis Rufi carta.
Dates proposées : 1080 - 1100

Commentaire : Notice. Arrangement entre Hugues Rufus de Castello et Grossa. Présence de Wichardus de Cavazola et de Hugo Rufus (Rubeo) de Castello, détectés depuis la décennie de 1060 et pendant la première période du prieur Hugues.
Personnages secondaires : ['H18', 'W4', 'H36']

Numéro CBMA : 7100 (Cartulaire de Paray-le-Monial, p. 45, n° 84.)
Titre : (cap. cix.) — Undradæ carta.
Dates proposées : 1080 - 1115
Commentaire : Nous n'avons pas trouvé de références chronologiques.

Numéro CBMA : 7101 (Cartulaire de Paray-le-Monial, p. 45-46, n° 85.)
Titre : (cap.) c.x. — Carta Eldigerii pro filio suo.
Dates proposées : 1080 - 1115
Commentaire : Impossible détecter la chronologie de cet acte. Tous les personnages portent seulement le nom de baptême.

Numéro CBMA : 7102 (Cartulaire de Paray-le-Monial, p. 46, n° 86.)
Titre : (cap.) c.xiiii. — Carta Girardi de Buxol.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation à Paray par Girardus de Buxol. Girardus, frère du prieur Hugues, signe avec les frères de Villa Urbana, témoins ici, un autre document à Paray (CBMA 7070) daté de cette même période où apparaissent aussi Dalmatius de Centarberent.
Personnages secondaires : ['G13']

Numéro CBMA : 7103 (Cartulaire de Paray-le-Monial, p. 46-47, n° 87.)
Titre : Cap. c.xv. — Carta Adeleydis comitissæ, Teudbaldi comitis filia.
Dates proposées : 1080 – 1085 ; 1085 - 1100
Commentaire : Notice. Document à date doublée. Dans la première action : donation d'Adelaïde, comtesse de Chalon avec son fils Wido de Tier et Hugo Dalmatius comme laudatores, Letbaldus de Digonia (I) et les frères Buxol parmi les témoins. Dans la deuxième action : arrangement peu de temps après entre Paray, sous le prieur Hugues, et Bernardus de Cacchiaco qui avait une *calumnia* sur une partie de cette donation. Hugues de Olsola et Stephanus de Parriniaco parmi les témoins. La donation est probablement faite pendant le temps le plus dur de l'interrègne (1080-1085) après la mort du comte Tetbaldus. La deuxième action est postérieure de quelques années étant donné que les signataires apparaissent pendant la première période du prieur Hugues.
Personnages principaux : Adelaïde, comtesse de Chalon ; Hugues, prieur de Paray ; Wido de Tiers.
Personnages secondaires : ['G6', 'H21', 'H26', 'A23', 'L1', 'H4', 'A26', 'B12', 'S7', 'L3', 'W11', 'H29']

Numéro CBMA : 7104 (Cartulaire de Paray-le-Monial, p. 48, n° 88.)
Titre : Carta domni Artaldi de Buxol.
Dates proposées : 1090 - 1100
Commentaire : Notice. Il s'agit d'un des très rares actes de Paray qui reprend le récit historique avec une formule : « *In præcedenti narratione hujus operis* ». L'acte ainsi cité est CBMA 7065 que nous avons daté de la première période du prieur Hugues. Cet acte reprend une affaire d'Atto de Buxol, neveux du prieur Hugues (ici présent), quelques années après étant donné la présence de deux témoins qui apparaissent en l'an 1100 environ.
Personnages secondaires : ['A26', 'B16', 'D6', 'S7', 'L3']

Numéro CBMA : 7105 (Cartulaire de Paray-le-Monial, p. 48, n° 89.)
Titre : Carta petri de castel.
Dates proposées : 1100 - 1115
Commentaire : Notice. Donation de Petrus de Castello lors de son entrée à la vie monastique. Petrus de Castello est détecté dans deux autres chartes de Paray pendant la deuxième période du prieur Hugues. Petrus est le fils de Hugo Rubeus de Castello dont l'existence est attestée depuis la décennie de 1060 jusqu'à 1105.

Personnages secondaires : ['P7']

Numéro CBMA : 7106 (Cartulaire de Paray-le-Monial, p. 49, n° 90.)
Titre : Carta Bernardi de Vals.
Dates proposées : 1094 - 1100
Commentaire : Notice. Donation de Bernardus de Vals lors de son entrée à la vie monastique. Témoins de l'acte trouvés depuis la dernière décennie du XIe (à partir 1094) et jusqu'à l'époque de l'abbé Pons. Cet acte fait un ensemble avec les deux suivantes puisque tous présentent les mêmes témoins.
Personnages secondaires : ['G23', 'A5', 'H20', 'A19', 'B15']

Numéro CBMA : 7107 (Cartulaire de Paray-le-Monial, p. 49, n° 91.)
Titre : Carta Arici militis Forensis.
Dates proposées : 1094 - 1098
Commentaire : Notice. Donation d'Aricus. Présence du prieur Hugues de Paray et du prieur Humbertus de Marcigny (1094-1098) .
Personnages principaux : Hugues, prieur de Paray ; Humbertus, prieur de Marcigny.
Personnages secondaires : ['G23', 'A19']

Numéro CBMA : 7108 (Cartulaire de Paray-le-Monial, p. 49-50, n° 92.)
Titre : Carta Boni Par.
Dates proposées : 1094 - 1098
Commentaire : Notice. Donation et arrangement de Bonus, fils de Tetardi Roenensis. Présence d'Hugues, prieur de Paray, et d'Humbertus, prieur de Marcigny (1094-1098).
Personnages principaux : Hugues, prieur de Paray ; Humbertus prieur de Marcigny.
Personnages secondaires : ['A23', 'H2', 'T1', 'A19', 'H23', 'L3']

Numéro CBMA : 7109 (Cartulaire de Paray-le-Monial, p. 50, n° 93.)
Titre : Carta Hugonis Iuvenis de larris.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Hugues de Larris. Témoins très connus qui apparaissent depuis la décennie de 1060 et coïncident dans la première période du prieur Hugues. L'un d'entre eux est détecté en 1108 (CBMA 5302).
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['H19', 'W5', 'G8', 'S7', 'H7', 'H16']

Numéro CBMA : 7110 (Cartulaire de Paray-le-Monial, p. 50-51, n° 94.)
Titre : Item alia carta de eodem.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Hugues de Larris. Présence du prieur Hugues. Acte complémentaire au précédent.
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['H19', 'H16', 'W4']

Numéro CBMA : 7111 (Cartulaire de Paray-le-Monial, p. 51, n° 95.)
Titre : (Carta Widonis de Corte militis).
Dates proposées : 1140 - 1156
Commentaire : Notice. Donation de Wido de Corte. Présence du prieur Girardus de Paray. On ne connait pas précisément les dates de succession entre les prieurs de Paray. Comme on peut voir dans la chronologie montrée auparavant, le prieur Girardus se trouve vers 1140/45 – 1156 sous la règence de Pierre le Vénerable à Cluny (1123-1157).
Personnage principal : Girardus, prieur de Paray.

Numéro CBMA : 7112 (Cartulaire de Paray-le-Monial, p. 52, n° 96.)
Titre : Cap. i. — Scammium cum Girardo et Hugone de Buxolio, de manso ad Sanctum Iustum cum fratribus Aureæ Vallis.

Dates proposées : 1030 - 1039
Commentaire : Notice. Présence du prieur Andraldus. Échange entre Hugues de Buxol (père du futur prieur Hugues) et la communauté de Paray qui cherche à compléter son domaine sur le territoire autrefois appelé Aurea Vallis. Le comte Hugues I apparait comme médiateur et parmi les témoins se trouve Tetbaldus, son neveu, portant aussi le titre de comte (ceci se produit aussi en CBMA 7117). Une date compatible au vu de cette association au comté de Tetbaldus serait la dernière décennie du période de comte Hugues I (voir ici CBMA 7117).
Personnages principaux : Andraldus, prieur de Paray ; Tetbaldus, comte de chalon ; Hugues I, comte-évêque, Odilon abbé de Cluny.
Personnages secondaires : ['A28', 'G13', 'H39']

Numéro CBMA : 7113 (Cartulaire de Paray-le-Monial, p. 52-53, n° 97.)
Titre : Cap. ii. — Carta Hugonis de Buxol.
Dates proposées : 1080 - 1090
Commentaire : Suite du document précédent, localisé chronologiquement après la mort d'Hugues I de Buxol (sa femme Aya est encore en vie). Leurs fils, Gerardus, Artaldus et Hugues, agissent en tant que témoins de l'acte comme d'autres personnages de la première période du prieur Hugues. Il est peu probable, si on considère l'acte précédent (qui nous propose une date la naissance d'Hugues I de Buxol vers la décennie de 1010), que Hugues soit vivant après la décennie de 1080, mais la date ne doit pas se situer beaucoup plus tard étant donné les témoins qui sont présents.
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['A28', 'H39', 'V2', 'B10', 'L4']

Numéro CBMA : 7114 (Cartulaire de Paray-le-Monial, p. 53, n° 98.)
Titre : Cap. iiii. — Carta Lamberti de Marciliaco.
Dates proposées : 1080 - 1100
Commentaire : Notice. *Donation pro anima* de Lambertus de Marciliaco. Quelques témoins apparaissent dans la charte précédente, témoins d'ailleurs très connus de la première période du prieur Hugues.
Personnages secondaires : ['H21', 'L2', 'L4', 'D4']

Numéro CBMA : 7115 (Cartulaire de Paray-le-Monial, p. 53-54, n° 99.)
Titre : (cap. v. —) carta Ilionis et uxoris ejus Rotrudis.
Dates proposées : 1080 - 1100
Commentaire : Notice. Ilius et sa femme Rotrudis offrent à Paray leur fils Achardus et quelques propriétés pour son entretien. Il s'agit d'Ylius de Chavasiget attesté en CBMA 7038 (1090-1100) dans la *dotatio pro anima* de sa femme et en CBMA 7026. Quelques témoins apparaissent dans la charte précédente et le donneur Lambertus de Marciliaco apparaît ici comme témoin. Souscripteurs très connus de la première période du prieur Hugues.
Personnages principaux : Durannus prepositus.
Personnages secondaires : ['L2', 'Y1']

Numéro CBMA : 7116 (Cartulaire de Paray-le-Monial, p. 54, n° 100.)
Titre : (cap. vi). — item eorumdem.
Dates proposées : 1080 - 1100
Commentaire : Notice. Acte complémentaire du précédent. Mêmes dates.
Personnages secondaires : ['L2', 'Y1']

Numéro CBMA : 7117 (Cartulaire de Paray-le-Monial, p. 54-55, n° 101.)
Titre : Carta Bertranni de Parriniaco.
Dates proposées : 1030 - 1039
Commentaire : Deuxième document (comme en CBMA 7112) signé par Hugues et Tetbaldus portant tous deux le titre de comte de Chalon. Selon nos estimations le comte-évêque Hugues est né vers 970 (on considère deux dates possibles de mort pour son père Lambertus :978 ou 988, cette dernière est la date proposée par le cartulaire ; Hugues aurait alors environ 18 ans) et son neveu Tetbaldus l'an 990

environ. Nous avons constaté que Tetbaldus ne porte le titre de comte avant de 1025 (voir CBMA 4265 et CBMA 7156) et dans d'autres chartes antérieures Tetbaldus est appelé *nepos* et avant *infans*. Donc l'association au titre de comte se produit vers la fin de la période du comte-évêque lorsqu'il est déjà âgé (60 ans environ). Parmi les témoins Hugo de Saliniaco, père de celui de CBMA 7095.
Personnages principaux : Hugues I, comte de Chalon ; Tetbaldus, comte de Chalon.
Personnages secondaires : ['H23']

Numéro CBMA : 7118 (Cartulaire de Paray-le-Monial, p. 55, n° 102.)
Titre : Item ejusdem.
Dates proposées : 1030 - 1039
Commentaire : Notice. Acte complémentaire du précédent. Donation pro anima de Bertrannus de Parriniaco pour son frère Ildinus qui apparait comme témoin dans la charte précédente. Nous considérons les mêmes dates étant donné que le comte-évêque Hugues apparait parmi les témoins.
Personnage principal : comte-évêque Hugues I

Numéro CBMA : 7119 (Cartulaire de Paray-le-Monial p. 55, n° 103.)
Titre : Item ejusdem.
Dates proposées : 1069 - 1089
Commentaire : Notice. Troisième acte de la série de Bertrannus de Parriniaco. *Donatio pro anima*. Le document date de la fin de la vie de Bertrannus et il est forcément postérieur à 1039 puisque le comte-évêque Hugues n'est plus présent. Walterius de Parreniaco, frère de Bertrannus et Ugo Mencioda, témoin ici, apparaissent ensemble pendant la première période d'Hugues (CBMA 7069), de même que Gaufredus de Varennas, témoin. Comme dans le cas de Hugues I de Buxol (CBMA 7113), chronologiquement il est improbable que Bertrannus, qui fait donation déjà vers 1030, soit vivant après la décennie de 1080, étant donné l'espérance de vie de l'époque.
Personnages secondaires : ['G5', 'H40']

Numéro CBMA : 7120 (Cartulaire de Paray-le-Monial p. 55, n° 104.)
Titre : Charte de M. Iosserand de Varennes et ses frères Hugues et Bernard, qui font donation du bois de corde (Cordensis), du temps de St Odile.
Dates proposées : 1080 - 1100
Commentaire : Texte manquant. Premier document d'une série de trois concernant la famille Varennas.

Numéro CBMA : 7121 (Cartulaire de Paray-le-Monial, p. 56, n° 105.)
Titre : Carta de Servis.
Dates proposées : 1080 - 1100
Commentaire : Notice. Opposition (*calumnia*) de Iotcerranus de Varennas et ses fils contre une donation de ses parents à Paray. Ses fils et tous les souscripteurs se trouvent régulièrement dans les chartes de la première période du prieur Hugues.
Personnages secondaires : ['A3', 'H3', 'D7']

Numéro CBMA : 7122 (Cartulaire de Paray-le-Monial p. 56, n° 106.)
Titre : Autre (charte) desd(its) de Varennes, qui suit.
Dates proposées : 1080 - 1100
Commentaire : Texte manquant. Troisième notice de la série à propos des Varennes. Probablement les mêmes dates.

Numéro CBMA : 7123 (Cartulaire de Paray-le-Monial, p. 56-57, n° 107.)
Titre : (cap.) lvii. — Carta donni Unberti Borbon.
Dates proposées : 1080 - 1083
Commentaire : Notice. *Donatio pro anima* de Hermengarda, sœur du comte Hugues II de Chalon et femme de Humbertus de Borbon, à Paray. Parmi les témoins se trouvent des membres des trois familles puissantes de la région : Buxol, Varennas et Copetra dont les membres apparaissent souvent dans la première période du prieur Hugues. Il existe une copie de cette charte (moins détaillée et sans

les témoins ici présents) daté de l'an 1083 dans le cartulaire de Cluny (CBMA 5027) inséré dans un récit mémoriel autour des donations du comte Tetbaldus et sa famille.
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['A23', 'G5', 'I2', 'I9', 'G1', 'A14', 'A9', 'A7']

Numéro CBMA : 7124 (Cartulaire de Paray-le-Monial, p. 57-58, n° 108.)
Titre : (cap.) lviii. — Carta Bernardi Senis de Angleduris.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Bernardus d'Angleduris et de ses fils. Témoins assez récurrents de la première période du prieur Hugues.
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['V2', 'A23', 'W4']

Numéro CBMA : 7125 (Cartulaire de Paray-le-Monial, p. 58, n° 109.)
Titre : Carta Wicardi de Vilers.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Wichardus de Vilers et Cecilia, sa femme ; probablement le « Richard de Vilers » de CBMA 7137 et le frère du Jotcerannus de Vilers qui apparait dans cette même série en CBMA 7123 et CBMA 7174.
Personnages secondaires : ['H12']

Numéro CBMA : 7126 (Cartulaire de Paray-le-Monial, p. 58, n° 110.)
Titre : (cap.) lxiii. — Carta Fulconis de Medens.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Fulco de Medens et de sa femme Raingardis. Gaufredus de Buxol, frère du prieur Hugues, apparait parmi les témoins.
Personnages secondaires : ['G3']

Numéro CBMA : 7127 (Cartulaire de Paray-le-Monial, p. 58-59, n° 111.)
Titre : Carta domni Rodberti de Montermenter.
Dates proposées : 1036, 1040-1050, 1080-1090.
Commentaire : Charte narrative à date triple. Narration du transfert du manse de Belmont à Paray. Le manse avait à l'origine été donné par le comte-évêque à Heldinus, seigneur du château du Mons Saint Vincent. Lors de son voyage à Jérusalem (en 1036 selon Canat de Chizy), le comte-évêque reçut la visite du prieur de Paray, Gunterius, qui lui demande la donation de ce manse. Hugues persuade finalement Heldinus de laisser le manse à Paray. Plus tard, après la mort du comte-évêque, le comte Tetbaldus reprend le manse pour récompenser Rotbertus de Montementer à la condition de le rendre à Paray après sa mort. Finalement à l'époque de l'abbé Hugues, Rotbertus rend la manse à Paray avant sa mort, accompagnant sa donation d'une forêt. Les témoins de cette dernière action apparaissent dans la première période du prieur Hugues. Le prieur Gunterius est enregistré à Paray depuis la décennie de 1030 environ et le comte Hugues est mort en 1039. La donation du comte Tetbaldus doit dater d'entre la décennie de 1040-50 et la donation de Robertus à Paray, à la fin de sa vie (au maximum à l'âge de 60 ans si on en croit l'espérance de vie de l'époque), doit avoir lieu effectivement peu après la décennie 1080.
Personnages principaux : Hugues I, comte de Chalon ; Hugues, prieur de Paray ; Gunterius, prieur.
Personnages secondaires : ['V2', 'B4', 'W16']

Numéro CBMA : 7128 (Cartulaire de Paray-le-Monial, p. 59-60, n° 112.)
Titre : Carta Wigonis de Varena.
Dates proposées : 1080 - 1109
Commentaire : Notice. Donation de Wido de Varennas. Témoins très connus de la première période du prieur Hugues.
Personnage principal : Rainerius archipresbiter
Personnages secondaires : ['G11', 'A23', 'W16', 'V1', 'D7']

Numéro CBMA : 7130 (Cartulaire de Paray-le-Monial, p. 60, n° 114.)
Titre : Carta Gaufredi canonici.
Dates proposées : 1070 - 1090
Commentaire : Notice. Donation par Gaufredus de Buxol d'un serf à Paray. Présence d'Aganon, évêque d'Autun (1055-1098). La mère de Gaufredus, Aya, et ici témoin, est probablement vivante (CBMA 7113) dans la décennie 1080.
Personnage principal : Aganon, évêque d'Autun
Personnages secondaires : ['A23','D6', 'A28', 'G3' 'W3']

Numéro CBMA : 7131 (Cartulaire de Paray-le-Monial, p. 60-61, n° 115.)
Titre : Carta domni Dalmatii de Borbon.
Dates proposées : 1115 - 1119
Commentaire : Notice. Premier document d'une série de trois à propos de la famille de Borbon-Lancy. Notice d'une querelle entre Dalmatius II de Borbon et Paray réglée peu de temps après par une donation pro anima sua par ses fils Foulques et Wichardus. Dalmas de Borbon est le petit-fils de Ansedeus I de Borbon déjà mentionné à Paray (CBMA 7156) et que nous avons daté entre 1025-1039. Nous sommes trois générations après cette date. Parmi les témoins apparait « Symonis ducis », que nous identifions comme Simon Ier de Lorraine, duc entre 1115-1139. Cette date, comme on l'a déjà mentionné, marque pour nous la limite de l'époque du prieur de Paray Hugues [de Buxol] (absent des actes après cette date), surtout parce que le remplaçant de Bernardus, qui lui succède dans ses fonctions, Artaldus est détecté de manière certaine en 1119 et on suppose qu'il est alors en poste depuis quelque temps.
Personnages principaux : Hugues, prieur de Paray ; Simon Ier, duc de Lorraine
Personnages secondaires : ['B3', 'D1', 'V2', 'R7']

Numéro CBMA : 7132 (Cartulaire de Paray-le-Monial, p. 61, n° 116.)
Titre : Carta dalmatii borbon.
Dates proposées : 1100 - 1115
Commentaire : Notice. Donation de Dalmatius II de Borbon et de son frère Humbertus II. Parmi les témoins figurent les frères d'Angleduris, détectés depuis la dernière décennie du XIe siècle et pendant la deuxième période du prieur Hugues.
Personnages secondaires : ['R7', 'B3', 'V2']

Numéro CBMA : 7133 (Cartulaire de Paray-le-Monial, p. 61, n° 117.)
Titre : Carta Umberti Borbon uxorisque ejus Ermengardis.
Dates proposées : 1080 - 1083
Commentaire : Notice de la donation que Humbertus de Borbon, père du dit Humbertus mentionné auparavant, et mari de Hermengarda de Chalon, fait en CBMA 7123 et qui est datée vers 1083.
Personnages secondaires : ['V2', 'H38', 'A8']

Numéro CBMA : 7134 (Cartulaire de Paray-le-Monial, p. 61, n° 118.)
Titre : Charta Widonis Florenzang'.
Dates proposées : 1080 - 1100
Commentaire : Notice très laconique. Donation de Widonis de Florenzang à la suite de son entrée dans la vie religieuse. Probablement le frère de Walterius de Florizangis (CBMA 7064) détecté dans la première période de Hugues.

Numéro CBMA : 7135 (Cartulaire de Paray-le-Monial, p. 62, n° 119.)
Titre : Donation de M. Anseau de Parriniaco, présens Iocerand Vilers, Guillaume de Velicourt, Geoffroy Digontii.
Dates proposées : 1080 - 1100
Commentaire : Texte manquant. Selon le titre y participent Ansedeus de Parriniaco (CBMA 7069) et Jocerannus Vilers (CBMA 7123) détectés depuis la décennie de 1080.

Numéro CBMA : 7136 (Cartulaire de Paray-le-Monial, p. 62, n° 120.)

Titre : Carta Emmonis militis et monachi de Foresta.
Dates proposées : 1100 - 1115
Commentaire : Impossible détecter la chronologie de cet acte. Tous les personnages portent seulement le nom de baptême.

Numéro CBMA : 7137 (Cartulaire de Paray-le-Monial, p. 62, n° 121.)
Titre : Charta Ansedei de la Fin.
Dates proposées : 1100 - 1115
Commentaire : Impossible détecter la chronologie de cet acte. Tous les personnages portent seulement le nom de baptême.

Numéro CBMA : 7138 (Cartulaire de Paray-le-Monial, p. 62, n° 122.)
Titre : Charta Stephani Goy.
Dates proposées : 1100 - 1115
Commentaire : Notice très laconique, impossible détecter la chronologie.

Numéro CBMA : 7139 (Cartulaire de Paray-le-Monial, p. 62-63, n° 123.)
Titre CARTA DE PONTONARIIS DE GRAVERIAS.
Dates proposées : 1080 - 1115
Commentaire : Notice. Affaire entre un pontenier (officier de pontenage) et Paray. La mention « *tunc temporis jam præfati loci prioris, domni Hugonis* » dénote une date indéterminée du priorat de Hugues.
Personnage principal : Hugues, prieur de Paray

Numéro CBMA : 7140 (Cartulaire de Paray-le-Monial, p. 63, n° 124.)
Titre : Carta Arnulfi Dulzoles.
Dates proposées : 1100 - 1115
Commentaire : Charte très laconique, impossible détecter la chronologie.

Numéro CBMA : 7141 (Cartulaire de Paray-le-Monial, p. 63-64, n° 125.)
Titre : Carta Maimbaldi.
Dates proposées : 1100 - 1115
Commentaire : Impossible de préciser la chronologie de cet acte. Tous les personnages portent seulement le nom de baptême.

Numéro CBMA : 7142 (Cartulaire de Paray-le-Monial, p. 64, n° 126.)
Titre : Donation par Hugues Dalmas, de choses sises à Orval ; cite son oncle Odilon parmi d'autres individus(sans date).
Dates proposées : 1090 - 1115
Commentaire : Texte manquant. Selon le titre il s'agit d'une donation de Hugues Dalmas (Dalmatius), fils de Dalmatius de Certarben (CBMA 7070), qui apparait faisant donation à Marcigny (CBMA 11291) dans une charte date en 1096.
Personnages secondaires : ['H29']

Numéro CBMA : 7143 (Cartulaire de Paray-le-Monial, p. 64, n° 127.)
Titre : Carta donni Hugonis de Saleniaco.
Dates proposées : 1090 - 1115
Commentaire : Notice. Donation, de Hugues de Saleniaco, au moment de sa mort. Hugues est détecté à Cluny et à Paray depuis 1087 et pendant la deuxième période du prieur Hugues, tout comme Letbaldus de Digonia, souscripteur.
Personnages secondaires : ['A23', 'H23', 'L3', 'A10']

Numéro CBMA : 7144 (Cartulaire de Paray-le-Monial, p. 64, n° 128.)
Titre : Carta Wilelmi monachi de Chanfeliz.
Dates proposées : 1090 - 1115

Commentaire : Notice très laconique, impossible de détecter la chronologie.

Numéro CBMA : 7145 (Cartulaire de Paray-le-Monial, p. 65, n° 129.)
Titre : Charta Bernardi monachi de Florenzangis.
Dates proposées : 1090 - 1115
Commentaire : Notice très laconique, impossible de détecter la chronologie.

Numéro CBMA : 7146 (Cartulaire de Paray-le-Monial, p. 65, n° 130.)
Titre : Carta Dalmatii de Borbon.
Dates proposées : 1100 - 1109
Commentaire : Notice d'un arrangement du temps de l'abbé Hugues de Cluny entre Dalmatius II de Borbon et le prieur Hugues à Marcigny. Falcon, fils de Dalmatius, apparait comme laudator. Dalmatius II apparait aussi dans un autre acte daté du même période (CBMA 7132).
Personnages principaux : Hugues, prieur de Paray ; Hugues, abbé de Cluny.
Personnages secondaires : ['D1']

Numéro CBMA : 7147 (Cartulaire de Paray-le-Monial, p. 65-66, n° 131.)
Titre : Carta domni Wichardi de Borbonio.
Dates proposées : 1055 - 1070
Commentaire : Notice. Donation de Wichardus de Borbon à Paray. Son fils Dalmatius I et son neveu Humbertus II (mari d'Ermengarde de Chalon, CBMA 7123) apparaissent comme laudatores. Wichardus est attesté dans une donation avec son grand-père vers 1030-1039 (CBMA 7156) et il fait une autre donation à Marcigny avec son fils dans une charte datée de 1055/65 (CBMA 11190). Ansedeus d'Angleduris, personnage de la première période d'Hugues figure parmi les témoins.
Personnages secondaires : ['A8', 'D1']

Numéro CBMA : 7148 (Cartulaire de Paray-le-Monial, p. 66, n° 132.)
Titre : Carta Aerderadi et Bernardi filii ejus.
Dates proposées : 1090 - 1110
Commentaire : Notice. Donation d'Aerderanus et sa famille. Durannus Merolus, témoin, est détecté depuis la dernière décennie du XIe siècle, et Stephanus de Parriciaco est attesté comme donneur pendant la première période d'Hugues.
Personnage principaux : Hugues, prieur de Paray ; Rodulfus archipresbiter.
Personnages secondaires : ['H28', 'D6', 'S7']

Numéro CBMA : 7150 (Cartulaire de Paray-le-Monial, p. 67, n° 134.)
Titre : Carta Ingeltrudis.
Dates proposées : 990 - 1039
Commentaire : Notice. Donation d'Ingeltrudis. Anselmus et Eldigerius, moines de Paray, apparaissent aussi comme témoins dans une charte datée par nous entre 992 et 1002 (CBMA 7209). Si on prend en compte cette information on peut identifier le « comte Hugues » nommé ici comme Hugues I de Chalon (987-1039).
Personnage principal : Hugues I, comte de Chalon

Numéro CBMA : 7151 (Cartulaire de Paray-le-Monial, p. 67, n° 135.)
Titre : ? ?
Dates proposées : 1075 - 1080
Commentaire : Notice très laconique. Présence du prieur Aymardus. Aymardus est probablement un prieur de transition entre Girbertus et Hugues, donc, vers 1075.
Personnage principal : Aymardus, prieur de Paray.

Numéro CBMA : 7152 (Cartulaire de Paray-le-Monial, p. 67, n° 136.)
Titre : ? ?
Dates proposées : 1080 - 1100

Commentaire : Notice. *Donatio pro anima* des frères Hugues et Bernardus de Buxol, neveux du prieur Hugues, en faveur de sa mère Helizabeth. Premier document de cette série de quatre à propos des Buxol. Hugues de Olsola, retrouvé dans la dernière décennie du XIe, parmi les témoins.
Personnages secondaires : ['H21', 'B11']

Numéro CBMA : 7154 (Cartulaire de Paray-le-Monial, p. 68, n° 138.)
Titre : Donum Girardi et item uxoris ejus Helisabet.
Dates proposées : 1060 - 1080
Commentaire : Notice. Donation de Girardus de Buxol, frère du prieur Hugues, et de sa femme Helisabet. Donation faite dans une date antérieure à la précédente. Girardus est détecté depuis la décennie de 1060, comme son frère Artaldus, signataire de la charte.
Personnages secondaires : ['A23', 'G13']

Numéro CBMA : 7155 (Cartulaire de Paray-le-Monial, p. 68, n° 139.)
Titre : ? ?
Dates proposées : 1080 - 1100
Commentaire : Notice qui répète l'action de CBMA 7152, mais avec d'autres souscripteurs.
Personnages secondaires : ['A23', 'G13', 'H34']

Numéro CBMA : 7156 (Cartulaire de Paray-le-Monial, p. 68, n° 140.)
Titre : Carta donni Hugonis comitis.
Dates proposées : 1025 - 1039
Commentaire : Notice. Donation du comte-évêque Hugues à Paray. Présence de Hugues I comte de Chalon (987-1039), et de Hermuinus, évêque d'Autun (1025-1055).
Personnages principaux : Hugues I, comte-évêque ; Hermuinus, évêque d'Autun ; Tetbaldus, comte de Chalon.
Personnages secondaires : ['A6']

Numéro CBMA : 7158 (Cartulaire de Paray-le-Monial, p. 69, n° 142.)
Titre : Charta Rainerii de Bhrat.
Dates proposées : 1030 - 1045
Commentaire : Notice très laconique. Déguerpissement d'une terre propriété de Raineirus. Présence du prieur Gunterius du temps du comte Tetbaldus.
Personnage principal : Gunterius, prieur de Paray.

Numéro CBMA : 7161 (Cartulaire de Paray-le-Monial, p. 70-71, n° 145.)
Titre : Carta de vivent.
Dates proposées : 999 - 1030
Commentaire : Notice de la concession de l'usufruit de différentes propriétés au comte-évêque Hugues. Présence du prieur Andraldus. Le remplaçant d'Adraldus, Gunterius, est détecté depuis la décennie du 1030.
Personnages principaux : le prieur Andraldus ; Odilon, abbé de Cluny ; Hugues I, comte-évêque.

Numéro CBMA : 7162 (Cartulaire de Paray-le-Monial, p. 71, n° 146.)
Titre : Carta Artaldi de Faltreriis.
Dates proposées : 1096 - 1115
Commentaire : Donation d'Artaldus de Faltrereeis, frère d'Antelmus et Jotcerannus. Tous trois apparaissent à Paray depuis la dernière décennie du XIe et pendant la deuxième période du prieur Hugues.
Personnages secondaires : ['A24']

Numéro CBMA : 7164 (Cartulaire de Paray-le-Monial, p. 71-72, n° 148.)
Titre : Carta de Monte Combroso.
Dates proposées : 1100 - 1123

Commentaire : Impossible de détecter une chronologie précise, mais cette notice est datée sans doute après l'an 1100 étant donné la terminologie. Les villes mentionnées appartiennent au Charollais et aux alentours d'Autun. On a constaté que l'église de Saint Didier, mentionné ici, est chef-lieu de paroisse depuis le début du Xe siècle (Perrecy n^{o}22) et que la *villa Balgeiacum* apparait nommé comme « *parrochia Balgiachensi*» dans le cartulaire de Marcigny dans une charte datée de l'an 1123 (CBMA 11412).

Numéro CBMA : 7165 (Cartulaire de Paray-le-Monial, p. 73, n° 149.)
Titre : Carta Willelmi de Castello.
Dates proposées : 1096 - 1115
Commentaire : Notice. Donation de Willemus de Castello. Willemus (de Lordono) et sa femme Emeldina sont attestés en Cluny (CBMA 4711), datés entre 1049 et 1109.

Numéro CBMA : 7166 (Cartulaire de Paray-le-Monial, p. 73, n° 150.)
Titre : Carta Wilelmi de Luurciaco.
Dates proposées : 1100 - 1115
Commentaire : Notice. Donation de Willemus de Lurciaco, lors de son entrée dans la vie religieuse, et après plusieurs litiges contre le monastère. Les souscripteurs sont détectés dans la deuxième période du prieur Hugues. Hugues de Scabellis et Aimone Jhavazole apparaissent aussi dans la charte suivante.
Personnages secondaires : ['A20', 'H16', 'A4']

Numéro CBMA : 7167 (Cartulaire de Paray-le-Monial, p. 74, n° 151.)
Titre : Carta domni Artaldi de Malereto.
Dates proposées : 1100 - 1115
Commentaire : Notice. Donation de Artaldus de Malereto lors de son entrée dans la vie religieuse. Souscripteurs détectés dans la deuxième période du prieur Hugues.
Personnages secondaires : ['G8', 'A20', 'R2', 'H16', 'A4']

Numéro CBMA : 7168 (Cartulaire de Paray-le-Monial, p. 74-75, n° 152.)
Titre : Carta donni Lebaudi de Digonia.
Dates proposées : 1105 - 1113
Commentaire : Notice. Donation *pro anima* de Letbaldus de Digonia. Letbaldus est né avant la régence du prieur Hugues. Les témoins de cette charte se trouvent réunis dans une autre donation faite par la comtesse Adelaïde que nous avons datée vers 1085. Cette charte-ci est datée de la deuxième période du prieur Hugues (Letbaldus fait sa dernière apparition datée en 1105, CBMA 5257) et avant l'an 1113, puisque dans la charte de fondation de la Fierté-sur-Grosne (daté de l'an 1113) Guide de Tiers, témoin ici, n'est plus le comte de Chalon et à sa place apparait déjà Guillemus (« *duorum comitum, Saberici videlicet et Guillelmi* », CBMA 10812). De plus, si on suit la chronologie d'Adelaïde, née probablement dans la décennie de 1020-30, il faudra rester plutôt autour de l'an 1105.
Personnages principaux : Adelaïde, comtesse de Chalon ; Wido de Tiers.
Personnages secondaires : ['A26', 'L3', 'W11', 'G1', 'A2']

Numéro CBMA : 7169 (Cartulaire de Paray-le-Monial, p. 75, n° 153.)
Titre : ? ?
Dates proposées : 1123 - 1147.
Commentaire : Notice d'un litige entre Paray et les fils de Hugues de Saliniaco, personnage détecté depuis la dernière décennie du XIe siècle et jusqu'à 1122. Son fils Hugues, ici présent, est détecté dans une charte de Cluny datée de 1147 (CBMA 5563).
Personnages secondaires : ['H23']

Numéro CBMA : 7170 (Cartulaire de Paray-le-Monial, p. 75-76, n° 154.)
Titre : Carta donni Dalmatii de Borbon.
Dates proposées : 1100 - 1115
Commentaire : Notice d'un arrangement entre Dalmatius II de Borbon et Paray à la suite d'un litige commencé par ce dernier. Il s'agit d'une donation *pro anima sua*, donc, probablement à la fin de sa

vie. Dans cette charte sont évoquées les actions notifiées en CBMA 7146 et CBMA 7131 que nous avons datées de la deuxième période du prieur Hugues. Parmi les souscripteurs, les frères Angleduris, qui sont détectés depuis la décennie 1090.
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['H4', 'A20', 'D1', 'B12', 'R7', 'B4']

Numéro CBMA : 7171 (Cartulaire de Paray-le-Monial, p. 76, n° 155.)
Titre : Carta Ioceranni de Centarben.
Dates proposées : 1080 - 1100, 1090 - 1100.
Commentaire : Notice à date doublée. *Donatio pro anima* de Jotcerannus de Centarben (voir CBMA 7070). Témoins très connus de la première période du prieur Hugues. Quelques temps après, Rotbertus Dalmatius, son neveu, détecté dans la période suivante, apparaît comme *laudator* de l'action. Les témoins de la deuxième action sont détectés depuis la décennie de 1090
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['H21', 'H26', 'A23', 'W7', 'H9', 'R8', 'H20', 'G14', 'W8', 'I6', 'L3', 'H7', 'H28']

Numéro CBMA : 7172 (Cartulaire de Paray-le-Monial, p. 77, n° 156.)
Titre : Carta Gaufredi de Varennas.
Dates proposées : 1080 - 1109
Commentaire : Notice. *Donatio pro salute anima* de Gaufredus de Varennas. Gaufredus et son frère, Guillaume, ici présent, sont repérés depuis 1083.
Personnages secondaires : ['R2', 'G5']

Numéro CBMA : 7173 (Cartulaire de Paray-le-Monial, p. 77-78, n° 157.)
Titre : Carta Bernardi de Angleduris.
Dates proposées : 1090 - 1097, 1097 - 1100
Commentaire : Notice. *Donatio pro redemptione anima* de Bernardus d'Angleduris et de sa femme Fulcreda. Son frère Ansedeus d'Angleduris, Falco de Borbon et Agnès, mère de ce dernier, apparaissent comme *laudatores*. Quelques temps après, Dalmatius II de Borbon fait laudatio de ce don à son retour de Jérusalem. Les frères d'Angleduris sont détectés depuis la décennie de 1090. Si on considère que le motif du voyage de Dalamtius II'était militaire et que la première Croisade a lieu en 1096, la deuxième date du document peut se trouver peu après, entre le 1097 et 1100, ce qui nous amène à dater la première action entre 1090 et 1097.
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['F1', 'B9', 'S5', 'R7', 'D5', 'B4']

Numéro CBMA : 7174 (Cartulaire de Paray-le-Monial, p. 78, n° 158.)
Titre : Carta donni Ansedei Angleduris.
Dates proposées : 1090 - 1097
Commentaire : Notice. Impignoration d'une terre, propriété d'Ansedeus d'Angelduris à Paray. Notice complémentaire de la précédente, mêmes dates.
Personnage principal : Hugues, prieur de Paray.
Personnages secondaires : ['G2', 'A23', 'I9', 'A8', 'H38', 'B3', 'B4', 'D5', 'H5']

Numéro CBMA : 7175 (Cartulaire de Paray-le-Monial, p. 79-80, n° 159.)
Titre : (carta domni Letbaldi de Digonia).
Dates proposées : 1100 - 1113
Commentaire : Notice de différents arrangements et donations entre Letbaldus de Digonia et Paray. La date de ce document est très proche de CBMA 7168 et probablement un peu antérieure, étant donné que ce dernier est le dernier document de Letbaldus de Digonia, dont on a déjà dit que la dernière apparition datée était en l'an 1105. Les souscripteurs sont aussi répétés.
Personnage principal : Hugues, prieur de Paray.
Personnages secondaires : ['G9', 'A23', 'I5', 'P8', 'L4', 'L3', 'G1']

Numéro CBMA : 7176 (Cartulaire de Paray-le-Monial, p. 80, n° 160.)
Titre : (carta domni Hugonis Dalmatii).
Dates proposées : 1090 - 1115
Commentaire : Notice d'un arrangement et donation de Hugues Dalmas à Paray. Hugues, fils de Dalmatius de Certarben (CBMA 7070), apparait depuis la décennie de 1090.
Personnage principal : Hugues, prieur de Paray.
Personnages secondaires : ['W6', 'D8', 'A19', 'L4', 'H29']

Numéro CBMA : 7177 (Cartulaire de Paray-le-Monial, p. 81, n° 161.)
Titre : (carta Willelmi de Varenas).
Dates proposées : 1080 - 1100
Commentaire : Notice. *Placitum* entre Willelmus de Varennas et Paray. Willelmus et son frère Gaufredus sont repérés depuis 1083. Letbaldus de Copetra et Artaldus de Buxol, trouvés dans la première période du prieur Hugues, figurent parmi les témoins.
Personnage principal : Hugues, prieur de Paray.
Personnages secondaires : ['A23', 'D8', 'S5', 'W16', 'R2', 'B14', 'L4']

Numéro CBMA : 7178 (Cartulaire de Paray-le-Monial, p. 81-82, n° 162.)
Titre : Carta de Luurciaco ecclesia.
Dates proposées : 1090 - 1109
Commentaire : Notice. Donation de Rodulfus de Turiaco et de son frère Rodulfus. Témoins trouvés principalement dans la deuxième période du prieur Hugues.
Personnages principaux : Hugues, prieur de Paray ; Hugues, abbé de Cluny.
Personnages secondaires : ['G2', 'R5', 'B14', 'P7', 'H16']

Numéro CBMA : 7180 (Cartulaire de Paray-le-Monial, p. 82, n° 164.)
Titre : ? ?
Dates proposées : 1090 - 1109
Commentaire : Notice. Donation d'Ildinus de Castello. Ildinus de Castello est trouvé en CBMA 7121 avec Ansedeus d'Angleduris et d'autres personnages de la deuxième période du prieur Hugues depuis la décennie de 1090.
Personnages principaux : Hugues, prieur de Paray ; Hugues, abbé de Cluny.
Personnage secondaire : ['I11']

Numéro CBMA : 7181 (Cartulaire de Paray-le-Monial, p. 82-83, n° 165.)
Titre : Incipiunt cartæ de Tolon et montis. — cap. i.
Dates proposées : 977 - 979
Commentaire : Notice de la consécration du monastère de Paray. Y participent le comte Lambertus et trois évêques signataires : Raoul, évêque de Chalon (977-986), Etienne (Stephanus) II, évêque de Clermont (942-984) et Ysardus (sans autre indication) ainsi que le frère de Lambertus Robert I, vicomte de Chalon.
Personnages principaux : Raoul, évêque de Chalon ; Etienne II, évêque de Clermont ; Robert vicomte de Chalon, Lambertus, comte de Chalon.

Numéro CBMA : 7182 (Cartulaire de Paray-le-Monial, p. 83, n° 166.)
Titre : Cap. ii. — (Placitum cum Walterio Vaslet).
Dates proposées : 1097 - 1115
Commentaire : Notice d'un *placitum* entre le décane de Tolon, Artaldus, et Walterius Vaslet. L'action est présidée par Gaufredus de Semur (deuxième période du prieur Hugues). Parmi les souscripteurs : Durannus Præpositus et Antelmus de Faltrieris (détecté depuis 1087). Artaldus est décane au début de l'époque de Berard de Châtillon, évêque de Mâcon (1097-1123).
Personnage principal : Artaldus, décane de Tolon.
Personnages secondaires : ['S2', 'A12', 'H14', 'G4', 'G18', 'W9']

Numéro CBMA : 7183 (Cartulaire de Paray-le-Monial, p. 83-85, n° 167.)

Titre : Carta de Sancto Benigno.
Dates proposées : 1097 - 1113
Commentaire : Notice. Donation commune à Paray de l'église de Sancto Benigno et ses propriétés par ses propriétaires dont Letbaldus de Digonia était le principal. Présence du prieur Hugues et d'Artaldus, décane de Tolon (vu dans la charte précédente). Letbaldus de Digonia a comme date de décès 1113 et Artaldus est décennie vers 1097.
Personnages principaux : Hugues, prieur de Paray ; Artaldus, décane de Tolon.
Personnages secondaires : ['G4', 'D1', 'D6', 'I4', 'G1', 'L3', 'H11']

Numéro CBMA : 7184 (Cartulaire de Paray-le-Monial, p. 85-86, n° 168.)
Titre : ? ?
Dates proposées : 1097 - 1115
Commentaire : Notice. Donation d'Ysiliardus, chevalier. Les témoins sont présentés en utilisant leur nom de baptême mais la série « Aymo, Gaufredus et Artaldus » correspond aux frères Fautrières.
Personnage principal : Hugues, prieur de Paray

Numéro CBMA : 7185 (Cartulaire de Paray-le-Monial, p. 86, n° 169.)
Titre : Charte de Hugues de Fautrières, chevalier.
Dates proposées : 1097 - 1115
Commentaire : Texte manquant. troisième document qui concerne la famille Fautrières.

Numéro CBMA : 7186 (Cartulaire de Paray-le-Monial, p. 86, n° 170.)
Titre : Carta Artaldi de Faltrieris.
Dates proposées : 1097 - 1115
Commentaire : Notice. Arrangement entre Artaldus de Faltrieris et Teuzam. Quatrième document concernant les Fautrières.
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['A24']

Numéro CBMA : 7189 (Cartulaire de Paray-le-Monial, p. 87, n° 173.)
Titre : Carta Agyæ.
Dates proposées : 1065 - 1079
Commentaire : Notice. Donation d'Agia à Paray. Les personnages ne portent pas de nom de famille mais leurs prénoms coïncident avec les Buxol : Agia (Aya) et ses fils : Gaufredus et Girardus, qui sont présents dans les actes depuis la décennie de 1070. La signature du comte Hugues II de Chalon (1065-1079) nous offre une tranche chronologique claire.
Personnage principal : Hugues II, comte de Chalon.
Personnages secondaires : ['A28', 'G3']

Numéro CBMA : 7190 (Cartulaire de Paray-le-Monial, p. 87, n° 174.)
Titre : Carta Alveræ.
Dates proposées : 1065 - 1079
Commentaire : Impossible détecter le moment chronologique, les personnages portent un seul nom.

Numéro CBMA : 7191 (Cartulaire de Paray-le-Monial, p. 87-88, n° 175.)
Titre : Donum Letbaldi de Digonia.
Dates proposées : 1080 - 1100
Commentaire : Notice. Donation de Letbaldus I de Digonia. Témoins très connus de la première période du prieur Hugues.
Personnages secondaires : ['H21', 'A23', 'H9', 'B12', 'L4']

Numéro CBMA : 7192 (Cartulaire de Paray-le-Monial, p. 88, n° 176.)
Titre : (Donum domni Guichardi de Digonia).
Dates proposées : 1123 (Date dans le document : 1149)

Commentaire : Notice datée dans le document de l'an 1149. Donation testamentaire *pro remedio anima* de Guichardus de Digonia avec l'accord de ses frères Letbaldus III et Jotcerannus II. La date que porte ce document (1149) ne se correspond pas avec la date de l'action qui doit avoir eu lieu pendant la décennie de 1120, selon la chronologie que nous avons exposée à propos des Digonia. Les témoins sont très connus de la deuxième période d'Hugues et apparaissent dans les nombreuses chartes signées par Letbaldus II de Digonia, père du donneur. Dans le document est consigné que la donation est faite « *in præsentia ejusdem et donni Hugonis abbatis et Artaldi prioris Cluniacensis* ». Artaldus est prieur de Paray sous l'abbé Pons après le prieur Hugues (vers 1118), et il continue à ce poste après la mort de l'abbé Hugues. Pourtant, comme le suppose Canat de Chizy, le « Hugonis abbatis » consigné ici pourrait faire référence l'abbé Hugues II (1123) qui passe quelques mois à l'abbatiat entre Pons et Pierre, ce qui est chronologiquement cohérent.
Personnages principaux : Hugues II, abbé de Cluny ; Artaldus, prieur de Cluny.
Personnages secondaires : ['G27', 'I5', 'H9', 'G10', 'P8', 'G1']

Numéro CBMA : 7193 (Cartulaire de Paray-le-Monial, p. 88-89, n° 177.)
Titre : (Carta d. Aymæ, uxoris Gulferii de Ialiniaco).
Dates proposées : 1090 - 1109
Commentaire : Notice. Donation *pro remedio anima* d'Ayma de Castello, fille de Hugues Rubeus de Castel et sœur de Petrus de Castello. Son père est repéré depuis la décennie de 1050 et jusqu'à la fin du siècle et son frère apparaît dans deux chartes de Paray depuis la décennie de 1090 (CBMA 7105), il faut imaginer qu'elle a une période chronologique similaire.
Personnage principal : Hugues, abbé de Cluny.
Personnages secondaires : ['H18', 'G24', 'P7']

Numéro CBMA : 7195 (Cartulaire de Paray-le-Monial, p. 89-90, n° 179.)
Titre : (Werpitio domni Letbaldi de Digonia).
Dates proposées : 1070 - 1079
Commentaire : Notice de déguerpissement de Letbaldus I de Digonia à Paray. La présence du prieur Girbertus est chronologiquement un peu problématique. Prédécesseur du prieur Hugues, on peut estimer qu'il a été prieur depuis la décennie de 1050 et jusqu'aux alentours de 1075. D'ailleurs Letbaldus de Digonia et Letbaldus de Copetra souscrivent à Paray et à Cluny plusieurs documents entre la décennie de 1090 et 1105, dont deux avec Bernardus Dolmont (Del Monte) qui apparaît ici comme témoin. Ils appartiennent à la génération du prieur Hugues, pourtant une chronologie vers la fin de la décennie de 1070 pour ces trois personnages, même si elle est parfaitement possible, peut sembler un peu fallacieuse.
Personnages principals : Girbertus prieur de Paray.
Personnages secondaires : ['L4', 'L3']

Numéro CBMA : 7196 (Cartulaire de Paray-le-Monial, p. 90, n° 180.)
Titre : Incipiunt Cartæ Baronenses.
Dates proposées : 988 - 1012
Commentaire : Notice. Donation du comte-évêque Hugues, de sa mère Adelaïde et de son frère Maurice. Maurice est probablement le demi-frère du comte Hugues, né du deuxième mariage d'Adelaïde, après la mort du comte Lambertus avec Geoffrey I d'Anjou mort en l'an 987, selon certaines généalogies. La date de mort de Maurice est estimée vers l'an 1012 dans nos bases de données mais nous n'avons pas trouvé la source. Il apparaît dans une autre donation à Marcigny, datée de l'an 1005.
Personnages principaux : Hugues I, comte-évêque ; Adelaïde de Chalon ; Maurice de Chalon.
Personnages secondaires : ['A2']

Numéro CBMA : 7198 (Cartulaire de Paray-le-Monial, p. 90-91, n° 182.)
Titre : Hugonis comitis carta de chamelgias.
Dates proposées : 988 - 1012
Commentaire : Notice. Donation à Paray par le comte Hugues I. Les signataires sont les mêmes frères de Paray que ceux trouvés en CBMA 7196.

Personnage principal : Hugues I comte-évêque.

Numéro CBMA : 7199 (Cartulaire de Paray-le-Monial, p. 91, n° 183.)
Titre : Carta Hugonis (comitis et episcopi).
Dates proposées : 999 - 1039
Commentaire : Notice. Donation à Paray par le comte-évêque Hugues I.
Personnage principal : Hugues I, comte-évêque

Numéro CBMA : 7200 (Cartulaire de Paray-le-Monial, p. 91-92, n° 184.)
Titre : (Carta Hugonis comitis Cabilonensium).
Dates proposées : 999 - 1019
Commentaire : Notice. Donation de plusieurs propriétés du comte Hugues à Paray. Parmi les signataires sa sœur Mathilde, son neveu Otton (comte de Mâcon) et Tetbaldus (futur comte de Chalon et fils de Mathilde). Mathilde fait une dernière donation à Cluny (CBMA 2693) datée par Bernard et Bruel vers 1015, puis elle disparaît des actes. Une charte à Cluny (CBMA 2722) datée vers 1019 notifie la confirmation par le comte Hugues de cette donation « *quam olim* » faite par sa sœur, ce qui pourrait être lu comme un *terminus ante quem* pour Mathilde.
Personnages principaux : Hugues I, comte-évêque ; Otton, comte de Macon ; Tetbaldus, comte de Chalon ; Mathilde de Chalon.

Numéro CBMA : 7201 (Cartulaire de Paray-le-Monial, p. 92, n° 185.)
Titre : Carta episcopi Hugonis, de Nova Villa.
Dates proposées : 999 - 1004
Commentaire : Notice. Donation du comte-évêque Hugues I. Parmi les souscripteurs son frère Robertus, vicomte et Otte-Guillaume, comte de Mâcon (982-1004).
Personnages principaux : Hugues I, comte-évêque ; Robertus, vicomte de Chalon ; Otte-Guillaume, comte de Mâcon.

Numéro CBMA : 7202 (Cartulaire de Paray-le-Monial, p. 92-93, n° 186.)
Titre : Carta Hugonis comitis, de manso qui vocatur Li chaux.
Dates proposées : 999 - 1012
Commentaire : Notice. Donation du comte-évêque Hugues I. Les signataires sont les mêmes frères de Paray trouvés en CBMA 7196 et 7198.
Personnage principal : Hugues I, comte-évêque

Numéro CBMA : 7203 (Cartulaire de Paray-le-Monial, p. 93, n° 187.)
Titre : Carta Vetus Milerias.
Dates proposées : 1024 (Daté de 1024)
Commentaire : Charte datée de l'an 28 du roi Robert (Robert II le Pieux, 996-1031). Donation *pro anima* de Deodatus à Paray.
Personnage principal : Robert II le Pieux roi des francs

Numéro CBMA : 7204 (Cartulaire de Paray-le-Monial, p. 93-94, n° 188.)
Titre : Carta de quibusdam terris, quas adquisivit domnus Artaldus decanus in monte Sancti Vincencii.
Dates proposées : 1090 - 1100
Commentaire : Notice. Donation *pro remedio anima* de Artaldus de Berziaco et de Richardus, homme à son service. Parmi les témoins figurent les frères d'Artaldus, Gaufredus et Seguinus. Artaldus de Berziaco est attesté dans deux chartes à Cluny datées de 1076 et 1090, et cette donation est faite à la fin de sa vie.
Personnages secondaires : ['A22']

Numéro CBMA : 7205 (Cartulaire de Paray-le-Monial, p. 94-95, n° 189.)
Titre : (carta domni Narjodi episcopi Aeduensis).
Dates proposées : 1109 (Daté de 1109)

Commentaire : Notice d'un arrangement après de multiples litiges entre l'évêché d'Autun et Cluny. Charte datée de l'an 1109 et signée à Nevers, alors que Pascal II (1099-1118) est pape et Norgaud de Toucy (1098-1112) évêque d'Autun
Personnages principaux : Norgaud de Toucy, évêque d'Autun, Pascal II, Pape.

Numéro CBMA : 7206 (Cartulaire de Paray-le-Monial, p. 95-96, n° 190.)
Titre : (Carta de scamnio de terris de Bosco et de Langiaco).
Dates proposées : 1116 - 1118
Commentaire : Notice. Arrangement entre Bernardus, prieur de Paray et Gauscerannus, décane de Prisiaco. Bernardus est prieur à Paray après le prieur Hugues et remplacé par Artaldus dont les chronologies attestent qu'il est déjà en poste en 1119.
Personnages principaux : Bernardus, prieur de Paray ; Gauscerannus, décane de Prisiaco ; Pons, abbé de Cluny.
Personnages secondaires : ['H30']

Numéro CBMA : 7208 (Cartulaire de Paray-le-Monial, p. 96-97, n° 192.)
Titre : Notitia placiti quod fecit domnus Hugo prior cum Fulcone et fratribus ejus de Castro Buxit. — xi.
Dates proposées : 1080 - 1093
Commentaire : Notice d'un arrangement, après divers conflits entre Paray et Gauslenus et d'autres frères de l'église de Chalon. Ce Gauslenus est détecté comme *decanus cabilonensis* dans deux chartes de l'an 1087 et 1093 (Gallia Christiana 888A). La comtesse Adelaïde intervient comme médiatrice et sa chronologie ne dépasse pas l'an 1100 (CBMA 7168). La charte est signé « *in Frantia Philippo regnante* » faisant référence à Philippe Ier, roi des Francs (1060-1108).
Personnages principaux : Hugues, abbé de Cluny ; Hugues, prieur de Paray ; Adelaïde, comtesse de Chalon.
Personnages secondaires : ['A2']

Numéro CBMA : 7209 (Cartulaire de Paray-le-Monial, p. 97-98, n° 193.)
Titre : Carta Hugonis comitis et Alaidis matris suæ.
Dates proposées : 992 - 1002
Commentaire : Notice. Donation de diverses propriétés du comte Hugues I et de sa mère Adelaïde à Paray. Parmi les souscripteurs figurent son demi- frère Maurice (mort vers 1012), Henri [Eudes] duc de Bourgogne (Henri Ier de Bourgogne) et sa femme Garlinda. Henri est mort en l'an 1002 et Garlinda (Gersenda) est sa deuxième femme, après la mort de sa première épouse Girberga (probablement fille de Lambertus, comte de Chalon) en 992. Henri et Garlinda se trouvent dans une autre charte du cartulaire de Saint-Symphorien d'Autun (nº15 et 17, pp. 37 et 42). D'après la chronique d'Autun, Gersenda est répudiée en 996, mais cette information reste incertaine.
Personnages principaux : Hugues I, comte de Chalon ; Adelaïde, comtesse de Chalon ; Henri Ier de Bourgogne.

Numéro CBMA : 7210 (Cartulaire de Paray-le-Monial, p. 98-99, n° 194.)
Titre : Carta Hugonis comitis de mercato Sancti Vincentii.
Dates proposées : 1010 - 1030
Commentaire : Notice. Donation par le comte Hugues du marché de Saint Vincent à Paray. Présence du comte Hugues I et de Tetbaldus. Telbaldus, né vers 990, porte encore la dénomination « *nepos* ». Comme on a vu en CBMA 7112 et 7117, Tetbaldus porte le titre de comte de Chalon probablement sur la fin de la vie de son oncle Hugues.
Personnages principaux : Hugues I, comte de Chalon ; Tetbaldus, comte de Chalon.

Numéro CBMA : 7211 (Cartulaire de Paray-le-Monial, p. 99, n° 195.)
Titre : Item, ejusdem comitis.
Dates proposées : 985 - 989
Commentaire : Notice. Donation du comte Hugues. Ses parents apparaissent comme témoins. Ce document contribue à confirmer que la mort de Lambertus se situe plus probablement vers 988 et

non en 978 (voir CBMA 7020), Hugues I étant alors déjà adulte (environ 18 ans). Ce document doit donc dater des dernières années de Lambertus.
Personnages principaux : Hugues I, comte de Chalon ; Lambertus, comte de Chalon ; Adelaïde, comtesse de Chalon.

Numéro CBMA : 7212 (Cartulaire de Paray-le-Monial, p. 99, n° 196.)
Titre : (carta) Lamberti comitis.
Dates proposées : 977 - 979
Commentaire : Notice. Donation de fondation par Lambertus et sa femme Adelaïde à Paray.
Personnages principaux : Lambertus, comte de Chalon ; Adelaïde, comtesse de Chalon.

Numéro CBMA : 7213 (Cartulaire de Paray-le-Monial, p. 99, n° 197.)
Titre : Carta Widonis de Rocca.
Dates proposées : 1050 - 1080
Commentaire : Notice. Donation à Paray par Wido de la Rocha. Parmi les témoins figurent sa femme et ses fils. Un de ses fils, Wido, témoin ici, apparait à Paray dans la première période du prieur Hugues.
Personnages secondaires : ['W12']

Numéro CBMA : 7214 (Cartulaire de Paray-le-Monial, p. 100, n° 198.)
Titre : ? ?
Dates proposées : 1050 - 1080
Commentaire : Notice laconique. Donation de Bernardus de la Porte, père du « Bernardus iuvenis de la Porta » et qui apparait dans CBMA 7104 daté de la première période du prieur Hugues, et probablement frère de Willelmus de la Porta détecté en CBMA 7061 vers 1065.
Personnages secondaires : ['B16']

Numéro CBMA : 7215 (Cartulaire de Paray-le-Monial, p. 100, n° 199.)
Titre : Carta de Curte Judea.
Dates proposées : 999 - 1039
Commentaire : Notice. Donation du comte-évêque Hugues à Paray.
Personnage principal : Hugues II comte-évêque.

Numéro CBMA : 7216 (Cartulaire de Paray-le-Monial, p. 100-101, n° 200.)
Titre : (Placitum cum domno Hugone de Borbon).
Dates proposées : 1130 - 1146
Commentaire : Notice d'un arrangement entre Hugues de Borbon et Paray. Hugues est le fils de Humbertus III de Borbon (CBMA 7146) qui apparait dans la deuxième période du prieur Hugues signant une charte avec son père. Le prieur Buchardus se trouve entre le prieur Bernardus et le prieur Girardus [de Copetra] (détecté déjà en 1147). Donc, il faut proposer pour Buchardus une chronologie sous l'abbé Pierre (1123-1147) vers la décennie de 1130/40.
Personnage principal : Burchardus, prieur de Paray
Personnages secondaires : ['H17', 'R6']

Numéro CBMA : 7217 (Cartulaire de Paray-le-Monial, p. 101, n° 201.)
Titre : (Placitum cum Dalmatio et Wichardo de Borbone).
Dates proposées : 1130 - 1146
Commentaire : Notice d'un arrangement entre Dalmacius III et Wichardus III de Borbon (fils du précédemment mentionné, Hugues de Borbon) et le prieur de Paray Burchardus. Parmi les signataires le comte Guillemus I de Chalon (1113-1166).
Personnages principaux : Burchardus, prieur de Paray ; Guillemus I de Chalon
Personnages secondaires : ['R4', 'P12', 'G1', 'P6']

Numéro CBMA : 7218 (Cartulaire de Paray-le-Monial, p. 101-102, n° 202.)
Titre : (Carta episcopi Eduensis de ecclesia Reniaci).

Dates proposées : 1151 (Daté de 1151)
Commentaire : Présence du prieur Girardus, successeur de Burchardus. Concession à Paray de la part de Henricus (Henri de Bourgogne), évêque d'Autun (1148-1170).
Personnages principaux : Girardus, prieur de Paray ; Henricus, évêque d'Autun.
Personnages secondaires : ['P13', 'P9', 'B13']

Numéro CBMA : 7219 (Cartulaire de Paray-le-Monial, p. 102, n° 203.)
Titre : Carta Guidonis de Bussul
Dates proposées : 1200 - 1228
Commentaire : Notice. *Donatio pro anima* de Guigundus de Buxol à Paray. Parmi les témoins se trouve « G. de Sancto Albino » dont quelques-uns de leurs apparentés se trouvent à Paray depuis la dernière décennie du XIIe siècle : Hugues (CBMA 7238, daté en 1180), Guillelmus (CBMA 7242, daté de 1205).
Personnages secondaires : ['G29']

Numéro CBMA : 7220 (Cartulaire de Paray-le-Monial, p. 102-103, n° 204.)
Titre : (Carta Gauceranni de Copetra).
Dates proposées : 1147 - 1156
Commentaire : Notice. Donation de Gauscerannus de Copetra, frère de l'actuel prieur de Paray, Girardus. Présence du prieur Girardus [de Copetra] et du comte Guillelmus (1113-1166). Parmi les témoins : Girardus de Digonia et Wido de Curte. Girardus est attesté comme prieur depuis au moins 1147 (CBMA 7222) jusqu'à 1156, dernière période de l'abbatiat de Pierre à Cluny.
Personnages principaux : Girardus [de Copetra] prieur de Paray ; Guillelmus, comte de Chalon.
Personnages secondaires : ['I8', 'A27', 'G20', 'B3', 'W10', 'L4', 'G1', 'P9']

Numéro CBMA : 7222 (Cartulaire de Paray-le-Monial, p. 104-105, n° 206.)
Titre : Hæc est carta de fine.
Dates proposées : 1147 (Daté de 1147)
Commentaire : Donation-vente par Guiccardus de Calvo Monte à Paray. Parmi les donneurs les frères Digonia, Gauscerannus II, Letbaldus III et Girardus détectés dans une donation commune dans le cartulaire de la Fierté-sur-Grosne (CBMA 11026) vers 1145.
Personnage principal : Girardus prieur de Paray.
Personnages secondaires : ['A27', 'G10', 'R8', 'G26', 'A11']

Numéro CBMA : 7223 (Cartulaire de Paray-le-Monial, p. 105-107, n° 207.)
Titre : (Sacramentum Karoli in manu cardinalium).
Dates proposées : 1119 - 1120
Commentaire : Jugement. Ce document assez rare dans le cartulaire notifie le procès d'excommunication, le pardon et la pénitence d'un certain Karolus et de ses complices, jugés par la communauté clunisienne pour leurs crimes contre l'ordre (probablement contre des biens matériels). Le document est postérieur au prieur Hugues, étant donné la présence du prieur Artaldus, et antérieur à l'arrivée de l'abbé Pierre. Canat de Chizy propose comme date 1119 étant donné la présence parmi les témoins de deux cardinaux (« *duo cardinales Romani, domnus Conradus et domnus Comes* ») présents à Cluny à l'occasion de l'élection du pape Calixte II dans l'abbaye de Cluny. En fait, ayant récupéré la liste des électeurs nous avons facilement identifié ces cardinaux comme : Conradus de Suburra, futur pape Anatase IV (1153-1154), et Kuno (Cunon) von Urach, cardinal allemand légat de la papauté en France.
Personnages principaux : Artaldus, prieur de Paray ; Bernardus, prieur de Cluny ; Pons, abbé de Cluny.
Personnages secondaires : ['P5', 'R9', 'G9', 'H25', 'G17', 'P2', 'A21', 'H17', 'G7', 'P1', 'G1', 'H6']

Numéro CBMA : 7224 (Cartulaire de Paray-le-Monial, p. 107-108, n° 208.)
Titre : Carta domni widonis de tierno atque comitis cabilonensis.
Dates proposées : 1096 - 1100

Commentaire : Charte mémorielle du temps du prieur Hugues. Wido de Tiers, comte de Chalon, confirme les « *bonas consuetidines* » de Paray, dans le cas contraire il renonce aux droits fonciers en faveur du monastère. Selon la charte, Wido vient au monastère avant son départ à Jérusalem lors de la première Croisade (1096) et jure devant le prieur Hugues, ce qui devrait marquer notre terminus a quo pour cet acte.
Personnages principaux : Wido de Tiers, comte de Chalon ; Hugues, prieur de Paray
Personnages secondaires : ['G20', 'A12', 'S7', 'L4', 'W11']

Numéro CBMA : 7225 (Cartulaire de Paray-le-Monial, p. 108-109, n° 209.)
Titre : (Carta domni Willelmi comitis).
Dates proposées : 1116 - 1118
Commentaire : Notice. Guillelmus comte de Chalon, assure devant l'abbé Pierre avoir terminé avec toutes les « *malas consuetudines* » qu'ils auraient pris pendant les campagnes à Jérusalem. Le prieur de Paray, Bernardus, ici présent, est à Paray vers 1116-1118 et Letbaldus II de Digonia, témoin, dans plusieurs chartes de la période 2 du prieur Hugues.
Personnages principaux : Guillelmus, comte de Chalon ; Pierre, abbé de Cluny ; Artaldus, prieur de Paray ; Bernardus, prieur de Cluny.
Personnages secondaires : ['L3', 'A3']

Numéro CBMA : 7226 (Cartulaire de Paray-le-Monial, p. 109-110, n° 210.)
Titre : Carta de dono curdiaci.
Dates proposées : 1100 - 1110, 1147 - 1157
Commentaire : Notice à date doublée. Donation *pro anima* par Witburgis à Paray, alors que Hugues est prieur. Longtemps après (« *Post multum vero temporis* »), sous le prieur Girardus, son fils, Petrus de Nucibus, fait une donation *pro anima sua* complémentaire à celle de sa mère à Curdiaco. La présence de Jotcerannus de Vilers et du *presbiter* Durannus parmi les témoins à la fin de la charte nous renvoie aux décennies de 1100-1110. La donation de Petrus de Nucibus a eu lieu en tout cas après 1147.
Personnages principaux : Hugues, prieur de Paray ; Girardus, prieur de Paray.
Personnages secondaires : ['I9', 'H15', 'H6', 'P11']

Numéro CBMA : 7227 (Cartulaire de Paray-le-Monial, p. 110, n° 211.)
Titre : (carta domni Hugonis de Borbon).
Dates proposées : 1116 - 1130
Commentaire : Notice. Concession et donation à Paray des propriétés que Hugues de Borbon avait à Pulcra Spina. Hugues de Borbon, Gaufredus de Palfol et Hugues de Maniaco, témoins, sont repérés depuis la décennie 1110. Le prieur Artaldus, successeur de Hugues, se trouve à Paray jusqu'à l'époque de l'abbé Pierre.
Personnages principaux : Artaldus, prieur de Paray.
Personnages secondaires : ['H33', 'H17', 'H23', 'G7']

Numéro CBMA : 7228 (Cartulaire de Paray-le-Monial, p. 110-111, n° 212.)
Titre : Carta Petri de Roccha.
Dates proposées : 1080- 1100
Commentaire : Notice. Arrangement et donation entre Petrus de Rochia (*infans*) et Paray. Parmi les témoins le *preaepositus* Durannus et Wido *presbiter* (CBMA 7041) qui nous amènent à la première période du prieur Hugues. Les témoins ici présents, bien que nombreux, sont complètement inconnus dans le cartulaire.
Personnages secondaires : ['D9']

Numéro CBMA : 7232 (Cartulaire de Paray-le-Monial, p. 116, n° 215.)
Titre : Carta de manso Hugonis, prioris de Paredo, ad Avengum.
Dates proposées : 1063 - 1065
Commentaire : Charte. Donation à Marcigny par Hugues [de Buxol] lors de son entrée dans la vie religieuse dans le monastère de Paray-le-Monial. Parmi les souscripteurs se trouvent sa mère Aya et

ses frères Gerardus, Artaldus y Gaufredus. L'écrivain est le prieur Durannus qui selon la chronologie de Marcigny avait occupé le poste entre 1063 et 1065 , ce qui est très cohérent par rapport aux dates proposées pour le prieur Hugues. Il existe une copie de cet acte à Marcigny (CBMA 11240).
Personnages principaux : Hugues, prieur de Paray ; Durannus, prieur de Marcigny.
Personnages secondaires : ['A28', 'G3']

Numéro CBMA : 7233 (Cartulaire de Paray-le-Monial, p. 117, n° 216.)
Titre : (carta Adaleidæ, uxoris Petri de Chucy).
Dates proposées : 1080 - 1100
Commentaire : Charte. Donation *pro anima* pour son mari par Adelaïde de Buxol. Ses fils, Galterius et Guillelmus, apparaissent comme *laudatores*, ses frères Hugues et Artaldus parmi les témoins. Très probablement première période du prieur Hugues.
Personnage principal : Hugues, prieur de Paray
Personnages secondaires : ['A23', 'H1', 'D3', 'P3', 'C1']

Numéro CBMA : 7234 (Cartulaire de Paray-le-Monial, p. 117, n° 217.)
Titre : (carta Mariæ, filiæ Albuini Grossi).
Dates proposées : 1115 - 1130
Commentaire : Notice. Donation de Maria, belle-sœur de Bernardus de Buxolio (neveu du prieur Hugues) lors de son entrée dans la vie religieuse à Marcigny. Maria est la fille d'Albuinus Grossus, qui apparait dans deux chartes entre 1123-1130. Par ailleurs Bernardus est repéré depuis le deuxième période du prieur Hugues. La chronologie de cette charte nous amène à la période suivante, étant donné que Maria est probablement très jeune au moment d'entrer à Marcigny.
Personnages secondaires : ['B11', 'A5']

Numéro CBMA : 7235 (Cartulaire de Paray-le-Monial, p. 117-118, n° 218.)
Titre : Carta Hugonis de Buxolio.
Dates proposées : 1112 - 1115
Commentaire : Notice. Arrangement privé entre Hugues de Buxol et les frères de Marcigny. Stephanus (Étienne Ier de Baugé) évêque d'Autun (1112-1139) apparait ici comme juge. Cet acte est une copie d'un original déposé dans le cartulaire de Marcigny (11284). Hugues de Buxol apparait comme « Hugues de Buxolio » et « Hugues prior Paredi » ce qui nous amène à penser que cet Hugues de Buxol est le neveu de Hugues, le prieur de Paray, fils de son frère Gerardus, mais que le Hugues prieur est aussi mentionné puisqu'il apparait comme témoin de l'affaire.
Personnages principaux : Hugues, prieur de Paray ; Stephanus, évêque d'Autun.
Personnages secondaires : ['H9']

Numéro CBMA : 7236 (Cartulaire de Paray-le-Monial, p. 118, n° 219.)
Titre : Carta Hugonis de Sivignon.
Dates proposées : 1098 - 1109
Commentaire : Notice. Donation à Marcigny d'Elisabeth, mère de Petrus et Hugues de Sivignon. Parmi les témoins les frères Hugues et Bernardus de Buxol. Tant les témoins que les frères Sivignon sont détectés depuis la décennie 1100 et jusqu'à la décennie 1130. Il existe une copie de cet acte dans le cartulaire de Marcigny (CBMA 11403).
Personnages secondaires : ['H9', 'H37', 'B11']

Numéro CBMA : 7237 (Cartulaire de Paray-le-Monial, p. 118, n° 220.)
Titre : Carta Petri de Sivignon.
Dates proposées : 1098 - 1109
Commentaire : Notice complémentaire à la précédente. Donation *pro redemptione anima* de Petrus de Civignon. Présence de Seguinus, prieur de Marcigny . Différents membres de la famille Buxol et Robertus Dalmatius apparaissent comme témoins. Selon la chronologie des prieurs de Marcigny, Seguinus se trouve entre l'an 1096 et l'an 1109 environ. Il existe une copie de cet acte dans le cartulaire de Marcigny (CBMA 11302).
Personnage principal : Seguinus, prieur de Marcigny

Personnages secondaires : ['A23', 'H9', 'I1', 'R8', 'P8', 'H37']

Numéro CBMA : 7238 (Cartulaire de Paray-le-Monial, p. 119-121, n° 221.)
Titre : Comes Cabilonensis, de Paredo.
Dates proposées : 1180 (Daté de 1180)
Commentaire : Charte. Arrangement après multiples litiges et confrontations entre Willemus (Guillaume III de Chalon), comte de Chalon, et Cluny, signé à Lourdon.
Personnages principaux : Guillaume III, comte de Chalon ; Johannes, prieur de Paray ; Tetbaldus, abbé de Cluny ; Beraldus prior de Cluny.
Personnages secondaires : ['P10', 'H24', 'H22']

Numéro CBMA : 7239 (Cartulaire de Paray-le-Monial, p. 122-124, n° 222.)
Titre : CartaPphilippi regis, de domo Paredi.
Date proposée : 1180 (Daté de 1180)
Commentaire : *Praeceptum* de confirmation par le roi Philippe (Philippe II Auguste) de l'arrangement entre le comte Guillelmus II de Chalon et Cluny notifié dans la charte précédente.
Personnages principaux : Philippe II Auguste, roi des francs ; Guillelmus II, comte de Chalon

Numéro CBMA : 7241 (Cartulaire de Paray-le-Monial, p. 125, n° 224.)
Titre : C(arta Philippi) regis pro Paredo.
Date proposée : 1204 (Daté de 1204)
Commentaire : Notice d'un diplôme. Le roi Philippe II Auguste prend en custodie toutes les villes de la Bourgogne, Paray inclus. Daté à Paris.
Personnage principal : Philippe II Auguste, roi des francs

Numéro CBMA : 7242 (Cartulaire de Paray-le-Monial, p. 125-126, n° 225.)
Titre : Carta comitissæ cabilonensis, de paredo.
Dates proposées : 1205 (Daté de 1205)
Commentaire : Charte. Confirmation par Beatrix (Beatrix de Chalon ou de Thiers, femme d'Étienne II d'Auxonne), comtesse de Chalon (1202-1227), du contenu de la charte (de commune) que son père (Guillaume II de Chalon) comte de Chalon (1167-1202) avait concédé à Paray et Tolon. Présence de l'évêque d'Auxerre, Hugues de Noyers (1183-1206).
Personnages principaux : Beatrix de Chalon ; Guillaume II de Chalon ; Hugues de Noyers, évêque d'Auxerre.
Personnages secondaires : ['H31', 'B6', 'R1', 'S3', 'W19', 'S4', 'B7', 'G25', 'H22']

Numéro CBMA : 7243 (Cartulaire de Paray-le-Monial, p. 126-127, n° 226.)
Titre : Carta et compositio pacis facte inter ecclesiam Cluniacensem et comitissam Cabilonensem, de Paredo, per manus G. Eduensis et R. Cabil(onensis) et P. Masticonensis episcoporum.
Dates proposées : 1205 (Daté de 1205)
Commentaire : Charte complémentaire de la précédente. Confirmation et notification de l'accord de paix entre la comtesse de Chalon, et par extension les villes de Paray et Tolon, prises sous sa protection, et l'ordre clunisien. Dans le compromis les deux parties s'accordent sur la cessation des hostilités, l'exonération fiscale et la restitution de l'aumône.
Personnages principaux : Beatrix de Chalon ; Guillaume II de Chalon.
Personnages secondaires : ['G27', 'H31', 'B6', 'R3', 'W19', 'S4', 'B7', 'G26', 'G25', 'H22']

Numéro CBMA : 7244 (Cartulaire de Paray-le-Monial, p. 128, n° 227.)
Titre : Carta ducis Burgundie, de Paredo et de Tolun.
Dates proposées : 1243 (Daté de 1243)
Commentaire : Charte. Présence de l'abbé Hugues [de Rochecorbon] (1236-1244) et de Hugues [IV] duc de Bourgogne (1218-1272) et comte de Chalon (depuis 1237). Confirmation par Hugues qu'il respectera et fera respecter le contenu des chartes concédées par ses prédécesseurs (CBMA 7238, 7242, 7243) aux villes de Paray et Tolon.
Personnages principaux : Hugues [de Rochecorbon] abbé de Cluny ; Hugues [IV] duc de Bourgogne.

Numéro CBMA : 7246 (Cartulaire de Paray-le-Monial, p. 128-137, n° 229.)
Titre : Hec est visitacio facta in provincia Lugdun(ensi) per donnum Stephanum (socium) in ordine.
Dates proposées : 1242 (Daté de 1242)
Commentaire : Visitation Cluniacense.

Numéro CBMA : 7247 (Cartulaire de Paray-le-Monial, p. 138-144, n° 230.)
Titre : Visitatio provin[cie Lug]d(unensis) facta per donnum Stephanum elemosinarium cluniaci et per priorem de k[adr]e[lla].
Dates proposées : 1248 (Daté de 1248)
Commentaire : Visitation Cluniacense.

Numéro CBMA : 7248 (Cartulaire de Paray-le-Monial, p. 144-145, n° 231.)
Titre : ? ?
Dates proposées : 1271 (Daté de 1271)
Commentaire : Charte. Présence de Girardus [de Beauvoir], évêque d'Autun (? - 1276). Vente par Girardus de Sinemuro (Sémur) d'une partie ses propriétés à Paray au duc Hugues [IV] de Bourgogne.
Personnages principaux : Hugues [IV] duc de Bourgogne, Girardus [de Beauvoir], évêque d'Autun
Personnages secondaires : ['G15']

INDEX PERSONARUM

G18 Girardus Vriols
G19 Girardus de Cahic
G20 Girardus de Copetra
G21 Girardus de Fracto Puteo
G22 Girbertus de Parriniaco
G23 Gotardus Bargi
G24 Gulferius de Ialiniaco
G25 Gvicardo de Sancto Albano
G26 Gviccardum de Calvo Monte
G27 Gvichardus de Digonia
G28 Gvido de Sancto Privato
G29 Gvigundus de Busol
G30 Gelinus Meschins

H

H1 Hebradus de Paret
H2 Heldinus Tisions
H3 Heldinus de Castel
H4 Helgodus Bers
H5 Heynricus de Maldelgo
H6 Hugo Beral
H7 Hugo Blanchus
H8 Hugo Bochars
H9 Hugo Buxul
H10 Hugo Larr'
H11 Hugo Letbals
H12 Hugo Morelion
H13 Hugo Parriniaco
H14 Hugo Ratbal
H15 Hugo Rodulfus
H16 Hugo Scabellus
H17 Hugo de Borbon
H18 Hugo de Castel
H19 Hugo de Larris
H20 Hugo de Laval
H21 Hugo de Olsola
H22 Hugo de Petra Campi
H23 Hugo de Saliniaco
H24 Hugo de Sancto Albino
H25 Hugo de Sancto Preiecto
H26 Hugo de Vals
H27 Hugo de la Tor
H28 Hugo del Paschet
H29 Hugone Dalmatio
H30 Hugo Xartines
H31 Hugo de Digonia
H32 Hugo de Giverze
H33 Hugo Marciniaco
H34 Hugo de Chialet
H35 Hugo Gaufredi
H36 Hugo Rufi
H37 Hugo de Sivignon
H38 Humbertus Borbon
H39 Hugo I Buxol
H40 Hugo Mencioda

I

I1 Ildricus Hisperons
I2 Ilius Paganus
I3 Iocerannus Valestinas
I4 Iocerannus de Faltrierias
I5 Iocerannus Marcile
I6 Iocerannus de Centarben
I7 Iocerannus de Maringis
I8 Iotcerannus de Petra Campi
I9 Iotcerannus de Vilers
I10 Iotcerannus de Digonia
I11 Ildinus de Castello

L

L1 Lambertus Descal
L2 Lambertus Marciliaco
L3 Letbaldus Digonia
L4 Letbaldus de Copetra
L5 Lethaldus de Capella

M

M1 Martinus de Vineis

P

P1 Petrus Blainus
P2 Petrus Capreolus
P3 Petrus de Chucy
P4 Petrus de Chanziaco
P5 Petrus Rufus
P6 Petrus de Bosco
P7 Petrus de Castel
P8 Petrus de Civinon
P9 Petrus de Frigido Puteo
P10 Petrus de Marciaco
P11 Petrus de Nucibus
P12 Petrus de Varenis
P13 Petrus de Vitriaco
P14 Pontius Rufo

R

R1 Radulfo de Marniaco
R2 Rainerius de Poli
R3 Reinaudus Dalmatius
R4 Robertus de Moneta
R5 Rodulfus de Turiaco
R6 Rodulfus de Vitriaco
R7 Rotbertus Anglediuris
R8 Rotbertus Dalmatius
R9 Rotbertus de Vigiaco

S

S1 Segvinus de Colmines
S2 Segvinus Rungifers
S3 Stephano de Bosco
S4 Stephano de Castello de Montana
S5 Stephanus Anglediuris
S6 Stephanus de Capella
S7 Stephanus Parriciacus

T

T1 Tetardus Renensis

V

V1 Vvalbertus de Laval
V2 Vvilelmus de Vetul'

W

W3 Walterius Florinzangis
W3 Walterius de Parriniaco
W4 Wichardus Cavazola
W5 Wichardus Lvurci
W6 Wichardus de Marciniaco
W7 Wichardus de Saliniaco
W8 Wido Forestarius
W9 Wido de Colchis
W10 Wido de Curte
W11 Wido de Tier
W12 Wido de la Rochia
W13 Wigo Meschins
W14 Wigo de Cunziaco
W15 Wilehmus de Maringis
W16 Wilelmus de Varenas
W17 Wilelmus de Valestinas
W18 Wilelmus de la Porta
W19 Willelmo de Sancto Albino
W20 Willelmus de Laval
W21 Walterius de Parriniaco

Y

Y1 Ylius de Chavasiget

Bibliographie

Abbé, Jean-Loup. "Arpenter et borner les terroirs de l'Europe méridionale au Moyen Âge : savoir et savoir-faire". In : *Annie Rousselle (éd.), Monde rural et histoire des sciences en Méditerranée. Du bon sens à la logique, Perpignan, Presses universitaires de Perpignan* (1998), p. 51-62.

— "Permanences et mutations des parcellaires médiévaux". In : *G. Chouquer, sd, Les formes du paysage* 2 (1996), p. 223.

Agerri, Rodrigo et German Rigau. "Robust multilingual Named Entity Recognition with shallow semi-supervised features". In : *Artif. Intell.* 238 (2016), p. 63-82.

Ambrosio, Antonella et al. *Digital diplomatics : the computer as a tool for the diplomatist ?* Böhlau, 2014.

Andrea K. Thomer and Nicholas M. Weber. "Using Named Entity Recognition as a Classification Heuristic". In : *iConference 2014 Proceedings*. 2014.

Anheim, Étienne et Pierre Chastang. "Les pratiques de l'écrit dans les sociétés médiévales (VIe-XIIIe siècle)". In : *Médiévales. Langues, Textes, Histoire* 56 (2009), p. 5-10.

Anis, Jacques. *Texte et ordinateur : les mutations du lire-écrire*. université de Paris X-Nanterre, 1993.

Ansani, Michele. "Edizione digitale di fonti diplomatiche : esperienze, modelli testuali, priorità". In : *Reti Medievali Rivista* 7.2 (2006), p. 1-1.

— *Le carte del monastero di Santa Maria di Morimondo, 2 vol (1010-1200)*. Spoleto, 1992.

Anton, Hans Hubert. *Pagus und Comitatus in Niederlothringen : Untersuchungen zur politischen Raumgliederung im früheren Mittelalter (Bonner Historische Forschungen, Bd. 49)*. 1986.

Ardanuy, Mariona Coll et Caroline Sporleder. "Clustering of novels represented as social networks". In : *LiLT (Linguistic Issues in Language Technology)* 12 (2015).

Atdag, Samet et Vincent Labatut. "A comparison of named entity recognition tools applied to biographical texts". In : *2nd International Conference on Systems and Computer Science*. 2013.

Atsma, Harmut et Jean Vezin. "Autour des actes privés du chartrier de Cluny (Xe-XIe siècles)". In : *Bibliothèque de l'Ecole des Chartes* 155.1 (1997), p. 470-471.

— "Gestion de la mémoire à l'époque de saint Hugues (1049-1109) : la genèse paléographique et codicologique du plus ancien cartulaire de l'abbaye de Cluny". In : *Histoire et archives* 7 (2000), p. 16-22.

— *Les responsables de la transcription des actes juridiques et les services de l'écriture au Xe siècle : l'exemple de Cluny*, p. 10-20.

Augenstein, Isabelle et al. "Generalisation in named entity recognition : A quantitative analysis". In : *Comput. Speech Lang.* 44 (2017), p. 61-83.

Aurell, Martin. "La parenté en l'an mil". In : *Cahiers de civilisation médiévale* vol. 43 (2000), p. 125-142.

Azpeitia, Andoni et al. “NERC-fr : Supervised Named Entity Recognition for French”. In : *Lecture Notes in Computer Science*. 2014, p. 158-165.

Azzam, Wagih et al. “Les manuscrits littéraires français : pour une sémiotique du recueil médiéval”. In : *Revue belge de philologie et d'histoire* 83.3 (2005), p. 639-669.

Bamman, David et Gregory Crane. “The Ancient Greek and Latin Dependency Treebanks”. In : *Language Technology for Cultural Heritage*. 2011, p. 79-98.

Bange, François. “L'ager et la villa : structures du paysage et du peuplement dans la région mâconnaise à la fin du Haut Moyen Age (IX e-XI e siècles)”. In : *Annales. Histoire, Sciences Sociales*. T. 39. 3. Cambridge University Press. 1984, p. 529-569.

Barbieri, Ezio et al. *La carte del monastero di San Pietro in Ciel d'Oro di PAvia*. Fontes, 1984.

Barret, Sébastien. *La mémoire et l'écrit : l'abbaye de Cluny et ses archives (Xe-XVIIIe siècle)*. T. 19. LIT Verlag Münster, 2004.

— “La mémoire et l'écrit : l'abbaye de Cluny et ses archives (Xe-XVIIIe siècle)”. In : *Bulletin du centre d'études médiévales d'Auxerre| BUCEMA* 13 (2009), p. 387-390.

— “Un avocat au service du Cabinet des chartes : les travaux de Louis-Henri Lambert de Barive dans les archives de Cluny (v. 1770-v. 1790)”. In : *Histoire et archives* 15 (2004), p. 29-64.

Barret, Sébastien et al. *Les plus anciens documents originaux de l'abbaye de Cluny, t. II : Documents nos 31 à 60*. 2000.

Barrière, Caroline. “Searching for Named Entities”. In : *Natural Language Understanding in a Semantic Web Context*. 2016, p. 23-38.

Barthélemy, Dominique. “La mutation féodale at-elle eu lieu ?(Note critique)”. In : *Annales. Histoire, Sciences Sociales*. T. 47. 3. Cambridge University Press. 1992, p. 767-777.

Bautier, Robert-Henri. “Caractères spécifiques des chartes médiévales”. In : *Publications de l'École Française de Rome* 31.1 (1977), p. 81-96.

— “Les demandes des historiens à l'informatique [La forme diplomatique et le contenu juridique des actes]”. In : *Publications de l'Ecole Française de Rome* 31.1 (1977), p. 179-186.

— “Olivier Guyotjeannin, Jacques Pycke et Benoît-Michel Tock.—Diplomatique médiévale. Turnhout, Brepols, 1992”. In : *Cahiers de civilisation médiévale* 38.152 (1995), p. 285-310.

Beck, Patrice et al. “Nommer au Moyen Âge : du surnom au patronyme”. In : *Le patronyme. Histoire, anthropologie, société* (2001), p. 13-38.

Bernard, Auguste. *Cartulaire de l'abbaye de Savigny : suivi du petit cartulaire de l'abbaye d'Ainay*. T. 2. Impr. impériale, 1853.

Birnbaum, David J. et al. “The Digital Middle Ages : An Introduction”. In : *Speculum* 92.S1 (2017), S1-S38.

Blanke, Tobias et Conny Kristel. “Integrating Holocaust Research”. In : *International Journal of Humanities and Arts Computing* 7.1-2 (2013), p. 41-57.

Blanke, Tobias et al. “Information Extraction on Noisy Texts for Historical Research”. In : *Digital Humanities* (2012).

Bollacker, Kurt et al. “Freebase : a collaboratively created graph database for structuring human knowledge”. In : *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM. 2008, p. 1247-1250.

Bornet, Cyril et Frédéric Kaplan. “A Simple Set of Rules for Characters and Place Recognition in French Novels”. In : *Frontiers in Digital Humanities* 4 (2017).

Boüard, Alain de. *Manuel de diplomatique française et pontificale*. Picard, 1948, p. 12-35.

Bourin, Monique. “Délimitation des parcelles et perception de l’espace en Bas-Languedoc aux 10è et 11è siècles”. In : *Campagnes médiévales : l’homme et son espace. Etudes offertes à Robert Fossier* (), p. 73-85.

Bourin, Monique et Pascal Chareille. *Genèse médiévale de l’anthroponymie moderne*. T. 2. Université de Tours, 1990.

Bourin, Monique et Elisabeth Zadora-Rio. “Pratiques de l’espace : les apports comparés des données textuelles et archéologiques”. In : *Actes des congrès de la Société des historiens médiévistes de l’enseignement supérieur public* 37.1 (2006), p. 39-55.

Bruel, Alexandre. “Note sur la transcription des actes privés dans les cartulaires antérieurement au XIIe siècle”. In : *Bibliothèque de l’École des chartes* 36 (1875), p. 445-456.

Bruel, Alexandre et Auguste Joseph Bernard. *Recueil des chartes de l’abbaye de Cluny : 802-954*. T. 49. Impr. Nat., 1876.

Brunner, Thomas. “Le passage aux langues vernaculaires dans les actes de la pratique en Occident”. In : *Le Moyen Age* 115.1 (2009), p. 29-72.

Budassi, Marco et Marco Passarotti. “Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon”. In : *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. 2016.

Burrows, John. “Textual Analysis.” In : *A Companion to Digital Humanities* (2004), p. 323-347.

Byrne, Kate. “Nested Named Entity Recognition in Historical Archive Text”. In : *International Conference on Semantic Computing (ICSC 2007)*. 2007.

Cai, Zhiyuan et al. “Wikification via link co-occurrence”. In : *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM. 2013, p. 1087-1096.

Canat de Chizy, M. *Origines du prieuré de Notre-Dame de Paray le Monial*. Saone-et-Loire, 1876.

Carreras, Xavier et al. “Named entity recognition for Catalan using Spanish resources”. In : *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - EACL ’03*. 2003.

Castagnetti, Andrea. *L’organizzazione del territorio rurale nel Medioevo : circoscrizioni ecclesiastiche e civili nella" Langobardia" e nella" Romania"*. T. 3. Pàtron, 1982.

Cavnar, William, John Trenkle et al. “N-gram-based text categorization”. In : *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*. T. 161175. Citeseer. 1994.

Chastang, Pierre. “Du locus au territorium. Quelques remarques sur l’évolution des catégories en usage dans le classement des cartulaires méridionaux au XIIe siècle”. In : *Annales du Midi*. T. 119. 260. Privat. 2007, p. 457-474.

Chaume, Maurice. *Les Origines du duché de Bourgogne : 2ème partie, Géographie historique*. E. Rebourseau, 1927.

Chevalier, Ulysse. *Cartulaire du prieuré de Paray-le-Monial, ordre de Saint-Benoît*. A. Picard, 1890.

Chizy, Marcel Canat de et Paul Canat de Chizy. *Cartulaire du prieuré de Saint Marcel lès-Châlon*. L. Marceau, 1894.

Chouquer, Gérard. “Aux origines antiques et médiévales des parcellaires”. In : *Histoire & sociétés rurales* 4 (1995), p. 11-46.

— “La forme juridique et cadastrale des actes “notariés” de Cluny en 870-935”. In : *www.formesdufoncier.org* ().

Chouquer, Gérard. “Une année d’exception pour l’archéogéographie”. In : *Études rurales* 173-174 (2005), p. 297-324.

Ciaramita, Massimiliano et Yasemin Altun. “Named-entity recognition in novel domains with external lexical knowledge”. In : *Proceedings of the NIPS Workshop on Advances in Structured Learning for Text and Speech Processing.* (2005).

Clanchy, Michael T. *From memory to written record : England 1066-1307.* John Wiley & Sons, 2012.

Cloppet, Florence et al. “ICDAR2017 Competition on the Classification of Medieval Handwritings in Latin Script”. In : *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR).* T. 1. IEEE. 2017, p. 1371-1376.

Contamine, Geneviève. “Traitement des textes diplomatiques : les problèmes de la lemmatisation”. In : *Publications de l’École Française de Rome* 31.1 (1977), p. 265-275.

Crehange, Marion et Lucie Fossier. “Essai d’exploitation sur ordinateur des sources diplomatiques médiévales”. In : *Annales* 25.1 (1970), p. 249-284.

Cucherat, François. *Cluny au onzième siècle : son influence religieuse, intellectuelle et politique.* Dejussieu, 1873.

Curran, James et Stephen Clark. “Language independent NER using a maximum entropy tagger”. In : *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003.* 2003, p. 164-167.

Cursente, Benoît. “Autour de Lézat : emboîtements, cospatialités, territoires (milieu X-milieu XIII siècle)”. In : *Les territoires du médiéviste. Rennes, Presses universitaires de Rennes* (2005), p. 151-167.

De Wilde, Max. “Semantic enrichment of a multilingual archive with linked open data”. In : *Digital Humanities Quarterly* (2017).

DeBats, Donald A. “A Tale of Two Cities : Using Tax Records to Develop GIS Files for Mapping and Understanding Nineteenth-Century U.S. Cities”. In : *Historical Methods : A Journal of Quantitative and Interdisciplinary History* 41.1 (2008), p. 17-38.

DeBats, Donald. A. et I. N. Gregory. “Introduction to Historical GIS and the Study of Urban History”. In : *Soc. Sci. Hist.* 35.4 (2011), p. 455-463.

Déléage, André. *La Vie économique et sociale de la Bourgogne dans le haut moyen âge : thèse pour le doctorat ès lettres présentée à la Faculté des lettres de l’Université de Paris, par André Déléage.* Protat frères, 1941.

Delisle, Léopold. *Le cabinet des manuscrits de la bibliothèque nationale : étude sur la formation de ce dépôt [...] avant l’invention de l’imprimerie.* T. 1. Imprimerie nationale, 1868, p. 563-564.

Denooz, Joseph. “L’ordinateur et le latin, Techniques et méthodes”. In : *Revue de l’Organisation Internationale pour l’Etude des Langues Anciennes par Ordinateur* (1978), p. 1-36.

Depauw, Mark et Tom Gheldof. “Trismegistos : An Interdisciplinary Platform for Ancient World Texts and Related Information”. In : *Communications in Computer and Information Science.* 2014, p. 40-52.

Drucker, Johanna. “Humanities approaches to graphical display”. In : *Digital Humanities Quarterly* 5.1 (2011), p. 1-21.

Dubois, Henri. “Population et fiscalité en Bourgogne à la fin du Moyen Âge”. In : *Comptes rendus des séances de l’Académie des Inscriptions et Belles-Lettres* 128.4 (1984), p. 540-555.

Duby, Georges. *La Societe aux Xle et Xlle siecles dans la region maconnaise.* A. Colin, 1953, p. 9-18.

— *Qu'est-ce que la société féodale ?* Flammarion, 2002, p. 1459-1465.

Ehrmann, Maud. et al. "Diachronic evaluation of NER systems on old newspapers". In : *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)* (2016), p. 97-107.

Ehrmann, Maud et al. "Building a multilingual named entity-annotated corpus using annotation projection". In : *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011* (2011), p. 118-124.

Elsebai, Ali et Farid Meziane. "Extracting person names from Arabic newspapers". In : *2011 International Conference on Innovations in Information Technology*. 2011.

Elson, David K et al. "Extracting social networks from literary fiction". In : *Association for Computational Linguistics. En Proceedings of the 48th annual meeting of the association for computational linguistics*. (2010), p. 138-147.

Erdmann, A. et al. "Challenges and solutions for Latin named entity recognition". In : *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)* (2016), p. 85-93.

Evert, Stefan. "The statistics of word cooccurrences". Thèse de doct. Dissertation, Stuttgart University, 2005.

Faruqui, Manaal et al. "Training and Evaluating a German Named Entity Recognizer with Semantic Generalization". In : *KONVENS* (2010), p. 129-133.

Favory, François et al. "Le territoire : un modèle de l'organisation de l'espace en archéologie rurale ; étude de cas dans la cité antique de Nîmes". In : *Habitat et société, actes des XIXe rencontres internationales d'archéologie et d'histoire d'Antibes* (1998), p. 499-518.

Fichtenau, Heinrich. *Arenga. Spätantike und Mittelalter im Spiegel von Urkundenformeln.* Mitteilungen des Institus für Ósterreichische Geschichtsforschung, Ergbd. XVIII, Böhlaus, 1957.

— "Das Urkundenwesen in Österreich vom 8. bis zum frühen 13". In : *Jahrhundert. MIÖG Ergbd* 23 (1971).

Filannino, Michele et Marilena Di Bari. "Gold standard vs. silver standard : the case of dependency parsing for Italian". In : *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015*.

Fitzgerald, Christina M. "Mapping the Medieval City : Space, Place and Identity in Chester c. 1200–1600". In : *J. Hist. Geogr.* 46 (2014), p. 133-134.

Fixot, Michel et Élisabeth Zadora-Rio. "L'environnement des églises et la topographie religieuse des campagnes médiévales". In : *Documents d'archéologie française* 46 (1994).

Foley, IV et J John. "Poetry : Identification, Entity Recognition, and Retrieval". In : (2019).

Fossier, Robert. *Cartulaire chronique du prieuré Saint-Georges d'Hesdin.* T. 32. Éditions du Centre national de la recherche scientifique, 1988.

— "La démographie médiévale : problèmes de méthode (Xe - XIIIe siècles)". In : *Annales de démographie historique*. JSTOR. 1975, p. 143-165.

— "Peuplement de la France du nord entre le Xe et le XVIE siècles". In : *Annales de démographie historique*. JSTOR. 1979, p. 59-99.

Gasse-Grandjean, Marie-José. "Les « Chartae Burgundiae Medii Aevi » (CBMA) et le numérique". In : *Francia* 40 (2011), p. 255-263.

Gazeau, Véronique. "Recherches autour de la datation des actes normands aux X e–XII e siècles". In : *Dating medieval undated charters* (2000), p. 61-79.

Geary, Patrick. "Entre gestion et gesta". In : *Ecole des chartes (éd), Les Cartulaires, Paris* (1993), p. 13-26.

Genet, Jean-Philippe. “L’informatique au service de la prosopographie : Prosop”. In : *Mélanges de l’École française de Rome* 100.1 (1988), p. 247-263.

Gervers, Michael. *Dating undated medieval charters*. Boydell & Brewer Ltd, 2002.

Gibbs, F et T Owens. *Building Better Digital Humanities Tools : Toward broader audiences and user-centered designs. Digital Humanit. Q. 6 (2)(2012)*.

Giry, Arthur. *Manuel de diplomatique : Diplomes et chartes.-Chronologie technique.-Éléments critiques et parties constitutives de la teneur des chartes.-Les chancelleries.-Les actes privés*. T. 1. Paris, Hachette, 1894.

Giuliano, Claudio. “Fine-grained classification of named entities exploiting latent semantic kernels”. In : *Proceedings of the Thirteenth Conference on Computational Natural Language Learning - CoNLL ’09*. 2009.

Gomaa, Wael H. et Aly Fahmy. “A survey of text similarity approaches”. In : *International Journal of Computer Applications* 68.13 (2013), p. 13-18.

Graesse, Johann Georg Theodor et Friedrich Benedict. *Orbis latinus : Lexikon lateinischer geographischer Namen des Mittelalters und der Neuzeit*. T. 1. Klinkhardt & Biermann, 1972.

Grésillon, Almuth. *Eléments de critique génétique. Lire les manuscrits modernes : Lire les manuscrits modernes*. Cnrs, 2016.

Grévin, Benoît. *Le Parchemin des cieux. Essai sur le Moyen Age du langage : Essai sur le Moyen Age du langage*. Le Seuil, 2013.

Grishman, Ralph et Beth Sundheim. “Message Understanding Conference-6”. In : *Proceedings of the 16th conference on Computational linguistics -*. 1996.

Grover, C., Givon, S., Tobin, R., & Ball, J. “Named Entity Recognition for Digitised Historical Texts”. In : *LREC*. (2008).

Grover, Claire et al. “Use of the Edinburgh geoparser for georeferencing digitized historical collections”. en. In : *Philos. Trans. A Math. Phys. Eng. Sci.* 368.1925 (août 2010), p. 3875-3889.

Guerreau-Jalabert, Anita. “Caritas y don en la sociedad medieval occidental”. In : *Hispania* 60 (2000), p. 27-62.

Guerreau, Alain. “Analyse factorielle et analyses statistiques classiques : le cas des ordres mendiants dans la France médiévale”. In : *Annales. Histoire, Sciences Sociales*. T. 36. 5. Cambridge University Press. 1981, p. 869-912.

— “Le champ sémantique de l’espace dans la vita de saint Maieul (Cluny, début du XIe siècle)”. In : *Journal des savants* 2.1 (1997), p. 363-419.

Guyotjeannin, Olivier. “La diplomatique médiévale et l’élargissement de son champ”. In : *La Gazette des archives* 172.1 (1996), p. 12-18.

Guyotjeannin, Olivier et al. *Diplomatique médiévale*. Brepols, 2006.

Guyotjeannin, Olivier et al. *Les cartulaires : actes de la Table ronde organisée par l’École nationale des chartes et le GDR 121 du CNRS*. T. 39. École nationale des chartes, 1993.

Hassner, Tal et al. “Digital Palaeography : New Machines and Old Texts (Dagstuhl Seminar 14302)”. In : *Dagstuhl Reports*. T. 4. 7. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2014.

Haug, Dag T T et Marius Jøhndal. “Creating a parallel treebank of the old Indo-European Bible translations”. In : *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)* (2008), p. 27-34.

Henriet, Patrick. “Chronique de quelques morts annoncées : Les saints abbés clunisiens (X e-XII e seècles)”. In : *Médiévales* (1996), p. 93-108.

Herren, Michael W. “Latin and the vernacular languages. Medieval Latin : An Introduction and Bibliographical Guide”. In : (1996), p. 122-130.

Hill, Mark J et Simon Hengchen. “Quantifying the impact of dirty OCR on historical text analysis : Eighteenth Century Collections Online as a case study”. In : *Digital Scholarship in the Humanities* (2019).

Hillebrandt, Maria et al. “À la recherche de personnes perdues...” In : *Médiévales* (1991), p. 21-25.

Hockey, Susan. *Electronic Texts in the Humanities : Principles and Practice*. en. OUP Oxford, nov. 2000.

Hoffart, Johannes et al. “Robust disambiguation of named entities in text”. In : *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2011, p. 782-792.

Houidaille, J. “Mortalité masculine dans les familles regnantes au Moyen Age”. In : *Population* 27 (1972), p. 1131-1134.

Hubert, Étienne. “Quelques considérations sur l’organisation de l’espace, la propriété foncière et la géographie du peuplement dans la vallée du Turano (IXe-XIIIe siècle)”. In : *Quelques considérations sur l’organisation de l’espace, la propriété foncière et la géographie du peuplement dans la vallée du Turano (IXe-XIIIe siècle)* (2000), p. 1000-1025.

Iogna-Prat, Dominique. “La confection des cartulaires et l’historiographie à Cluny (XIe–XIIe siecles)”. In : *Les cartulaires. Actes de la Table ronde, op. cit* (1993), p. 27-44.

— “La geste des origines dans l’historiographie clunisienne des XIe-XIIe siècles”. In : *Revue bénédictine* 102.1-2 (1992), p. 135-191.

Isaksen, Leif et al. “Pelagios and the emerging graph of ancient world data”. In : *Proceedings of the 2014 ACM conference on Web science - WebSci ’14*. 2014.

Joachims, Thorsten. “Text categorization with support vector machines : Learning with many relevant features”. In : *European conference on machine learning*. Springer. 1998, p. 137-142.

Kang, Ning et al. “Training text chunkers on a silver standard corpus : can silver replace gold ?” en. In : *BMC Bioinformatics* 13 (jan. 2012), p. 17.

Kennedy, Graeme. *An introduction to corpus linguistics*. Routledge, 2014, p. 201-230.

Kestemont, Mike et Jeroen De Gussem. “Integrated sequence tagging for medieval Latin using deep representation learning”. In : *arXiv preprint arXiv :1603.01597* (2016).

Kettunen, K. et al. “Old Content and Modern Tools-Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771-1910”. In : *arXiv preprint arXiv :1611.02839* (2016).

Koistinen, Mika et al. “How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine”. In : *Proc. of LTC* (2017), p. 279-283.

Korchagina, Natalia. “Building a Gold Standard for Temporal Entity Extraction from Medieval German Texts”. In : *Conference on Language Technologies & Digital Humanities Ljubljana* (2016).

Kripke, Saul. “Identity and necessity”. In : *Perspectives in the Philosophy of Language* (1971), p. 93-126.

Lafferty, John et al. “Conditional random fields : Probabilistic models for segmenting and labeling sequence data”. In : (2001).

Lauwers, Michel et Laurent Ripart. “Représentation et gestion de l’espace dans l’Occident médiéval”. In : *Actes du colloque «Rome et l’État moderne européen : une comparaison*

typologique». In : J.-P. Genêt (dir.), Rome et l'État moderne européen. Rome : Collection de l'École française de Rome. T. 377. 2007, p. 115-171.

Lavergne, Thomas et al. "Practical Very Large Scale CRFs". In : *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Uppsala, Sweden : Association for Computational Linguistics, 2010, p. 504-513. url : `http://www.aclweb.org/anthology/P10-1052`.

Lazer, David et al. "Computational social science". In : *Science* 323.5915 (2009), p. 721-723.

Lehmann, Jens et al. "DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia". In : *Semantic Web* 6.2 (2015), p. 167-195.

Mac Kim, Sunghwan et Steve Cassidy. "Finding names in trove : named entity recognition for Australian historical newspapers". In : *Proceedings of the Australasian Language Technology Association Workshop 2015* (2015), p. 57-65.

Magnani, Eliana. "Le don au moyen âge". In : *Revue du MAUSS* 1 (2002), p. 309-322.

— "Les CBMA en corpus structuré. Atelier 2. Le corpus hagiographique bourguignon. Débats et recherches. LaMOP-Sorbonne, 19 juin 2018". In : *Bulletin du centre d'études médiévales d'Auxerre/ BUCEMA* (2018).

— "Un corpus structuré et hétérogène de textes latins médiévaux (Bourgogne, Ve-XVe siècle)". In : *Bulletin du CERCOR-Centre Européen de recherches sur les congrégations et ordres religieux* 41 (2017), p. 59-65.

Mailloux, Anne. "Perception de l'espace chez les notaires de Lucques (VIIIe-IXe siècles)". In : (1997).

Manning, Christopher D. et Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. en. MIT Press, 1999.

Marchello-Nizia, Christiane. *Grammaticalisation et changement linguistique*. De Boeck-Duculot, 2006, p. 304.

Marrier, Martin. *Bibliotheca cluniacensis*. Sumptibus Roberti Fovet Via Iacobaea, sub insigni Temporis et Occasionis, 1915.

Maynard, D., Tablan, V., Ursu, C., Cunningham, H., & Wilks, Y. "Named entity recognition from diverse text types. En 2001. p. 257-274". In : *Recent Advances in Natural Language Processing 2001 Conference* (2001), p. 257-274.

Mazel, Florian. *Encore les "mauvaises coutumes…Considérations sur l'Église et la seigneurie à partir de quelques actes des cartulaires de Saint-Victor de Marseille"*. 2010.

— *La réforme" grégorienne" dans le Midi, milieu XIe-début XIIIe siècle*. 2013.

McEnery, Tony et Andrew Hardie. *Corpus linguistics : Method, theory and practice*. Cambridge University Press, 2011.

Mcgillivray, Barbara et al. "The Index Thomisticus Treebank Project : Annotation, Parsing and Valency Lexicon". In : *TAL, 2009* 50.2 (), p. 103-127.

McKenzie, Donald Francis. *Bibliography and the Sociology of Texts*. Cambridge University Press, 1999.

Meeks, Elijah et Karl Grossner. "ORBIS : An interactive scholarly work on the Roman world". In : *Journal of Digital Humanities* 1.3 (2012), p. 1-3.

Méhu, Didier. "Paix et communautés autour de l'abbaye de Cluny (Xe-XVe siècle)". Thèse de doct. Lyon 2, 1999, p. 17-41.

Menant, François. "L'anthroponymie du monde rural". In : *Publications de l'École Française de Rome* 226.1 (1996), p. 349-363.

Mikolov, Tomas et al. "Distributed representations of words and phrases and their compositionality". In : *Advances in neural information processing systems*. 2013, p. 3111-3119.

Morelle, Laurent. “Instrumentation et travail de l’acte : quelques réflexions sur l’écrit diplomatique en milieu monastique au xie siècle”. In : *Médiévales. Langues, Textes, Histoire* 56 (2009), p. 41-74.

— “Pratiques médiévales de l’écrit documentaire”. In : *Annuaire de l’École pratique des hautes études (EPHE), Section des sciences historiques et philologiques. Résumés des conférences et travaux* 139 (2008), p. 368-371.

Mostern, R. et I. Johnson. “From named place to naming event : creating gazetteers for history”. In : *Int. J. Geogr. Inf. Sci.* 22.10 (2008), p. 1091-1108.

Nadeau, David et Satoshi Sekine. “A survey of named entity recognition and classification”. In : *Benjamins Current Topics*. 2009, p. 3-28.

Neiske, Franz. “Les débuts du prieuré clunisien de Paray-le-Monial”. In : *Paray-le-Monial actes du colloque* (1992), p. 1-12.

Nothman, Joel et al. “Learning multilingual named entity recognition from Wikipedia”. In : *Artif. Intell.* 194 (2013), p. 151-175.

Nouvel, Damien et al. *Named Entities for Computational Linguistics*. 2016.

O’Keeffe, Anne et Michael McCarthy. *The Routledge handbook of corpus linguistics*. Routledge, 2010, p. 345-359.

Ortí, M. Milagros Cárcel. *Vocabulaire international de la diplomatique*. T. 28. Universitat de València, 1997.

Oursel-Quarre, Madeleine. “A propos du chartrier de Cluny”. In : *Annales de Bourgogne Dijon*. T. 50. 198. 1978, p. 103-107.

Overell, Simon et Stefan Rüger. “Using co-occurrence models for placename disambiguation”. In : *International Journal of Geographical Information Science* 22.3 (2008), p. 265-287.

Packer, Thomas L et al. “Extracting person names from diverse and noisy OCR text”. In : *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*. ACM. 2010, p. 19-26.

Palmer, David D et David S Day. “A statistical profile of the named entity task”. In : *Fifth Conference on Applied Natural Language Processing*. 1997.

Panfili, Didier. *L’évolution des repères spatiaux en Bas-Quercy et Haut-Toulousain de 930 à 1130 : une approche des transformations sociales et des paysages agraires*. 2004.

Passarotti, Marco. “From Syntax to Semantics. First Steps Towards Tectogrammatical Annotation of Latin”. In : *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and humanities (LaTeCH)*. 2014, p. 100-109.

Passmann, Johannes et Asher Boersma. “Unknowing algorithms : On transparency of unopenable black boxes”. In : (2017).

Patel, Chirag et al. “Optical Character Recognition by Open source OCR Tool Tesseract : A Case Study”. In : *Int. J. Comput. Appl. Technol.* 55.10 (2012), p. 50-56.

Perreaux, Nicolas. “L’écriture du monde (I).. Les chartes et les édifices comme vecteurs de la dynamique sociale dans l’Europe médiévale (viie-milieu du xive siècle)”. In : *Bulletin du centre d’études médiévales d’Auxerre| BUCEMA* 19.2 (2015).

Pettersson, Eva et al. “Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting”. In : *Proceedings of the 19th Nordic conference of computational linguistics (Nodalida 2013)*. 2013, p. 163-179.

Peytremann, Édith. “Les structures d’habitat rural du haut Moyen Age en France (Ve-Xe s.). Un etat de la recherche”’. In : *Lorren and Perin (eds), l-28* (1995).

Pierazzo, Elena. “A rationale of digital documentary editions”. In : *Literary and linguistic computing* 26.4 (2011), p. 463-477.

Porro, Giulio. *Codex diplomaticus Longobardiae, Augustae Taurinorum.* Regio Typographeo, 1873.

Porter, Dot. "Medievalists and the scholarly digital edition". In : *Scholarly Editing* 34 (2013), p. 1-26.

Portet, Pierre. "Les techniques du bornage au moyen âge : de la pratique à la théorie". In : *Sfruttamento tutela e valorizzazione del territorio. dal diritto romano alla regolamentazione europea e internazionale.* T. 18. Jovene, Napoli. 2007, p-195.

Prévost, Sophie. "Corpus informatisés de français médiéval : contraintes sur leur constitution et spécificités de leurs apports". In : *Corpus* 7 (2008), p. 35-64.

Price, Kenneth. "Edition, project, database, archive, thematic research collection : What's in a name ?" In : *Faculty Publications–Department of English* (2009), p. 69.

Ragut, Camille et Théodore Chavot. *Cartulaire de Saint-Vincent de Mâcon : connu sous le nom de Livre enchaîné.* Impr. d'É. Protat, 1864.

Ramsay, S. "Special Section : Reconceiving Text Analysis : Toward an Algorithmic Criticism". In : *Literary and Linguistic Computing* 18.2 (2003), p. 167-174.

Ramshaw, L. A. et M. P. Marcus. "Text Chunking Using Transformation-Based Learning". In : *Text, Speech and Language Technology.* 1999, p. 157-176.

Reveyron, Nicolas. "Marcigny, Paray-le-Monial et la question de la chapelle mariale dans l'organisation spatiale des prieurés clunisiens au XIe–XIIe siècle". In : *Viator (English and Multilingual Edition)* 41 (2010), p. 63-94.

Al-Rfou, Rami et al. "Polyglot : Distributed word representations for multilingual nlp". In : *arXiv preprint arXiv :1307.1662* (2013).

Richard, Jean. "La publication des chartes de Cluny". In : *A Cluni : congrès* (1950).

— *Le cartulaire de Marcigny-sur-Loire, 1045-1144 : essai de reconstitution d'un manuscrit disparu.* Sociéte des Analecta Burgundica, 1957.

Richards, Julian. "Text Mining in Archaeology : Extracting Information from Archaeological Reports". In : *Mathematics and Archaeology.* 2015, p. 240-254.

Rigault, Jean. *Dictionnaire topographique du département de Saône-et-Loire : comprenant les noms de lieux anciens et modernes.* T. 38. Comité des travaux historiques et scientifiques-CTHS, 2008.

Rio, Alice. "Les formulaires et la pratique de l'écrit dans les actes de la vie quotidienne (vie-xe siècle)". In : *Médiévales. Langues, Textes, Histoire* 56 (2009), p. 11-22.

Ripley, Brian D. *Pattern recognition and neural networks.* Cambridge university press, 2007, p. 354-360.

Ritter, Alan et al. "Named entity recognition in tweets : an experimental study". In : *Proceedings of the conference on empirical methods in natural language processing.* Association for Computational Linguistics. 2011, p. 1524-1534.

Rosé, Isabelle. "Panorama de l'écrit diplomatique en Bourgogne : autour des cartulaires (XIe-XVIIIe siècles)". In : *Bulletin du centre d'études médiévales d'Auxerre| BUCEMA* 11 (2007).

Rosenwein, Barbara. *Cluny's immunities in the tenth and eleventh centuries : images and narratives.* Lit, 1998.

— *To be the neighbor of Saint Peter : the social meaning of Cluny's property, 909-1049.* Cornell University Press, 2006.

Sainte-Marthe, Denis de et Barthélemy Hauréau. *Gallia Christiana in provincias ecclesiasticas distributa.* T. 14. Coignard, 1751.

Santos, Cicero dos et Victor Guimarães. "Boosting Named Entity Recognition with Neural Character Embeddings". In : *Proceedings of the Fifth Named Entity Workshop.* 2015.

Sapin, Christian. “Saint-Marcel-lès-Chalon (Saône-et-Loire), église Saint-Marcel”. In : *Bulletin du centre d'études médiévales d'Auxerre| BUCEMA* 10 (2006).

Schmid, Helmut. “Treetagger| a language independent part-of-speech tagger”. In : *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart* 43 (1995), p. 28.

Schnapp, Jeffrey et al. “Digital humanities manifesto 2.0”. In : *Hentet* 10 (2009), p. 2016.

Schneider, Laurent. “Du pagus aux finages castraux, les mots des territoires dans l'espace oriental de l'ancienne Septimanie (IXe-XIIe siècle)”. In : *Les territoires du médiéviste* (2005), p. 109.

Schreibman, Susan et al. *A Companion to Digital Humanities*. John Wiley & Sons, 2008.

Schreibman, Susan et al. *A new companion to digital humanities*. John Wiley & Sons, 2015.

Sekine, Satoshi et Chikashi Nobata. “Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy”. In : *LREC* (2004), p. 1977-1980.

Sickel, Theodor. *Beiträge zur Diplomatik I-VIII*. Georg Olms Verlag, 1975.

Siemens, Ray et al. “Toward modeling the social edition : An approach to understanding the electronic scholarly edition in the context of new and emerging social media”. In : *Literary and Linguistic Computing* 27.4 (2012), p. 445-461.

Simon, Eszter. “Approaches to hungarian named entity recognition”. In : (2013).

Simon, Pierre. *Bullarium sacri ordinis cluniacensis*. Lyon : Antonium Jullieron, 1680.

Simon, Rainer. “Towards semi-automatic annotation of toponyms on old maps”. In : *e-Perimetron* 9.3 (2014), p. 105-128.

Smith, David A et Gregory Crane. “Disambiguating Geographic Names in a Historical Digital Library”. In : *Lecture Notes in Computer Science*. 2001, p. 127-136.

Spence, Paul. “La investigación humanística en la era digital : mundo académico y nuevos públicos”. In : *Humanidades Digitales : una aproximación transdisciplinar*. SIELAE. 2014, p. 117-131.

Steinberger, Ralf et Bruno Pouliquen. “Cross-lingual Named Entity Recognition”. In : *Benjamins Current Topics*. 2009, p. 137-164.

Suchanek, Fabian M. et al. “Yago : a core of semantic knowledge”. In : *Proceedings of the 16th international conference on World Wide Web*. ACM. 2007, p. 697-706.

Tessier, Georges. *La diplomatique (3 e éd.)* 1966.

Tock, Benoît-Michel. “L'acte privé en France, VIIe siècle-milieu du Xe siècle”. In : *Mélanges de l'école française de Rome* 111.2 (1999), p. 499-537.

Toubert, Pierre. *Les structures du Latium médiéval : le Latium méridional et la Sabine du IXe siècle à la fin du XIIe siècle*. École française de Rome, 1973.

Tsai, Chen-Tse et al. “Cross-Lingual Named Entity Recognition via Wikification”. In : *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. 2016.

Van Uytfanghe, Marc. *Le latin et les langues vernaculaires au Moyen Âge : un aperçu panoramique*. na, 2003, p. 2-18.

Vicente, Montserrat Rangel. “La glose comme outil de désambiguïsation référentielle des noms propres purs”. In : *Corela. Cognition, représentation, langage* HS-2 (2005).

Vigneron, Bernard. “La vente dans le Mâconnais du IX e au XIII e siècle”. In : *Revue historique de droit français et étranger (1922-)* 36 (1959), p. 17-47.

Wallach, Hanna M. “Conditional random fields : An introduction”. In : *Technical Reports (CIS)* (2004), p. 22.

Wollasch, Joachim. “Parenté noble et monachisme réformateur. Observations sur les «conversions» à la vie monastique aux XI e et XII e siècles”. In : *Revue historique* 264.Fasc. 1 (535 (1980), p. 3-24.

Won, Miguel et al. “ensemble named entity recognition (ner) : evaluating ner Tools in the identification of Place names in historical corpora”. In : *Frontiers in Digital Humanities* 5 (2018), p. 2.

Yadav, Vikas et Steven Bethard. “A Survey on Recent Advances in Named Entity Recognition from Deep Learning models”. In : *Proceedings of the 27th International Conference on Computational Linguistics.* (2018), p. 2145-2158.

Yujian, Li et Liu Bo. “A normalized Levenshtein distance metric”. In : *IEEE transactions on pattern analysis and machine intelligence* 29.6 (2007), p. 1091-1095.

Zeumer, Karl. *Monumenta Germaniae historica : Formulae Merowingici et Karolini aevi accedunt ordines iudiciorum Dei*. Impensis Bibliopolii Hahniani, 1886.

Zhang, Boliang et al. “Name Tagging for Low-resource Incident Languages based on Expectation-driven Learning”. In : *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*. 2016.

Zimmermann, Michel. “Glose, tautologie ou inventaire ? L’énumération descriptive dans la documentation catalane du Xe au XIIe siècle”. In : *Cahiers d’Études Hispaniques Médiévales* 14.1 (1989), p. 309-338.

— “Protocoles et Préambules dans les documents Catalans du Xe au XIIe siècle : évolution diplomatique et signification spirituelle I Les protocoles”. In : *Mélanges de la Casa de Velázquez* 10.1 (1974), p. 41-76.

Titre : Un modèle de reconnaissance automatique des entités nommées et des structures textuelles pour les corpus diplomatiques médiolatins.

Mots clés : reconnaissance des entités nommées, apprentissage automatique, TAL, humanités numériques, latin médiéval, diplomatique numérique

Résumé : Nous présentons dans cette thèse deux modèles informatiques développés pour délivrer de l'information structurée et applicables à de grandes bases de données de textes médiévaux. Les deux modèles, l'un appliqué à la reconnaissance des entités nommées, l'autre à la détection des parties du discours diplomatique, ont suivi un apprentissage supervisé utilisant la méthode des Champs aléatoires conditionnelles (CRF) sur un corpus manuellement annoté de actes médiévaux (*Corpus Burgundiae Medii Aevi* ou CBMA).
Notre modèle principal de reconnaissance d'entités nommées a prouvé sa robustesse lorsqu'il a été appliqué sur des échantillons de corpus de taille, chronologie et origine très variés. Le modèle secondaire détectant les parties du discours diplomatique, bien que moins performant, s'est montré valide comme outil de structuration. Ils peuvent à présent être utilisés pour l'indexation et l'étude d'une grande variété de sources diplomatiques.
Nous avons développé différentes solutions destinées à trouver un juste équilibre entre la dépendance du modèle à son corpus d'origine et sa capacité à être appliqué à d'autres corpus. De même, différents ajouts et corrections ont été opérés sur le corpus de référence à partir de plusieurs observations de type historique et linguistique concernant les documents utilisés, ce qui a permis d'améliorer la performance initiale.
Nous avons ensuite appliqué les outils ainsi générés à la reconnaissance de noms de personnes, de lieux et de parties du discours diplomatique sur des milliers d'actes du CBMA afin d'étudier différentes questions intéressant la science historique et la diplomatique. Ces études concernent la datation semi-automatique d'un cartulaire qui en était dépourvu ; l'évolution du vocabulaire spatial dans les actes du Moyen Âge Central; et l'indexation des documents à partir des modules les intégrant, notamment les formules du protocole des actes. Par ces études nous poursuivons un double objectif: illustrer différentes stratégies permettant d'abstraire et d'adapter au traitement automatique des données des méthodes de recherche classiques en Histoire ; démontrer que nos outils de traitement massif permettent la génération de connaissances pertinentes pour la science historique.

Title : A model for automatic named entities recognition and textual structures for Latin medieval diplomatic corpora.

Keywords : named entities recognition, NLP, medieval latin, digital humanities, digital diplomatics

Abstract : In this thesis, we present two computer models to structure textual information for large databases of medieval charters. The two models, one applied to the recognition of named entities, the other to the detection of parts of the diplomatics discourse, are supervised Conditional random fields (CRF) models trained on a hand-annotated corpus of medieval charters. (*Corpus Burgundiae Medii Aevi* or CBMA).
The main Named Entity Recognition model has proven to be robust in its application to widely varying corpora in size, chronology and origin. The secondary model detecting parts of the diplomatic discourse, although less efficient, remains valid as a structuring tool. At the moment both can be used for indexing and studying a wide variety of diplomatics sources, thus saving huge human efforts.
We have developed different solutions to overcome the gap between model's dependence on its original training-set and its ability to be applied to other corpora. Similarly, various corrections and additions were made to the golden-corpus from several historical and linguistic analysis concerning writing phenomena in charters, which greatly helped to improve the initial performance.
In a later step we applied our automatic tools in the recognition of names of people, places and parts of the diplomatics discourse on thousands of charters from the CBMA corpus in order to study different questions concerning historical science and diplomatics. These studies concern the semi-automatic dating of a non-dated cartulary; the evolution of the spatial vocabulary in the charters of the central Middle Ages and the indexing of charters from their scriptural modules, in particular formulae of the charter protocols. This studies has a twofold purpose: on the one hand have shown different strategies for abstracting and adapting to the automatic processing well-known methods of research in history; on the other hand, seek to provide us tools with an applicative framework to obtain relevant knowledge to the historical science using massive processing.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France