# Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ?

Dominique Stutzmann

Schriften des Instituts für Dokumentologie und Editorik — Band 3

# Kodikologie und Paläographie im digitalen Zeitalter 2

---

# Codicology and Palaeography in the Digital Age 2

herausgegeben von | edited by

Franz Fischer, Christiane Fritze, Georg Vogeler

unter Mitarbeit von | in collaboration with

Bernhard Assmann, Malte Rehbein, Patrick Sahle

2010

BoD, Norderstedt

**Bibliografische Information der Deutschen Nationalbibliothek**:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über **http://dnb.d-nb.de/** abrufbar.



Online-Fassung
Herstellung und Verlag der Druckfassung: Books on Demand GmbH, Norderstedt 2010
ISBN: 978-3-8423-5032-8
Einbandgestaltung: Johanna Puhl, basierend auf dem Entwurf von Katharina Weber
Satz: Stefanie Mayer und LaTeX

# Paléographie statistique pour décrire, identifier, dater… Normaliser pour coopérer et aller plus loin ?

Dominique Stutzmann

**Résumé**

Les initiatives TEI (Text Encoding Initiative) et MUFI (Medieval Unicode Font Initiative) multiplient les potentialités des transcriptions et invitent à interroger le processus même d'une opération pourtant commune dans les études médiévales. Transcrire, c'est décrire un texte et sa mise en forme graphique. Cela ouvre des voies aux études paléographiques, notamment par la description des diverses variantes morphologiques des lettres ou abréviations. Des recherches entreprises sur la production écrite de l'abbaye cistercienne de Fontenay démontrent le potentiel considérable des transcriptions dites « allographétiques » et des statistiques descriptives pour comprendre les évolutions de l'écriture médiévale. La fragmentation du paysage actuel de la recherche amène pourtant à souhaiter une harmonisation des pratiques, tant pour décrire les phénomènes paléographiques que pour structurer les données et formuler des ontologies.

**Zusammenfassung**

Elektronische Texte auf Basis der Text Encoding Initiative (TEI) und der Medieval Unicode Font Initiative (MUFI) erhöhen den Wert von Transkriptionen mittelalterlicher Handschriften so immens, dass dieser in der Mediävistik so gängige Prozess neu zu bewerten ist. Transkribieren heißt einen Text und dessen graphische Inszenierung beschreiben. Durch die Beschreibung der Buchstaben- und Abkürzungsvarianten werden der Erforschung der mittelalterlichen Schriften neue Wege geöffnet. Die Untersuchung der Buch- und Urkundenschrift der Zisterzienserabtei Fontenay zeigt, dass sogenannte allographetische Transkriptionen und die deskriptive Statistik einen beachtlichen Wert für das Verständnis der Schriftentwicklung besitzen. Die Vielfalt der heutigen Forschung in diesem Feld lässt eine Vereinheitlichung ihrer Praxis, sowohl für die Beschreibung der Formen als auch für die Datenstrukturierung und den Aufbau der notwendigen Ontologien, wünschen.

**Abstract**

Electronic texts employing the Text Encoding Initiative (TEI) and the Medieval Unicode Font Initiative (MUFI) provide added value for transcriptions of medieval manuscripts.

*Kodikologie und Paläographie im Digitalen Zeitalter 2 – Codicology and Palaeography in the Digital Age 2.* Hrsg. Franz Fischer, Christiane Fritze, Georg Vogeler, unter Mitarbeit von Bernhard Assmann, Malte Rehbein, Patrick Sahle. Schriften des Instituts für Dokumentologie und Editorik 3. Norderstedt: Books on Demand, 2010. 247–277.

This is such a significant improvement that it requires rethinking the established process of transcribing. Transcribing is describing: with the encoding of variants for letters or abbreviations, palaeography can explore new horizons. A research on the scriptorium of Fontenay (O. Cist.) proves that graphetic transcriptions and statistical analysis improves our understanding of medieval scripts and their evolutions. Nonetheless, the field is diverse, and an elaboration of good practices for describing, structuring and organizing palaeographical data and relevant ontologies is urgently needed.

## 1. Introduction

Les humanités numériques et les éditions électroniques sont une chance pour les paléographes. La philologie étant de plus en plus attentive à la matérialité du texte, de nombreuses informations pouvant intéresser l'histoire de l'écriture sont consignées par les éditeurs de texte. Pourtant, malgré le temps passé et les nombreux projets dans le domaine philologique et paléographique[1], les transcriptions disponibles ayant un véritable intérêt pour une analyse paléographique sont rares. Il y a près de vingt ans, en 1993, P. Robinson et E. Solopova tiraient des conclusions critiques sur leur édition numérique de textes chaucériens (Robinson et Solopova 1993) ; les questions posées par ces deux auteurs demeurent d'actualité. L'amélioration des techniques ainsi que l'évolution des recommandations de la TEI (TEI Consortium) pour la structuration de l'information graphique ont profondément modifié les pratiques et les outils, mais trop de développements spécifiques empêchent encore une pleine collaboration.

Les enjeux scientifiques et industriels sont pourtant majeurs : du point de vue disciplinaire, il s'agit d'améliorer les classifications d'écritures et les identifications de mains, de créer de nouveaux critères de datation ; du point de vue de l'ingénierie, l'appréhension des écritures anciennes et de leur variabilité pourrait bénéficier de l'existence de sources préparées et aboutir dans le futur à la reconnaissance optique des écritures manuscrites anciennes.

Le présent article essaie de définir les étapes nécessaires pour tirer bénéfice des possibilités d'analyse offertes par l'ordinateur : normaliser les transcriptions afin de favoriser les coopérations entre chercheurs et améliorer les outils d'analyse. Nous questionnerons dans un premier temps une dimension théorique de l'activité de transcription (que signifie transcrire ? à quel niveau transcrire ?), avant d'exposer la méthodologie et les résultats d'un projet spécifique de paléographie statistique. En confrontant différentes approches, nous proposerons un modèle et un cadre formalisé pour asseoir les pratiques futures.

---

[1] Une liste des corpus en langue française est présentée par C. Guillot *et al.* (2008).

## 2. L'image et la description du texte

Dans leur introduction, Robinson et Solopova abordaient la question théorique de la *signification de la transcription*, en affirmant que la transcription ne génère pas un substitut, mais constitue une suite d'actes de traduction d'un système sémiotique à un autre. La traduction, toujours incomplète et interprétative, est ainsi une pratique dont il s'agit d'interroger l'essence. Reposons la question : que signifie transcrire ?

### 2.1. Transcrire, c'est décrire

Il ne faut pas, selon nous, réfléchir à la transcription, même selon un encodage fin avec les éléments de la TEI, comme à la restitution ou à la « traduction » d'un texte, mais plutôt comme à sa description. Lors de la transcription, un premier document textuel donne naissance à un second ainsi qu'à de multiples descriptions : c'est ce que les sciences de l'information nomment « redocumentarisation » (Salaün 2008). Celle-ci intervient aussi bien lors du passage d'un « document » ou unité documentaire de l'analogique au numérique, avec la création de nouvelles métadonnées (descriptives, administratives ou techniques, en particulier sur la granularité, les relations entre documents et leur gestion) que lors de la constitution *a posteriori* d'un ensemble d'informations autonomes en « document ». Dans l'univers numérique, le texte même d'une ressource est indexable et exploitable : il double donc les indexations descriptives traditionnelles et devient sa propre métadonnée, ou, pour mieux dire, sa propre description à l'échelle 1:1.

Une structure de balisage telle que celle proposée par la TEI permet de décrire un « texte » en créant un nouveau document où l'opération descriptive va plus loin que la transcription, en explicitant des implicites du texte (par exemple, structure linguistique ou présence d'entités nommées, noms géographiques et anthroponymes). Cette possibilité nouvelle d'établir une description formalisée d'un texte à des échelles plus grandes que l'échelle 1, où le document descriptif dit davantage que l'original décrit, modifie profondément l'univers documentaire actuel et la conception des opérations documentaires : *transcrire* et *encoder* ne sont pas des opérations de traduction, mais des opérations descriptives et d'explicitation. Aussi de nouvelles questions se posent-elles : comment formaliser la description et ses différents niveaux ? Quel niveau de description choisir ? Faut-il étendre les règles internationales de description bibliographique (ISBD) pour rendre compte d'un texte et de sa description ?

### 2.2. Le texte est une image comme les autres

Robinson et Solopova proposaient quatre niveaux de transcriptions : « *graphic* » rendant toute la forme ; « *graphetic* » distinguant chaque type de chaque lettre ; « *graphemic* » préservant la suite de lettres ; « *regularized* » unifiant les suites de lettres attestées à une

forme normalisée. Si encoder, c'est décrire, l'on peut considérer sous un angle différent ces quatre niveaux allant des transcriptions « graphiques » aux éditions « régularisées » : les deux premiers décrivent l'image, le troisième décrit le texte-objet soumis aux accidents et le quatrième, le texte-idée. C'est la division fondamentale entre texte et image que nous retenons ici principalement. Dans un texte transmis par écrit, il y a toujours deux informations : le texte et l'image, le sens et la forme du signe[2].

La description de l'image, ou restitution formalisée des informations de formes, peut se diviser en deux niveaux. La transcription « graphique », tout d'abord, est définie comme restituant tous les caractères graphiques de l'original. Ainsi formulé, il s'agit d'une illusion, car nulle « transcription » ne peut donner accès à toutes les caractéristiques graphiques de l'original. En effet, la transcription porte sur le texte, alors que l'information décrite relève de l'image. Le document qui peut rendre compte de toutes les caractéristiques graphiques est la photographie, de sorte que la « transcription graphique » correspond à un ensemble descriptif constitué de métadonnées descriptives, d'une transcription du texte, de la photographie, avec association lettre à lettre des signes et de leur forme graphique (ensemble des coordonnées des points formant la lettre), voire les mesures effectuées par un logiciel d'analyse des formes. C'est, du reste, sur une « analyse graphique » plutôt que sur une « transcription graphique » que se fonde une part substantielle de la paléographie numérique, cherchant à reconnaître des mains et, partiellement, à améliorer la reconnaissance optique des caractères (Brink, Bulacu et Schomaker 2008 ; Bulacu et Schomaker 2007 ; Aussems et Brink 2009 ; Stokes 2009).

La transcription « graphétique » (*graphetic transcription*) ou, mieux, « allographétique » vise à donner accès à toutes les formes de chaque lettre ou signe[3]. C'est à elle que nous nous intéressons ici, car elle permet d'explorer des systèmes graphiques. Elle impose, pour ce faire, une réflexion sur les « types » et la réduction des variantes à des classes que l'on puisse désigner (cette nécessité confirme que la transcription allographétique n'est pas une transcription, mais une description). Il faut donc disposer d'un vocabulaire normalisé pour donner accès à certaines informations et asseoir sa pratique sur des règles qui permettent de prioriser les éléments à décrire, de hiérarchiser les informations selon les contextes et de traiter correctement les cas intermédiaires[4].

---

[2] Cette distinction est une réalité largement problématisée dans les opérations de numérisation du patrimoine littéraire pour la constitution des bibliothèques numériques (André 2007).

[3] La « graphétique » recouvrant l'étude des signes dans leur substance (formation, perception, lisibilité) et la « graphémique » celle des formes, l'étude « allographétique » est, malgré son nom, une branche de la graphémique (Catach 2001a ; Catach 2001b). C'est en raison de la polysémie des termes et des confusions possibles que nous retenons l'expression « transcription allographétique ».

[4] Voir ci-dessous. Trois traitements sont en théorie possibles pour les cas intermédiaires : redoublement de l'information pour conserver toutes les interprétations possibles ; l'utilisation de valeurs prédéfinies permettant de trancher arbitrairement ; élaboration d'ontologies intégrant la double appartenance de valeurs mixtes.

La transcription allographétique pose ainsi des questions plus difficiles, car moins habituelles, que la *graphemic transcription* qui réduit chaque forme à son sens dans un système alphabétique et où la tradition philologique a déjà clarifié les problèmes (Ecole nationale des Chartes 2001–2003). L'influence de ce que l'on peut appeler la « nouvelle philologie », attentive à la matérialité du texte (Cerquiglini 2000), oblige à remettre une partie de l'ouvrage sur le métier. La solution choisie pour les *Contes de Canterbury* est une description graphémique hybride qui consiste à transcrire les lettres, en ajoutant les abréviations composées d'une lettre avec un signe (comme les différentes formes de *p* barré) ou hors système alphabétique, ainsi que la ponctuation (Robinson et Solopova 1993). Ce système ne se justifie qu'avec la réduction du système abréviatif dans les langues vernaculaires (cf. Cottereau 2005 625–26). Si, dans les *Contes*, il n'y a que onze caractères abréviatifs codés hors lettres suscrites, ce type de transcription perd son sens dans des textes aux abréviations nombreuses et ambiguës : il accorde une place peut-être exagérée au système autonome de la ponctuation et pose implicitement l'hypothèse que la ponctuation et la capitalisation constituent une caractéristique plus structurante que les différentes formes des lettres, voire que les abréviations (approche commune dans les philologies non latines, cf. Lavrentiev 2007). Ce système hérite des préoccupations philologiques (savoir quelles parties du texte sont attestées ou restituées) et de la conception moderne de l'orthographe ; il étudie en conséquence le système graphique à partir du point de vue d'un alphabet unifié et décide *a priori* de ce qui est structurant et empêche par conséquent l'étude et la découverte de structures cachées.

La description allographétique est cependant possible : les réalisations actuelles semblent riches de promesses pour l'avenir. En exposant les résultats d'une analyse fondée sur de telles transcriptions, nous analyserons leur apport tout en insistant sur la nécessaire définition de « bonnes pratiques » et de normes communes qui assurent la pérennité des documents.

## 3. Un essai d'analyse de système graphique : corpus et méthode

Dans les paragraphes qui suivent, nous décrivons un projet et sa méthodologie, afin d'expliciter nos réflexions sur la structuration des données et réévaluer *a posteriori* la valeur heuristique de l'encodage allographétique utilisé. Trois exemples illustreront l'intérêt d'une analyse statistique des écritures sur la base d'un encodage limité ; ils permettront également de souligner les besoins actuels de normalisation et de tracer des pistes pour favoriser les enquêtes futures.

L'analyse allographétique et les résultats présentés ci-dessous ont été élaborés dans le cadre d'une thèse de doctorat où les différentes écritures attestées dans la production de l'abbaye bourguignonne de Fontenay aux XII$^{e}$ et XIII$^{e}$ siècles sont étudiées d'un point de vue statistique (Stutzmann 2009a). L'un des objectifs était d'identifier la production

par l'impétrant sédimentée dans le chartrier afin de confronter les caractéristiques paléographiques de la production pragmatique à celles de la production livresque.

### 3.1. Le corpus étudié

Le corpus initial était composé de 173 chartes originales antérieures à 1214 et l'analyse assistée par ordinateur a porté sur un sous-ensemble de 83 chartes et 40 livres manuscrits attribuables au scriptorium. Pour les livres, il a été procédé par sondage tandis que les chartes ont été encodées *in extenso*. Tous les écrits analysés sont en latin et en écriture gothique textuelle ou *textualis libraria* selon la typologie de Derolez (2003)[5]. L'exploration statistique a eu pour base un encodage suivant les recommandations de la TEI. Les résultats de l'analyse statistique sont particulièrement riches en regard des faibles moyens techniques mis en œuvre. Dans un système d'écriture très stable (écriture en langue latine, posée et dépourvue des phénomènes de cursivité permettant à l'individualité des scribes de s'exprimer), il apparaît tout de même possible d'étudier l'évolution des écritures ainsi que les divergences entre différentes mains, autant dans les taux d'utilisation des différentes formes que dans leur répartition et leur éventuelle régularisation.

L'étude de l'écriture dans une abbaye présente des caractéristiques particulières qui obligent à multiplier les approches et à élargir le champ d'investigation, puisqu'il ne faut pas se limiter à un auteur, mais en considérer plusieurs, dont le nombre et la qualité varient au fil du temps.

Au XII$^{e}$ siècle et au début du XIII$^{e}$ siècle, en Bourgogne du Nord, l'existence de plusieurs chancelleries peut être envisagée, dont un reflet pourrait se trouver dans le chartrier de l'abbaye cistercienne de Fontenay : celles des évêques d'Autun et de Langres tout d'abord, dont le diocèse s'étend sur une partie de la Bourgogne ducale (Richard 1954). La production documentaire des évêques d'Autun n'a pas encore été étudiée, tandis que celle des évêques de Langres fait l'objet des travaux d'Hubert Flammarion (cf. Flammarion 1982 ; Flammarion 2004). La troisième chancellerie dont on attend que le chartrier de Fontenay porte la trace est celle des ducs de Bourgogne où un bureau d'écritures s'organise sous Hugues III (1162–1192), même si les actes émanant des ducs ne se distinguent guère de ceux des barons bourguignons, et où une chancellerie se met réellement en place sous Eudes III avant 1218, avec des officiers et un formulaire imité de celui des rois de France (Richard 1984 381–84).

---

[5] La classification élaborée pour les écritures minuscules gothiques par G. Lieftinck et augmentée par P. Gumbert et A. Derolez se fonde sur les morphologies de la lettre *a* (à simple ove ou à crosse), des hastes (bouclées ou non) et des lettres *f* et *s* long (sur la ligne ou plongeant) pour proposer des distinctions de types : *textualis* avec *a* à double panse, hastes non bouclées et *f* sur la ligne ; *cursiva* à l'opposé et *hybrida*, une *cursiva* sans boucle.

Ces trois autorités, disposant peut-être déjà de chancellerie, se retrouvent effectivement dans le chartrier, en compagnie de quelques autres[6] :

- 81 actes, soit 46,8%, donnés sous le nom des évêques d'Autun ou de Langres, seuls ou en compagnie d'autres évêques ou abbés (53 pour l'évêque d'Autun et 30 pour celui de Langres, avec deux actes donnés en commun),
- 17 actes donnés par les ducs de Bourgogne (9,8%)[7],
- 13 actes donnés par les archiprêtres de Touillon (7,5%)[8],
- 7 actes donnés par les abbés de Fontenay (4%),
- 4 actes donnés par les archidiacres de Flavigny[9],
- 3 actes donnés par les abbés de Flavigny[10].

La répartition des autorités en diachronie n'est pas du tout homogène. Si les premiers actes, peu nombreux, sont placés sous le nom d'abbés, ceux rédigés entre 1150 et 1189 sont très majoritairement au nom d'évêques. À partir de 1190, l'importance numérique des autres auteurs croît fortement : les ducs de Bourgogne et l'abbé de Fontenay lui-même apparaissent de plus en plus fréquemment, ainsi que l'archiprêtre de Touillon, mais uniquement dans la décennie 1189–1199.

Si les autorités épiscopales et ducale sont bien attestées par le chartrier de Fontenay, leurs chartes ne forment cependant pas des ensembles cohérents et nulle trace de chancellerie constituée ne se reflète dans ce miroir. En revanche, plusieurs groupes d'actes se distinguent, qui dépassent les limites posées par le critère d'auteur. Une étude diplomatique et paléographique que nous ne reprendrons pas ici permet d'acquérir la conviction que leur élaboration est intervenue au sein de l'abbaye (Stutzmann 2009a 164–444). C'est sur les actes les plus anciens et ceux du scriptorium que se fondent les résultats présentés ci-dessous.

### 3.2. Méthodologie : encodage allographétique

En parallèle à l'étude morphologique, l'enquête paléographique a été menée sur le système graphique des scribes. La base en est un encodage allographétique respectant

---

6 La structure du fonds est similaire si l'on inclut les actes copiés, exception faite des bulles pontificales dont il ne subsiste qu'un original pour 17 actes connus par des copies.

7 Soit dix-huit actes si l'on compte l'un donné en commun avec les évêques de Lyon, Autun et Langres.

8 Le premier acte est donné en 1189 et les suivants entre 1194 et 1199, mais aucun entre 1200 et 1213. L'un de ces actes est donné sous les noms conjoints de l'archiprêtre de Touillon et l'abbé de Fontenay (ADCO 15 H 199/3).

9 Tous dans les quatre dernières années de notre période d'étude (un en 1210 et trois en 1213).

10 Actes répartis sur toute notre période (le premier, ADCO 15H257/ 1 datant de 1126–1149, le second 15 H 130 / 4 de 1180 et le dernier ADCO 15 H 58 / 1 de 1202) ; le second étant toutefois coémis par l'évêque d'Autun.

les directives de la TEI-P5, effectué avec le logiciel Oxygen[11]. L'encodage s'est fait sans souci de révéler des individus, mais avec l'objectif de voir des évolutions collectives. C'est un encodage de type générique qui a été choisi et ne portant que sur des phénomènes graphiques très largement répandus et observables y compris dans des textes courts.

Tous les actes ont été analysés à partir d'un seul fichier XML (chaque acte ou feuillet de manuscrit est encodé dans un élément ). L'analyse a été effectuée grâce à une transformation XSLT qui établit la liste des mots abrégés, calcule la largeur moyenne des lettres et décompte le nombre d'occurrences des allographes et des abréviations[12].

A partir des résultats chiffrés de la transformation XSLT, des analyses statistiques, essentiellement en composante principale, ont été réalisées avec logiciel R (Gentleman, Ihaka et R Development Core Team 2007).

Il faut bien noter ici que, comme toutes les études nécessitant un encodage des « caractères » d'une « population », celle-ci a opéré des choix qui ne sont pas neutres et la situent dans un champ où les tensions sont multiples : latin/vernaculaire ; écriture formelle/cursivité ; diplomatique/livresque ; individuel/collectif ; précision/généricité ; phénomènes ordinaires et fréquents/extraordinaires et rares.

En traitant de textes latins du XII$^{e}$ siècle, l'alphabet présent est réduit : il y a peu de signes diacritiques (hormis la cédille de *e*) ; l'écriture ne présente pas de traits de cursivité dont le signalement modifierait substantiellement le travail de transcription et d'analyse : les ligatures canoniques *ct*, *et* et *st* apparaissent, mais sont presque seules, avec la fusion des oves contraposées qui apparaît sporadiquement. Les divergences morphologiques entre écritures de la pratique et écritures livresques sont en apparence minimes et ne peuvent pas être encodées directement avec MUFI (balancement des hastes et hampes).

La précision de l'encodage a été calibrée et mise en relation avec l'objectif. Les choix faits lors de cet encodage sont les suivants[13] :

---

[11] Au moment de commencer cet encodage, la troisième version de MUFI (2009) n'avait pas encore paru, de sorte que certaines solutions d'encodage des signes abréviatifs sont personnelles.

[12] Les deux premières versions de cette transformation ont été programmées par Florence Clavaud, de l'École nationale des Chartes, que je tiens à remercier ici.

[13] Le choix des encodages est justifié plus à plein dans le travail d'origine. C'est un choix qui s'insère dans une tradition paléographique ancienne. Dès les premiers traités diplomatiques, la forme des lettres comme critère de datation est expliquée, y compris pour les lettres *d* et *s*, mais les descriptions sont encore très sommaires (cf. Tassin et Toustain 1750–1765 : II, 167–73, étude de « *d* » avec une longue note sur la domination de la forme onciale à partir du milieu du XII$^{e}$ siècle, et p. 260–72, étude de « s », en particulier à la p. 65 sur l'emploi des différentes formes dans les manuscrits). La première étude sur la forme du *d* est celle de Wilhelm Meyer qui, outre ses deux « lois » portant sur le *r* courbe après *o* et l'assimilation des oves contraires, constate une loi inconstante sur l'emploi des *d* : *d* droit devant lettres verticales *i*, *u*, *n*, (*m* et *r*) et *d* courbe devant lettre à ove *a*, *e* et *o* (cf. Meyer 1897 17–19). Bischoff prend soin d'indiquer la présence de *d* onciaux dans l'écriture des gloses (1954 8). Dans le scriptorium de Cluny, tous les scribes identifiés dans la deuxième moitié du X$^{e}$ siècle utilisent les deux formes, mais leur utilisation n'est pas étudiée de façon différenciée (Garand 1978). Dans la bibliographie ultérieure, Petrucci affirme que l'emploi

- distinction de trois formes de la lettre *d* : capitale (« D »), *d* droit (« d »), *d* oncial (« ᵭ »)
- distinction de trois formes de la lettre *s* : capitale (« S »), s rond (« s ») et s long (« ſ »)
- encodage de toutes les abréviations avec leur résolution (par exemple : suite de balises <choice><expan>anima</expan><abbr>ai~a</abbr></choice> pour le mot « anima ».

Outre cet encodage allographétique, des caractéristiques externes du texte (fin de ligne, espace blanc pour assurer la justification) ainsi que des éléments extra-paléographiques (noms de personnes et de lieu, mots en langue vernaculaire) ont été également enregistrés afin de pouvoir étudier leur influence sur le comportement des scribes. Des phénomènes ont été encodés aussi qui n'ont finalement pas été retenus dans le périmètre de l'étude (e.g. : degré d'abréviation et influence des fins de ligne).

### 3.3. Résultats obtenus

Au-delà des conclusions sur les pratiques paléographiques du scriptorium de Fontenay, le principal résultat de l'étude est que l'encodage de caractères limités (allographes *d* et *s*) a suffi à mettre en évidence des groupes homogènes et des évolutions, ainsi qu'à ouvrir la réflexion sur la perception de l'écriture dans le contexte médiéval. Plus

---

du *s* courbe en fin de ligne se constate en France dès les années 1140, puis en Germanie vers 1150 et à la fin du siècle en Italie, avant d'être d'usage régulier dans la seconde moitié du XII$^{e}$ siècle en fin de mot (Petrucci 1968 1121–25). C'est aussi comme cela que nous interprétons la mention « (final) » sous le dessin de *s* courbe qui apparaît sur certains relevés de lettres de Gasparri (1973 28, 30, 38–39 etc). L'étude est poursuivie dans le contexte italien avec une approche systémique (Zamponi 1989 326–27), ainsi que dans le domaine allemand (Heinemeyer 1982 32–34 et 42–44). Dans le contexte espagnol, des exemples anciens de *d* oncial en finale précèdent la progression à partir de l'initiale et parachevée à la mi-XIII$^{e}$ siècle. L'emploi des formes du *s* semble, plus que tout autre, répondre à la liberté des copistes et ne pas suivre d'évolution linéaire. Les autres allographes étudiés par Torrens sont *r* droit et rond, *i* et *j*, *u* et *v*, et *z* et *ç* (Millares Carlo et Ruiz Asencio 1983 I, 111, 85 et 94 ; Torrens 1995 355–59 pour d, 60–62 pour s). Si les planches sont représentatives, c'est la même évolution sur le plan local que nous constatons chez G. Nicolaj (1987 pl. XV) : apparition de *s* courbes dans la minuscule notariale qui en est dépourvue, en fin de mot et seulement après 1170. L'étude sur Gerhoh de Reichersberg, dont le scriptorium semble plus évolué et dont l'attention se porte sur les signes de ponctuation, ne fait pas le point complet sur la forme onciale de *d*, pourtant évoquée, (cf. Frioli 1999 207). A Brescia, la forme onciale, rare au début du siècle devient majoritaire aux alentours de 1150 pour s'évanouir ensuite et disparaître presque complètement dans les manuscrits liturgiques (Pantarotto 2005 5, note 25). L'approche morphologique est renouvelée par l'analyse des formes rondes comme élément d'un système graphique permettant de faciliter la lecture et d'assurer la compréhension en *lecture globale* (Frioli 2000 22–23). Il n'y a cependant pas d'évaluation statistique et l'interprétation pour la lettre *d* est très problématique : D. Frioli veut que le *d* oncial, après avoir marqué la fin d'une préposition (*ad*, *apud*) ou d'une forme pronominale (*quid*, *quod*), vienne à signaler le début ou la fin d'une syllabe, surtout si le mot est composé. Or la lettre *d* n'est, dans le système linguistique latin, jamais au milieu d'une syllabe. On en déduit donc qu'elle constate l'accroissement de la proportion de formes onciales.

concrètement, voici trois exemples choisis parmi les résultats positifs que cette méthode a permis d'obtenir.

**Exemple 1**

Le premier cas étudié est celui de la forme onciale de *d* dans un groupe d'actes écrits par le scribe principal des années 1150–1170. L'emploi des allographes de *d* est d'une nature dont ni la logique ni l'évolution n'apparaissent d'évidence. L'examen de tous les critères susceptibles d'influencer le scribe est trop complexe : il faut étudier la position dans le mot (initiale, médiane ou finale), les lettres précédentes et suivantes, ainsi que la présence de signes abréviatifs sur une lettre précédente ou suivante, voire sur la lettre *d* elle-même.

Les tableaux de chiffres rassemblant les cotes et les pourcentages selon chaque critère sont difficilement lisibles, d'autant qu'il faut pouvoir, dans chaque cas, avoir le nombre d'occurrences pour évaluer la représentativité d'un pourcentage. La table 1 indique les pourcentages d'emploi de la forme onciale selon la position dans le mot, sans le nombre d'occurrences, dans un tableau où les actes, majoritairement non datés, sont ordonnés selon leur *terminus ante quem*.

La lecture de ce tableau ne permet pas de tirer des conclusions immédiates. Il en est de même pour les pourcentages de formes onciales de *d* devant les voyelles *a*, *e*, *i*, *o* et *u*. Une analyse en composante principale, en revanche, fait apparaître une cohérence qui dépasse celle des séries linéaires. La figure 1 ci-dessous, qui est une représentation graphique de l'analyse en composante principale, met en évidence une évolution du système graphique au sein du scriptorium de Fontenay.

Les chartes originales sont dispersées selon plusieurs axes en fonction des pourcentages de formes onciales de *d* après les voyelles *a*, *e*, *i* et *u*[14]. Deux groupes apparaissent clairement et, dans la projection graphique, toutes les chartes antérieures à 1169 se retrouvent dans la partie gauche, tandis que celles postérieures se retrouvent à la droite. Cette distinction est extrêmement nette, au point qu'elle semble pouvoir être un critère pour proposer une datation de deux actes datables dans un intervalle assez long allant de 1163 à 1179 : d'un côté l'acte 15H 203/1 qui situe dans la partie gauche du graphique est plus vraisemblablement des années 1163–1169 ; de l'autre, l'acte 15H 249/1 se retrouve dans la partie droite du graphique, ce qui incite à le dater d'après 1169 sur des critères paléographiques, qui viennent ici corroborer un indice textuel ténu, puisqu'un témoin de l'acte n'est attesté que pour la période après 1171. Cette évolution concorde également avec d'autres évolutions de l'écriture dans le scriptorium (par exemple, multiplication de l'emploi de la forme ronde de *s* à partir d'environ 1170).

Cette mise en évidence est cruciale si l'on compare les chiffres bruts aux résultats de l'analyse en composante principale et à leur représentation graphique. L'analyse en

[14] Le faible nombre d'occurrences de la séquence « do » empêche de tenir compte de l'influence de cette voyelle sur les formes de *d*.

| Cote aux Archives dép. de Côte d'Or | Pourcentage de ꝺ initial | Pourcentage de ꝺ médian | Pourcentage de ꝺ final |
|---|---|---|---|
| 15 H 199 / 2 (s.d. [1154–1162]) | 14,3 | 23,1 | 100 |
| 15 H 156 / 2 (s.d. [1154–1162]) | 9,1 | 25 | 66,7 |
| 15 H 163 / 1 (s.d. [1154–1162]) | 25 | 46,2 | 100 |
| 15 H 163 / 3 (s.d. [1154–1162]) | 25 | 22,2 | 100 |
| 15 H 190 / 7 (s.d. [1154–1162]) | 18,5 | 37,2 | 72,7 |
| 15 H 193 / 2 (s.d. [1154–1162]) | 22,2 | 8,3 | 66,7 |
| 15 H 243 / 2 (daté 1162) | 11,1 | 47,1 | 73,1 |
| 15 H 190 / 2 (s.d. [1163–1169]) | 25 | 37,5 | 66,7 |
| 15 H 190 / 4 (s.d. [1163–1169]) | 37,5 | 11,1 | 50 |
| 15 H 357 / 2 (daté 1169) | 25 | 40 | 64,3 |
| 15 H 148 / 1 (s.d. [1148–1170]) | 15,4 | 17,6 | 66,7 |
| 15 H 163 / 2 (s.d. [1148–1170]) | 15,4 | 6,7 | 60 |
| 15 H 199 / 1 (s.d. [1162–1170]) | 6,3 | 0 | 66,7 |
| 15 H 249 / 2 (daté 1171) | 51,4 | 61,4 | 87,5 |
| 15 H 249 / 3 (daté 1178) | 39,1 | 46,4 | 91,7 |
| 15 H 193 / 1 (s.d. [1163–1179]) | 14,3 | 0 | 100 |
| 15 H 203 / 1 (s.d. [1163–1179]) | 29,4 | 20 | 68,8 |
| 15 H 249 / 1 (s.d. [1163–1179]) | 26,3 | 40 | 90 |
| 15 H 203 / 2 (s.d. [1171–1189]) | 40 | 52,6 | 100 |

Table 1. Pourcentage des formes onciales de la lettre *d* selon la position dans le mot.

composante principale offre un nouveau point de vue sur l'écriture et révèle l'influence du contexte (ici, la voyelle subséquente) ; elle permet d'affiner l'étude statistique et de rapprocher des écritures dont les pourcentages moyens divergent fortement. Ainsi, les deux actes 15 H 163/1 et 15 H 163/3, très proches l'un de l'autre sur les critères tant internes qu'externes, se retrouvent très proches sur le graphique alors que leurs pourcentages de *d* onciaux en position médiane varient du simple au double.

Ce premier exemple montre qu'une analyse statistique du matériau paléographique est possible et pertinente, y compris pour obtenir de nouveaux indices de datation. Il montre également que l'enquête statistique permet d'approcher de façon neuve des mécanismes d'écriture invisibles ou impossibles à étudier autrement, tels que l'influence des lettres subséquentes sur la graphie d'une lettre précédente.

**Exemple 2**

Le deuxième exemple choisi est l'évolution de l'emploi de la forme onciale de *d* entre 1150 et 1189, période durant laquelle le scribe principal des années 1150–1180 se voit doter de deux collègues actifs dans les années 1170–1189. Le procédé est le même : une

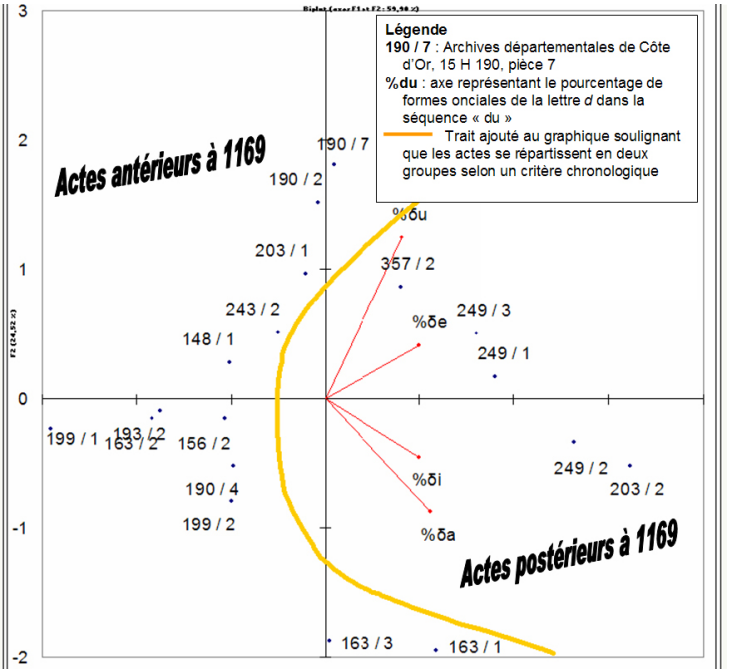


Figure 1. Analyse en composante principale mettant en lumière l'évolution du système graphique.

analyse en composante principale est effectuée, en tenant compte de la voyelle qui suit immédiatement la lettre *d*, et les résultats de cette analyse sont représentés en deux dimensions.

La frontière déjà observée vers 1170 se retrouve, mais surtout l'ensemble permet de constater une évolution plus générale.

Le premier graphique ci-dessous montre que ces deux scribes (ici appelés « 2a » et « 2b »), plus jeunes, ont des habitudes graphiques globalement distinctes de celles de leur aîné, mais qui ne se distinguent pas entre elles. Dans ce cas-ci, ce n'est pas l'évolution dans la production d'un même scribe que l'on fait apparaître, mais des évolutions générationnelles et partagées par plusieurs individus.

L'examen étendu à la production des années 1190–1215 fait, lui aussi, apparaître des divergences entre les pratiques de différents scribes, mais les séparations sont plus progressives et moins tranchées, alors que du point de vue morphologique, des caractéristiques distinctives permettent d'isoler les écritures tardives. Nous pouvons ainsi conclure de l'analyse statistique des phénomènes paléographiques que des

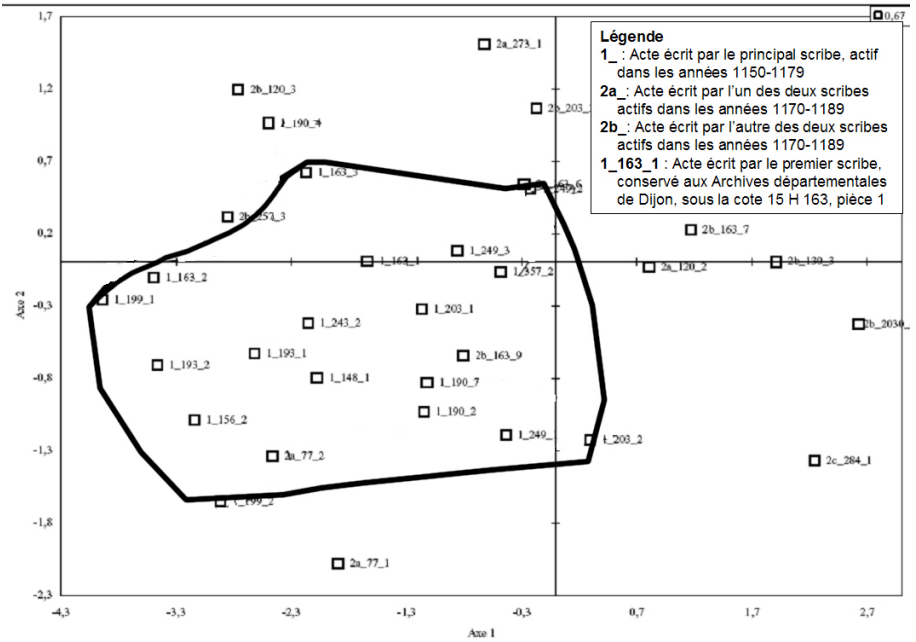


Figure 2. Analyse en composante principale montrant la différence entre le système graphique d'un scribe d'une génération antérieure et celui partagé par deux scribes d'une nouvelle génération.

glissements générationnels interviennent, affectant séparément la morphologie des lettres et le système graphique dans son ensemble.

### Exemple 3

Le troisième exemple montre la liaison possible entre des caractéristiques d'écriture et des réalités extra-paléographiques, en particulier décoratives. En effet, dans les livres manuscrits de l'abbaye de Fontenay, plusieurs relèvent du style « monochrome », déjà étudié en détail pour Cîteaux par Y. Załuska (1989).

Ce style monochrome, qui est une création cistercienne, impose que les initiales soient d'une seule couleur. Il n'interdit pas l'emploi de plusieurs couleurs sur une même page et n'entraîne pas d'économie de pigment : au contraire, il se développe une technique graphique qui joue avec le blanc du parchemin, autorise de somptueuses initiales comme dans la Bible de saint Bernard et permet même un jeu multicolore d'initiales enclavées.

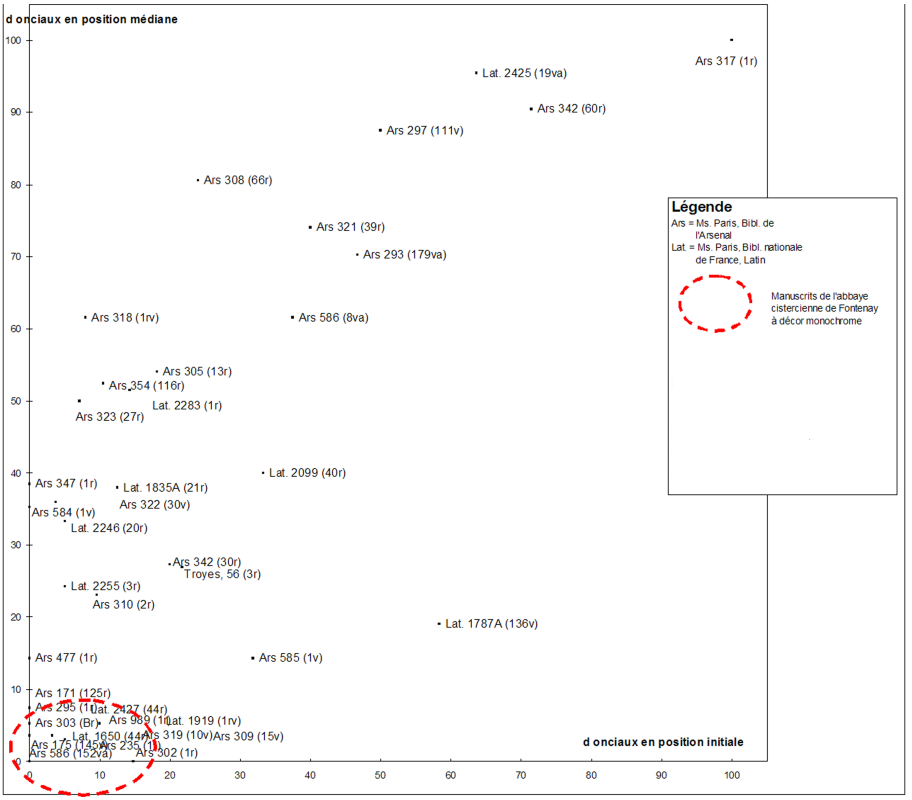


FIGURE 3. Régularisation de l'écriture dans les manuscrits à décor monochrome (spécialisation de la forme onciale de *d* pour la fin de mot).

A Fontenay, l'analyse paléographique et codicologique démontre que ces manuscrits partagent de nombreuses spécificités outre leur décoration. En particulier l'emploi des formes droite et onciale de la lettre *d* montre une spécialisation des formes de cette lettre *d* dans les manuscrits à décor monochrome, où la forme onciale est exclue des positions initiales et médianes et strictement réservée à la finale.

L'utilisation d'une forme spécifique en position finale permet d'améliorer la lisibilité d'un texte où les espaces entre les mots sont irréguliers en soulignant, par la morphologie, un phénomène linguistique (Frioli 2000). Cependant il ne faut pas réduire ce processus de spécialisation à une seule recherche de lisibilité : en effet, si celle-ci était le seul

enjeu, elle eût pu être maintenue hors des modes décoratives. Or, non seulement, ce phénomène est transitoire, mais il intervient précisément dans les manuscrits qui se distinguent par leur décor soumis à la norme nouvelle de la monochromie ainsi que par d'autres caractéristiques matérielles, tels largeur des marges ou grandeur de l'unité de réglure (Stutzmann 2009a). Une telle règle ou régularisation graphique sur les formes de la lettre *d* est en outre totalement inconnue à la production documentaire. Cela vient, à notre sens, confirmer qu'il s'agit de l'effet d'une mise en ordre délibérée des réalités paléographiques et pratiques décoratives dans les livres manuscrits. De même que l'interdiction de la polychromie peut être interprétée comme une conséquence d'ordre esthétique de la théorie des sens et du salut chez Bernard de Clairvaux (Stutzmann 2009b), il apparaît que la régularisation de l'écriture, qui est soumise à une contrainte nouvelle ou à une règle, correspond à une volonté de maîtrise des sens et à une réflexion sur l'ascèse.

### 3.4. Conclusions méthodologiques

Les trois exemples allégués ci-dessus, exploitant des données simples et des ensembles à faible population, appellent des conclusions méthodologiques. Des analyses statistiques, notamment factorielles, permettent d'exploiter des informations dont la pertinence n'est pas prédictible et de mettre au jour des cohérences de pratiques scripturales malgré la forte divergence des données. L'encodage allographétique permet d'interroger sous un angle supplémentaire les données paléographiques et de découvrir des évolutions substantielles de l'écriture.

L'analyse statistique, fût-ce d'un unique caractère, constitue un critère déterminant pour l'enquête paléographique, car le processus d'évolution des écritures ne porte pas seulement sur la morphologie des lettres, mais aussi sur l'agencement des allographes. C'est le système graphique en son entier qui évolue, y compris au sein d'une production limitée dans le temps et l'espace.

## 4. L'analyse des écritures : normaliser pour coopérer

La conclusion positive issue du processus analytique précédemment exposé donne naissance à de nouvelles interrogations. Est-il envisageable de traiter de façon similaire des corpus à la fois plus vastes et moins homogènes ? Quel rôle attribuer à la machine dans l'analyse des écritures ? Dans le cas présenté ci-dessus, l'analyse statistique permet de souligner les phénomènes de cohérence ou de divergence au sein d'un ensemble qui se présente comme homogène et peu variant. Les conclusions plus profondes s'obtiennent en croisant les informations sur les pratiques d'écriture, les réalités codicologiques, l'enluminure et la connaissance de l'univers intellectuel médiéval.

Pour étendre l'enquête et affiner la méthodologie, il faut surmonter le défi de la masse des données et exploiter de façon adaptée et pertinente des données hétérogènes. En conséquence se pose la question des pratiques d'encodage utilisées et des fonctionnalités requises pour les outils de traitement de l'information paléographique.

Encore rares actuellement, les données paléographiques s'accroissent en effet rapidement. A l'heure actuelle, chaque projet mené par des équipes aux objectifs différents élabore ses propres solutions spécifiques, même si TEI et MUFI offrent un canevas commun. Des oppositions tendent à disparaître, notamment celles consistant à désigner des phénomènes graphiques différents par un même descripteur : par exemple « s » soit pour *s* rond (solution la plus fréquente), soit pour *s* long, l'autre forme pouvant être déclarée contradictoirement sous la même entité « &s ; » (Uitti 1997 ; McGillivray 1994–2010). Il est en revanche naturel qu'un projet portant sur un corpus spécifique où un allographe est nettement dominant, voire d'utilisation exclusive, tende à transcrire celui-ci sous la forme équivalente de l'alphabet latin actuel, indépendamment de son inclusion dans l'histoire de l'écriture médiévale : ainsi l'on trouvera les caractères « d » et « ſ » pour transcrire une forme onciale bouclée ou un *s* plongeant sous la ligne, y compris dans des projets avec une granularité descriptive forte, allant jusqu'à avoir plusieurs types de *i* court (Hofmeister, Hofmeister-Winter et Thallinger 2009). La lourdeur de préparation des fichiers et la diversité des objectifs poursuivis exigent des solutions économiques et adaptables aux besoins de chaque équipe.

Or, dans notre première partie, nous avons évoqué la possibilité de décrire les informations à une échelle supérieure à 1. Si l'analyse et les transcriptions allographétiques portent, selon les projets, à la fois sur les écritures livresques et celles de la pratique et qu'elles couvrent plusieurs des siècles médiévaux, le nombre de documents concernés peut rapidement atteindre des proportions hors des capacités d'analyses individuelles, proportions où chaque signe devient une information à analyser et exploiter pour interpréter le système graphique d'un document. Le nombre en est potentiellement infini. Même si les informations paléographiques sont homogènes et nonobstant les difficultés d'analyse, l'on court le risque de ne pouvoir interpréter les similitudes entre écritures relevant de familles graphiques différentes. Des similitudes démontrées par l'analyse statistique seront une donnée supplémentaire à interpréter, et non une preuve démonstrative. Il est en effet difficile d'établir des arguments et des valeurs causales dans un domaine strictement paléographique et sans donnée extérieure pour appuyer l'interprétation.

L'amélioration des analyses exige ainsi un plus grand nombre de fichiers structurés avec des informations de nature allographétique et la possibilité de disposer de corpus de comparaison pour valider les hypothèses et résultats. La pluralité des choix d'encodage nécessite d'établir des passerelles, des normalisations ou des bonnes pratiques (valeur des caractères, structuration, granularité descriptive) et une formalisation des choix

précisant les pratiques adoptées, afin de pouvoir traiter de grandes masses de données partiellement hétérogènes.

### 4.1. Valeurs

Le premier niveau auquel on pense est naturellement celui des valeurs utilisées : quel caractère utiliser pour quelle forme graphique. Paradoxalement, ce sont les signes non-alphabétiques ou de l'alphabet non latin qui font l'objet de la plus grande attention : signes graphiques adventices, abréviations, capitales, après un essai infructueux avec *r* et *s* pour Robinson et Solopova (1993), lettres anglo-saxonnes pour Rumble (2004), abréviations, *s* et ligatures pour McGillivray (2005). L'étude minutieuse des pointages de *i* et de plusieurs autres caractéristiques a pu être menée sur des corpus restreints et constitue encore un cas particulier (Hofmeister, Hofmeister-Winter et Thallinger 2009). Certaines listes sont particulièrement longues (Uitti 1997 ; McGillivray 1994–2010). Les projets portant sur des corpus vernaculaires anglo-saxons, germaniques ou nordiques, dont les particularités graphiques imposent l'usage de signes graphiques hors de l'alphabet latin de 22 lettres ont déjà abouti à des réalisations fondatrices et bien documentées.

La question de la normalisation et de l'évolution d'Unicode est parfois explicitement mentionnée, en particulier dans des projets liés à MUFI (Haugen 2008 ; Anderson et al. 2007 ; MUFI 2009). Cette initiative répond en effet à de nombreuses questions, mais pas encore toutes, et, parfois, engendre de nouvelles difficultés. Ainsi la forme plongeante de la lettre *s* : MUFI permet d'encoder une forme agrandie (code EEDF) ou la forme insulaire (code A785), mais pas la forme fréquente du *s* de la bâtarde bourguignonne. Cela pourrait être amélioré dans les prochaines versions à condition de considérer ces allographes comme des caractères et non comme des formes ; une réflexion approfondie et une enquête nouvelle doivent être menées pour définir l'allographie selon ce nouveau point de vue.

Outre la question des entités et valeurs à utiliser pour décrire les différentes morphologies d'une lettre, la principale question concernant les valeurs tient à la double possibilité d'utiliser un encodage spécifique ou des caractères génériques. Les lettres suscrites, notamment, donnent presque toutes occasion à variantes d'encodage : les voyelles et semi-voyelles (*a*, *e*, *i*, *o*, *u*, *w*) avec lettre suscrite ont un code particulier, alors qu'existent par ailleurs les lettres suscrites seules (par exemple 0363 pour *a* suscrit, autrement dit *combining latin small letter a*). Les ligatures, au contraire des lettres suscrites, sont systématiquement traitées par MUFI comme un unique caractère spécifique et ne donnent pas lieu à un codage générique, avec pour principaux inconvénients l'oubli inévitable de cas rares (par exemple, ligature ou plutôt lettres conjointes *nt* avec forme capitale de *n*) et la réduction du décompte du nombre de lettres — la solution choisie n'est pas indifférente pour des projets où les lettres sont décomptées

pour calculer la densité graphique ou le degré d'abréviation d'un texte (Cottereau 2005 ; Bozzolo et al. 1997)[15].

Les artefacts de présentation tels qu'agrandissement ou étirement d'une lettre sont aussi traités comme caractères spécifiques, de sorte que un *a* carolin et un *a* carolin agrandi seront deux caractères distincts. Or, d'un point de vue paléographique, on peut considérer qu'il s'agit de la même morphologie. Pour traiter une première ligne d'un acte carolingien en lettres étirées ou un mot en lettres agrandies, deux possibilités contradictoires apparaissent : soit coder chaque caractère spécifiquement, soit marquer une chaîne de caractères avec un qualificatif de mise en forme (le cas échéant, la valeur de celui-ci devrait également être normalisée, par exemple « enlarged », « elongated », comme dans <hi rend="enlarged">), de façon à rendre possible les conversions vers les fontes qui nécessitent effectivement que chaque caractère ait un code.

Un cas limite du respect de la morphologie et de la nécessité du sens est soulevé par Robinson et Solopova (1993) : l'existence de formes identiques pour des lettres différentes, telles que *c*, *e*, *o*. Deux formes différentes de codage sont possibles :

- soit transcrire avec un caractère explicitant uniquement la forme (par exemple « c » pour un *e* non bouclé) et indiquer la forme normalisée (en TEI, <c rend="c">e</c> ou, plus complexe, <choice><orig>c</orig><reg>e</reg></choice>),
- soit transcrire avec un caractère spécifique, à déterminer, la forme de *e* qui ressemble à un « c ».

Dans ces deux solutions ne se pose plus seulement un problème de valeur de caractère, mais de structuration de l'information et de granularité descriptive.

### 4.2. Structuration

Les mises en formes ou les morphologies ambiguës ne sont pas les seules réalités paléographiques qui peuvent être rendues soit par des caractères spéciaux, soit par un balisage et une structuration alternative. Ce sont les abréviations qui posent le mieux le problème de la structuration de l'information. Certains projets y ont consacré une grande attention, mais la diversité est encore de mise (Lavrentiev 2002 ; Heiden, Guillot et Lavrentiev 2002–2008). Le débat n'est pas clos et les réflexions, menées principalement pour les langues vernaculaires, doivent encore être affinées (Mazziota 2008).

---

[15] Pour l'analyse de l'espace écrit, ligatures et abréviations constituent des cas d'études distincts. La lettre suscrite n'est pas omise, mais ne modifie pas l'espace d'écriture occupé par la lettre qu'elle surplombe ; la présence d'une ou plusieurs lettres adscrites, au contraire, est à évaluer dans l'analyse de l'espace écrit, mais leur module réduit rend l'évaluation difficile ; les ligatures, enfin, posent aussi problème puisque la longueur du digramme est susceptible d'être modifiée, mais ne l'est pas toujours (les ligatures *st*, *fi* ou *ti* sont-elles plus courtes que les deux lettres séparées ?) et il est délicat de ne le tenir que pour un seul caractère.

La TEI-P5 prévoit un système complet pour déclarer une abréviation (<g>, <char>, <glyph>) et pour décrire le mot : soit <abbr> et <expan> pour le mot entier (abrégé ou avec abréviations résolues), soit <am> et <ex> pour chacune des abréviations. Ce système pose des problèmes de cohérence et d'exploitation. Certaines abréviations par contraction syllabique apparaissent construites sur un radical puis déclinées, et autorisent ainsi plusieurs encodages divergents : dans le mot *anima* abrégé *a*, *i*, tilde, *a*, l'abréviation est-elle composée de *a*, *i* tilde, ou de l'ensemble du mot ? et auquel cas, peut-on encore utiliser <am> ou bien doit-on utiliser <abbr> ? et alors faut-il considérer que *anima* et *anime* sont le sujet de deux abréviations différentes, et quel lien établit-on avec *animus* ?

Dans le cas d'autres abréviations par contraction, si l'on fait le choix de mettre en évidence le radical, celui-ci prend lui-même les différents degrés : *frater* abrégé sur le radical *fr* se décline en *fratrum* marqué « frm » et tilde, ou encore *homo*, *-inis* avec les abréviations tildées *ho*, *hoie*, *hois*, *hoim* où les radicaux *ho* et *hoi* signifient respectivement *homo* et *homin-*, *homini* et *hominu-*. Un très grand nombre de mots courants sont l'objet des abréviations par contraction et suppriment un nombre variable de lettres (par exemple : *abbas*, *dominus*, *gratia*, *martir*, *nomen*, *noster*, *omnis*, *pater*, *peccatum*, *sanctus*, *seculum*, *spiritus*, *vester*, ou encore *ecclesia* aussi bien *eccla* ou *ecclia*). Ces abréviations, à la fois courantes et simples, obligent à donner systématiquement la résolution. Des mots peuvent connaître plusieurs abréviations apparentées : faut-il en considérer certaines comme par suspension et d'autres par sigle abréviatif (exemples : *presbiter*, abrégé soit *pbr* par contraction et *p* macron, *s*, *b* barré, *r* par sigles abréviatifs) ? D'autres abréviations par contraction, notamment avec lettre suscrite, sont ambiguës et exigent leur développement (ex. : *m* avec *o* suscrit pour *modo* et *monacho*) : évidemment les abréviations par suspension, mais aussi les signes abréviatifs les plus courants peuvent être ambigus (*d* barré pour *de*, *-dem* ou *–ud* ; *p* barré pour *per*, *par* et *por*, etc.), tandis que le simple encodage par résolution n'est pas toujours suffisant pour savoir quelle abréviation a été utilisée (« q*ue* » peut valoir pour « q ; », « qz », « q » avec *e* suscrit, ou « q » barré ; « b*us* » pour « b ; », « bz » ou « $b^9$ » ; « c*om* » pour le neuf tironien, le *c* tildé ou le *c* retourné ; « e*sse* » pour « êê » ou « eê » ; « r*um* » pour *r* rond barré et *r* tildé). L'étude des abréviations doit, par ailleurs, pouvoir tenir compte des alternances allographétiques (par exemple *dictus* abrégé *d*, *c*, tilde avec *d* droit ou oncial).

Les solutions proposées par Mazziota pour les abréviations sont complexes et s'opposent aux tendances de MUFI et de l'Unicode (par exemple il ne faudrait pas considérer *p* barré comme un signe abréviatif autonome ou « logogramme », mais le décomposer en lettre *p* et signe distinctif ou « cénégramme »). Elles s'enferment aussi dans des barrières de modélisation qui nous semblent contraire avec la perception globale des mots, en refusant le traitement des abréviations par contraction et en imaginant des « périgrammes » (sous-type des cénégrammes) à « portée discontinue » (Mazziota 2008 : §46). Ces propositions ont pourtant des avantages : elles permettent de

rendre compte du déplacement du signe abréviatif ou de superpositions et de décrire très finement l'abréviation tout en unissant structurellement les lettres restituées au signe abréviatif.

A des fins de coopération, il serait souhaitable de pouvoir unifier la structuration, ou, pour le moins, de formaliser les choix d'encodage, car celui-ci dépend en effet de la compréhension du système médiéval et l'unanimité demeure incertaine. Aussi les pratiques adoptées doivent-elles être documentées (en TEI dans l'élément <encodingDesc>) et formalisées.

### 4.3. Granularité descriptive et explicitation du référentiel utilisé

Au-delà du choix des valeurs utilisées et de la structuration de l'information paléographique, un choix encore plus fondamental doit être précisé : la granularité descriptive, ou degré de précision de l'encodage. Celle-ci porte d'une part sur la globalité du système d'écriture, d'autre part sur les morphologies des lettres.

Pour analyser de grandes masses de données, deux logiques, en apparence opposées, coexistent : d'une part, celle de « l'exploration des données » et des moteurs de recherche sémantiques, approche microscopique cherchant à analyser les informations dans leur plus grand détail afin d'établir *a posteriori* une interprétation hors d'un modèle probabiliste et hors des modèles d'interprétation *a priori* ; les statistiques descriptives et analyses factorielles se rattachent à cette famille d'analyse. D'autre part, la logique descriptive catalographique, perspective de télescope, qui cherche à rendre accessible et visible des ensembles vus dans leur globalité et ne s'appréhende que par des points d'accès prédéfinis et un vocabulaire contrôlé, éventuellement coordonné et hiérarchisé, formant une ontologie.

Dans le domaine documentaire, plusieurs travaux montrent que les meilleurs résultats sont obtenus en confrontant les référentiels structurés, élaborés par analyse intellectuelle, et la fouille de données, fondée largement sur l'analyse statistique et portant sur des documents non structurés (Mane 2010 17–27 ; Beneventano et Bergamaschi 2006 ; Criado Fernández et Martínez-Tomás 2009). Dans le domaine purement paléographique, c'est également la double approche qui semble être la plus prometteuse (Muzerelle 2009 6–9 ; Hofmeister, Hofmeister-Winter et Thallinger 2009) : apporter une description télescopique globale (« textualis », « hybrida », etc.) et examiner les caractéristiques comme au microscope (*s* rond, *i* pointés, hastes bouclées, orientation des traits etc.).

Pour que chaque corpus puisse servir de point de comparaison aux autres, de façon neutre et sans préjuger des recherches de chaque projet, il est indispensable de pouvoir déclarer non seulement les caractères qui sont encodés, mais aussi ceux pour lesquels aucune supposition ne peut être faite. A cette fin, l'utilisation de typologies d'écritures, autrement dit d'ontologies, apparaît la meilleure solution. En effet, pour comparer des données en très grand nombre et structurées selon des codes hétérogènes, deux solutions

sont envisageables en théorie : convertir et réduire chacun des documents produits pour le soumettre à une norme unique, ou élaborer un système qui permette au document de déclarer selon quel système est réalisé l'encodage. La conversion de données signifie souvent la réduction au plus petit commun dénominateur et engendre une grande perte d'informations. C'est donc plutôt la seconde solution qu'il faut encourager et favoriser : définir plusieurs niveaux descriptifs, ayant chacun ses caractéristiques

L'emploi d'une classification telle que celle de Lieftinck-Gumbert-Derolez (Derolez 2003) permet de préciser la pratique d'encodage au niveau global du document et évite de préciser systématiquement la forme présente dans l'original. Ainsi, l'emploi des dénominations « *textualis* » ou « *hybrida* » permet d'interpréter les caractères *b*, *d*, *l*, *h* comme désignant des morphologies dépourvues de boucle (inversement pour la *cursiva*), indépendamment de leur encodage au sein du document traité et indépendamment de l'acceptation faite par ailleurs du système de classification de l'écriture. En revanche ces mêmes dénominations distinguent que partiellement les formes de *s* si la différence n'est pas faite entre *s* longs et *s* ronds. La précision du système doit être insérée soit au niveau du document, soit à l'intérieur même de la transcription allographétique. Imaginons, en effet, le cas des manuscrits dont les débuts de chapitres sont en *textualis libraria* et le corps en *cursiva* : sans une granularité et une explicitation suffisante, les caractères « a » ou « ſ » ne pourront être simplement pas être exploités si l'on veut distinguer les allographes *a* à simple boucle et *s* plongeant.

La granularité descriptive joue également un rôle dans l'encodage allographétique pour chaque lettre à formes multiples. Robinson et Solopova justifient le rejet d'une transcription allographétique par la multiplicité des morphologies et renvoient aux huit formes de *s* de Benskin (1990), en insistant sur le développement pyramidal des graphèmes : « *It is assumed by both McIntosh and Benskin that the relationship of graphemes to graphetes is hierarchical : so many graphetes of* s *; so many sub-types of each graphete ; even sub-sub-types, and so on.* » (Robinson et Solopova 1993).

C'est bien là, pourtant, que réside une solution : adopter une hiérarchie et encoder, au besoin, à niveau élevé, en acceptant la part d'arbitraire qui réside dans la nature même de la description par encodage. Pour des comparaisons à large échelle visant à comprendre les évolutions du système graphique, des distinctions objectives simples pourraient sans doute suffire — et seule la mise en œuvre réalisera la preuve de concept —, même si la hiérarchie devrait, en bonne théorie, être construite par le bas, en repérant toutes les formes et leurs contextes d'apparition pour pouvoir établir des types. Les travaux d'Oeser sur les formes de la lettre *a* et des jambages montrent bien la signification que peuvent avoir des variations morphologiques (Oeser 2001–2002 ; Oeser 1994), mais n'invalident pas la distinction fondamentale entre *a* à simple ove et *a* à double panse. L'on peut, à l'envi, raffiner sa description ou la simplifier, indépendamment de l'absence de consensus parmi les paléographes sur ce qui fonde un type. Il s'agit là de choix

d'encodages et de granularités descriptives différentes : ceux-ci doivent être réfléchis et explicités de façon à préserver l'interopérabilité sémantique des données.

## 5. Les outils

Dans les paragraphes qui précèdent, nous avons décrit les besoins de normalisation et d'harmonisation concernant les valeurs, la structuration et la granularité, de façon à obtenir des fichiers de travail dont le contenu soit explicite et utilisable par un traitement automatisé d'informations nombreuses sans générer de bruit. L'accent y a été mis délibérément sur la structuration des fichiers contenant des informations paléographiques fines. Néanmoins la question des outils est tout sauf accessoire, aussi bien en amont, pour la constitution de bases de données paléographiques, qu'en aval, pour leur exploitation. Les outils génériques n'existent pas encore, qui permettraient de proposer une nouvelle perspective sur l'ensemble des écritures médiévales. Examinons tout de même les fonctionnalités souhaitables des logiciels d'aide à l'analyse paléographique.

### 5.1. Constituer des bases de données paléographiques

En amont, dans la phase d'enregistrement des informations paléographiques et dans la structuration même, tout le labeur ne doit pas être pris en charge par les chercheurs et la machine ne sert pas qu'aux calculs. L'équipe de DA*mal*S a exprimé à quel point l'encodage allographétique et la mise en relation du texte transcrit avec l'image correspondante sont aussi lourde qu'importante (Hofmeister, Hofmeister-Winter et Thallinger 2009 270, 276). La création d'un logiciel de prétraitement des textes manuscrits médiévaux serait très précieuse. En offrant une segmentation (par mots et caractères, ou par groupes de caractères liés), éventuellement des fonctions de reconnaissance des formes à l'intérieur d'une même page (pour repérer les mots répétés) ainsi qu'une interface encourageant la saisie des caractéristiques paléographiques pour chaque zone repérée, un tel outil permettrait non seulement de faciliter le travail des chercheurs, mais également d'assurer la normalisation et de diffuser les bonnes pratiques pour disposer de données interopérables, tant pour les valeurs et descripteurs utilisés, la structuration et l'explicitation du niveau d'encodage utilisé.

Lors de la réalisation d'une transcription allographétique, il serait *primo* souhaitable de disposer d'une interface présentant une lettre isolée, un unique mot ou une seule ligne à transcrire. Du point de vue logiciel, cela correspond aux problématiques de segmentation, c'est-à-dire de reconnaissance des unités graphiques et des espaces entre les lettres, mots et lignes. Les algorithmes de segmentation sont complexes (Gatos, Stamatopoulos et Louloudis 2010) ; ce sont des boucles récursives qui analysent la probabilité de chaque hypothèse de segmentation et les confrontent d'un côté à

des dictionnaires et de l'autre aux résultats de l'analyse des formes effectuée selon l'hypothèse considérée (Tzadok et Walach 2009). Des essais effectués à la Bibliothèque nationale de France montrent que la fonction de segmentation du logiciel FineReader (société ABBYY) obtient déjà des résultats corrects, même si elle n'est pas tout à fait mûre pour les manuscrits médiévaux en raison de l'impossibilité de confronter efficacement l'hypothèse de segmentation à un dictionnaire.

*Secundo*, des hypothèses de transcription pourraient être proposées. Dans des corpus suffisamment homogènes, une interaction entre les algorithmes de segmentation et des fonctionnalités d'apprentissage serait bénéfique. Des logiciels peuvent même déjà se vanter de résultats corrects d'apprentissage, tant pour les mots, ou « word-spotting », que pour les lettres (Tomasi et Tomasi 2009 ; Leydier, Duong et Ouji 2006–2009).

*Tertio*, face à une image ou à une proposition de transcription, le paléographe a besoin d'une interface adéquate, rendant aisée la tâche d'analyse paléographique. Cela signifie notamment que le chercheur doit disposer d'un dictionnaire des entités et des abréviations facilement utilisable et lié à l'éditeur XML.

*Quarto*, afin de préparer l'exploitation des données, la constitution d'une base de données paléographiques implique de conserver les informations reliant la transcription allographétique réalisée et l'image originelle.

La situation est insatisfaisante. D'un côté les outils d'annotations actuels permettent d'enregistrer l'information, mais sans offrir ni aide à la saisie, ni structuration suffisante, de l'autre les outils de segmentation ou de reconnaissance de texte ne permettent pas l'insertion d'informations allographétiques. Prenons l'exemple d'un logiciel comme *Image Markup Tool* (Holmes et University of Victoria HCMC 2010), qui permet d'associer une partie d'image à sa transcription et enregistre l'information dans des fichiers XML conformes à la TEI : il pourrait servir de base à des développements ultérieurs de sorte que les annotations de transcription puissent être structurées hiérarchiquement et correspondre au texte, c'est-à-dire qu'il faudrait pouvoir encoder des transcriptions comme des mots, voire des caractères (lettre ou signe abréviatif) et simplifier la transcription des abréviations. A l'heure actuelle, toutes les annotations sont de niveau égal, même si leur type diffère, et il n'y a pas d'imbrication des annotations et des zones[16]. De nouveaux projets en cours font naître de grands espoirs, tel Text-Image Linking Environment (TILE).

De l'autre côté, les logiciels d'OCR mériteraient d'être mieux intégrés, afin de limiter la perte d'information lors du passage des données graphiques primaires, analysées par le logiciel, à la sortie en ALTO ou TEI avec caractères Unicode, qu'il s'agisse de segmentation, ou, à l'intérieur d'un ensemble homogène, de reconnaissance de formes. Ces logiciels, même ceux qui reconnaissent certaines abréviations canoniques (*per*, *pre*,

[16] On peut certes créer des catégories telles que « w » et « c » puis programmer une transformation pour inclure les éléments typés « c », mais c'est une étape supplémentaire, et la gestion des différentes zones devrait être adaptée en conséquence si l'on veut disposer d'un affichage par zones imbriquées.

*que*) n'offrent pas d'interface de modification manuelle pour enregistrer une analyse de type paléographique (résolution, typologie, allographes). La société Isako fournit un logiciel pour contrôler et modifier le résultat de la reconnaissance optique (logiciel OCRView), mais celui-ci, bien que couplé avec des logiciels livrant des fichiers ALTO, n'est pas adapté pour un ajout d'informations structurées, allographétiques ou non. Le projet européen Impact (Improving Access to Text)[17] prévoit certes un nouvel outil de correction du texte reconnu, mais il n'est pas encore certain qu'il offrira la possibilité de récupérer la segmentation pour ajouter une information allographétique.

L'ensemble de ces outils ne peut donc pas, à l'heure actuelle, pallier directement la complexité d'un processus de transcription allographétique, ni prétraiter les images en repérant l'espace de la forme (segmentation qui évite aux chercheurs de doubler l'opération d'identification de la forme à celle, ô combien ingrate de sélection des formes). L'intégration de ces fonctionnalités des outils de production forme une étape indispensable afin de faciliter la transcription et l'établissement de liaison avec l'image de l'original.

### 5.2. Créer des outils d'analyse génériques

De l'autre côté, en aval de la production de fichiers à encodage allographétique, il faut améliorer les outils d'analyse, en offrant la possibilité de généraliser le questionnement et l'analyse. Cette dernière consiste à parcourir les bases de données paléographiques : en l'occurrence, en utilisant la TEI, cela revient à *parser* ou procéder à l'analyse syntaxique d'un fichier XML[18].

Une interface générique doit être conçue pour paramétrer l'analyse des sources et obtenir, *in fine*, des informations chiffrées. Cette interface doit permettre non seulement de sélectionner les sources disponibles, mais aussi la nature de l'analyse : ainsi l'utilisateur doit être mesure de préciser sur quels allographes, quels paramètres, quelles abréviations porte son analyse.

Dans l'idéal, il faudrait être en mesure d'exploiter les fichiers existants selon leur pertinence et leur degré d'encodage. Pour cela, nous avons évoqué ci-dessus l'idée d'inscrire de façon formelle dans l'en-tête des fichiers le type de données exploitables. Cette formalisation n'existant pas à l'heure actuelle, une telle fonctionnalité ne peut être envisagée pour faire interopérer des fichiers hétérogènes. Si la situation s'améliore, le logiciel devrait également pouvoir tenir compte des ontologies utilisées et restreindre l'enquête aux sources pertinentes. Par exemple, une enquête à grande échelle sur

[17] Le projet européen Impact (Improving Access to Text) rassemble 26 partenaires du privé et du public (bibliothèques nationales et de recherche) pour 4 ans (2008–2011), afin d'améliorer les techniques de reconnaissance de caractères et la mise à disposition du texte.

[18] C'est avec une démarche de ce type que Florence Clavaud et moi avons travaillé : l'analyse se fait par XSLT, mais l'outil générique prévu au début n'a pas été développé jusqu'au bout.

l'influence de la fin de ligne ou des noms propres et des langues vernaculaires sur le travail des scribes (degré et type d'abréviation) ne peut se faire que sur des fichiers qui intègrent les notions d'abréviation, de lignes, de changements linguistiques et d'anthroponymes.

Nous ne considérons pas ici qu'il soit du ressort d'une telle application de préparer une édition électronique. Édition et encodage allographétique ont en effet des nécessités différentes, même si les deux peuvent se compléter réciproquement et ont en commun une longue histoire d'interdisciplinarité : l'attention accordée à l'une ne suffit pas nécessairement à l'autre[19].

Il serait également bon de programmer des sorties sous un format directement importable ou exploitable par des logiciels standard d'analyse des données, c'est-à-dire non pas seulement sous un format HTML. Ainsi parviendrait-on à créer un environnement de travail où toutes les opérations automatisables seraient prises en charge par des outils adaptés, réservant au paléographe ce qui relève de son expertise : analyse des formes graphiques, élaboration des hypothèses, formalisation du questionnement et analyse des résultats à l'aide des outils de statistiques descriptives.

Que les principes présidant à l'établissement d'une édition ou description numérique dictent les possibilités d'exploitation du texte produit, cela est clair et a déjà donné lieu au jeu de mot « ce que tu (pré)vois est ce que tu obtiens » (Bradley 2005). C'est doublement vrai : aussi bien en amont de la préparation des sources qu'en aval, où l'on ne peut analyser qu'avec les outils à disposition, qu'il faut donc concevoir le plus largement possible. Une compréhension claire de ce que signifie aborder les écritures médiévales et transcrire est un préalable indispensable à une analyse paléographique assistée par ordinateur. Il nous semble que l'opération nommée « transcrire » signifie « décrire » dès lors que l'on s'intéresse aux formes graphiques et à l'image d'un texte autant qu'à son contenu. Aussi faut-il raisonner en termes de description et de structuration des informations, tant celles décrites que celles permettant de désigner les phénomènes décrits (ontologies, vocabulaires, descripteurs, granularité…). Une recherche en paléographie, menée avec des moyens réduits, nous a permis de faire la preuve de concept : une description allographétique minimale ouvre la voie à de nouvelles études sur l'histoire de l'écriture et des systèmes graphiques. Elle suffit à

[19] Notre projet nous mène à une appréciation mitigée sur la complémentarité de l'encodage allographétique et de l'édition : il nous est apparu plus facile de réinsérer les abréviations et allographes étudiés *a posteriori*. Le problème a déjà été signalé par Robinson et Solopova : devoir veiller à trop de paramètres en même temps (abréviations, ponctuations, graphies) nuit à la qualité de la transcription. Dans notre cas, l'édition traditionnelle complète, avec apparat et notes, a précédé systématiquement l'encodage. La production des fichiers avec encodage allographétique n'exige pas l'analyse des noms de lieu et de personne ou de la syntaxe, analyse qui permet les capitalisations et ponctuation modernes que la machine peut restituer. L'étude des comportements scripturaux face aux noms et surnoms est un domaine pourtant fascinant qui peut inciter à étendre le champ d'encodage, surtout si les outils d'encodage rendent la tâche plus aisée.

éclairer des pratiques scripturales et des évolutions qui permettent de dater et d'identifier des écritures.

Pour avancer sur cette voie, il devient nécessaire de coopérer et de faire interopérer les bases de donnée paléographiques. Il faut donc, en amont, modéliser l'information pour pouvoir l'enregistrer dans sa complexité (normalisation et bonnes pratiques) et prévoir des outils qui favorisent une formalisation des choix d'encodage et une saisie aussi complète que possible, puis, en aval, des outils d'analyse. La coopération et le partage sont absolument nécessaires : d'une part entre les paléographes, qui pourront tirer grand profit de l'existence de données de comparaison, en évitant l'enfouissement et la déperdition de leurs propres données après aboutissement d'un projet ; d'autre part, entre les paléographes et les ingénieurs. Ceux-ci reconnaissent déjà l'intérêt de disposer de textes corrigés pour améliorer leurs algorithmes ; la mise en lumière de nouvelles règles correspondant aux différents types d'écritures pourra à son tour nourrir les paramètres exploités par les algorithmes, combinant segmentation, formes des lettres et règles de répartitions des allographes selon la position dans le mot.

Le travail conjoint des paléographes et des développeurs des outils d'analyse créera ainsi une masse d'informations paléographiques exploitable qui, en retour, pourra améliorer les algorithmes de reconnaissance des écritures en fournissant un corpus d'apprentissage. Seules la modélisation, la normalisation et l'élaboration de bonnes pratiques permettront de répondre aux défis industriels et scientifiques posés par les écritures manuscrites médiévales.

## Bibliographie

ABBYY : ABBYY FineReader 10. 2009. <http://finereader.abbyy.com/>.

Anderson, Lisa, et al. *EpiDoc : Guidelines for Structured Markup of Epigraphic Texts in TEI.* Stoa Consortium, 2007. <http://www.stoa.org/epidoc/gl/5/>.

André, Jacques. « Numérisation et codage des caractères de livres anciens. » *Document numérique* 7.3–4 (2007) : 127–142. <http://www.cairn.info/article.php?ID_ARTICLE=DN_073_0127>.

Aussems, Mark, et Axel Brink. « Digital palaeography. » *KPDZ 1.* 293–308.

Beneventano, Domenico, et Sonia Bergamaschi. *Semantic search engines based on data integration systems.* International workshop on distributed agent-based retrieval tools. The future of search Engines' technologies. June 26, 2006 - PULA (CA – Italy). Cagliari : Center for Advanced Studies, Research and Development in Sardinia, 2006. <http://www.crs4.it/ict/dart06/slides/bergamaschi.pdf>.

Benskin, Michael. « The Hands of the Kildare Poems Manuscript. » *Irish University Review* 20 (1990) : 163–193.

Bischoff, Bernhard. « La nomenclature des écritures livresques du IXe au XIIIe siècle. » *Nomenclatures des écritures livresques du IXe au XVIe siècle : premier colloque international de paléographie latine, Paris, 28–30 avril 1953.* Eds. Bernhard Bischoff, Gerard Isaac Lieftinck

et Giulio Battelli. Colloques internationaux du C.N.R.S. – Sciences humaines (Vol. 4). Paris : Édition du C.N.R.S., 1954. 7–14.

Bozzolo, Carla, et al. « Les abréviations dans les livres liturgiques du XV^e^ siècle : pratique et théorie. » *La face cachée du livre médiéval : l'histoire du livre vue par Ezio Ornato, ses amis et ses collègues.* Actas del VIII coloquio del Comité internacional de paleografia latina (Madrid-Toledo, set.–oct. 1987) [Madrid, 1990, 17–27.] Rééd. Ezio Ornato. I libri di Viella 10. Roma : Viella, 1997. 555–565.

Bradley, John. « What You (Fore)see is What You Get. Thinking about usage paradigms for computer assisted text analysis. » *TEXT Technology* 14.2 (2005) : 1-19. <http://texttechnology.mcmaster.ca/pdf/vol14_2/bradley14-2.pdf>.

Brink, Axel, Marius Bulacu, et Lambert Schomaker. « How much handwritten text is needed for text-independent writer verification and identification. » *Proceedings of 19th International Conference on Pattern Recognition (ICPR 2008), 8–11 December, Tampa, Florida.* Los Alamitos : IEEC Computer Society, 2008. <http://figment.cse.usf.edu/sfefilat/data/papers/WeBT6.2.pdf>.

Bulacu, Marius, et Lambert Schomaker. « Automatic handwriting identification on medieval documents. » *14th International Conference on Image Analysis and Processing (ICIAP 2007). Proceedings. 11–13 September, Modena, Italy.* Los Alamitos : IEEE Computer Society, 2007. 279–284. <http://www.ai.rug.nl/%7Ebulacu/iciap2007-bulacu-schomaker.pdf>.

Catach, Nina. « Graphémique. » *Lexikon der romanistischen Linguistik (LRL). Band I,1, Geschichte des Faches Romanistik, Methodologie : das Sprachsystem.* Eds. Michael Metzeltin, Christian Schmitt et Günter Holtus. Tübingen : Niemeyer, 2001a. 736–747.

Catach, Nina. « Graphétique. » *Lexikon der romanistischen Linguistik (LRL). Band I,1, Geschichte des Faches Romanistik, Methodologie : das Sprachsystem.* Eds. Michael Metzeltin, Christian Schmitt et Günter Holtus. Tübingen : Niemeyer, 2001b. 725-735.

Cerquiglini, Bernard « Une nouvelle philologie ? » *Philology in the Internet Era / Philologie à l'ère de l'Internet. International Colloquium / Colloque international.* Budapest : Eötvös Loránd University, 2000.

Cottereau, Emilie. « La copie et les copistes français de manuscrits aux XIV^e^ et XV^e^ siècles. Etude sociologique et codicologique. » Thèse de doctorat. Université Paris 1 – Panthéon-Sorbonne, 2005.

Criado Fernández, Luis, et Rafael Martínez-Tomás. « The problem of constructing general-purpose semantic search engines. » *Methods and models in artificial and natural computation. A homage to Professor Mira's scientific legacy.* Lecture Notes in Computer Science. Vol. 5601. Berlin, Heidelberg : Springer, 2009. 366–74.

Derolez, Albert. *The Palaeography of Gothic Manuscript Books From the Twelfth to the Early Sixteenth Century.* Cambridge studies in palaeography and codicology. Vol. 9. Cambridge : Cambridge University Press, 2003.

Ecole nationale des Chartes. *Conseils pour l'édition de textes médiévaux (Fascicule I, Conseils généraux. Fascicule II, Actes et documents d'archives. Fascicule III, Textes littéraires).* Orientations et méthodes. 3 vols. Paris : Éd. du CTHS - École des chartes, 2001–2003.

Flammarion, Hubert. « Une équipe de scribes au travail au XIII$^{e}$ siècle : le grand cartulaire du chapitre cathédral de Langres. » *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde* 28 (1982) : 271–305.

Flammarion, Hubert. *Cartulaire du chapitre cathédral de Langres*. Nancy : ARTEM, 1995, 449 p. ARTEM Atelier de recherches sur les textes médiévaux. 2$^{e}$ ed. Turnhout : Brepols, 2004. 525 p.

Frioli, Donatella. « La 'Grammatica della Leggibilità' nel manoscrittto cisterciense. L'esempio di Aldersbach. » *Liturgie und Buchkunst der Zisterzienser im 12. Jahrhundert : Katalogisierung von Handschriften der Zisterzienserbibliotheken*. Ed. Charlotte Ziegler. Frankfurt am Main : Peter Lang, 2000. 17–47.

Frioli, Donatella. « Per una storia dello *scriptorium* di Reichersberg. Il prevosto Gerhoch e i suoi 'segretari'. » *Scrittura e civiltà* 23 (1999) : 177–212.

Garand, Monique-Cécile. « Copistes de Cluny au temps de saint Maieul (948–994). » *Bibliothèque de l'École des Chartes* 136.1 (1978) : 5–36.

Gasparri, Françoise. *L'écriture des actes de Louis VI, Louis VII et Philippe Auguste*. Hautes études médiévales et modernes. Vol. 20. Paris, Genève : Minard, Droz, 1973.

Gatos, Basilis, Nikolaos Stamatopoulos, et Georgios Louloudis. *ICDAR2009 Handwriting Segmentation Contest*. 10$^{th}$ International Conference on Document Analysis and Recognition. Athens : Institute of Informatics and Telecommunications. National Center for Scientific Research « Demokritos », 2010. <http://users.iit.demokritos.gr/%7Ebgat/HandSegmCont2009/HandSegmCont2009.pdf>.

Gentleman, Robert, Ross Ihaka, et R Development Core Team. *R*. version 2.5.0 ed : The R Foundation for Statistical Computing, 2007.

Guillot, Céline, et al. « Constitution et exploitation des corpus d'ancien et de moyen français. » *Corpus* 7 (2008). <http://corpus.revues.org/index1495.html>.

Haugen, Odd Einar (ed.). *The Menota handbook : Guidelines for the electronic encoding of Medieval Nordic primary sources*. version 2.0 ed. Bergen : Medieval Nordic Text Archive, 2008. <http://www.menota.org/guidelines>.

Heiden, Serge, Céline Guillot, et Alexei Lavrentiev. *Manuel d'encodage XML-TEI des textes de la Base de Français Médiéval*. 2002–2008. <http://ccfm.ens-lsh.fr/IMG/pdf/Manuel_Encodage_TEI.pdf>.

Heinemeyer, Walter. *Studien zur Geschichte der gotischen Urkundenschrift*. Archiv für Diplomatik. Beiheft 4. Köln : Böhlau Verlag, 1982.

Hofmeister, Wernfried, Andrea Hofmeister-Winter, et Georg Thallinger. « Forschung am Rande des paläographischen Zweifels : Die EDV-basierte Erfassung individueller Schriftzüge im Projekt DA*mal*S. » *KPDZ 1*. 261–292.

Holmes, Martin, et University of Victoria HCMC. *Image Markup Tool. Tool for annotating images using TEI*. version 1.8.1.7 ed. 2010. <http://tapor.uvic.ca/mholmes/image_markup/>.

IMPACT. Den Haag : Koninklijke Bibliotheek, 2008-2010. <http://www.impact-project.eu/index.php>.

KPDZ 1 : *Kodikologie und Paläographie im digitalen Zeitalter – Codicology and Palaeography in the Digital Age*. Eds. Malte Rehbein, Patrick Sahle et Torsten Schaßan. Schriften des

Instituts für Dokumentologie und Editorik 2. Norderstedt : Books on Demand, 2009. En ligne : <urn:nbn:de:hbz:38-29393>, <http://kups.ub.uni-koeln.de/volltexte/2009/2939/>.

Lavrentiev, Alexei. *Proposal for XML markup of Old French text corpora.* The Princeton Charrette Project. Princeton, 2002. <http://www.princeton.edu/%7Elancelot/ss/media/docs/Transcription-Proposal.doc>.

Lavrentiev, Alexei. *Systèmes graphiques de manuscrits médiévaux et incunables français : ponctuation, segmentation, graphies : actes de la journée d'étude de l'ENS LSH, 6 juin 2005.* Langages. Vol. 3. Chambéry : Université de Savoie, 2007.

Leydier, Yann, Jean Duong, et Asma Ouji. *Ulysse 0.3g09.* 2006–2009. <http://liris.cnrs.fr/graphem/?p=73>.

Mane, Laure. *TELplus. WP3 Task 1 – Indexing for usability. A prototype of semantic full-text search engine indexing multilingual OCRed corpus from European digital libraries. Feasibility assessment report.* Den Haag : The European Library, 2010. <http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/documents/TELplus_D3.3_04012010.pdf>.

Mazziota, Nicolas. « Traiter les abréviations du français médiéval. Théorie de l'écriture et pratiques d'encodage. » *Corpus* 7 (2008). <http://corpus.revues.org/index1517.html>.

McGillivray, Murray. « The Cotton Nero A.x. Project. » 1994–2010. <http://people.ucalgary.ca/Scriptor/cotton/>.

McGillivray, Murray. « Statistical analysis of digital paleographic data : what can it tell us ? » *TEXT Technology* 14.1 (2005) : 47–60. <http://texttechnology.mcmaster.ca/pdf/vol14_1_05.pdf>.

Meyer, Wilhelm. *Die Buchstaben-Verbindungen der sogenannten gothischen Schrift.* Abhandlungen der königlichen Gesellschaft der Wissenschaften zu Göttingen, Phil.-hist. Klasse. Vol. Neue Folge, 1,6. Berlin : Weidmannsche Buchhandlung, 1897.

Millares Carlo, Agustin, et José Manuel Ruiz Asencio. *Tratado de paleografía española.* 3$^{e}$ ed. 3 vols. Madrid : Espasa-Calpe, 1983.

MUFI : *Medieval Unicode Font Initiative. MUFI character recommendations.* Bergen, 2009. <http://www.mufi.info/specs/>.

Muzerelle, Denis. « Graphem for Dummies. » *The Manuscript Triangle France-England-Scandinavia. 1100-1300.* Bergen : University of Bergen, 2009. <http://www.uib.no/filearchive/graphemfords.pdf>.

Nicolaj, Giovanna. « Alle origini della minuscola notarile italiana e dei suoi caratteri storici. » *Scrittura e civiltà* 10 (1987) : 49–82.

Oeser, Wolfgang. « Beobachtungen zur Strukturierung und Variantenbildung der Textura. Ein Beitrag zur Paläographie des Hoch- und Spätmittelalters. » *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde* 40 (1994) : 359–439.

Oeser, Wolfgang. « Beobachtungen zur Differenzierung in der gotischen Buchschrift. Das Phänomen des Semiquadratus. » *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde* 47–48 (2001–2002) : 223–83.

Pantarotto, Martina. « La scrittura delle carte bresciane nel sec. XII. » *Scrineum – Rivista* 3 (2005) : 1–20. <http://scrineum.unipv.it/rivista/3-2005/pantarotto.pdf>.

Petrucci, Armando. « Censimento dei codici dei secoli XI–XII. » *Studi medievali,* Ser. 3, 9.2 (1968) : 1115–1194. <http://dida.let.unicas.it/links/didattica/palma/testi/petrucci1.htm>.

Richard, Jean. *Les ducs de Bourgogne et la formation du duché du* XI^e^ *au* XIV^e^ *siècle*. Publications de l'université de Dijon 12. Paris : Les Belles Lettres, 1954.

Richard, Jean. « La chancellerie des ducs de Bourgogne de la fin du XIIème au début du XVème siècle. » *Landesherrliche Kanzleien im Spätmittelalter. Referate zum VI. Internationalen Kongreß für Diplomatik, München 1983*. Vol. 1. Ed. Gabriel Silagi. München : Arbeo, 1984. 381–413.

Robinson, Peter, et Elizabeth Solopova. « Guidelines for the transcription of the manuscripts of the *Wife of Bath*'s Prologue. » *The Canterbury Tales Project Occasional Papers*. Vol. 5. Ed. Norman Blake. Oxford : Office for Humanities Communication, 1993. 19–52. <http://www.canterburytalesproject.org/pubs/transguide-MI.pdf>.

Rumble, Alexander. *The palaeographical material in the C11 Database [Dating and describing eleventh-century vernacular script]*. MANCASS C11 Database Project. Manchester : Manchester Centre for Anglo-Saxon Studies, [2004]. <http://www.arts.manchester.ac.uk/mancass/C11database/data/PalaeogIntro.pdf>.

Salaün, Jean-Michel. « Web, texte, conversation et redocumentarisation. » *JADT 2008 : actes des 9es Journées internationales d'Analyse statistique des Données Textuelles, Lyon, 12–14 mars 2008 : proceedings of 9^th^ International Conference on Textual Data statistical Analysis, Lyon, March 12–14, 2008*. Eds. Serge Heiden et Bénédicte Pincemin. Lyon : Presses universitaires de Lyon, 2008. <http://jadt2008.ens-lyon.fr/spip.php?article197>.

Stokes, Peter A. « Computer-aided palaeography, present and future. » *KPDZ 1*. 309–338.

Stutzmann, Dominique. « Écrire à Fontenay. Esprit cistercien et pratiques de l'écrit en Bourgogne (XIIe–XIIIe siècles). » Thèse de doctorat. Université Paris 1 Panthéon-Sorbonne, 2009a.

Stutzmann, Dominique. « La sobriété ostentatoire : l'esthétique cistercienne d'après les manuscrits de Fontenay. » *Culture et patrimoine cisterciens. Colloque du vendredi 12 juin 2009*. Vol. 4. Parole et Silence / Cours, colloques et conférences. Paris : Collège des Bernardins, 2009b. 46–86.

Tassin, René Prosper, et Charles-François Toustain. *Nouveau traité de diplomatique, où l'on examine les fondemens de cet art […] par deux religieux bénédictins de la Congrégation de S. Maur*. 6 vols. Paris : G. Desprez, 1750–1765.

TEI Consortium. « TEI P5. Guidelines for electronic text encoding and interchange ».TEI consortium, 2007-2010. <http://www.tei-c.org/release/doc/tei-p5-doc/html/>.

TILE : *Text-Image Linking Environment*. Maryland Institute for Technology in the Humanities, 2009. <http://mith.info/tile/>.

Tomasi, Gilbert, et Roland Tomasi. « Approche informatique du document manuscrit. » *KPDZ 1*. 197–218.

Torrens, Maria Jesus. « La paleografía como instrumento de datación. La escritura denominada "littera textualis". » *Cahiers de linguistique hispanique médiévale* 20 (1995) : 345–380.

Tzadok, Asaf, et Eugeniusz Walach. « Adaptive OCR for Books. » Armonk (NY) : International Business Machines Corporation, 2009. <http://www.freepatentsonline.com/7627177.html>.

Uitti, Karl D. *Charrette Project SGML Codes*. 1997. <http://www.princeton.edu/lancelot/ss/materials.shtml#ms-transcriptions>.

Załuska, Yolanta. *L'enluminure et le scriptorium de Cîteaux au XIIe siècle*. Studia et documenta. Vol. 4. Abbaye de Cîteaux (Saint-Nicolas-les-Cîteaux) : Cîteaux Commentarii cistercienses, 1989.

Zamponi, Stefano. « La scrittura del libro nel Duecento. » *Civiltà comunale. Libro, scrittura, documento. Atti del Convegno (Genova, 8–11 novembre 1988)*. Nuova Serie, Vol. 29, fasc. 2. Atti della Società Ligure di Storia Patria. Genova : Società Ligure di storia patria, 1989. 317–346.