De l'accumulation à l'exploitation ? Expériences et propositions pour l'indexation et l'utilisation des bases de données diplomatiques

Nicolas Perreaux

Digital diplomatics

The computer as a tool for diplomatist?

Edited by

ANTONELLA AMBROSIO, SÉBASTIEN BARRET,
GEORG VOGELER

ELEKTRONISCHER SONDERDRUCK

2014

BÖHLAU VERLAG KÖLN · WEIMAR · WIEN

Table of Content

Preface 7

Antonella Ambrosio, Sébastien Barret, Georg Vogeler
Digital Diplomatics. Expertise between computer science and diplomatics 9

I. Technical and theoretical models

Benoît-Michel Tock
La diplomatique numérique, une diplomatique magique?..... 15

Camille Desenclos, Vincent Jolivet
Diple, propositions pour la convergence de schémas XML/TEI dédiés à l'édition de sources diplomatiques 23

Francesca Capochiani, Chiara Leoni, Roberto Rosselli Del Turco
Codifica, pubblicazione e interrogazione sul web di *corpora* diplomatici per mezzo di strumenti *open source* 31

Serena Falletta
Dalla carta al bit. Note metodologiche sull'edizione digitale di un cartulario medievale 61

Gunter Vasold
Progressive Editionen als multidimensionale Informationsräume 75

Luciana Duranti
The return of diplomatics as a forensic discipline 89

II. Projects for the edition of texts and the publication of information

Daniel Piñol Alabart
Proyecto ARQUIBANC – Digitalización de archivos privados catalanes. Una herramienta para la investigación 99

Antonella Ghignoli
Sources and persons of public power in 7th–11th-century Italy. The idea of *Italia Regia* and the *Italia Regia* project 109

Richard Higgins
The Repository view. Opening up medieval charters 123

Žarko Vujošević, Nebojša Porčić, Dragić M. Živojinović
Das serbische Kanzleiwesen. Die Herausforderung der digitalen Diplomatik ... 133

Aleksandrs Ivanovs, Aleksey Varfolomeyev
Some approaches to the semantic publication of charter corpora.
The case of the diplomatic edition of Old Russian charters 149

III. Digital diplomatics in the work of the historian

Els De Paermentier
Diplomata Belgica. Analysing medieval charter texts (dictamen) through a quantitative approach. The case of Flanders and Hainaut (1191–1244) 169

Nicolas Perreaux
De l'accumulation à l'exploitation? Expériences et propositions pour l'indexation et l'utilisation des bases de données diplomatiques 187

Gelila Tilahun, Michael Gervers, Andrey Feuerverger
Statistical methods for applying chronology to undated English medieval documents 211

Michael Hänchen
Neue Perspektiven für die Memorialforschung. Die datenbankgestützte Erschließung digitaler Urkundencorpora am Beispiel der Bestände von Aldersbach und Fürstenzell im 14. Jahrhundert 225

Martin Roland
Illuminierte Urkunden im digitalen Zeitalter. Maßregeln und Chancen 245

Dominique Stutzmann
Conjuger diplomatique, paléographie et édition électronique. Les mutations du XII^{e} siècle et la datation des écritures par le profil scribal collectif 271

Jonathan Jarrett
Poor tools to think with. The human space in digital diplomatics 291

Appendices

Abstracts 303
Colour plates 321
Glossary of technical terms 337
The Authors 345

De l'accumulation à l'exploitation? Expériences et propositions pour l'indexation et l'utilisation des bases de données diplomatiques[1]

NICOLAS PERREAUX

«Indépendamment même de toute éventualité d'application à la conduite, l'histoire n'aura donc le droit de revendiquer sa place parmi les connaissances vraiment dignes d'effort, seulement dans la mesure où, au lieu d'une simple énumération, sans liens et quasiment sans limites, elle nous promettra un classement rationnel et une progressive intelligibilité.»
MARC BLOCH, *Apologie pour l'histoire ou métier d'historien* (1943)

Depuis maintenant plusieurs décennies, les diplomatistes disposent de bases de données numérisées remarquables, dont le contenu est désormais propre à révolutionner nos connaissances concernant le Moyen Âge et ceci à plus ou moins court terme. Ainsi, sans même évoquer Google Books et ses milliers d'éditions d'actes diplomatiques numérisés, les médiévistes sont de plus en plus nombreux à connaître l'existence de sites tels que celui des *Chartes originales* de l'Artem[2],

[1] De nombreuses pistes et réflexions proposées dans cet article doivent beaucoup à des échanges avec Alain Guerreau: qu'il en soit tout d'abord chaleureusement remercié. Nous tenions aussi à remercier ceux qui ont contribué à sa structuration, soit par une relecture et des conseils circonstanciés – Eliana Magnani, notre directrice de thèse avec Daniel Russo, Marie-José Gasse-Grandjean et Sébastien Barret –, soit par des discussions, à la suite de notre intervention à Naples – Georg Vogeler, Michael Gervers, Benoît-Michel Tock, Dominique Stutzmann, Olivier Canteaut et Frédéric Glorieux.

[2] B.-M. Tock (dir.)/M. Courtois/M.-J. Gasse-Grandjean/P. Demonty, La diplomatique française du Haut Moyen Âge: inventaire des chartes originales antérieures à 1121 conservées en France, (1) «Introduction générale, album diplomatique, table chronologique, table des auteurs», (2) «Table des destinataires, table des genres diplomatiques, table des états de la tradition manuscrite, table des sceaux, table des chirographes, table des cotes d'archives ou de bibliothèques» (2001). En dernier lieu, pour une présentation globale du *corpus*, voir l'article introductif de B.-M. Tock, La diplomatique française du Haut Moyen Âge vue à travers les originaux, ibid. p. 1–37. Depuis 2011, en dehors des reproductions photographiques, le *corpus* est désormais entièrement consultable en ligne: <http://www.cn-telma.fr/originaux/colophon/>.

Monasterium[3], le *Deeds Project*[4], les *Chartae Burgundiae Medii Aevii*[5], le

[3] Monasterium.net est sans aucun doute l'un des sites hébergeant des documents diplomatiques qui génère le plus de réflexions abstraites, méthodologiques, mais aussi techniques. Voir entre autres: Th. Aigner, MOnasteriuM –Die mittelalterlichen Urkunden der Klöster des Landes Niederösterreich (A) im Internet (www.mom.archiv.net), in: Archivpflege in Westfalen und Lippe 58 (2003) p. 43–44; id., Digitale Bereitstellung historischer Quellen aus Ordenstiftsarchiven, in: Der Archivar 8 (2003) p. 295–306; K. Heinz, Monasterium.Net – Das virtuelle Urkundenarchiv niederösterreichischer Klöster, in: Österreich in Geschichte und Literatur 49 (2005) p. 48–49; id., Monasterium.net – Auf dem Weg zu einem europäischen Urkundenportal, in: Regionale Urkundenbücher. Die Vorträge der 12. Tagung der Commission Internationale de Diplomatique [Sankt Pölten, 2009], éd. T. Kölzer/W. Rösener/R. Zehetmayer (2010) p. 139–145; A. Krah, Monasterium.net – das virtuelle Urkundenarchiv Europas. Möglichkeiten der Bereitstellung und Erschliesung von Urkundenbeständen, in: Archivalische Zeitschrift 91 (2009) p. 221–246; id., Monasterium.net: Auf dem Weg zu einem mitteleuropäischen Urkundenportal, in: Digitale Diplomatik. Neue Technologien in der historischen Arbeit mit Urkunden, dir. G. Vogeler (AfD Beiheft 12, 2009) p. 70–77. Le site est disponible à l'adresse suivante: <http://www.monasterium.net/>

[4] L'origine du projet remonte à 1975 et la base actuelle contient environ 10 000 documents. Présentation du site et expériences dans: M. Gervers, The DEEDS Project and the Development of a Computerized Methodology for Dating Undated English Private Charters of the Twelfth and Thirteenth Centuries, in: Dating Undated Medieval Charters, éd. M. Gervers (2000) p. 13–36; id., The dating of medieval English private charters of the twelfth and thirteenth centuries, in: A Distinct Voice. Medieval Studies in Honour of Leonard E. Boyle, O. P., dir. J. Brown/ W. P. Stoneman (1997) p. 455–505; M. Gervers/M. Margolin, Managing Meta-data in a Research Collection of Medieval Latin Charters, in: Digitale Diplomatik. Neue Technologien p. 271–282. En ligne: <http://deeds.library.utoronto.ca/>

[5] La base des CBMA est sans doute celle qui propose, à l'heure actuelle et avec la base des originaux, le plus d'expériences/d'études concrètes. Voir en particulier I. Rosé, À propos des *Chartæ Burgundiæ Medii Ævii* (CBMA). Éléments de réflexion à partir d'une enquête sur la dîme en Bourgogne au Moyen Âge, in: Bulletin du Centre d'études médiévales d'Auxerre (Collection CBMA. Les cartulaires, Études, 2008), version en ligne <http://cem.revues.org/index8412.html>; E. Magnani, L'échange dans la documentation diplomatique bourguignonne: autour d'un champ sémantique, in: L'acte d'échange, du VIII^e^ au XII^e^ siècle/Tauschgeschäft und Tauschurkunde vom 8. bis zum 12. Jh., Limoges 11–13 mars 2010, dir. Ph. Depreux/I. Fees (AfD Beiheft 13, 2013), p. 403–426); ead., Les moines et la mise en registre des transferts. Formules textuelles, formules visuelles, in: Cluny, le monachisme et la société au premier âge féodal (880–1050), éd. D. Iogna-Prat et al. (2013) p. 199–214; N. Perreaux, L'eau, l'écrit et la société (IX^e^-XII^e^ siècle). Étude statistique sur les champs sémantiques dans les bases de données [CBMA et autres], in: Bulletin du Centre d'études médiévales d'Auxerre 15 (2011) p. 439–449, version en ligne: <http://cem.revues.org/index12062.html>; id., La production de l'écrit diplomatique en Bourgogne. Hypothèses sur les dynamiques sociales inégales et les aires de scripturalité (IX^e^-XIII^e^ siècle) au regard des bases des données, in: Productions, remplois, mises en registre: la pratique sociale de l'écrit à travers

Codice Diplomatico della Lombardia medievale[6],

---

la documentation médiévale bourguignonne Auxerre, Abbaye Saint-Germain, 24 et 25 septembre 2009, dir. E. MAGNANI, à paraître en 2013; ID., Dynamique sociale et écriture documentaire (Cluny, X^e-XII^e siècle). Observations statistiques sur le champ sémantique de l'eau, in: Cluny, le monachisme et la société, p. 111–127.
Présentation de la base en elle-même (plus de 11 000 documents), in: E. MAGNANI/M.-J. GASSE-GRANDJEAN, *Chartae Burgundiae Medii Aevi* (CBMA), in: Bulletin du Centre d'études médiévales d'Auxerre 9 (2005), p. 179–181; EAED., CBMA – *Chartae Burgundiae Medii Aevi*.1. Les fonds diplomatiques bourguignons, in: ibid. 11 (2007) p. 163–169; EAED., CBMA. – *Chartae Burgundiae Medii Aevi*. 2. Cartulaires, éditions, base de données, in: ibid. 12 (2008) p. 237–244; EAED., CBMA – *Chartae Burgundiae Medii Aevi*. 3. Systèmes d'interrogation et recherches sur les fonds diplomatiques bourguignons, in: ibid. 13 (2009) p. 245–251; EAED., CBMA. – *Chartae Burgundiae Medii Aevi*. 4. Études, éditions, historiographie, in: ibid. 14 (2010) p. 197–209; EAED., CBMA. – *Chartae Burgundiae Medii Aevi.* 5. Actes cisterciens et prémontrés, in: ibid. 14 (2010) p. 273–275, version en ligne <http://www.artehis-cnrs.fr/CBMA-Chartae-Burgundiae-Medii-Aevi,964> et <http://www.artehis-cbma.eu/> pour le site de la base de données.
Pour le corpus des originaux de l'Artem, une liste d'expériences concrètes avait déjà été donnée par B.-M. TOCK, L'apport des bases de données de chartes pour la recherche des mots et des formules, in: Digitale Diplomatik. Neue Technologien, p. 283–293, à la p. 284. Quelques jalons importants: M. PARISSE, À propos du traitement automatique des chartes: chronologie du vocabulaire et repérage des actes suspects, in: La lexicographie du latin médiéval et ses rapports avec les recherches actuelles sur la civilisation du Moyen Âge (Actes du colloque de Paris, 1978) (1981) p. 241–249; ID., Premiers résultats d'un traitement automatique des chartes, in: Le Moyen Âge 2 (1978) p. 337–343; ID., Remarques sur les chirographes et les chartes-parties antérieurs à 1120 et conservés en France, in: AfD 32 (1986) p. 546–568; ID., Croix autographes et souscriptions dans l'Ouest de la France au XI^e siècle, in: Graphische Symbole in mittelalterlichen Urkunden. Beiträge zur diplomatischen Semiotik, éd. P. RÜCK (Historische Hilfswissenschaften 3, 1996) p. 143–155; ID., Écriture et réécriture des chartes: les pancartes aux XI^e et XII^e siècles, in: Pratiques de l'écrit documentaire au XI^e siècle, éd. O. GUYOTJEANNIN/L. MORELLE/M. PARISSE (in: BECh 155 [1997] p. 5–347]) p. 247–265; ID., Les pancartes. Étude d'un type diplomatique, in: Pancartes monastiques des XI^e et XII^e siècles, éd. M. PARISSE/P. PÉGEOT/B.-M. TOCK (1998) p. 11–62; B.-M. TOCK, La diplomatique sans pancarte. L'exemple du diocèse d'Arras et de Thérouanne, 1000–1120, in: Pancartes monastiques p. 131–157; ID., *Altare* dans les chartes françaises antérieures à 1121, in: Roma, Magistra Mundi. Itineraria culturae medievalis. Mélanges offerts au Père L. E. Boyle à l'occasion de son 75^e anniversaire (2), éd. J. HAMESSE (1998) p. 901–926; ID., L'étude du vocabulaire et la datation des actes: l'apport des bases de données informatisées, in: Dating Undated Medieval Charters p. 81–96; ID. Scribes, souscripteurs et témoins dans les actes privés en France (VII^e-début du XII^e siècle) (ARTEM 9, 2005) (ne porte pas exclusivement sur la base de l'Artem); ID., Les actes entre particuliers en Bourgogne méridionale (IX^e-XI^e siècles), in: Die Privaturkunden der Karolingerzeit, éd. P. ERHART/K. HEIDECKER/B. ZELLER (2009) p. 121–134.

[6] La base actuelle contient environ 5 000 documents. Présentation complète du projet à l'adresse suivante: <http://cdlm.unipv.it/progetto/>. Voir aussi M. ANSANI, Il

les dMGH[7], ou encore les *Cartulaires numérisés d'Île de France*[8], etc. Des corpus remarquables, dont le dénominateur commun est d'avoir choisi l'option du libre accès: un point essentiel sur lequel il est parfois encore nécessaire d'insister[9].

Pourtant, malgré cette profusion, ainsi que le faisait encore récemment remarquer Benoît-Michel Tock[10], force est de constater que l'exploitation de ces vastes corpus – ou plutôt devrions-nous dire de *ce* vaste corpus – reste encore largement à faire, les entreprises dans le domaine restant pour le moment embryonnaires, ceci malgré la bonne volonté et l'intérêt notable (mais aussi, il faut bien l'admettre, ponctuel) des médiévistes pour les nouvelles technologies. Souvent rebutés par la difficile gestion d'une telle masse documentaire, les chercheurs qui tentent l'expérience des bases de données retournent souvent à des méthodes plus traditionnelles, employant certes, en parallèle, ces corpus, mais presque toujours comme de simples carrières de données. Il ne s'agit pas d'un phénomène nouveau: le *Thesaurus Diplomaticus* aura en effet bientôt 15 ans et les résultats de son exploitation systématique sont pour le moment largement en deçà de ce que nous aurions pu attendre d'un tel «trésor»[11]. Le constat est donc simple: il existe un décalage important entre la

---

Codice diplomatico digitale della Lombardia medievale, in: Comuni e memoria storica: alle origini del comune di Genova; atti del Convegno di studi, Genova, 24–26 settembre 2001, dir. D. Puncuh (2002) p. 23–49; M. Ansani/V. Leoni, Experiment einer digitalen Edition urkundlicher Quellen. Der «Codice diplomatico della Lombardia medievale» (8.–12. Jahrhundert), in: QUFIAB 86 (2006) p. 538–561; V. Leoni, Der Codice diplomatico della Lombardia medievale, in: Editionswissenschaftliche Kolloquien 2005/2007: Methodik – Amtsbücher, digitale Edition – Projekte, dir. M. Thumser (2008) p. 219–228. Site en ligne: <http://cdlm.unipv.it/>.

[7] C. Radl, Die Urkundeneditionen innerhalb der dMGH, in: Digitale Diplomatik. Neue Technologien p. 101–115. En ligne: <http://www.dmgh.de/>.

[8] Site en ligne: <http://elec.enc.sorbonne.fr/cartulaires/>.

[9] A. Guerreau, Textes anciens en série. Outils informatiques d'organisation et de manipulation de bases de données textuelles, in: Bulletin du Centre d'études médiévales d'Auxerre (Collection CBMA, 2012), version en ligne <http://cem.revues.org/index12177.html>; id., Pour un corpus de textes latins en ligne, in: ibid. (Collection CBMA, 2011), version en ligne <http://cem.revues.org/index11787.html>.

[10] «Mais curieusement, celles [les bases de données de textes diplomatiques] qui existent ne sont pas encore très utilisées. Disons-le franchement: elles sont sous-utilisées, et n'ont pas, ou pas encore, révolutionné la diplomatique comme elles auraient dû le faire», in: B.-M. Tock, L'apport des bases de données p. 283.

[11] C'était ce même exemple qu'avait déjà choisi B.-M. Tock dans son article pour les actes du colloque Digital Diplomatics (ibid.) Sur CD-ROM: *Thesaurus Diplomaticus,* version 1.0, 1997 (environ 6 000 textes en mode texte); G. Declercq/P. Demonty, L'informatisation de la Table chronologique d'A. Wauters. Méthodologie du nouveau répertoire des documents diplomatiques belges antérieurs à 1200, in: Bulletin de la Commission royale d'Histoire 153 (1987) p. 223–302; P. Demonty, Le *Thesaurus*

création de bases de données – activité en quoi les médiévistes excellent, et c'est une bonne chose – et l'exploitation de celles-ci. Dans cet article, on propose l'hypothèse que cette situation relève d'une inadaptation de la méthode historique traditionnelle à ce nouveau matériau[12], inadaptation qui ne pourra être dépassée que par une mise en ordre globale – aussi bien technique qu'historique – du matériau diplomatique, à l'échelle européenne. De fait, ainsi que le faisait remarquer à juste titre Alain Guerreau en 2001[13], si les médiévistes disposent d'une méthode solide afin de gérer quelques dizaines/centaines d'occurrences de *molendinum*, force est d'admettre qu'une difficulté, non pas seulement différente au plan quantitatif mais d'abord et avant tout au plan qualitatif, émerge lorsqu'il s'agit de donner du sens à plusieurs milliers ou plusieurs dizaines de milliers d'occurrences d'un même vocable. Pourtant, il fait pour nous peu de doutes que seule une approche globale de cette documentation pourrait aider à avancer radicalement dans la compréhension des phénomènes sociaux consignés dans les chartes[14]. Ainsi, nous faisons nôtre la logique systématique d'une partie de l'École des Annales, exprimée par Lucien Febvre en 1933: «Pas une concession à l'esprit de spécialité, qui est l'es-

Diplomaticus, un instrument de travail pour une nouvelle approche en diplomatique médiévale, in: La diplomatique urbaine en Europe au Moyen Âge, éd. W. Prevenier/Th. de Hemptinne (2000) p. 123–132. Lors du colloque de Naples, plusieurs communications en rapport avec cette base ont cependant été présentées, montrant que sa grande richesse attire toujours beaucoup les chercheurs; voir la contribution d'Els De Paermentier au présent volume.

[12] Une enquête sur les convergences et divergences entre ce «nouveau» matériel et les textes marquants de l'historiographie traditionnelle serait sans doute des plus instructives. On pense en particulier à J. G. Droysen, Grundriss der Historik (1875); C. Seignobos, Introduction aux études historiques (1898); A. Comte, Discours sur l'esprit positif (1842); M. Bloch, Apologie pour l'histoire ou métier d'historien (1943). Dans le domaine diplomatique aussi, une telle confrontation épistémologique serait, sans doute, tout aussi intéressante (en considérant les textes classiques de Mabillon, Giry, Bresslau, etc.).

[13] A. Guerreau, L'avenir d'un passé incertain. Quelle histoire du Moyen Âge au XXI[e] siècle? (2001).

[14] Par là, nous n'entendons évidemment pas dénigrer les études visant à exploiter un *corpus* géographiquement circonscrit, par ailleurs tout à fait indispensables; on pense en particulier à H. Fichtenau, Das Urkundenwesen in Österreich vom 8. bis zum frühen 13. Jahrhundert (1971). Simplement, nous pensons que la compréhension du phénomène «charte», dans son unité, ne peut venir que d'études prenant en compte l'échelle de cette production: *a minima* l'échelle européenne. La réflexion sur cette question de l'échelle, conjointe avec celle de seuil, nous semble être une étape essentielle pour voir naître une histoire médiévale renouvelée. Voir J. Le Goff, la civilisation de l'Occident médiéval (1964); id., L'Europe est-elle née au Moyen Âge? (2003); A. Borst, Lebensformen im Mittelalter (1973); M. Mitterauer, Warum Europa? Mittelalterliche Grundlagen eines Sonderwegs (2003).

prit de mort dans l'état actuel du travail humain»[15]. Le but de cet article est ainsi de montrer que cette mise en ordre est non seulement possible, mais aussi nécessaire, afin d'exploiter les bases de données diplomatiques, et ceci même si elle soulève d'importants problèmes aussi bien au niveau épistémologique que technique. Tenter de comprendre comment s'agencent ces textes à l'échelle européenne, tant au plan typologique que chronologique ou géographique, nous semble en effet être un préalable afin de pouvoir réaliser des études lexicales comparées[16] réellement fondées. Sur l'ensemble de cet espace, les différences dans la production textuelle sont en effet telles que tout rapprochement semble parfois audacieux. Ici, nous ne prétendons pas donner une solution générale au problème, mais simplement le mettre en lumière, tout en proposant quelques pistes techniques envisageables pour le dépasser. Notre hypothèse, encore une fois, est qu'il faudra nécessairement traverser cette «étape de masse» pour passer à une exploitation raisonnée des corpus diplomatiques numériques, autrement dit de l'accumulation à l'exploitation.

Au moins depuis l'important livre de Thomas Kuhn, *La structure des révolutions scientifiques*, en 1962[17], nous savons en effet que le progrès en science ne s'effectue pas d'une manière linéaire, mais à travers une série de ruptures, au sein desquelles l'apparition de nouveaux paradigmes – souvent

[15] L. Febvre, Contre l'esprit de spécialité. Une lettre de 1933, in: Combats pour l'Histoire (1953) p. 104–106. Voir aussi A. Guerreau, Fief, féodalité, féodalisme. Enjeux sociaux et réflexion historienne, in: Annales ESC 1 (1990) p. 137–166, en part. p. 137–138.

[16] «Certains aussi concevaient qu'au-delà de l'étude des chancelleries particulières devait naître, ou plutôt renaître, une diplomatique comparative, condition de nouveaux développements de secteurs entiers des sciences historiques», R.-H. Bautier, La Commission internationale de diplomatique. Sa genèse. Son organisation. Son programme de travail, in: BECh 129 (1971) p. 421–425, ici p. 421.

[17] T. S. Kuhn, La structure des révolutions scientifiques (1983) (trad. de: The structure of Scientific Revolutions (1962). Sur la question, très débattue, du lien entre production du savoir scientifique, paradigmes scientifiques et structures sociales, on renvoie aux travaux clés de G. Bachelard, La formation de l'esprit scientifique. Contribution à une psychanalyse de la connaissance objective (1947); P. Bourdieu, Science de la science et réflexivité (2001). Dans le domaine de l'histoire médiévale, sur ce même problème, voir L. Kuchenbuch, Zwischen Lupe und Fernblick: Berichtspunkte und Anfragen zur Mediävistik als historischer Anthropologie, in: Mediävistik im 21. Jahrhundert. Stand und Perspektiven der internationalen und interdisziplinären Mittelalterforschung, dir. H. W. Goetz/J. Jarnut (2003) p. 269–293; Guerreau, L'avenir d'un passé incertain; J. Morsel, L'histoire (du Moyen Âge) est un sport de combat ...: réflexions sur les finalités de l'histoire du Moyen Âge destinées à une société dans laquelle même les étudiants d'histoire s'interrogent (2007); J. Demade, Pourquoi étudier l'histoire (médiévale) au XXIe siècle?, version en ligne <http://lamop.univ-paris1.fr/IMG/pdf/Demade_1.pdf>; id., Produire un fait scientifique. La méthodologie de l'histoire des prix entre structures académiques et enjeux intellectuels (milieu XIXe-milieu XXe), version en ligne <http://lamop.univ-paris1.fr/IMG/pdf/Demade_2–2.pdf>.

concomitante à l'apparition de nouveaux outils – joue un rôle fondamental. La situation de blocage face à ces bases de données, même si tous ne s'accordent pas sur ce sujet, la situation que vivent actuellement les médiévistes lorsqu'il s'agit de reconstruire la logique globale du système de l'Occident médiéval (d'abord, nous l'admettons, par désintérêt de cette perspective globale) – deux phénomènes dont le parallélisme est plus que troublant –, constituent, à en croire Kuhn, une situation typique de la science en crise. Est-ce l'accumulation qui apportera une solution à ces difficultés? Il est possible d'en douter[18]... Ainsi, si à première vue, ces bases de données diplomatiques ne font que numériser et compiler des séries de documents déjà abordables, soit sous la forme de parchemins, soit sur papier, elles proposent probablement bien plus que cela. Elles ne sont pas uniquement un changement d'échelle, une vitesse supplémentaire lors des recherches, elles ne proposent pas seulement un changement d'objet mais bien davantage: la numérisation des données textuelles du Moyen Âge occidental invite à notre sens à une double rupture, d'ordre méthodologique et conceptuelle. En d'autres termes, il s'agit d'inventer un questionnaire scientifique et un dispositif technique – certes, ce n'est pas chose facile –, en s'appuyant sur des outils qui n'ont peu ou pas retenu l'attention des médiévistes jusqu'ici: dans le cadre du présent exposé, nous pensons en particulier au data/text mining[19]. Cependant, la mise en place de tels outils nécessite une restructuration des données textuelles livrées par le Moyen Âge, et ceci à grande échelle. Cette formalisation est essentielle, puisque c'est d'elle dont dépend l'ensemble du travail d'exploitation

[18] «La même recherche historique qui met en lumière combien il est difficile d'isoler les inventions et découvertes individuelles nous amène à douter profondément du processus cumulatif par lequel, pensait-on, ces contribution individuelles s'étaient combinées pour constituer la science», in: Kuhn, La structure des révolutions scientifiques p. 19.

[19] Les exceptions sont récentes et montrent, dans ce domaine précis, un certain retard non seulement par rapport aux sciences dites «dures», mais aussi en regard des autres sciences sociales (psychologie, sociologie, etc.). La bibliographie du data/text mining est devenue, en quelques années seulement, à proprement parler gigantesque. Mentionnons simplement quelques ouvrages considérés comme canoniques ou relativement simples d'accès pour un non-mathématicien: C. D. Manning/H. Schütze, Foundations of Statistical Natural Language Processing (1999); I. H. Witten/E. Frank/M. A. Hall, Data Mining. Practical Machine Learning Tools and Techniques (³2011); C. D. Manning/H. Schütze/P. Raghavan, An Introduction to Information Retrieval (2008); R. Feldman/J. Sanger, The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data (2007); H. A. Do Prado/E. Ferneda, Emerging Technologies of Text Mining: Techniques and Applications (2008); L. Torgo, Data Mining with R. Learning with Case Studies (2011); R. Bilisoly, Practical Text Mining with Perl (2008); The Handbook of Computational Linguistics and Natural Language Processing, éd. A. Clark/C. Fox/S. Lappin (2010).

qui devrait découler de ces bases de données. Il apparaît donc que cette mise en forme, autant technologique qu'abstraite, est une étape cruciale et non une banalité formelle, afin de passer de l'accumulation documentaire brute à l'exploitation des documents[20]. Au plan historique, celle-ci passe par une mise en ordre typologique, géographique et chronologique du matériau disponible. Afin de nous insérer dans cette discussion, nous aimerions présenter une série d'outils et d'expériences, réalisée dans le cadre d'une thèse en cours à l'Université de Bourgogne, sous la direction d'Eliana Magnani et de Daniel Russo, menée avec l'aide d'Alain Guerreau. Cette thèse vise à exploiter systématiquement une base de données de documents diplomatiques, en réalisant plusieurs investigations statistiques sur les champs sémantiques, en particulier ceux dits de l'«environnement» ou, mieux, de l'espace[21] (*aqua*, *terra*, *mundus*, *silva*, *arbore*, *campus*, *villa*, etc.), ceci afin de faire ressortir les inégalités régionales, dans le domaine de la scripturalité tout d'abord, et, par là, s'interroger sur la dynamique de l'Occident médiéval. Concrètement, on présentera deux expériences visant à montrer qu'un dispositif technique global et adapté permettrait de mieux mettre en ordre la documentation numérisée, ceci afin de rendre accessible une information pour le moment seulement disponible *en puissance*.

[20] Cela implique aussi que la mise en forme des données dépend, au moins dans une certaine mesure, du questionnaire que l'on souhaite leur appliquer. Concrètement, cela exclut toute prétention à un formatage définitif de la documentation.

[21] En histoire médiévale, les investigations portant sur les champs sémantiques restent, pour le moment, relativement rares. Première approche de la question dans les textes d'un des fondateurs du concept: J. Trier, Über Wort- und Begriffsfelder (1931), in: Wortfeldforschung. Zur Geschichte und Theorie des Sprachlichen Feldes, éd. L. Schmidt (1973) p. 1–38; id., Das sprachliche Feld. Eine Auseinandersetzung (1934), in: Wortfeldforschung p. 129–161; id., Der heilige Jodocus, sein Leben und seine Verehrung, zugleich ein Beitrag zur Geschichte der deutschen Namengebung (1924); plus récemment: A. Guerreau, Le champ sémantique de l'espace dans la *vita* de saint Maïeul (Cluny, début du XI^{e} siècle), in: Journal des Savants 1997:2 (1997) p. 363–419; A. Guerreau-Jalabert/B. Bon, *Pietas*: réflexions sur l'analyse sémantique et le traitement lexicographique d'un vocable médiéval, in: Médiévales 42 (2002) p. 73–88; iid., Le trésor au Moyen Âge: étude lexicale, in: Le trésor au Moyen Âge: discours, pratiques et objets, dir. L. Burkart/P. Cordez/P.-A. Mariaux (2010) p. 11–32; L. Kuchenbuch/U. Klein, 'Textus' im Mittelalter: Erträge, Nachträge, Hypothesen, in: 'Textus' im Mittelalter: Komponenten und Situationen des Wortgebrauchs im schriftsemantischen Feld, éd. L. Kuchenbuch/U. Klein (2006) p. 416–453. Sur la question, seulement évoquée ici, du lien entre structure sociale, langage et milieu: C. Lévi-Strauss, La pensée sauvage (1962); M. Godelier, L'idéel et le matériel – Pensée, économies, sociétés (1986); P. Descola, Par-delà nature et culture (2005).

Typologie(s), data mining et catégorisation (AI::Categorizer; Text-to-CSV)

Notre présente réflexion s'appuie sur une base de données en cours de réalisation, base de données dans laquelle une majeure partie du matériau diplomatique numérisé disponible en ligne a été collecté, puis entièrement reformaté afin de devenir une méta-source à part entière. Nous avons fait le choix de réencoder l'ensemble des chartes ainsi obtenues, principalement en employant des séries de scripts Perl *ad hoc*[22], en un format XML-«TEI light», acceptable par le seul logiciel actuellement capable de gérer efficacement autant de documents: Philologic[23]. Ce dernier est en effet à notre connaissance l'unique solution, en dehors de l'austère (mais puissant) CQP/CWB[24], capable d'absorber une telle masse de textes: 64 000 unités, soit à l'origine 64 000 chartes. Grâce à une série d'échanges avec son créateur Mark Olsen, nous avons pu modifier cette limite et amener le logiciel à gérer non plus seulement 64 000 documents, mais 64 000 volumes d'éditions. De fait, notre collecte, basée avant tout sur les bases déjà disponibles, mais aussi sur des numérisations personnelles[25], aidée par de nombreux chercheurs, a abouti à une base de données contenant près de 150 000 chartes en mode texte, soit

[22] Le langage Perl s'impose, avec Python, comme l'une des voies possibles pour les médiévistes souhaitant traiter des corpus. Créé par Larry Wall en 1987, celui-ci est simple d'apprentissage, tout en étant particulièrement adapté au traitement de fichiers textuels. Surtout, sa gestion puissante mais aussi très souple des *regex* (expression régulière) permet d'écrire des scripts de conversion et de manipulation de textes avec une grande facilité.

[23] Logiciel développé par Mark Olsen et son équipe à l'Université de Chicago, disponible en ligne: <http://sites.google.com/site/philologic3/> (version 3.2). Il s'agit du choix initialement fait par l'équipe du projet CBMA.

[24] Disponible en ligne, <http://cwb.sourceforge.net/>.

[25] Relues ou non. Quelques exemples: Chartes et documents concernant l'abbaye de Cîteaux: 1098–1182, éd. J. Marilier (1961); Le livre des serfs de l'abbaye de Marmoutier suivi de chartes sur le même sujet et précédé d'un essai sur le servage en Touraine, par Ch.-L. de Grandmaison (*Liber de servis majoris monasterii*), éd. A. Salmon (1865); Cartulaire de Brioude. *Liber de honoribus Sancto Juliano collatis*, éd. H. Doniol (1863); Codex diplomaticus Fuldensis, éd. E. Dronke (1850); etc. Ces numérisations ont été réalisées en fonction des «vides» géographiques laissés par les bases de données actuellement disponibles. D'autre part, on a ajouté à ce noyau initial une série d'autres fichiers pour lesquels la relecture n'avait pas été systématique. Ces fichiers en «dirty-OCR» permettent d'ajouter plusieurs dizaines de milliers de chartes à cette collection, sans grand effort de relecture, et offrent des possibilités de comparaisons accrues. Pour les comptages fiables, il suffit alors de réaliser deux versions de la base: l'une comprenant les fichiers en dirty-OCR, l'autre les excluant. Rappelons au passage que l'exhaustivité documentaire est une chimère et que c'est plutôt un souci d'homogénéité (très relative, en l'occurrence) qui a guidé notre démarche.

plus de 570 éditions, pour une période allant du VII^e siècle au début du XIV^e siècle, et ceci pour une large part de l'Europe chrétienne. D'une manière générale cependant, la base a aussi été générée dans d'autres formats, en particulier celui employé par le logiciel Textométrie[26]: cette flexibilité voulue, aux antipodes d'un modèle fixiste, nous semble être une considération centrale dans un domaine où la réactualisation technologique est constante et difficilement prévisible.

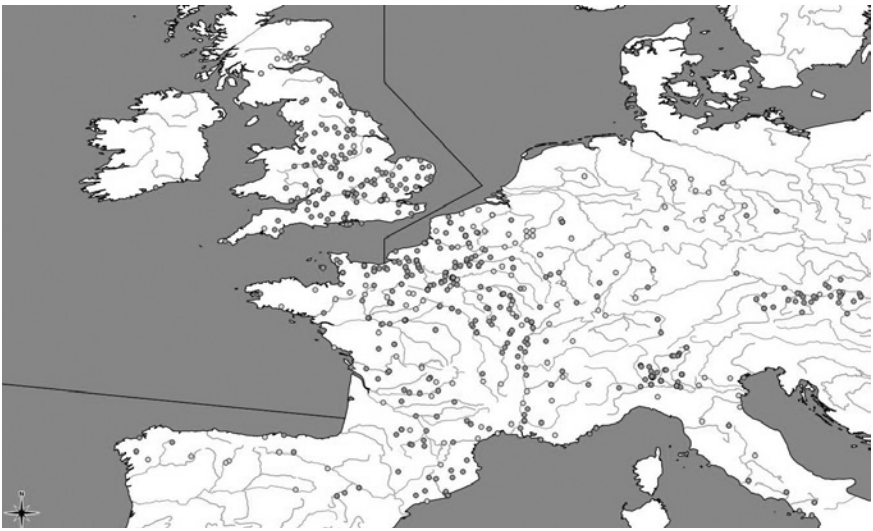

Fig. 1: Carte des corpus présents dans notre base. Légende: en gris, les corpus relus; en blanc: les corpus en dirty-OCR. (Carte générée grâce à Q-GIS).

[26] Le projet Textométrie, développé à Lyon par l'équipe de Serge Heiden et qui vient tout juste de sortir dans sa version 0.6 (6 avril 2012), semble être une alternative importante à Philologic: tout d'abord parce que le logiciel gère la lemmatisation des textes (formes / lemmes / POS). Le logiciel est en outre basé sur le moteur de CQP, ce qui est un gage de sa solidité. Il fonctionne sur trois systèmes d'exploitation (Linux/Windows/Mac), ce qui le rend accessible au plus grand nombre. Enfin, *last but not least*, le logiciel possède une interface – encore en travail – le liant au logiciel de statistique R. À l'heure actuelle, Textométrie apparaît donc comme un challenger de premier ordre, même s'il connaît encore quelques difficultés face aux plus larges corpus. Voir S. Heiden/J.-P. Magué/B. Pincemin, TXM: Une plateforme logicielle open-source pour la textométrie – conception et développement, in: Proc. of 10th International Conference on the Statistical Analysis of Textual Data – JADT 2010, éd. I. C. Sergio Bolasco (2010) p. 1021–1032 (2); S. Heiden, The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme, in: 24th Pacific Asia Conference on Language, Information and Computation – PACLIC24, éd. K. I. Ryo Otoguro (2010) p. 389–398. Logiciel disponible en ligne: <http://textometrie.ens-lyon.fr/>.

Rapidement cependant s'est imposée la question de l'indexation de cette masse: un critère déterminant pour passer de l'accumulation à une exploitation globale, en particulier statistique. C'est aussi, évidemment, une opération qu'il est improbable de réaliser en considérant les actes un à un. Or, nous savons tous, en effet, qu'il est important de distinguer les catégories diplomatiques, afin de ne pas tomber dans des effets de corpus redoutables[27]. Le text mining offre une série d'algorithmes d'intelligence artificielle, largement utilisés dans d'autres disciplines, permettant d'attribuer automatiquement – moyennant un entraînement –, des catégories typologiques ou même chronologiques à un ensemble documentaire. Une de nos problématiques visait donc à établir un programme non seulement capable de reconnaître, dans une masse de documents, les bulles, les diplômes, les actes épiscopaux, de distinguer les chartes des notices, mais surtout à même de réinjecter par la suite ces catégories dans un formatage XML lisible et indexable par Philologic (et, par extension, par d'autres logiciels). Les obstacles sont cependant nombreux: outre la sensibilité de ces algorithmes au bruit (difficulté technique), certaines catégories diplomatiques possèdent des contours flous (difficulté disciplinaire/historiographique)[28].

La stratégie retenue ici consiste à déterminer, dans un premier temps et de manière globale, sur un échantillon documentaire aléatoire, les typologies/catégories diplomatiques les plus distinctes les unes des autres. Pour ce faire, nous avons employé un second logiciel développé dans le cadre de cette thèse, intitulé Text-to-CSV, et qui permet de transformer une série de textes en matrices afin d'en évaluer la proximité. Cette estimation de la distance d'un texte à un autre est obtenue grâce à une comparaison d'une partie (sélectionnée de manière non a priori) ou de la totalité du lexique de ceux-ci. Pour cela, on

[27] La base des CBMA offre plusieurs bons exemples de biais liés à des corpus factices: on pense en particulier à l'édition du Cartulaire général de l'Yonne. Sur ce point, voir le très pertinent article de I. Rosé, À propos des *Chartæ Burgundiæ Medii Ævii*. Sur cette thématique de la production documentaire et de la gestion/interprétation des biais/effets structuraux dans les corpus diplomatiques numérisés, on se permet de renvoyer à: N. Perreaux, La production de l'écrit diplomatique en Bourgogne.

[28] Beaucoup de clarifications utiles dans O. Guyotjeannin/J. Pycke/B.-M. Tock, Diplomatique médiévale (³2006), en particulier le chapitre 4: «Brève typologie des actes médiévaux», p. 103–222. On est bien conscient que, pour prendre un exemple, la totalité des «actes des papes ne sont pas, loin s'en faut, des bulles». Néanmoins, la majorité des actes papaux dans les recueils d'actes diplomatiques le sont: c'est ce qui compte dans le cadre de notre démarche. L'algorithme n'a donc pas pour but d'effectuer un classement «parfait», qui, s'il est louable comme objectif, n'est pas réalisable par un tri automatique. Les «niveaux de confiance», décrits plus bas, servent à palier – mais seulement en partie – cette difficulté. La machine ne remplace donc pas le diplomatiste/l'historien, elle aide à construire un dispositif, avec ses limites, propre à l'exploitation de questionnaires renouvelés.

applique différents traitements à la matrice, d'abord en la transformant – pour réduire sa taille, à travers un échantillonnage, et le «bruit» qu'elle contient – (par exemple *via* des coefficients tels que le TF-IDF[29], ou encore *via* des AFC/ACP[30]), puis en tentant de donner une modélisation graphique ou un résumé de l'information qu'elle renferme (AFC / Clustering)[31]. Or, si l'on applique ce logiciel à un échantillon de textes provenant de toutes les catégories diplomatiques («auteurs» et typologies) à la fois, et que nous réalisons ensuite une analyse factorielle[32] sur le tableau obtenu, ceci afin de visualiser la distance entre les différents textes considérés, nous voyons que le bruit de fond est extrêmement fort. Autrement dit, il sera impossible, dans une telle configuration et avec un test unique, d'obtenir une classification automatique valable. Dans ces conditions, seuls les diplômes seront, dans presque tous les cas de figure, facilement identifiés par les algorithmes (fig. I, voyez les planches en couleur, pages 321–322). En revanche, lors d'une seconde phase de nos expériences, on a pu observer que la discrimination entre un nombre réduit de catégories était beaucoup plus efficace. Prenons un exemple: si une analyse considérant en même temps les bulles, les documents épiscopaux et les diplômes semble vouée à se fourvoyer très fréquemment (fig. II)[33], une analyse sur un nombre de catégories plus réduit, par exemple considérant d'une part les bulles et les documents épiscopaux et, d'autre part, les diplômes, a une probabilité nettement plus faible de se tromper (fig. III).

---

[29] Pour Term Frequency-Inverse Document Frequency. Cette méthode a pour but d'évaluer le poids relatif d'un terme ou d'un ensemble de termes dans un document, par rapport à un ensemble documentaire. Schématiquement, l'indice TF-IDF d'un terme augmente proportionnellement avec le nombre d'occurrences de celui-ci dans le document considéré, mais diminue si le mot se trouve aussi dans les autres documents du corpus. On peut ainsi distinguer les termes qui caractérisent le mieux un document donné au sein d'un ensemble documentaire.

[30] AFC: Analyse factorielle des correspondances; ACP: Analyse en composantes principales.

[31] Littérature en français sur la question: A. Salem/P. Lafon, L'inventaire des segments répétés d'un texte, in: Mots 6 (1983) p. 161–177; A. Salem, Segments répétés et analyse statistique des données textuelles, Histoire & Mesure 1 (1986) p. 5–28; E. Brunet, Peut-on mesurer la distance entre deux textes?, in: Corpus 2 (2003), version en ligne: <http://corpus.revues.org/index30.html>; L. Lebart/A. Salem, Statistique textuelle (1994).

[32] Sur cette méthode devenue standard dans de nombreuses sciences sociales, un classique et une introduction: J.-P. Benzécri, L'analyse des données, 3 vol. (1984); P. Cibois, L'analyse factorielle: analyse en composantes principales et analyse des correspondances (2000).

[33] Dans la figure II, on note en effet un fort recouvrement entre les catégories «bulles» et «actes d'évêques», ce qui était prévisible. Il s'agit donc d'éviter ce recouvrement, afin de faciliter la tâche de l'algorithme de reconnaissance.

Construction du schéma de catégorisation

Partant d'une série d'observations de ce type, on a pu construire un schéma, puis un programme permettant de tester les actes de notre base un à un, pour enfin «réinjecter» le résultat de la catégorisation dans un fichier XML Philologic. Pour créer la bibliothèque d'entraînement nécessaire au logiciel pour l'apprentissage des catégories diplomatiques, nous sommes partis de deux bases: les CBMA et les originaux de l'Artem, auxquelles nous avons adjoint des diplômes carolingiens et ottoniens des MGH. Ces deux premières bases possèdent en effet un nombre conséquent de champs d'ores et déjà renseignés (auteur, bénéficiaire, genre ou typologie, fiabilité de l'acte, etc.), et existent toutes deux dans des formats relativement simples à manipuler *via* des scripts Perl ou XSLT (pour *Extensible Stylesheet Language Transformations*). Une fois ce corpus d'entraînement organisé, il s'agit de trouver un logiciel dédié ou adapté à la catégorisation textuelle. Dans ce domaine en pleine émergence, les possibilités de qualité, gratuites voire libres/open sources, ne manquent pas: Weka[34], le logiciel de statistique R avec une myriade de *packages* adaptés[35], RapidMiner[36], Tanagra[37], Knime[38], etc. Pour notre part, tout en continuant de travailler avec ces différents logiciels, nous avons retenu, pour cette expérience, la bibliothèque Perl AI::Categorizer, développée par Ken Williams[39]. Celle-ci inclut la plupart des algorithmes utiles à la classification textuelle et dispose d'une interface pour envoyer et recevoir des informations vers/depuis Weka[40]. D'autre part, étant une bibliothèque Perl, elle convenait parfaitement

[34] M. Hall et al., The WEKA Data Mining Software: An Update, in: SIGKDD Explorations 11,1 (2009) p. 1–18. Le manuel de Witten, Frank et Hall, déjà mentionné, est basé sur Weka. En ligne: <http://www.cs.waikato.ac.nz/ml/weka/>.

[35] R, qui constitue le standard en matière de statistiques open source, ne se présente plus <http://www.r-project.org/>. Liste de packages dédiés au «machine learning» à cette adresse: <http://cran.r-project.org/web/views/MachineLearning.html>. A noter que R communique facilement avec Weka, via le package *RWeka*.

[36] <http://rapid-i.com/content/view/181/190/>.

[37] <http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html>.

[38] <http://www.knime.org/>.

[39] Le package et ses algorithmes font l'objet de la thèse de l'auteur: K. Williams, A Framework for Text Categorization (2003), version en ligne: <http://search.cpan.org/~kwilliams/AI-Categorizer-0.09/>.

[40] Naive Bayes, SVM, k-Nearest-Neighbor, arbres de décisions, etc. La bibliographie, même de synthèse, sur les algorithmes de catégorisation/d'apprentissage supervisé est énorme et se développe à une vitesse surprenante. Voir, entres autres, Y. Yang/X. Liu, A re-examination of text categorization methods, in: SIGIR '99 Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (1999) p. 42–49; F. Sebastiani, Machine Learning

à notre problématique qui était d'intégrer une catégorisation dans une série de traitements (extraction du texte, puis réinjection de celui-ci, accompagné des catégories, dans un nouveau fichier). Autre avantage, enfin: ce choix nous permettait de traiter facilement les résultats, de les comparer et de les manipuler, voire d'adjoindre au test un algorithme extérieur à la bibliothèque, tout cela sans interfaçage complexe, puisque l'ensemble des opérations se déroulent alors sous Perl. Après de nombreux essais, le principe du modèle retenu consiste à combiner un arbre de décision à une série de tests (cross-validation) qui interviennent à chaque nœud de l'arbre: le Support Vector Machine (Machine à vecteurs de support – SVM)[41], l'algorithme Naive Bayes (ou Classification naïve bayésienne – NB)[42], mais aussi un test développé pour l'occasion et qui se base sur une analyse des syntagmes ou morceaux de

in Automated Text Categorization, in: ACM Computing Surveys 34,1 (2002) p. 1–47; Manning/Schütze/Raghavan, An Introduction to Information Retrieval p. 234–320 (particulièrement intéressant sur ces points). En outre, la plupart des manuels récents concernant le data mining contiennent un chapitre sur l'apprentissage (supervisé / non supervisé) et la catégorisation: Torgo, Data Mining with R. Learning with Case Studies p. 185–186. Cf. plus bas pour plus de détails sur les algorithmes retenus.

[41] Méthode d'apprentissage supervisé, inspirée des travaux de Vladimir Vapnik et introduite en 1995; V. Vapnik, The Nature of Statistical Learning Theory (1995). D'une grande efficacité mais assez lente, elle se base sur une séparation des données via une méthode à noyaux (Kernel). Witten/Frank/Hall, Data Mining. Practical Machine Learning Tools and Techniques p. 192: «Support vector machines select a small number of critical boundary instances called support vectors from each class and build a linear discriminant function that separates them as widely as possible. This instance-based approach transcends the limitations of linear boundaries by making it practical to include extra nonlinear terms in the function, making it possible to form quadratic, cubic, and higher-order decision boundaries». D'une manière générale, le Wikipedia anglais propose des articles de grandes qualités en ce qui concerne les algorithmes de data mining. Voir aussi: V. Kecman, Learning and Soft Computing – Support Vector Machines, Neural Networks, Fuzzy Logic Systems (2001); bonne synthèse dans: K. P. Bennet/C. Campbell, Support Vector Machines: Hype or Hallelujah?, in: SIGKDD Explorations 2,2 (2000) p. 1–13; pour une application concrète sous R: A. Karatzoglou/D. Meyer, Support Vector Machines in R, in: Journal of Statistical Software 15,9 (2006), disponible en ligne <http://www.jstatsoft.org/v15/i09/paper>

[42] Les méthodes bayésiennes sont plus anciennes que le SVM: elles sont basées sur les théories du mathématicien et pasteur anglais, Thomas Bayes [1702–1761]. Souvent d'une bonne efficacité (cependant moindre que le SVM), en particulier dans le cas de données nombreuses ou complexes – ce qui est le cas avec les actes diplomatiques –, elles se montrent d'abord particulièrement rapide. Ces caractéristiques sont principalement dues à son algorithme qui suppose une indépendance des caractéristiques analysées pour une classe donnée. H. Zhang, The Optimality of Naive Bayes, in: FLAIRS2004 (2004), disponible en ligne <http://courses.ischool.berkeley.edu/i290-dm/s11/SECURE/Optimality_of_Naive_Bayes.pdf>; D. J. Hand/Y. Yu, Idiots

formules présents dans les actes[43]. L'avantage qu'il y a à combiner ces techniques est évident: s'il arrive fréquemment qu'un test attribue une mauvaise catégorie à un acte donné, il est en revanche peu probable que l'erreur se répète sur trois tests différents, d'abord car leur «sensibilité» porte sur des éléments statistiques différents[44]. Le second avantage de cette technique est qu'elle nous permet d'intégrer la probabilité d'erreur d'attribution dans notre logiciel de fouille textuelle. L'idée est la suivante: si les trois tests proposent une seule et même catégorie – disons par exemple une bulle –, non seulement celle-ci sera donnée au document, mais aussi le niveau «1», ce qui signifie une très forte probabilité d'attribution réussie. Ce sont les modules de décision, visibles sur le schéma ci-dessous, qui se chargent de cette tâche visant à attribuer jusqu'à 3 niveaux de confiance. L'intérêt est évident: le chercheur peut ainsi sélectionner un échantillon plus ou moins large d'un type donné, et espérer travailler soit un échantillon très représentatif (en prenant seulement les documents classés en niveau 1), soit un échantillon plus large mais contenant plus de bruit (en retenant aussi les documents de niveau 2 ou 3).

Bayes – not so stupid after all?, in: International Statistical Review 69 (2001) p. 385–389; G. I. Webb/J. R. Boughton/Z. Wang, Not So Naive Bayes: Aggregating One-Dependence Estimators, in: Machine Learning 58,1 (2005) p. 5–25.

[43] Cet algorithme, codé en Perl par nos soins, est rudimentaire. Il s'agit, à partir de corpus déjà indexés (par exemples les CBMA ou l'Artem), d'extraire les termes ou groupes de termes qui ne se rencontrent que dans une catégorie diplomatique donnée. On obtient ainsi une liste d'environ 34 500 éléments, qui contient à la fois des formes, des bi-formes, et des tri-formes. Lorsque le document est examiné par l'algorithme, celui-ci attribue des points en fonction des groupes de termes rencontrés: *imperio atque pro*, ajoutera, par exemple, 4 points à la variable $count_dip (pour les diplômes donc), car c'est un indice qui pousse fortement à penser qu'il s'agit d'un acte de souverain. Au terme de l'examen de l'ensemble des mots et syntagmes du document, on compare les scores obtenus pour des différentes variables ($count_dip; $count_bul; $count_cha; etc.), après les avoir pondérés. Une pondération en effet nécessaire car le nombre de termes typiques n'est pas équivalent en fonction des «genres» documentaires. Le score pondéré permet enfin d'attribuer le document à une catégorie.

[44] Cette affirmation est avant tout un résultat empirique: c'est à la suite de tests successifs et avec des séries de comparaisons que nous sommes arrivés à cette conclusion. L'examen des algorithmes tend par ailleurs à confirmer théoriquement ce que montrent les expériences. Ces différentes méthodes sont en effet toutes fondées sur le calcul puis la comparaison de distances, directes ou indirectes, entre les éléments à analyser. Mais ces distances ne sont pas calculées de la même manière, d'où l'instabilité fondamentale des résultats, qui sont en fait complémentaires: pour le dire autrement, en utilisant différents algorithmes, on ne regarde pas l'objet (i.e. les chartes, ou plutôt le tableau de mots qui en résulte) sous le même angle. C'est la combinaison comparative des visions obtenues depuis ces différents plans qui donne une tendance générale.

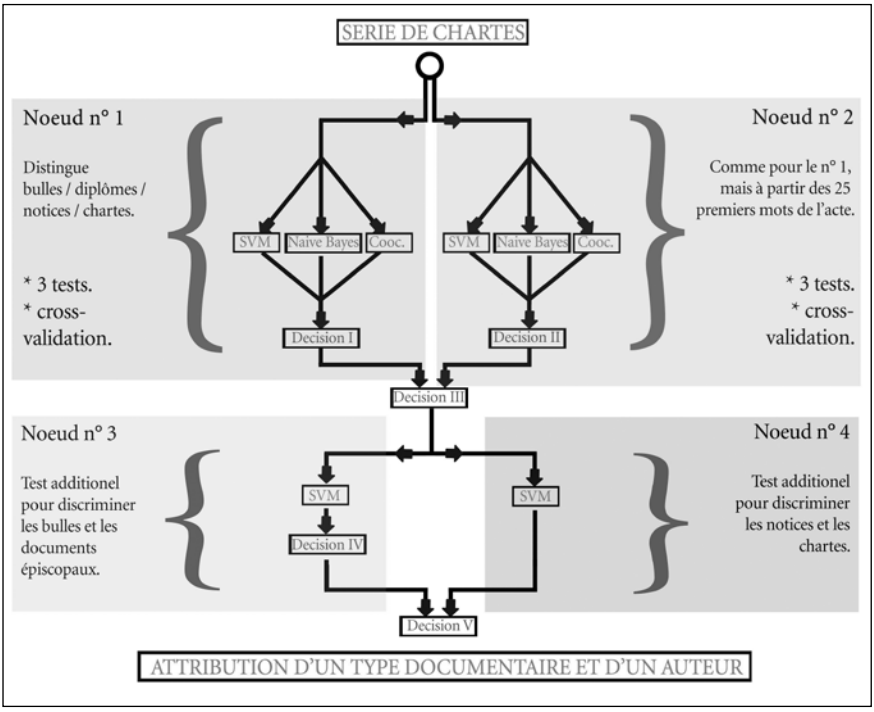

Fig. 2: Schéma du *«classifier»*, combinant arbre de décision et, à chaque nœud, un ou plusieurs algorithmes (SVM, NB, Cooc).

Validité de la méthode et possibilités d'extensions

Qu'en est-il maintenant des résultats obtenus grâce à ce schéma? Afin de tester la qualité de cette classification automatique, nous avons opté pour des matrices de confusion[45], outil standard en apprentissage supervisé et qui permet d'évaluer non seulement le taux d'erreurs positives mais aussi le taux d'erreurs négatives (autrement dit ce que notre modèle n'a pas réussi à identi-

[45] J.-P. Benzécri, Sur l'analyse des matrices de confusion, in: Revue de statistique appliquée 18,3 (1970) p. 5–63. D'une manière générale, on peut résumer la matrice de confusion pour une classe donnée – ici, pour l'exemple, les bulles – aux cases suivantes: a: vrais positifs (bulles correctement catégorisées comme telles) / b: faux positifs (autres documents catégorisés comme bulles) / c: faux négatifs (bulles catégorisées comme autre chose qu'une bulle) / d: vrais négatifs (tous les documents «non-bulles» correctement catégorisés comme tels). On possède ensuite différents indices pour calculer la précision (= ((a+d)/(a+b+c+d)), le rappel (=a/(c+a)), le taux de faux positifs (=b/(d+b)), le taux de vrais négatifs (=d/(d+b)), le taux de faux négatifs (=c/(c+a)) et enfin l'exactitude de la classification (=a/(b+a)).

fier). Pour ce faire, il faut tester la validité du modèle obtenu sur des documents non renseignés mais dont on connaît la typologie, et, surtout, qui n'ont pas été inclus dans la base d'entraînement. Ce dernier point est fondamental, car réaliser un test de validité sur des textes déjà contenus dans la base d'entrainement fausse totalement les résultats, donnant des scores largement favorables au modèle. En comparant les résultats obtenus aux résultats attendus (encore une fois: il faut partir de documents dont on connaît la typologie, mais qui n'ont pas été inclus dans le lot destiné à l'apprentissage du logiciel), on peut connaître la précision de son modèle et savoir s'il est plutôt précis ou plutôt fautif. Dans ce cas, il s'agit d'examiner aussi bien les résultats obtenus pour les vrais positifs, que pour les faux positifs, les faux négatifs ou encore les vrais négatifs[46]: on connait ainsi le taux de reconnaissance, mais aussi de rejet, pour chaque catégorie. C'est à partir de ces résultats de tests que nous avons progressivement et empiriquement amélioré notre modèle, par exemple en ajoutant des coefficients de pondération pour le troisième algorithme, décrit plus haut. Néanmoins, il est très probable que si l'expérience devait être réitérée, il faudrait chercher d'autres combinaisons d'algorithmes et de méthodes, car il nous semble qu'il en existe, après examen, de plus puissantes, offrant des résultats plus stables. Dans cette optique, les classifications hiérarchiques ascendantes (type arbre de décision), réalisées sur les coordonnées obtenues suite à une analyse factorielle des correspondances, elle-même effectuée sur une matrice de groupes de formes, sont connues pour donner des résultats robustes[47] et semblent fournir, dans le cas des chartes, des scores prometteurs. À l'heure actuelle, sans entrer dans des détails superflus, le modèle élaboré permet de reconnaître correctement environ 95 % des bulles, environ 95 % de diplômes – ces deux éléments constituant notre objectif initial, étant donné qu'il s'agissait du point le plus crucial pour une exploitation efficace de notre base de données –, plus de 85 % des documents épiscopaux, et une large majorité des chartes et des notices. Pour arriver à ces taux de reconnaissance, il nous a bien entendu fallu tester toute une série de modèles et créer une longue liste de fichiers d'entraînement pour chaque nœud, le programme final ne comptant pas moins de 42 000 documents indexés, ajoutés par itérations successives[48].

[46] Voir note précédente.

[47] Cf. plus haut les travaux déjà mentionnés de J.-P. Benzécri.

[48] Un point clé pour une évaluation correcte du modèle consiste à réaliser ces tests sur des documents choisis aléatoirement et non inclus dans les fichiers d'entraînement. Si ce dernier point n'est pas respecté, on obtient, en toute logique, des résultats très supérieurs à l'efficacité réelle du schéma.

En outre, il ne faut pas concevoir cette expérience de statistique linguistique[49] – puisqu'après tout, c'est bien de ça qu'il s'agit avec le text mining – comme un simple outil de classification typologique. Une telle méthode est bien entendu extensible à une vaste série de problématiques, en premier lieu à l'épineuse question de la datation des actes[50]. Dans cette optique, il s'agirait de constituer une autre base d'entraînement, contenant cette fois des actes dont la chronologie est assurée, et qui deviendraient alors des références pour les documents à la datation floue ou inconnue.

Régionalisation et géographie des pratiques de l'écrit (Text-to-CSV)

Cependant, il nous a aussi semblé que ces questions de mise en ordre et de classification ne pourraient progresser qu'en créant des corpus d'entraînement régionaux. De fait, on peut supposer que les spécificités locales des actes diplomatiques soient telles que les résultats du classement dépendraient aussi, pour une large mesure, du rapport géographique entre documents à classer et groupe d'entraînement. Existe-t-il des particularités régionales, plus ou moins fixes dans le temps, en ce qui concerne l'écriture documentaire? C'est là le cœur de notre problématique: les spécificités des pratiques de l'écrit et la dynamique de l'Occident médiéval. Sans en venir à la question des champs sémantiques, qui nécessiterait des développements beaucoup plus longs, il est fort probable que de «simples» études de lexicographie statistique pourraient apporter des éléments favorisant notre compréhension de l'Occident médiéval. Un peu plus haut, nous avons évoqué un programme développé lors de cette première année de thèse, programme nommé provisoirement

[49] C'est un point à ne jamais oublier: les textes et les mots possèdent des spécificités statistiques qui les rendent très complexes à étudier et à manipuler: «Similarly, the parameters of many models for word frequency distribution are highly dependent on the sample size. This property sets lexical statistics apart from most other areas in statistics [...]», in: H. Baayen, Word Frequency Distributions (2001) p. 1. Les distributions lexicales suivent en effet un modèle que ce dernier auteur à nommé, à juste titre, LNRE (Large Number of Rare Events). Quelques autres ouvrages clés sur ce sujet: Manning/Schütze, Foundations of Statistical Natural Language Processing; S. Evert, The statistics of word cooccurrences: Word pairs and collocations (2004); B. Mandelbrot, Les objets fractales. Forme, hasard et dimension. Troisième édition, révisée par l'auteur et augmentée d'un «Survol du langage fractal» (³1990); id., Fractales, hasard et finance (1997); dans le domaine historique, une exception de par son contenu avancé: A. Guerreau, Statistique pour historiens, cours professé à l'École des chartes (2003–2004) (2004), disponible en ligne <http://elec.enc.sorbonne.fr/statistiques/stat2004.pdf>.

[50] Outre les travaux de M. Gervers déjà mentionnés plus haut, on renvoie à son article donné dans le présent volume.

Text-to-CSV, et qui permet de convertir une série de textes en tableau. Aucun programme en effet ne proposait une conversion de ce type en dehors de tout logiciel tiers, à la fois flexible, transparente, nous offrant un contrôle total et clair sur toutes les étapes de la procédure. Le logiciel développé pour l'occasion gère donc la tokenization des fichiers[51], le traitement des mots isolés mais aussi des syntagmes (bi/tri/quadri et penta-formes), le clustering en interne (algorithme Fuzzy C-means/FCM, par Mizuki Fujisawa[52]), plusieurs types de pruning, et différents traitements sur les tableaux tels que le coefficient TF-IDF[53], la binarisation, le traitement par rangs, etc. Ces sorties sont directement utilisables sous R, mais aussi sous Weka, qui accepte les formats CSV «classiques».

Afin de prendre un exemple concret, nous avons ainsi passé en tableau une série de corpus à notre disposition, pour la période allant de 900 à 1050. Cette première approche sur à peine plus d'un siècle nous a semblé intéressante, car

[51] Cette étape est essentielle afin de normaliser des traditions d'éditions divergentes. Écrite en Perl, cette routine permet de résoudre les ligatures (œ/æ), en séparant les lettres (oe/ae), de convertir des j en i et les v en u, de retirer certaines ponctuations ou éléments propres aux choix éditoriaux. En procédant ainsi, on perd sans doute une part d'information, mais l'étape est inévitable pour ne pas confondre groupes de textes et groupes d'éditeurs. Dans la dernière version du programme, nous avons intégré certaines parties du script de tokenisation réalisé par Renaud Alexandre et Alain Guerreau dans le cadre de l'ANR OMNIA (IRHT/École nationale des Chartes/UMR 6298 Artehis; Alain Guerreau, Anita Guerreau-Jalabert, Eliana Magnani, Marie-José Gasse-Grandjean, Bruno Bon, Renaud Alexandre, Olivier Canteaut, Frédéric Glorieux et nous-même). Ce programme, aussi écrit en Perl, est disponible en ligne, sur le site de l'ANR: <http://www.glossaria.eu/>.

[52] Il s'agit d'un autre package Perl, et par là-même, très facile à implémenter dans notre script; disponible en ligne, <http://search.cpan.org/~fujisawa/Algorithm-FuzzyCmeans-0.02/>. À l'origine, le Fuzzy C-means est une méthode de «soft clustering»: c'est-à-dire que les points/données ne sont pas associés à un cluster de manière binaire (oui/non), mais par un certain degré (une variable). Par rapport à un K-means classique, cet algorithme permet de jouer sur les seuils afin d'observer les liens multiples qu'entretient un document/corpus avec un cluster. Sur cette «nouvelle» tendance en clustering, voir: R. Nock/F. Nielsen, On Weighting Clustering, in: IEEE Transaction on Pattern Analysis and Machine Intelligence 28 (2006) p. 1–13.

[53] Pour Term Frequency-Inverse Document Frequency. Cette méthode a pour but d'évaluer le poids relatif d'un terme ou d'un ensemble de termes dans un document, par rapport à un ensemble documentaire. Schématiquement, l'indice TF-IDF d'un terme augmente proportionnellement avec le nombre d'occurrences de celui-ci dans le document considéré, mais diminue si le mot se trouve aussi dans les autres documents du corpus. Le but est de distinguer le groupe de mots qui caractérise le mieux un document donné au sein d'un ensemble documentaire.

cette fourchette est souvent considérée comme relativement homogène. Le but était évidemment de détecter des différences et des similitudes entre nos corpus, sans faire de choix a priori sur le vocabulaire étudié[54]. Une fois le tableau obtenu, passé en codage logique, on réalise une série d'analyses factorielles afin d'observer comment se distribuent les corpus explorés. Au passage, on rappelle que l'analyse factorielle des correspondances n'est pas une technique nouvelle, mais, découverte par Jean-Paul Benzécri dans les années 1970, elle est désormais intégrée au corpus des outils du data mining, dont, *a posteriori*, elle fait indubitablement partie. Dans le cas présent, outre ses capacités heuristiques liées à ses propriétés graphiques, cette technique possède l'insigne l'avantage de ne pas (ou peu) être soumise aux effets de corpus[55].

[54] Voir A. Salem/P. Lafon, L'inventaire des segments répétés d'un texte, in: Mots 6 (1983) p. 161–177; A. Salem, Segments répétés et analyse statistique des données textuelles, in: Histoire & Mesure 1 (1986) p. 5–28.

[55] Or, ces effets sont une difficulté structurelle lorsqu'on examine des corpus. Par «structurelle», on entend ici que ces effets sont en fait le résultat de structures historiques: en particulier les différences de quantité de production documentaire d'un établissement à un autre. Il ne s'agit donc pas de les effacer, ce qui est par ailleurs impossible, mais de les contourner. Or, depuis les travaux de Zipf/Mandelbrot, mais plus encore ceux d'Harald Baayen (cités plus haut), nous savons que les distributions lexicales n'ont rien de commun avec les distributions normales. Les difficultés liées à l'inégale production documentaire sont ainsi multipliées par la nature complexe et évolutive des distributions lexicales, car les distributions de mots sont sensibles à la taille de l'échantillon. Quand les corpus n'ont pas la même taille – ce qui est, en fait, la situation standard –, les proportions relatives des mots au sein de ces ensembles ne sont pas les mêmes. L'analyse factorielle ne solutionne pas totalement ce lourd problème, loin s'en faut, car elle se base elle aussi sur des calculs à partir d'écarts. Mais parce qu'elle considère plus ou moins indépendamment les lignes et les colonnes du tableau, elle «isole», en quelque sorte, chaque profil/élément du corpus et le rend comparable avec les autres. C'est pour cela qu'un traitement du tableau est nécessaire, car il renforce cette propriété de l'AFC, en l'éloignant d'une analyse basée sur des concepts mathématiques tels que la moyenne, le pourcentage, la variance, l'écart-type, tous parfaitement impropres dans le cas de distributions lexicales. Sur les effets de corpus, voir: I. Rosé, À propos des *Chartæ Burgundiæ Medii Ævi* (C.B.M.A.). Éléments de réflexion à partir d'une enquête sur la dîme en Bourgogne au Moyen Âge, in: Bulletin du Centre d'études médiévales d'Auxerre 12 (2008), disponible en ligne: <http://cem.revues.org/document6822.html>; ead., Enquête sur le vocabulaire et les formulaires relatifs à la dîme dans les chartes bourguignonnes (IX^{e}-XII^{e} siècle), in: La dîme, l'Église et la société féodale, éd. M. Lauwers (2012) p. 191–234; nous nous permettons de citer l'article que nous avons presque essentiellement consacré à cette question: N. Perreaux, La production de l'écrit diplomatique en Bourgogne.

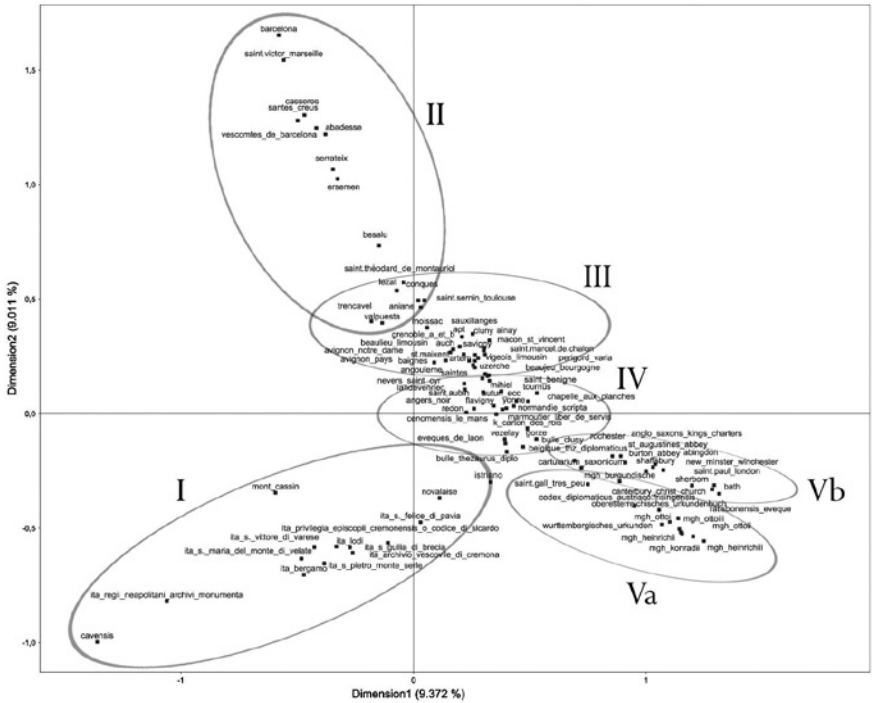

Fig. 3: Analyse factorielle (axe 1–2) du tableau généré par Text-to-CSV, à partir d'une sélection de corpus présents de notre base (900–1050).

Ainsi, sur cette analyse automatique du vocabulaire des actes, nous voyons plusieurs groupes se former. Tout d'abord en bas à gauche, on peut remarquer que l'ensemble des documents provenant de la péninsule italique se sont regroupés: actes de la Trinité de Cava[56], chartes lombardes du Codice Diplomatico della Lombardia Medievale[57], de Naples[58], ou encore du Mont-Cassin[59] (groupe I). À l'opposé, en haut sur l'axe 2, mais toujours à gauche sur l'axe 1, émerge un groupe formé les documents catalans: Cathédrale de Barcelone[60],

[56] Codex diplomaticus Cavensis, éd. M. Morcaldi/M. Schianni/S. De Stephano (1873–1970).

[57] Il s'agit de documents issus du CDLM (voir note 6).

[58] Monumenta regii Neapolitani archivi edita ac illustrata (1845–1861).

[59] Documents inédits ou peu connus des archives du Mont-Cassin, éd. E. Cuozzo/J.-M. Martin, in: Mélanges de l'École française de Rome – Moyen Âge 103 (1991) p. 115–210; Les premiers contrats agraires du Mont-Cassin. Les *livelli* de l'abbé Aligerne dans les Abruzzes (948–985), éd. L. Feller, in: Histoire et sociétés rurales (2004) p. 133–185.

[60] Diplomatari de l'Arxiu Capitular de la Catedral de Barcelona – Segle XI, éd. J. Baucells i Reig et al. (2006).

San Pere de Casseres[61], Serrateix[62]... mais aussi Saint-Victor de Marseille[63] (groupe II). Cet ensemble est d'ailleurs contigu à un autre, un peu plus en bas à droite, et qui contient cette fois les documents de la France du sud-ouest (Montauriol[64], Aniane[65], Saint-Sernin de Toulouse[66], Moissac[67]), mais aussi, plus loin, du centre (Sauxillanges[68], Cluny[69], Mâcon[70], Savigny[71], ainsi que Baignes[72], Angoulême[73]) (groupe III). Le paquet central dérive ensuite encore un peu plus vers le nord, et nous voyons ainsi se regrouper automatiquement Saint-Aubin d'Angers[74], le Cartulaire noir de la cathédrale d'Angers[75], Marmoutier[76], Le Mans[77], etc., ainsi que des régions telles que la Normandie (groupe IV). Enfin, à l'opposé des deux premiers groupes sur l'axe 1, apparaît un dernier ensemble qui en contient, en définitive, deux autres: d'une part les documents impériaux ainsi que ceux qui proviennent, d'une manière plus générale, de l'Empire[78] (groupe Va). D'autre part, les documents anglo-saxons:

---

[61] Col·lecció diplomàtica de Sant Pere de Casserres, éd. I. Llop (2009).

[62] Diplomatari del monestir de Santa Maria de Serrateix (segles X-XV), éd. J. Bolòs i Masclans (2006).

[63] Cartulaire de l'abbaye de Saint-Victor de Marseille, éd. B. Guérard (1857).

[64] Cartulaire de Saint-Théodard de Montauriol, éd. D. Panfili. Nous remercions chaleureusement l'auteur de nous avoir confié une copie de son travail.

[65] Cartulaires des abbayes d'Aniane et de Gellone: publiés d'après les manuscrits originaux, éd. L. Cassan/E. Meynial (1900).

[66] Cartulaire de l'abbaye de Saint-Sernin de Toulouse (844–1200), éd. C. Douais (1887).

[67] Recueil des actes de l'abbaye de Moissac [680]–1175, éd. R. La Haye (2001).

[68] Cartulaire de Sauxillanges, éd. H. Doniol (1864).

[69] Recueil des chartes de l'abbaye de Cluny, éd. A. Bernard/A. Bruel (1876–1903).

[70] Cartulaire de Saint-Vincent de Mâcon: connu sous le nom de Livre enchaîné, éd. M.-C. Ragut (1864).

[71] Cartulaire de l'Abbaye de Savigny suivi du petit cartulaire de l'Abbaye d'Ainay, éd. A. Bernard (1853).

[72] Cartulaire de l'abbaye de St.-Étienne de Baigne, (en Saintonge), éd. P. F. Cholet (1868).

[73] Cartulaire de l'église d'Angoulême, éd. J. Nanglard (1900).

[74] Cartulaire de l'Abbaye de Saint-Aubin d'Angers, éd. A. Bertrand de Broussilon (1903).

[75] Cartulaire noir de la Cathédrale d'Angers, éd. C. Urseau (1908).

[76] Le livre des serfs de l'abbaye de Marmoutier (1865).

[77] Cartulaire de l'abbaye de Saint-Vincent du Mans (Ordre de Saint-Benoît), Premier cartulaire, 572–1188, éd. C. Charles (1886).

[78] Codex chronologico-diplomaticus episcopatus Ratisbonensis, éd. T. Ried (1816); Württembergisches Urkundenbuch (1849–1913), version en ligne: <www.wubonline.de>; documents du site Monasterium.net et des dMGH, etc.

Rochester[79], Christ Church de Canterbury[80], Saint-Paul de Londres[81], etc. (groupe Vb). Cette expérience tout à fait grossière, très facilement perfectible, fait donc immédiatement ressortir la géographie des corpus et indique, par là même, que leurs spécificités lexico-sémantiques se répartissent plus ou moins géographiquement. Schématiquement, on voit s'opposer une Europe du sud, et une Europe du nord, même si, à une échelle plus fine, il y a fort à parier que des ruptures beaucoup plus complexes et nuancées, des gradients inattendus, des isolats aussi, apparaissent. Précisons au passage, sans pouvoir pour autant développer ce point clé, que ce n'est pas, dans notre expérience, le vocabulaire toponymique ou anthroponymique qui «classe» les documents en ensembles géographiques, mais bien des parties de formules, des termes médiolatins endémiques, ou encore du lexique à forte consonance vernaculaire[82].

[79] Charters of Rochester, éd. A. Campbell (1973).

[80] Charters of Christ Church, Canterbury, éd. N. P. Brooks/S. E. Kelly (2010).

[81] Charters of St Paul's, London, éd. S. E. Kelly (2004).

[82] Notre thèse en cours, intitulé *l'Ecriture du monde. Dynamique du féodalisme et perception du* mundus*: essais statistiques sur les champs sémantiques (*aqua/terra*) dans les bases de données (750–1350)*, reviendra très largement sur cette question. Par ailleurs, nous avons co-organisé, avec Coraline Rey, dans le cadre des journées du programme CBMA, une journée d'étude consacrée au «Vocabulaire courant en diplomatique» (CBMA 7). Au cours de celle-ci, nous sommes largement revenus sur cette question de l'ancrage géographique des textes diplomatiques, à partir du vocabulaire dit «courant» donc. On a alors pu formuler les trois hypothèses suivantes: a) Le stock de vocabulaire très courant – dans le cas où le qualificatif est entendu au sens d'une omniprésence spatiale –, est relativement réduit. b) Dans ce domaine, il existe bel et bien des emplois régionaux. c) Ce sont les associations entre les termes courants qui jouent un rôle dans la formation des entités géographiques. D'autres analyses, portant sur des termes courants à l'échelle d'une période précise, dans les chartes de Cluny (par exemple *increpitus*), mais presque totalement absents du reste de la production diplomatique, ont permis de montrer qu'une partie du lexique ne se rencontre que dans certaines zones: en s'inspirant de la biogéographie, nous avons proposé de les qualifier d'«endémiques». En définitive, l'examen des zones ainsi révélées nous mène à une interrogation sur le lien entre la fréquence d'un terme et sa sémantique, cette dernière émergeant d'abord d'une pratique d'écriture (donc d'une pratique sociale) régionale. Nous nous permettons de renvoyer à l'annonce et au compte-rendu détaillé de cette journée: N. Perreaux/C. Rey, CBMA. – *Chartae Burgundiae Medii Aevi.* 7. Le «vocabulaire courant» en diplomatique: techniques et approches comparées, in: Bulletin du Centre d'études médiévales d'Auxerre 16 (2012), disponible en ligne: < http://cem.revues.org/12513>; iid., CBMA – *Chartae Burgundiae Medii Aevi.* 7. «Le "vocabulaire courant" en diplomatique: techniques et approches comparées», in: ibid. 17 (2013), à paraître en ligne.

Une meilleure formalisation est nécessaire pour passer de l'éclatement à une exploitation concrète

Cette étude sommaire – et, encore une fois, aisément perfectible –, portant sur une large part du vocabulaire contenu dans ces corpus pour la période 900–1050, semble montrer que les spécificités lexicales sont avant tout régionales, c'est-à-dire liées à des zones géographiques. Les zones observées ici, même si elles restent très largement à déterminer, pourraient être relativement vastes. Surtout, à une échelle plus fine, il y a fort à parier que de nouvelles oppositions apparaissent. Schématiquement, on voit s'opposer une Europe du sud, – contenant elle-même des groupes foncièrement différents, la Péninsule Ibérique s'opposant par exemple à la Péninsule Italique –, et une Europe du nord, toutes deux très différentes. Bien entendu, nous ne pouvons en rester à cette exploitation sommaire. Il s'agirait maintenant d'affiner ces remarques très rapides, mais aussi de comprendre ce qui oppose sémantiquement ces différentes zones, dans une perspective d'histoire sociale. C'est ce second point qui nous ferait entrer dans une étude beaucoup plus poussée au plan historique. On reprendrait alors les classements réalisés dans notre seconde partie, afin d'en tirer des résultats. Reste que, pour nous, ces deux expériences, sur la typologie documentaire et sur sa régionalisation, forment en fait un tout. Il s'agit en effet de mieux saisir comment se structure la production documentaire, typologiquement, chronologiquement, spatialement. C'est à notre sens une étape préalable si l'on souhaite construire une diplomatique numérique solide: de l'accumulation à l'indexation, puis enfin à l'exploitation, il faut penser la totalité des étapes comme un dispositif technique et méthodologique englobant. Ainsi, les études sur les champs sémantiques, les études recourant à la linguistique de corpus, si elles semblent extrêmement prometteuses, ne pourront être réalisées qu'à partir d'ensembles de chartes plus ou moins homogènes, au plan chronologico-spatial. Il s'agirait donc, dans un premier temps, de mieux saisir la structuration interne de la documentation médiévale conservée, et surtout de considérer cette étape comme un préalable indispensable qui ne peut se résumer à une présentation des «sources» disponibles. Dans l'optique de l'exploitation maintenant, bien qu'ils soient d'ores et déjà capables de nous rendre d'importants services, il y fort à parier que les logiciels disponibles ne pourront suffire aux besoins, plutôt spécifiques, des historiens. Notre jeune expérience de la *digital diplomatics* montre que le médiéviste doit bien souvent s'armer de courage et modifier/programmer lui-même ses outils propres s'il souhaite exploiter la masse documentaire numérisée d'une manière plus efficace.

Appendices

Abstracts

I. Technical and theoretical models

BENOÎT-MICHEL TOCK

La diplomatique numérique, une diplomatique magique?

Les nouvelles possibilités offertes par les techniques modernes ont donné à la diplomatique un aspect que l'on pourrait qualifier de «magique». L'on accède facilement à un nombre de documents encore impensable il y a quelque années de cela, l'on dispose d'outils renouvelés pour leur étude ... Pour autant, l'on peut encore améliorer la chose, en affinant les instruments ainsi mis à notre disposition, et en profitant de ces nouvelles possibilités pour ne pas hésiter à donner des éditions et des outils différenciés, avec des niveaux d'élaboration correspondant aux situations et aux besoins. Il convient en outre de s'assurer de la pérennité des données ainsi produites, et de ne pas négliger pour autant les compétences de base qui sont traditionnellement celles du diplomatiste.

Digital Diplomatics: Magical Diplomatics?

The new possibilities which modern technologies offer have given diplomatics what might be called a "magic touch". One has easy access to a number of documents which would have been absolutely inconceivable a few years ago. We now have renewed tools at our disposal for their study ... Yet, things could still be improved by making these tools more efficient; the opportunity these new possibilities give us could be taken to resolutely create differentiated editions and tools, whose degree of elaborations would correspond to the situations and needs encountered. Moreover, the durability of the data which is thus produced is a question which should be considered; and for all this, the basic competencies which are traditionally those of a diplomatist should not be neglected.

CAMILLE DESENCLOS, VINCENT JOLIVET

Diple, propositions pour la convergence de schémas XML/TEI dédiés à l'édition de sources diplomatiques

La complexité des normes de l'édition critique et la singularité de chaque projet éditorial semblent faire obstacle aux tentatives de standardisation que l'informatisation exige. La TEI, davantage adaptée aux sources littéraires que diplomatiques, permet de définir des schémas très différents et parfois difficilement interopérables pour des projets similaires. Solution alternative, la CEI (Charters Encoding Initiative) promeut des éléments spécifiques, au risque d'empêcher l'interopérabilité avec d'autres types de sources. L'École des chartes, à travers le développement de Diple, s'est engagée dans une voie médiane qui consiste à définir des schémas XML conformes à la TEI et dédiés à l'édition critique et à ses spécificités (tableau de la tradition, apparat critique, etc.).

Diple, en associant systématiquement à ses schémas des fonctionnalités pour les exploiter (affichage HTML, exports divers), facilite la mise en ligne des corpus, résout les problèmes de présentation des structures textuelles récurrentes et permet à l'éditeur de se concentrer sur les particularités de son corpus (développement des abréviations, repérage des parties du discours, etc.). Par ses schémas partagés et ses outils d'édition, Diple pourrait permettre une convergence des usages de la TEI pour l'édition de sources diplomatiques. Une telle convergence ouvrirait des perspectives scientifiques en facilitant l'interrogation croisée des corpus édités.

Diple, a proposal for the convergence of XML/TEI schemas dedicated to the edition of diplomatic sources

The complexity of norms for a critical edition and the singularity of each editorial project seem to prevent any standardisation required by the computerisation. TEI, better adapted to literary texts than to diplomatic sources, allows, even for projects similar to each other, the building of very different schemas which are thus hardly interoperable. Another option, CEI (Charters Encoding Initiative) enhances specific elements but could impair interoperability with other kind of sources. The École nationale des chartes, through Diple's development, has chosen a middle path: defining XML schemas which are consistent to TEI and dedicated to specific elements of critical edition (tradition, critical apparatus, etc.).

The systematical association between Diple's schemas and the functionalities intended to exploit them (HTML display, various exports) facilitates corpora uploading, resolves the display problems of recurrent text structures and allows the editor to concentrate on his corpus' specificities (abbreviations development, diplomatic formulas, ...). Its shared schemas and edition tools could allow a convergence of TEI practices for the edition of diplomatic sources. Such a convergence would make cross-searching within edited corpora easier, thus opening new scientific perspectives.

FRANCESCA CAPOCHIANI, CHIARA LEONI,
ROBERTO ROSSELLI DEL TURCO

Codifica, pubblicazione e interrogazione sul web di *corpora* diplomatici per mezzo di strumenti *open source*

La consultazione di testi diplomatici costituisce uno strumento di lavoro insostituibile per gli storici e gli archivisti. La loro disponibilità online offre il massimo della flessibilità e della diffusione, permettendo allo studioso di accedere a questo materiale prezioso senza barriere spaziali o temporali: alcuni progetti, come *The Electronic Sawyer* (http://www.esawyer.org.uk/) e l'attività della *École nationale des chartes* (http://www.enc.sorbonne.fr), mostrano come sia possibile offrire testi di alta qualità scientifica sul web usando una codifica XML delle fonti. La loro creazione, tuttavia, richiede risorse non indifferenti: è possibile digitalizzare e mettere online questo materiale, per le proprie ricerche e per il beneficio della comunità accademica, in maniera (relativamente) semplice ed efficace? Inoltre una pubblicazione sul web è incompleta se non consente un'agevole consultazione e il *data mining* delle risorse offerte: come facilitare l'accesso e la ricerca all'interno dei testi?

Questo articolo si propone di mostrare come, grazie all'uso di software *open source*, il singolo studioso o un piccolo team di ricercatori possa digitalizzare un corpus di documenti usando il formato TEI (http://www.tei-c.org/), pubblicarlo sul web e inserire nell'interfaccia un motore di ricerca come eXist (http://exist-db.org/) o XTF (http://xtf.cdlib.org/) per effettuare ricerche complesse.

Open source tools for online publication of charters

Diplomatic texts consultation is an indispensable tool for historians and archivists. Their online availability offers maximum flexibility and dissemination, allowing the scholar to access this valuable material unimpeded by spatial or temporal barriers: some projects, such as *The Electronic Sawyer* (http://www.esawyer.org.uk/) and the encoding activities of the *École nationale des chartes* (http://www.enc.sorbonne.fr), show how it is possible to offer high quality scientific texts on the Web based on an XML mark-up of the archival sources. Their creation, however, requires substantial resources: is it possible to digitise and to put online this type of document collections for personal research, or to benefit the whole Academic community, in a (relatively) simple and effective way? Furthermore, Web publishing is only effective if it allows easy document browsing and a way to perform powerful data mining of the resources it offers: which methods are best to allow easy text access and search?

This paper aims to show how, through the use of open source software, single scholars or a small team of researchers can encode a corpus of documents using the TEI (http://www.tei-c.org/) standard, publish it on the Web, and provide a search engine such as eXist (http://exist-db.org/) or XTF (http://xtf.cdlib.org/) for complex queries.

SERENA FALLETTA

Dalla carta al bit. Note metodologiche sull'edizione digitale di un cartulario medievale

Il testo presenta un modello ipertestuale e sperimentale di edizione di fonti storiche codificate adottando la sintassi XML e uno schema di marcatura calibrato sulle caratteristiche specifiche della documentazione e sulle classiche esigenze e categorie di analisi critica della ricerca storica, nel tentativo di superare i problemi concettuali e metodologici riscontrati dagli storici nell'utilizzo dei database, inadeguati ad esprimere la complessità di dati qualitativi.

L'esperimento condotto, basato sull'edizione quattrocentesca del *Liber Privilegiorum Sanctae Montis Regalis Ecclesiae*, è stato pensato come un laboratorio attraverso cui comprendere con più precisione le profonde implicazioni epistemologiche dell'applicazione delle tecnologie informatiche nel trattamento delle fonti storiche. Superando il tradizionale discorso storico, caratterizzato da linearità, l'ipertesto codificato è stato infatti in grado di incorporare la fonte e gli strumenti d'indagine, riconfigurando così strumentazione tecnica e prassi accademica. Si tratta, chiaramente, di una *sperimentazione sostenibile,* basata su un'ampia ragnatela ipertestuale i cui nodi sono sia i documenti prodotti da uno specifico soggetto istituzionale che la loro storia.

From paper to bit. Methodological notes on a medieval cartulary digital edition

The paper introduces a model of an historical sources hypertext edition encoded adopting the XML syntax and a markup pattern tailored to the specific characteristics of the documentation and the categories of critical analysis of historical research. The aim of the project is the attempt to overcome the conceptual and methodological problems experienced by historians in the use of databases, inadequate to express the complexity of qualitative data.

The experiment, based on the issue of *Liber Privilegiorum Sanctae Montis Regalis Ecclesiae*, was designed as a workshop for the comprehension of the deep epistemological implications of computer technology in the treatment of historical sources. Going beyond the traditional historical discourse, characterised by linearity, the encoded hypertext has been able to incorporate source and survey instruments, thus reconfiguring the technical equipment of academic practice. It is, of course, a sustainable experiment, based on extensive web hypertext whose nodes are both the documents produced by a specific institution and their story.

GUNTER VASOLD

Progressive Editionen als multidimensionale Informationsräume

Dynamischen Editionen ermöglichen Änderungen nach dem Zeitpunkt ihrer Publikation. Dadurch werden einige Grundprinzipien von Edition in Frage gestellt, etwa dass eine Edition eine stabile und zitierbare Ressource darstellt, oder dass es nur einen oder einige wenige Editoren gibt, die autoritative Textversionen produzieren. Einen möglichen Ausweg bietet die progressive Edition. Diese ist grundsätzlich offen und ohne definiertes Ende gedacht, kann also jederzeit verändert und erweitert werden. Voraussetzung dafür ist, dass die Edition nicht nur als Produkt gesehen wird, sondern als Prozess, der zu einer stetig wachsenden Menge von Teilresultaten und Versionen führt. Daraus lassen sich statische (und damit zitierbare) Editionen ableiten; diese sind aber immer nur Momentaufnahmen, die den Status zu einem bestimmten Zeitpunkt festhalten. Resultate, Versionen und die vielfältigen Beziehungen zwischen diesen bilden einen komplexen Informationsraum, der sich mit jeder Änderung der Edition erweitert. Die Beherrschung dieses Raumes ist nur möglich, wenn zusätzlich Prozessdaten generiert werden, die alle Änderungen dokumentieren. Auf Basis dieser Prozessdaten können Beiträge überprüft, bewertet, verwaltet und für bestimmte Ansprüche aufbereitet werden. Prozessdaten sind somit wesentlicher Bestandteil einer progressiven Edition, weil sie deren Dynamik für (Mit)Editoren und Benutzer transparent machen.

Progressive editions as multidimensional information space

Dynamic forms of critical editions allow the modification of texts after the edition has been published. But how can we implement dynamic editions without breaking basic principles of critical editions, such as the assumption that an edition does not change after publication or that an edition has only one or very few editors who produce a single authoritative text version? The concept of progressive editions might be a solution, because it does no longer conceive of critical editions as products, but as perpetual processes which lead to accumulating sets of distinct results, versions, and relations between these. It is still possible to generate traditional editions, but these are merely snapshots, each representing a certain (citeable) state of the steadily growing and shifting progressive edition. Such an edition forms a multidimensional information space, which becomes more complex with every change made to the edition. Keeping track of a progressive edition requires putting a stronger focus on the processes involved. Therefore process data documenting the changes has to be considered a relevant part of the edition, as it provides opportunities to organise, analyse, and evaluate data, and to make progressive editions transparent to fellow editors and users.

LUCIANA DURANTI

The return of diplomatics as a forensic discipline

Research has proven that digital documents/records, whether born digital or digitised, cannot be preserved. It is only possible to maintain the ability to reproduce them time after time. The most complex aspects of this ongoing preservation involve those activities that aim to counteract system and format obsolescence or to extract documents from their original environment when obsolescence has occurred before any measure could be taken to avoid it. To maintain and assess the authenticity of entities that no longer exist in their native environment requires the strong theoretical and methodological framework which, for traditional documents, has been provided by diplomatics. Although such framework is still valid when examining documents in digital form, it is no longer sufficient, and needs to be integrated with a tested robust practice that allows the certain authentication of what we keep in digital form, such as the digital objects – metadata – we link to digitised medieval documents to make them accessible and analyse them. This article will discuss the integration of digital diplomatics with digital forensics, a discipline that originated a decade ago and developed into a rigorous body of concepts, principles and procedures used internationally to fight cybercrime and identify, retrieve, and make accessible authentic digital objects as evidence of the facts and acts they reveal or to which they attest.

Il ritorno della diplomatica come scienza forense

La ricerca ha dimostrato che i documenti digitali non si possono conservare. È solo possibile mantenere la capacità di riprodurli. L'aspetto più complesso di questo tipo di conservazione è estrarre i documenti dal loro ambiente nativo quando esso è diventato obsoleto. Per determinare o mantenere l'autenticità di oggetti digitali che non esistono più nel loro ambiente originale è necessario fare uso dei concetti e del metodo di analisi della diplomatica, che tuttavia devono essere integrati con le pratiche sviluppate da una nuova area di conoscenze, digital forensics. Questo articolo discute l'integrazione della diplomatica con digital forensics, una disciplina che ha avuto origine dalla necessità di combattere il cybercrime e rendere accessibili oggetti digitali come prova autentica degli atti e fatti che essi attestano.

II. Projects for the edition of texts and the publication of information

DANIEL PIÑOL ALABART

Proyecto ARQUIBANC – Digitalización de archivos privados catalanes: una herramienta para la investigación

En Cataluña se conserva un notable patrimonio documental privado que es fundamental para estudiar la historia del país. La mayoría de archivos privados se conservan en manos privados, aunque algunos están depositados en archivos públicos. Pero uno de los mayores problemas que revisten estos fondos documentales es el difícil acceso para los investigadores. Atendiendo a esta situación el proyecto ARQUIBANC, gestionado desde el Departamento de Historia Medieval, Paleografía y Diplomática de la Universitat de Barcelona tienen como objetivo facilitar el acceso a algunos de estos archivos. Para ello el proceso de investigación sigue diferentes pasos entre los que destaca la digitalización de algunos documentos. Las imágenes de éstos se introducen en unas bases de datos accesibles desde internet con la descripción de cada documento. Para poder llevar a cabo la descripción archivística es necesario un estudio individualizado de cada documento, incluyendo estudios diplomáticos y paleográficos, así se puede dotar de información cada una de las fichas introducidas en las bases de datos.

The ARQUIBANC project – Digitisation of private archives in Catalonia: a research tool

Catalonia preserves a significant private documentary heritage which is essential for studying the country's history. Most of the private archives remain in private hands, although some are placed in public archives. One of the biggest problems of these documents is the difficult access for researchers. In response to this situation, the ARQUIBANC project, managed by the Department of Medieval History, Palaeography and Diplomatics of the University of Barcelona, intends to give access to some of these files. The research process follows different steps, among which is the digitisation of some documents. These images are entered into a database available on the Internet with the description of each document. To create the archival description, and in order to provide each of the cards introduced in the databases with the necessary information, an individualised study of each document is necessary, including diplomatic and palaeographical study.

ANTONELLA GHIGNOLI

Sources and persons of public power in 7th–11th century Italy
The idea of *Italia Regia* and the *Italia Regia* project

Italia Regia is an on-line system whose purpose is the digitalisation and restitution on the web of the documents issued by the rulers (kings, emperors) in the Italian kingdom between the 7th and the end of the 11th century. Its elaboration is the result of an international cooperation involving French, Italian and German researchers. The core of the work is a relational and dynamic database enabling corrections and amendements at any time, providing four types of forms: royal charters, records of legal disputes, individuals, and bibliography. Since the technical infrastructure was created in recent years of the project and is already operational (tested successfully on the documentation of Tuscany), the project now intends to cover the whole of the north of Italy (the main region of the Italian kingdom), in order to give access to new information about all public documents during the period. By using this tool, researchers will be able to discern a number of phenomena simultaneously: 1) the global distribution, on the territory of the Italian kingdom, of the documentation issued by the royal power or by public agents; 2) the recipients as a whole of the entirety of the preserved public documents; 3) the individuals as a whole involved in the genesis of public writings, at any level, with or without an official title. Finally, *Italia Regia* offers a technical and scientific model which could be extended to other European regions.

Les sources et les personnages de la puissance publique dans l'Italie des VIIe-XIe siècles. L'idée et le projet de l'*Italia regia*

Italia Regia est un système en ligne pour la numérisation et la restitution sur le web des documents émis par le pouvoir royal ou impérial dans le royaume d'Italie entre le VIIe et la fin du XIe siècle. Conçu dans le cadre d'une collaboration entre des équipes française, italienne et allemande, il fonctionne grâce à une banque de données relationnelle dont les informations sont en permanence modifiables, mettant en jeu quatre catégories de fiches: les diplômes des souverains, les comptes rendus judiciaires, les personnes, la bibliographie. Disposant déjà d'une infrastructure technique et ayant déjà couvert la Toscane, le projet compte étendre le traitement du matériel à l'ensemble de l'Italie du Nord (c'est-à-dire la région la plus importante du royaume), de manière à donner accès à des informations renouvelées sur l'ensemble de la documentation publique pour la période considérée. Comme instrument de recherche, il permettra de percevoir plusieurs phénomènes en même temps: 1) la distribution globale, sur le territoire du royaume d'Italie, de la documentation émise par le pouvoir royal ou délégué; 2) l'ensemble des destinataires de toute la documentation publique conservée; 3) l'ensemble des personnes qui, à un niveau ou un autre, avec ou sans titre, sont impliquées dans la genèse des écritures publiques. *Italia Regia* fournit enfin un modèle technique et scientifique applicable à d'autres régions européennes.

RICHARD HIGGINS

The Repository view. Opening up medieval charters

As custodians of a broad range of collections we require a system that enables cataloguing and presentation of related digital material that is flexible enough to cope with all materials. We have been using EAD as a data storage format since 1996, and in combination with our Fedora-Commons digital repository we have a powerful, adaptable tool. It is imperative that one system includes all our collections, so that enhancements and migrations apply to the whole and do not break or drop the more complex data. EAD has proven hugely adaptable and scalable, ranging from brief description of collections to the calendaring of individual charters. As one of hundreds of collections in our care, the archive of Durham Priory and Cathedral includes thousands of charters, as well as cartularies and a full range of medieval documents. EAD has been able to accommodate descriptions of all of these – even 3,000 seals. The digital repository also stores images and transcripts of the documents. It enables the association of description and image using index terms, hyperlinks, and RDF, producing a more permanent linking between data within the catalogue, and offering researchers the ability to make reliable citations of online representations of individual documents. This enables investigation of not just additional versions of the charter, but also other documents witnessed by the same parties and other common features.

ŽARKO VUJOŠEVIĆ, NEBOJŠA PORČIĆ,
DRAGIĆ M. ŽIVOJINOVIĆ

Das serbische Kanzleiwesen. Die Herausforderung der digitalen Diplomatik

Die zunehmende Anwendung von Informationstechnologien bei der wissenschaftlichen Bearbeitung und Darstellung der Urkunden findet die serbische Diplomatik in der Lage vor, immer noch keine einheitliche Edition ihres nicht allzu umfangreichen Bestands von mittelalterlichen Dokumente zu haben. In diesem Zusammenhang stellt sich die Frage, ob die Entscheidung zugunsten der Herstellung eines digitalen Urkundenkorpus statt einer gedruckten Edition die bessere Lösung wäre. Hinsichtlich der Einfachheit, Schnelligkeit und Zuverlässigkeit des Datenzugangs würde ein solches Ergebnis die Benutzbarkeit der diplomatischen Quellen bei der Forschung aller Aspekte des serbischen Mittelalters erheblich erhöhen. Der wichtigste Beitrag des *born digital* Vorhabens ist in der Anregung zu sehen, die in der serbischen Diplomatik bisher oberflächlich geklärten Schlüsselfragen des Kanzleiwesens, d. h. des Entstehungsprozesses von Urkunden, leichter erforschbar zu machen. Durch eine Darstellung der Daten in nach einzelnen Urkundenelementen aufgegliederten Feldern der Datenbank und ihre kombinierte Durchsuchbarkeit könnte induktiv die bürokratische Formalitätsstufe der serbischen diplomatischen Produktion untersucht werden. Als Ergebnis wäre festzulegen, inwiefern dabei von einem System die Rede sein kann, also von einer Einrichtung, die die Forschung „Kanzlei" zu nennen pflegt, oder es sich um für die mittelalterliche Gesellschaft typische Improvisation und *ad hoc* Lösungen handelte.

The medieval Serbian chancery. Challenges of digital diplomatics

At a time when digital information technologies are increasingly entering the field of processing and presentation of documentary heritage, Serbian diplomatics still has not published its comparatively modest corpus of medieval documents within one all-encompassing collection. The question therefore arises as to whether it would be better to shift the efforts aimed at producing a traditional printed edition towards producing a digital one. Simple, quick and reliable access to digitised data would considerably enhance the use of diplomatic sources in research of almost all aspects of medieval Serbian past. Most importantly for Serbian diplomatics as a scholarly discipline, this *born digital* project would give strong encouragement to dealing with the inadequately studied key issues relating to the functioning of the chancery, that is, the process of document creation. Fitted into separate cross-searchable fields of a digital database, the ample and diverse information about diplomatic features of Serbian charters would enable a comprehensive inductive study of the actual degree of bureaucratic formalisation of Serbian documents. Through this, it should be possible to establish whether they were indeed products of a system, represented by an institution commonly termed 'the chancery', or improvisation and *ad hoc* solutions typical of medieval society.

ALEKSANDRS IVANOVS, ALEKSEY VARFOLOMEYEV

Some approaches to the semantic publication of charter corpora
The case of the diplomatic edition of Old Russian charters

The paper discusses the basic principles that can be applied to semantic publications of electronic scholarly editing of charter corpora. In order to reveal the advantages of such editions in diplomatics and historical research on medieval charters, it presents a multifunctional prototype of a semantic publication of the 13th-century Old Russian charter corpus – a constituent part of the vast collection of medieval and early modern records "Moscowitica – Ruthenica" in the Latvian State Historical Archives (Riga). It uses the «Semantic MediaWiki» software which provides a special markup for semantic links. In addition the paper explores the possiblities of translating the tenor of the charters into Attempto Controlled English. The prototype of the semantic publication is designed as a comprehensive diplomatic edition of Old Russian charters; the transcription of the texts represents palaeographic features of the charters. At the same time, the semantic edition provides the texts with additional information. Firstly, information about persons, sites, events, etc. provided by the charters is given and linked with corresponding data reflected in different specialised ontologies. As a result, the semantic publication creates a specific model of historical reality. Secondly, the semantic publication provides appropriate tools for an in-depth pattern analysis of the charters, because it is based on a detailed markup of the texts. Thus, the semantic edition is designed as a specific Web information system that incorporates medieval charters, research tools, and research results into a knowledge-based system, which is specially created for a network community.

Quelques approches de la publication sémantique des corpus des chartes
Le cas d'une édition diplomatique des anciennes chartes russes

Cet article discute des principes de base des publications sémantiques et leur application à l'édition savante électronique d'un corpus des chartes. Afin de révéler les potentialités de ces éditions, l'article présente un prototype multifonctionnel d'une publication sémantique du corpus des anciennes chartes russes du XIII^{e} siècle. Ces chartes font partie intégrante de la vaste collection de documents historiques, médiévaux et plus modernes, 'Moscowitica–Ruthenica', conservée aux Archives Historiques d'État de Lettonie (Riga). Le prototype de la publication sémantique est conçu comme une édition diplomatique globale des chartes. Dans cet article, une attention particulière est accordée au système «Semantic MediaWiki», qui fournit des outils de balisage spéciaux pour la création de couches sémantiques dans des publications sémantiques. Dans le même temps, les auteurs affirment que la teneur des chartes peut être pleinement représentée dans le Web sémantique par le biais d'Attempto Controlled English. Le prototype développe une approche orientée sur le document de la représentation des données historiques récupérées à partir des chartes. Par conséquent, dans ce prototype, les relations entre les divers objets (lieux, personnes, événements, etc.) reflètent réellement les liens entre les documents qui fournissent des informations sur les objets. La publication sémantique, qui révèle ces relations, peut être utilisée comme base pour un système d'information Web spécifique qui intègre des textes de chartes, des outils et des résultats de recherche dans un système fondée sur connaissance, et fournit aux chercheurs des outils appropriés de critique des sources historiques.

III. Digital diplomatics in the work of the historian

ELS DE PAERMENTIER

Diplomata Belgica. Analysing medieval charter texts (dictamen) through a quantitative approach. The case of Flanders and Hainaut (1191–1244)

The ongoing process of digitisation has obviously not bypassed the auxiliary science of diplomatics. This contribution focuses on the opportunities offered by the digital source collection *Diplomata Belgica* to analyse medieval charter texts through a quantitative approach. Starting from a case study on the charters and chancery of the count(esse)s of Flanders and Hainaut during the period between 1191 and 1244, this contribution will explain how the diplomatic method of analysis elaborated by L. Delisle and Th. von Sickel (the so-called *Stilvergleichung*), and refined by Walter Prevenier in the early 1960s, was extended with a whole new dimension of word statistics within a corpus of over 16,000 digitised charter texts. Until now, the existing method of diplomatic analysis has been limited to a 'manual' comparison of the Latin protocol formulas, and only juxtaposed the text of the charters issued by the count(esse)s with other texts from the archives of their recipients. However, the new research criteria and standards which were developed, gathered into a so-called 'three step action plan of determination', made it possible for the first time also to draw the dispositive text parts into the analysis, and to examine them from a much more comparative and creative perspective. Consequently, this 'modern' methodology was not only elaborated in order to find out the editorial origin of the comital charter texts. Gradually, it also offered new insights on the editorial traditions and 'innovations' within the chancery of the count(esse)s, on the extent to which this chancery tended to differentiate itself from other secular or ecclesiastical editorial centres, and on the direct influence some important chancery clerks had on the organisation and the editorial customs within the administrative entourage of the count(esse)s.

Diplomata Belgica: analyser le dictamen des chartes médiévales au moyen d'une approche quantitative

Grâce à l'«évolution numérique», de nouvelles approches du dictamen des actes médiévaux deviennent possibles. En se focalisant sur la chancellerie et les actes des comtes et comtesses de Flandre et de Hainaut (1191–1244) retenus dans la base *Diplomata Belgica*, cette contribution vise à montrer que grâce aux *Diplomata Belgica*, la méthode de la *Stilvergleichung*, élaborée au XIX^{e} siècle par L. Delisle et Th. von Sickel, et affinée dans les années 1960 par Walter Prevenier, a pu être approfondie et adaptée aux nouveaux moyens de recherche. Dans le cadre de l'analyse du dictamen des actes comtaux, la procédure d'identification se déroule en trois phases, à savoir la recherche de fréquences significatives, l'ajout de critères additionnels de nature «quantitative» (et la confrontation avec des contre-épreuves) et, enfin, l'identification finale d'un acte comtal en tant que produit rédactionnel de la chancellerie. En outre, il devient clair qu'une approche «quantitative» basée sur le calcul de la fréquence des mots dans un ensemble d'environ 16.000 textes d'actes de la période 1191–1244 n'aboutit pas seulement à des résultats

bien fondés concernant la production rédactionnelle à la chancellerie comtale; en outre, elle nous renseigne sur les traditions diplomatiques en usage pendant les règnes successifs, sur les conséquences possibles d'une union personnelle de deux comtés concernant les tâches administratives, et même sur la mesure dans laquelle certains cadres actifs au sein de la chancellerie, ou du moins dans l'entourage administratif des comtes, ont exercé une influence personnelle sur la production des actes et sur les habitudes dominantes. Enfin, elle nous permet d'appréhender la manière dont la chancellerie comtale, en utilisant «ses» actes comme instruments, a tenté de se profiler et de se différencier par rapport à la concurrence d'autres centres de rédaction contemporains.

NICOLAS PERREAUX

De l'accumulation à l'exploitation? Expériences et propositions pour l'indexation et l'utilisation des bases de données diplomatiques

Depuis maintenant plusieurs décennies, les diplomatistes disposent de bases de données remarquables, dont le contenu est propre à révolutionner nos connaissances concernant le Moyen Âge. Pour autant, force est de constater que l'exploitation de ces vastes *corpus* reste encore largement à faire, les entreprises dans le domaine restant pour le moment embryonnaires. Il est donc intéressant de s'interroger sur l'origine des blocages structurels qui empêchent encore, à l'heure actuelle, l'utilisation massive de ces ressources. L'article fait l'hypothèse qu'une partie de ces problèmes naît d'une difficulté à concevoir le dispositif numérique comme un tout, impliquant à la fois le matériau, la création de schémas abstraits et techniques, mais aussi le questionnaire scientifique. Une thèse en cours, visant à exploiter une base de 150 000 chartes, servira d'exemple concret. On présentera un dispositif d'indexation/de classification automatique des actes, basé sur le data/text mining et l'intelligence artificielle. Une exploitation plus efficace du matériau numérisé passe en effet par une «mise en ordre» des actes, aussi bien au plan typologique, que géographique ou chronologique. On présentera aussi une première expérience, concernant la dynamique de la production diplomatique à l'échelle européenne, visant à mettre en lumière, à terme, les liens entre les différentes zones productrices de documents. Le but de cette démarche globale est de faire apparaître des structures restées invisibles à l'œil nu.

From data accumulation to data exploitation: proposals and experiments for indexing and using digital diplomatic databases

For several decades now, diplomatists have had at their disposal remarkable digitised databases, whose contents are now about to revolutionise our knowledge of the Middle Ages. However, it is clear that the exploitation of this vast *corpus* remains largely to be done, while experiments in this field are still in an embryonic state. It is time now to question ourselves about the structural obstacles that still prevent, at least until now, the massive use of these resources. This article makes the assumption that a part of these problems stems from the difficulty to conceive the numeric device as a whole, involving at the same time charters, the creation of an abstract, but also technical schemas, as well as the scientific enquiries. A PhD thesis in progress on the exploitation of

a corpus made from 150 000 charters will be used as a concrete example. I will describe a system dedicated to the automatic indexation/classification of charters, based on data/text mining and artificial intelligence. Indeed, a more efficient use of digitised material implies an "arrangement" of it, on the typological, geographical as well as on the chronological levels. I will also present a first experiment, regarding the dynamics of the documentary production, at the scale of medieval Europe, aimed at finally highlighting the links and discrepancies between the different areas of this production. The purpose of this global approach is to show structures that remained invisible to the naked eye.

GELILA TILAHUN, MICHAEL GERVERS, ANDREY FEUERVERGER

Statistical methods for applying chronology to undated English medieval documents

A primary objective of ongoing research at the DEEDS Project (Documents of Early England Data Set) at the University of Toronto is to develop statistical methods for the dating of undated English private (as opposed to royal) charters. Of such documents issued between 1066 and 1307, only about five per cent were dated internally. Researchers at DEEDS have developed a database of over 10,000 dated charters from the period and used it to recognize chronological differences in word order and vocabulary which, through the application of statistical methodology, have enabled us to establish a temporal „footprint" for undated charter sources. In this paper, we present two such statistical methodologies that rely on usage patterns of words and phrases, and on the notion of „distances" between documents. Both methods are computer automated and use the DEEDS database as their source.

In the first method, we define a notion of „distance" between two documents. A kernel weight on the distance between an undated document and a dated document is defined, and the date of an undated document is estimated as a weighted sum of the dates from the dated documents. The second method is based on estimating from dated documents the probability of occurrence of words and phrases through time.

The procedure for dating an undated document involves combining the estimated probabilities of the occurrences of words and phrases from the document evaluated at every point in time. The time value that maximises the combined probabilities is taken to be the date estimate for the undated document. These methods could also be adapted to a setting in which we are estimating features of documents that are not necessarily dates. We could, for example, use these methods to identify the religious house which composed a charter or the geographical location from which a charter originates.

This paper is a review of the research which has led to the publication of the papers by A. Feuerverger, M. Gervers, P. Hall and G. Tilahun as they appear in the bibliography of the contribution.

MICHAEL HÄNCHEN

Neue Perspektiven für die Memorialforschung. Die datenbankgestützte Erschließung digitaler Urkundencorpora am Beispiel der Bestände von Aldersbach und Fürstenzell im 14. Jahrhundert

Die Diplomatik erfährt durch die Möglichkeiten der modernen Informationstechnologie einen wesentlichen Wandel. Die digitale Fotografie großer Urkundenbestände ermöglicht einen raum- und zeitunabhängigen Zugriff auf große Urkundencorpora mittels hochauflösender Digitalisate. Das Ziel des vorgestellten Projektes ist die datenbankgestützte Untersuchung von Seelgerätstiftungen bestimmter sozialer Gruppen während des „krisenhaften" 14. Jahrhunderts im Bistum Passau.

Eine Datenbank erlaubt es uns, digitale Urkundenbestände für spezifische Fragestellungen weiter zu fokussieren. Für das Forschungsprojekt werden die Urkunden betrachtet, gelesen und die Informationen der digitalen Abbildungen in eine Erfassungsmatrix übertragen. Die Matrix ist den Erkenntnisperspektiven entsprechend in elf Kategorien untergliedert und bündelt formale sowie inhaltliche Aspekte der Digitalisate in Form von abfragbaren Datensätzen. Hierbei ist es weiterhin vorgesehen, die Informationen der Digitalisate zu formalisieren und zu normieren, um die Abfragen stets unter gleichen terminologischen Voraussetzungen zu ermöglichen. Um dies zu illustrieren, stellte der vorliegende Beitrag Auswertungsmöglichkeiten anhand der Zisterzienserklöster Aldersbach und Fürstenzell vor.

Hinsichtlich der formalen Urkundenaspekte kann konstatiert werden, dass alle Stücke den voll entwickelten Urkunden des Spätmittelalters mit Intitulatio, Publicatio, Dispositio, Corrobatio und Datatio entsprechen. Die vorherrschende Sprache ist die Volkssprache, und es wurden nur geringe Abweichungen bei den äußeren und inneren Merkmalen durch die sozialen Gruppen ersichtlich. Bezüglich des historischen Kontextes konnte ein signifikantes Absinken der Stiftungszahlen gegenüber beiden Häusern bereits vor dem Ausbrechen der Pest festgestellt werden. Leicht unterscheiden sich die Stiftungsmaterien und die sozialen Schichten der Stifter bei beiden Klöstern. So wurde Aldersbach vornehmlich mit dauerhaften und einmaligen Geldzuwendungen bedacht und Fürstenzell mit Zehntrechten. Nahezu gleich sind die Gaben an Naturaleinkünften bei beiden Klöstern. Während Aldersbach sowohl durch den niederen und fürstlichen Hochadel als auch durch Bürger verschiedener Städte bestiftet wurde, ist Fürstenzell nur durch den landsässigen Adel und Bürger der Stadt Passau bedacht worden.

Mittels der Untersuchung großer Quantitäten kann die Nutzung und Auswertung einer solchen Datenbank Tendenzen und Entwicklungen entlang einer spezifischen Fragestellung aufzeigen, und es ist ein vielversprechender Weg, digitale Urkundencorpora für die Memorialforschung zu öffnen.

New perspectives for commemoration research. The database-assisted development of digital charter corpora: the examples of the 14th-century collections of Aldersbach and Fürstenzell

With the capacities of modern information technology, diplomatic studies have experienced a fundamental change. The digital photography of great charter collections enables immediate research using high resolution images. The main goal of the present database-

assisted project is the investigation of the mentality of benefactors from different social groups in the diocese of Passau having issued commemoration charters during the 14th century, the so-called "century of crisis".

A database allows us to concentrate on digital charters for further specified evaluations. Such evaluations assess, read, and transfer the information contained in the digital images into a database matrix. Structured in 11 categories, the matrix registers formal and contextual aspects of the images as consultable data sets. With the help of this matrix, it is intended to formalise and to normalise the images as much as possible in order to query the database under similar terminological conditions. In order to illustrate this, the article presents selected examples based on two Cistercian monasteries, Aldersbach and Fürstenzell.

Concerning the formal aspects we can say that all pieces constitute fully developed charters of the late Middle Ages with *Intitulatio*, *Publicatio*, *Dispositio*, *Corroboratio* and *Datatio*. The preferred language is German with a few variations of the inner and outer features depending upon the social groups. Regarding contextual aspects we can say that there was a significant decrease of donations before the plague, and we have differentiated the type of donations and the benefactors. Aldersbach was granted more permanent and singular endowments in currency than Fürstenzell, which received more donations in tithe. Nearly equal are the number of natural product donations to both monasteries. While both the lower and higher nobility and citizens from cities donated to Aldersbach, only lower nobility and citizens from the city of Passau donated to Fürstenzell.

With a great number of charters, the use and evaluation of a database can show tendencies and developments according to a defined question, and is a useful way to open up digital corpora for further historical research.

MARTIN ROLAND

Illuminierte Urkunden im digitalen Zeitalter – Maßregeln und Chancen

Illuminierte Urkunden standen bisher – abgesehen von der Dotalurkunde für Theophanu von 972, die als UNESCO-Weltkulturerbe vorgeschlagen wurde – nur vereinzelt im Fokus der Forschung. Der Beitrag stellt zuerst einige charakteristische Beispiele vor und behandelt eine 1028 in Bari ausgestellte Hochzeitsurkunde ausführlicher, um die Möglichkeiten der interdisziplinären Erforschung illuminierter Urkunden exemplarisch vorzuführen.

Kernstück ist eine Definition des Untersuchungsgegenstandes. Anschließend werden der Stand der Forschung und die Schwierigkeiten bei der Auffindung illuminierter Urkunden in den Archiven thematisiert. Die ungeheuren Möglichkeiten, die die Digitalisierung großer Urkundenbestände bietet, werden hervorgehoben und einige Maßregeln zur Erschließung des Materials formuliert. Anschließend wird ein praktischer Versuch auf Grundlage von *monasterium.net* beschrieben und der Nutzen für Kunstgeschichte, Archive und Diplomatik beleuchtet. Abschließend werden die Zukunftsperspektiven benannt: Hervorzuheben ist die gesamteuropäische Dimension des Phänomens und das kreative Potential, das sich aus der impliziten Spannung zwischen der Urkunde als Rechtsdokument und dem Bild bzw. Kunstwerk mit dessen medienspezifischer Ausrichtung auf den Betrachter ergibt.

Illuminated charters in the digital age – rules and opportunities

With the exception of the dowry charter for Theophanu from 972 (proposed as UNESCO-World Heritage), illuminated charters have only rarely been the subject of scientific research. This paper presents some characteristic examples, dealing more extensively with a marriage contract (issued in Bari in 1028) to demonstrate the interdisciplinary potential of the matter.

The central point is a definition of what "illuminated charter" exactly means. Subsequent topics focus on the current state of research, and the problems of finding illuminated charters in the archives. The substantial opportunities offered by mass digitisation are highlighted alongside a number of rules to be observed. In addition, a practical test using *monasterium.net* are described, revealing the numerous benefits for the history of art, archives and diplomatic research. As a final point some future prospects are discussed: illuminated charters have a pan-European dimension and a creative potential resulting from the implicit tension between the charter as legal document and the image/work of art with its media-specific orientation towards the observer.

DOMINIQUE STUTZMANN

Conjuger diplomatique, paléographie et édition électronique. Les mutations du XII^{e} siècle et la datation des écritures par le profil scribal collectif

L'apparente unité graphique retrouvée de l'Occident latin au XII^{e} siècle et l'hégémonie incontestée de la minuscule caroline «prégothique» ouvrent aux scribes la voie de l'inventivité par artifice. Ils développent des systèmes graphiques variés et des profils scribaux collectifs émergent au sein de groupes restreints (chancelleries, scriptoria). Ceux-ci se caractérisent par leur emploi variable, mais spécifique, des différents allographes d'une même lettre. L'édition électronique permettant de conserver et montrer différentes formes d'un même texte, il est désormais possible d'éditer les textes médiévaux en en conservant l'information graphique ou allographétique et, en même temps, d'étudier les cohérences et les variations des systèmes graphiques. Ce faisant, émergent de nouvelles pistes de datation et d'attribution des actes, dans un dialogue renforcé entre diplomatique et paléographie.

Diplomatik, Paläographie und digitale Edition. Nutzung von schriftlichen Profilen für die Datierung und Lokalisierung von Urkunden

Die scheinbare schriftliche Einheitlichkeit des lateinischen Abendlands im 12. Jahrhundert wurde durch den Sieg der „vorgotischen" karolingischen Minuskel vollzogen. Sie ebnete den Schreibern den Weg einer künstlichen Kreativität. Es wurden unterschiedliche graphische Systeme entwickelt, und gemeinsame schriftliche Profile bildeten sich in gewissen Milieus wie Kanzleien und Skriptorien auf. Diese Profile kennzeichnen sich durch den zwar unstabilen, aber spezifischen Gebrauch der verschiedenen Allographen der jeweiligen Buchstaben. Da die digitale Edition die Koexistenz verschiedener Textformen erlaubt, ist es möglich geworden, auch die graphetische Information

während der editorischen Phase zu speichern, und die graphischen Systeme zu analysieren. Dadurch öffnen sich neue Wege für die Datierung und Zuschreibung von Urkunden im Dialog zwischen Diplomatik und Paläographie.

JONATHAN JARRETT

Poor tools to think with. The human space in digital diplomatics

The huge potential of the many new digital resources for work in diplomatics, and for historical work of a less technical nature, has so far remained largely potential, while on the other hand many projects continue to come forth in which researchers constructed their own digital resource rather than use an existing one. As such a researcher, the author here presents an anecdotal account of his progress into the field of digital diplomatics, and uses several detailed cases from his own research to argue that, although some of his choices were bad ones and made on the basis of inadequate information, the relatively unstructured systems that he has come to use have the virtue of requiring minimal interpretation to create a database, meaning that the human task of interpretation can be kept outside the data itself. Since the basic requirement of much historical work using such corpora is simply to put the information before the researcher in a comprehensible form without making judgements on it, such low-structure resources may have wider applicability and use for researchers than more heavily-structured resources in which more interpretation has had to precede the data entry.

Poor tools to think with: Der menschliche Raum in der digitalen Diplomatik

Das große Potential der zahlreichen neuen digitalen Ressourcen für die Urkundenforschung und für die – weniger technische – historische Arbeit ist bisher größtenteils nur Potential geblieben, während zugleich jedoch viele neue Projekte entstehen, bei denen Wissenschaftler die Entwicklung eigener digitaler Ressourcen der Nutzung existierender Werkzeuge vorziehen. Als ein solcher Forscher stellt der Autor hier einen anekdotischen Bericht über sein Eindringen in die digitale Diplomatik vor und zeigt anhand einiger Beispiele aus seiner Arbeit, daß, obwohl einige seiner Entscheidungen auf unzulänglichen Informationen gegründet waren, die von ihm genutzten relativ unstrukturierten Systeme mit nur wenig Interpretationsaufwand in eine Datenbank überführt werden können. Das bedeutet, daß die Interpretationsleistung des Forschers außerhalb der eigentlichen Daten bleiben konnte. Weil es eine der Grundvoraussetzungen historischen Arbeitens mit derartigen Corpora ist, Daten in einer übersichtlichen Form bereitzustellen, ohne im Vorhinein zu Urteilen zu gelangen, könnten derartige kaum strukturierte Ressourcen breitere Anwendbarkeit und Nützlichkeit für die Wissenschaft haben als stärker strukturierte Systeme, bei denen der Dateneingabe mehr Interpretationsarbeit vorausgeht.

Colour plates

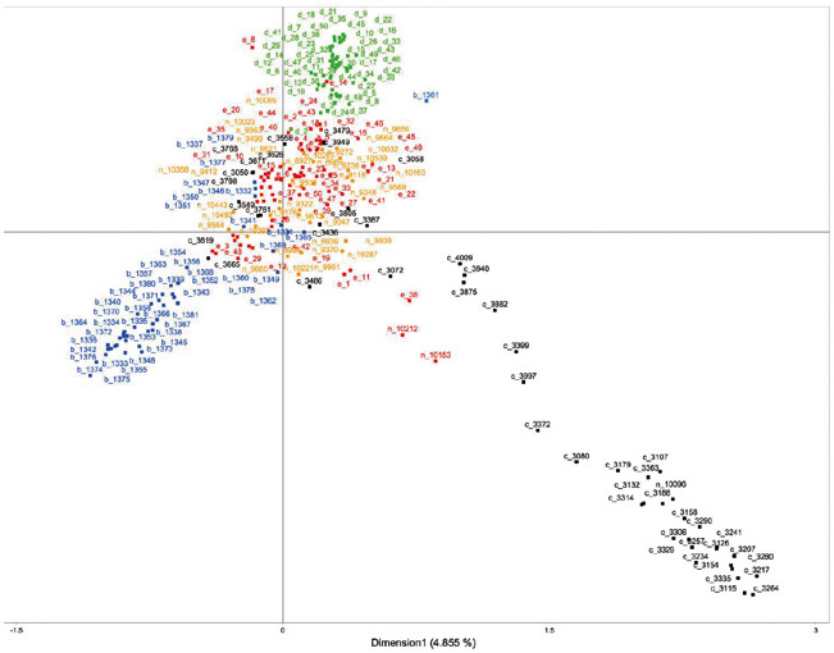

Fig. I: Analyse factorielle (AFC – Axe 1–2) de la matrice lexicale, réalisée grâce à Text-to-CSV, d'une série d'actes sélectionnés aléatoirement dans notre base. Légende: d_ (vert) = diplômes; b_ (bleu) = bulles; e_ (rouges) = actes épiscopaux; n_ (orange) = notices; c_ (noir) = chartes ne rentrant pas dans les catégories précédentes (hors pancartes)

Fig. II et III : (ci-dessus): Analyses factorielles (Axe 1–2) d'une matrice lexicale, réalisée grâce à Text-to-CSV, d'une série d'actes diplomatiques contenant a. Des bulles (b_). b. Des actes épiscopaux (e_). c. Des diplômes royaux et impériaux (d_).

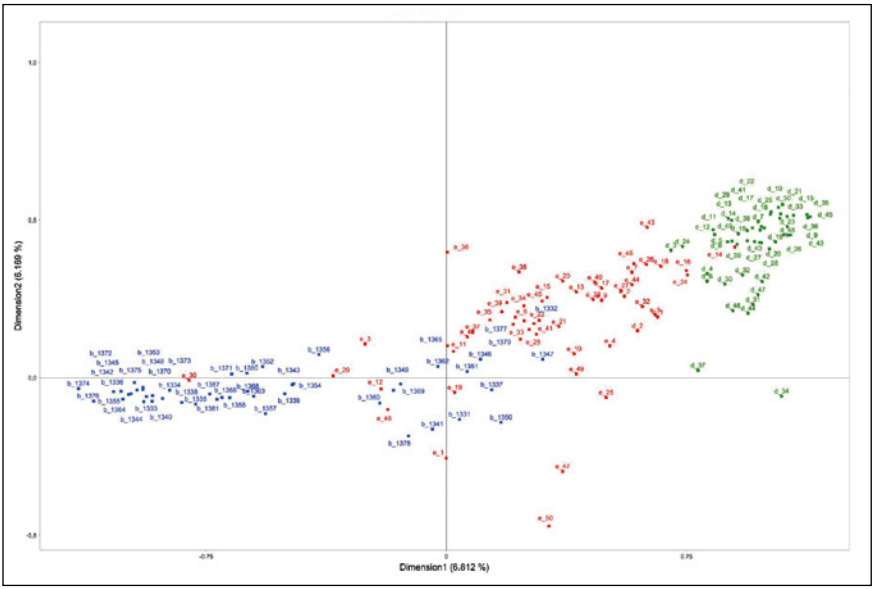

Fig. II

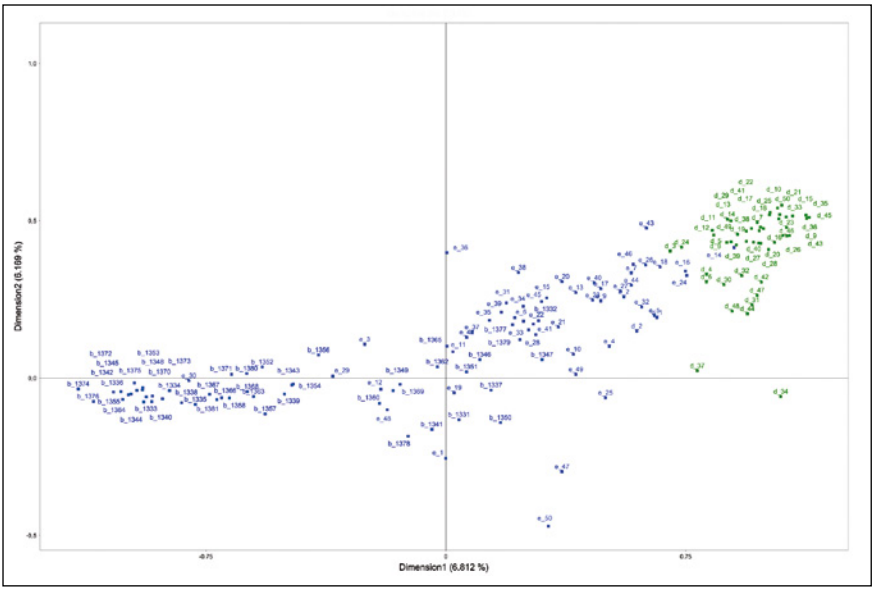

Fig. III