Ontologie des formes et encodage des textes manuscrits médiévaux. Le projet ORIFLAMMS

Dominique Stutzmann

D. Stutzmann, « Ontologie des formes et encodage des textes manuscrits médiévaux. Le projet ORIFLAMMS », *Document numérique*, 16/3, 2013, p. 81 95
Version auteur du 11 février 2014 (avant révision)

Ontologie des formes et encodage des textes manuscrits médiévaux

Le projet ORIFLAMMS (ANR- 12-CORP-0010)

Dominique Stutzmann[1]

1. *Institut de Recherche et d'Histoire des Textes, CNRS (UPR 841)*
40 avenue d'Iéna, F-75116 Paris, France
dominique.stutzmann@irht.cnrs.fr

RESUME. *L'approche historique des écritures du Moyen Âge soulève des questions complexes, (1) pour rendre compte, décrire et représenter le substrat graphique par le « codage », (2) pour analyser les formes en tant que telles par Vision par ordinateur, (3) pour procéder à l'analyse des phénomènes graphiques par le « balisage » ou « encodage ». Dans le cadre du projet ANR ORIFLAMMS, les sept partenaires (IRHT, A2iA, CESCM, ENC, ICAR, LIPADE, LIRIS) proposent des solutions nouvelles, en particulier pour l'analyse et l'encodage des abréviations et pour l'alignement texte-image afin de constituer une ontologie des formes écrites du Moyen Âge.*

ABSTRACT. *Medieval scripts are a challenge to historical analysis, as for (1) describing et representing the graphical evidence through "coding"; (2) analyzing and clustering letter forms and their features through Computer Vision; (3) analyzing historical phenomena through "encoding". All seven partners of the ANR funded research project ORIFLAMMS (IRHT, A2iA, CESCM, ENC, ICAR, LIPADE, LIRIS) develop new solutions, esp. for encoding abbreviations and aligning text and image, as a first step to build an ontology of medieval written forms.*

MOTS-CLES : *paléographie – encodage – TEI-P5 - ontologie.*

KEYWORDS: *palaeography – encoding – TEI-P5 - ontology*

DOI:10.3199/JESA.45.1-n © Lavoisier 2012 AR_DOI

D. Stutzmann, « Ontologie des formes et encodage des textes manuscrits médiévaux. Le projet ORIFLAMMS », *Document numérique*, 16/3, 2013, p. 81 95
Version auteur du 11 février 2014 (avant révision)

1. Introduction

Représenter et analyser les écritures anciennes sont deux objectifs distincts, mais pouvant concourir à une même finalité de compréhension historique des phénomènes textuels et graphiques. En effet, les outils informatiques pour étudier l'écriture et, en particulier, la reconnaissance d'écriture, sont actuellement très loin d'offrir des réponses satisfaisantes aux questions posées par l'écriture médiévale. Pour l'heure, malgré des avancées partielles, la variabilité des productions écrites du Moyen Âge a jusqu'à présent eu raison de toutes les tentatives de lecture automatisée et même des études systémiques et classificatoires.

Les questions primordiales que l'on pose au texte écrit (quel est le texte écrit ? quand ? où ? par qui ? comment ? pour qui ? pourquoi ? à quelles fins ?) réfèrent aux différents niveaux de la paléographie et de l'histoire de la culture écrite : paléographie de lecture ; expertise (datation, localisation, attribution) ; histoire sociale et culturelle. Or, si les travaux de codicologie et de paléographie quantitatives ont mis en évidence qu'une partie, au moins, de ces questionnements peut être abordée par les moyens informatiques, les réalisations actuelles ne projettent généralement un rayon de lumière que sur un secteur très délimité (étude d'un manuscrit, d'un auteur, d'une chancellerie).

Dans le cadre du projet ORIFLAMMS, financé par l'Agence nationale de la recherche et Cap Digital, une modification de l'approche nous semble pouvoir changer les conditions actuelles de l'antagonisme entre des écritures infiniment variables et des outils à paramétrer à neuf. La réflexion sur l'ontologie des formes et l'encodage doivent en effet permettre une approche plus souple, où la variabilité n'est plus l'obstacle, mais le cœur même de notre interrogation, ce qui nous permet d'ouvrir un nouveau dialogue avec des disciplines que la paléographie n'aurait jamais dû perdre des yeux, en particulier l'épigraphie et la linguistique.

Après la présentation du projet ORIFLAMMS, de ses origines et des partenaires impliqués, nous insisterons sur les enjeux scientifiques principaux d'un tel projet en discutant les notions d'ontologie et d'encodage : l'une et l'autre posant des questions de granularité et d'interprétation d'une réalité historique complexe et variable.

2. Le projet ANR Oriflamms

2.1. Concept et consortium

Avec l'idée que la variabilité de l'écriture est un concept clé pour avancer dans l'analyse historique (Stutzmann 2014a), plusieurs équipes se sont réunies pour monter un projet commun, le projet ORIFLAMMS (*Ontology Research, Image Features, Letterform Analysis on Multilingual Medieval Scripts*), soit, en français : Recherche en ontologie, Descripteurs d'images, Analyse des formes et lettres des écritures médiévales multilingues. Il est financé par l'Agence Nationale de la Recherche (ANR-12-CORP-0010) et labellisé par le pôle de compétitivité Cap Digital, pour une durée de 36 mois de février 2013 à janvier 2016 (Agence Nationale de la Recherche 2013). Il vise à étudier les écritures du Moyen Âge central et tardif, du XII^e^ au XV^e^ siècle, et le multilinguisme médiéval dans une approche interdisciplinaire et novatrice. À la rencontre d'enjeux épistémologiques, scientifiques, technologiques, industriels et sociétaux, ORIFLAMMS analyse l'évolution des systèmes et formes graphiques des écritures d'un temps long (le Moyen Âge) selon leur contexte de production (écritures usuelles, diplomatiques ou livresques) et leur langue (latin ou vernaculaire).

Ce projet rassemble sept partenaires, sous le pilotage de l'IRHT - Institut de Recherche et d'Histoire des Textes (CNRS, UPR 841). Les autres partenaires relèvent pour trois d'entre eux des Sciences humaines et sociales, à savoir l'École nationale des chartes, le Centre d'Études Supérieures de Civilisation Médiévale (Université de Poitiers, CNRS, UMR 7302), le laboratoire ICAR - Interactions, Corpus, Apprentissages, Représentations (Université Lyon 2, ENS de Lyon, CNRS, UMR 5191). Trois autres partenaires relèvent des Sciences de l'ingénieur : le LIRIS - Laboratoire d'Informatique en Images et Systèmes d'Information (INSA Lyon, Université Lyon 1 et 2, CNRS, UMR 5205), le LIPADE - Laboratoire d'Informatique de l'université Paris Descartes et la société industrielle A2iA, l'un des leaders mondiaux de la reconnaissance d'écriture manuscrite pour les chèques et les adresses postales.

2.2. Du pixel au trait, du trait à la lettre[1]

Plusieurs des partenaires d'ORIFLAMMS ont déjà travaillé ensemble dans le cadre du projet GRAPHEM (Graphem based Retrieval and Analysis for PaleograpHic Expertise of Middle Age manuscripts, 2007-2011), à savoir l'IRHT, l'École nationale des chartes, le LIRIS et le LIPADE. L'objectif consistait en l'amélioration des techniques de classification automatique des écritures, par la création de méthodes efficaces d'accès au contenu des manuscrits et d'expertise paléographique, afin d'améliorer notre compréhension de l'évolution des formes de l'écriture (Muzerelle 2011). Méthodes globales et locales ont été appliquées et plusieurs moteurs de recherche par le contenu développés (Cloppet et al. 2011; Daher et al. 2011; Joutel 2011; Lebourgeois et Moalla 2011; Siddiqi, Cloppet, et Vincent 2011).

Sans entrer dans le détail, il faut pourtant ici expliciter que l'évaluation des résultats a été une opération délicate et que les classifications obtenues étaient loin d'être exemptes de défaut. Cela a mené à des débats sur l'herméneutique dans le cadre de projets transdisciplinaires, ainsi qu'à de nouvelles interprétations et à de nouveaux outils, comme la décomposition en graphèmes, qui a été développée à la demande expresse des paléographes du projet.

Il est aussi apparu que la réalité graphique est trop complexe, car ce n'est pas une réalité discrète ; c'est au contraire un continuum évolutif. Nous avons acquis la certitude, également, que l'ordinateur ne parviendrait pas à faire ce que l'humain ne sait pas faire, et qu'il ne pourrait pas nous expliquer en quoi une cursive du XV^e^ siècle est plus proche historiquement d'une *textualis formata* du même siècle, que celle-ci ne pourrait l'être d'une onciale. Si l'on peut évidemment remettre en cause les raisonnements historicistes, les systèmes actuels ne sont pas suffisamment

---

[1] Définitions des mots techniques utilisés dans la suite :
Allographe (n.m.) : Réalisation concrète d'une lettre, c'est-à-dire l'une des différentes formes qui peuvent représenter la même lettre (par exemple « s », « S », « ſ », « ſ »). Les écritures médiévales sont généralement distinguées selon la présence d'allographes déterminés pour les lettres de l'alphabet.
Graphème (n.m.) : Unité graphique minimale entrant dans la composition d'un système d'écriture (Larousse). Ici, utilisé au sens littéral : trait élémentaire, entrant dans la composition d'un signe graphique.
Ligature (n.f.) : Ensemble de lettres liées et qui forme un caractère unique (« e » et « t » dans « & », « o » et « e » dans « œ »).
Ove (n.m.) : partie arrondie d'une lettre.

performants pour autoriser à bâtir des ponts solides au-delà de la chronologie et rejeter la primauté de la réalité historique.

Lors d'un séminaire à Dagstuhl en 2012, la nécessité est apparue clairement d'établir des « mid-level features », des descripteurs de niveau intermédiaire qui permettraient de décrire l'écriture à mi-chemin entre le pixel et l'image entière, entre les caractéristiques de chaque trait et celles de l'écriture qui se manifestent pour l'humain par une impression générale difficile à exprimer de façon objective (Hassner et al. 2013). Ces descripteurs de niveau intermédiaires pourraient être des formes, des caractéristiques du style, des allographes, des lettres, c'est-à-dire des concepts que les paléographes ont l'habitude de manier et discuter et qui peuvent en conséquence servir de truchement dans le dialogue avec la machine et ses programmeurs.

Nous avons souligné ailleurs combien les répertoires de « glyphes » tels que ceux proposés dans le monde anglo-saxon n'aident guère l'analyse, car ils sont à la fois trop riches et trop peu structurés (Stutzmann 2013b). En revanche, le dialogue entre ingénieurs et paléographes peut s'en nourrir et s'approfondir tant par l'aide à la modélisation des « formes » (Ciula 2004; Ciula 2005) que par le retour à des descripteurs de niveau intermédiaire. Cette approche est riche de promesses et apportera assurément une contribution originale aux enquêtes internes à la paléographie, par exemple sur les règles d'emploi des différentes formes, ouvertes avec les travaux de W. Oeser sur les formes de **a** dans la *textualis* (Oeser 1971; Oeser 1994; Oeser 2001), et auxquelles nous avons également contribué en proposant de porter le regard sur les règles d'emploi des autres allographes et en sortant de catégories binaires (Stutzmann 2009; Stutzmann 2010; Stutzmann 2013a; Stutzmann sous presse). Tandis que les travaux et classifications actuelles se fondent sur des oppositions deux à deux (majuscule/minuscule ; **a** rond / **a** fermé ou à crosse **; a** en boîte / **a** à tête en retrait, etc.), il apparaît désormais qu'il faut identifier les formes et les hiérarchiser, c'est-à-dire bâtir une « ontologie des formes », avant d'étudier leurs emplois[2]. C'est là qu'intervient le projet ORIFLAMMS.

2.3. Enjeux

C'est en suivant l'angle des systèmes graphiques, unissant la linguistique, la paléographie et l'épigraphie que le projet ORIFLAMMS entend avancer. En abolissant les frontières disciplinaires traditionnelles et en créant un corpus de référence pour toutes les études sur l'histoire de l'écriture et de la langue, ainsi qu'en créant des outils d'exploitation et en mettant en œuvre des standards et des bonnes pratiques, nos équipes entendent ouvrir des verrous scientifiques : dépasser les cloisonnements disciplinaires entre les disciplines SHS que sont la linguistique, l'histoire, l'épigraphie, la diplomatique et la paléographie, ou entre domaines linguistiques (latin / vernaculaire) et même entre champs de la science (STIC et analyse d'image, et SHS), et contribuer à libérer l'histoire de l'écriture de son statut de « science auxiliaire » pour la concevoir comme part intégrante de l'histoire culturelle et mentale. Les verrous technologiques à ouvrir sont la dispersion des corpus existants, la lourdeur de préparation des sources pour les analyses et la difficulté à appréhender la variabilité dans les outils d'analyse de l'écriture.

En matière d'histoire de l'écriture, ORIFLAMMS propose la création d'une ontologie des formes en considérant l'écriture comme objet unique, pour mener à une réflexion théorique et à une nouvelle vision des écritures, en étudiant spécifiquement la stabilité et variabilité de leurs systèmes graphiques selon les lieux, les temps et les

[2] Sur le concept de « ontologie des formes », voir ci-dessous (§ 3.2.).

langues. En linguistique, l'approche (paléo)graphique enrichira le concept déjà opérant de systèmes graphiques, pour aborder à nouveaux frais l'étude des abréviations, de la ponctuation, de la segmentation et des relations inter-langues.

Dans le domaine industriel et technologique, l'objectif est de remédier à l'absence de base d'entraînement et de validation, et d'étudier spécifiquement la cursivité et la variabilité des écritures, en améliorant les technologies d'alignement et d'enrichissement des transcriptions, puis en établissant des mesures de similarité. Dans les humanités numériques, ce projet propose un changement d'échelle et une modification des paradigmes, en prévoyant de passer d'un univers de corpus dispersés à un Corpus de référence unifié, comprenant textes et images, réutilisable et pérenne, doté des outils d'exploitation ergonomiques. L'objectif est aussi de répondre aux attentes sociétales (multilinguisme actuel et apprentissage des langues et écritures) et économiques (outils d'analyse de l'écriture dans un environnement numérique).

3. « Coder » et « encoder », ou : « représenter » et « analyser » ?

3.1. Constituer un corpus de référence aligné

La première étape du projet ORIFLAMMS consiste en la constitution d'un corpus de référence, large et représentatif des lieux, temps, langues et modes d'écritures du Moyen Âge, en mettant en commun des ensembles déjà élaborés dans le cadre de projets antérieurs et de procéder au repérage de sources complémentaires pour assurer la représentativité de l'ensemble.

En se positionnant contre ce qui est, aujourd'hui encore, la norme dans le domaine de l'édition numérique, à savoir la séparation de l'image et de l'édition, ORIFLAMMS s'inscrit volontairement à la suite d'initiatives rapprochant la transcription et l'édition de l'image, comme dans les dossiers documentaires de Thélème (École nationale des chartes 2007) ou les outils de transcription et d'annotation de T-PEN, de Monasterium.net et de l'*Album interactif de paléographie* (Ginther et Firey 2012; ICARUS – International Centre for Archival Research 2011; Burghart et UMR 5648 - Histoire, Archéologie, Littératures des Mondes Chrétiens et Musulmans Médiévaux 2011).

Aussi le corpus associera-t-il les images des écritures médiévales, les textes eux-mêmes, sous forme informatique, inclus dans un document XML-TEI pouvant associer chaque unité graphique du texte à sa représentation en image, et des métadonnées descriptives, comprenant date et lieu de production, ainsi que auteur, titre et typologie textuelle de l'œuvre.

Le Corpus de référence se subdivise en écritures diplomatiques et pragmatiques (XII^{e}-XV^{e} siècles : chartes latines et françaises de la France de l'Ouest, de Bourgogne et de Champagne ; registres de la chancellerie royale ; registre notarial du Sud de la France) ; écritures livresques (XII^{e}-XV^{e} siècles : manuscrits français et latins de dates et lieux représentatifs, manuscrits latins et français d'œuvres de large diffusion, manuscrits bilingues, corpus Queste du Graal, corpus Projet Charrette, corpus BFM – Manuscrits) ; écritures gravées (inscriptions latines et françaises) et imprimées.

Pour que ce corpus serve à l'analyse historique du développement des systèmes graphiques, il apparaît immédiatement que le « balisage », ou « encodage », est nécessaire et indispensable. Un simple « codage » visant un fac-simile numérique dépassant en fidélité les éditions « record type » de la tradition anglo-saxonne, ne permettrait pas une analyse complète, graphique et linguistique. Il ne rendrait en effet

pas compte de la nature complexe du système graphique médiéval, dont les variations peuvent, certes, parfois s'expliquer par la structure physique du document, telle que page, colonne, ligne, mais où la sémantique même des mots tient un rang majeur (Stutzmann 2013b).

3.2. L'encodage des abréviations

Dans le domaine latin encore plus que dans le domaine vernaculaire, il est en effet impossible d'étudier le système abréviatif en se contentant de « coder » les signes ou de « transcrire » la résolution des abréviations, car il n'y a pas de relation bijective entre les deux ensembles.

De ce point de vue, au demeurant, les solutions préliminaires et questions que nous avons posées à la suite de N. Mazziotta n'ont pas trouvé de réponse stricte (Mazziotta 2008; Stutzmann 2010; Stutzmann 2013b). La façon même de composer les entités qui décrivent les abréviations doit être normalisée, et tenir compte des impératifs d'ergonomie sans obérer la possibilité de restituer le texte (pour une plus grande acceptation par les chercheurs), ni, surtout, d'analyser la substance graphique. Comme toujours, les choix doivent être faits à quatre niveaux : identifiant ; valeur ; structuration ; granularité[3].

3.1.1. Identifiant

Ainsi, nous proposions de traiter les abréviations par des entités donnant la prééminence à la forme résolue pour des raisons d'ergonomie : c'est, entre autres, un moyen facile de retrouver l'entité par copier-coller lorsque l'on encode un texte déjà transcrit. Dans le cadre d'ORIFLAMMS, la tendance actuelle se fait jour, avec l'emploi d'entités nommées d'après la substance graphique dont elles rendent compte : par exemple « &e-tilde-est; » à la place de « &est-e-tilde; ».

3.1.2. Structuration

Pour la structuration même de l'entité et de son contenu, en lien avec les travaux du laboratoire ICAR, nous cherchons aussi à affiner la structuration de l'entité. En effet si, conformément aux traditions paléographique et diplomatique françaises (École nationale des chartes 2001), l'on désire indiquer les lettres restituées par une mise en forme telle que l'italique, plutôt que de marquer tout le lieu de l'abréviation, à comparer avec le texte restitué, il faut spécifier des entités plus grandement structurées. Ainsi, on distinguera à l'intérieur même de la restitution ce qui est présent sur l'image. Les deux exemples ci-dessous montrent la structuration nécessaire pour rendre correctement l'abréviation « p, i suscrit » pour « p*ri* » et « pp, tilde » pour « p*ropter* »[4] :

---

[3] Il nous faut ici remercier tout particulièrement et chaleureusement Mlle Irene Ceccherini et M. Alexey Lavrentev pour nos échanges sur ces questions.

[4] Nous utilisons ici les recommandations de la Text Encoding Initiative (http://www.tei-c.org/), en particulier les éléments suivants :
<abbr> (abréviation) : contient une abréviation quelconque ;
<choice> (choix) : regroupe un certain nombre de balisages alternatifs possibles pour un même endroit dans un texte
<ex> (développement éditorial) : contient une succession de lettres ajoutées par un éditeur ou un transcripteur pour développer une abréviation ;
<expan> (expansion) : contient l'expansion d'une abréviation ;

```
<!ENTITY p-i-suscrit-pri "
<choice>
  <expan>p<ex>r</ex>i</expan>
  <abbr>p&#x0365;</abbr>
</choice>">

<!ENTITY pp-tilde-propter "
<choice>
  <expan>p<ex>ro</ex>p<ex>ter</ex></expan>
  <abbr>p&#x0304;p</abbr>
</choice>">
```

Pour des abréviations portant sur un mot complet, la solution en TEI est claire. Elle l'est malheureusement moins si l'abréviation porte sur une partie de mot, et le problème est d'autant plus délicat que de nombreuses prépositions peuvent être considérées selon le contexte comme indépendantes ou comme partie de mot. Faut-il deux entités pour respecter la distinction de la TEI qui veut que <abbr> porte sur un mot entier ? Nous pensons que non.

3.1.3. Valeur

Enfin, il ne faut pas oublier que l'encodage est une interprétation. Préliminaire, certes, mais qui risque d'entraîner des biais d'analyse si l'on n'en prend garde. Par exemple, nous avions noté que des abréviations semblent construites sur un radical, tel que « ai », « dn », « hoi » pour « anima », « animus », « dominus » et « homo, -inis ». C'est une hypothèse sur le fonctionnement des abréviations qui n'est pas pleinement démontrée et s'oppose, en tout cas, à la tradition paléographique qui considère ces abréviations comme étant des contractions. Par prudence méthodologique, nous sommes revenus sur ce choix. Si l'on respecte l'analyse paléographique lors de l'encodage, les entités les décrivant doivent, par conséquent, englober également la dernière lettre. La perte en ergonomie peut être réelle, puisque ce choix impose de multiplier le nombre d'entités (par exemple &anima; et &animo; au lieu de &anim; +a ou +o). En revanche, le rôle de l'analyse intellectuelle est souligné, par une structuration correspondant mieux aux théories historiques actuelles. Du point de vue analytique *a posteriori*, la machine peut indifféremment traiter les deux chaînes de caractères et ouvrir toutes les voies à l'heuristique paléographique. Pourtant, il s'agit d'un choix de valeurs qui n'est pas sans conséquence. De même que la majuscule codée « A » et l'encodage <hi rend='capitale'>a</hi> ne sont pas strictement équivalents, car ils demandent des traitements différenciés, la valeur de l'abréviation, et sa conformité aux théories paléographiques communes, risquent d'obérer l'analyse, soit en l'enfermant dans des théories anciennes, soit, plus simplement, en empêchant un traitement automatisé en cas d'hétérogénéité des corpus. Dans le cadre d'ORIFLAMMS, nous nous orientons actuellement vers une homogénéisation des corpus avec des entités conformes aux théories actuelles de la paléographie.

3.1.4. Granularité : abréviations et allographes

Un problème, déjà maintes fois soulevé, est celui des allographes, de leur définition et de leur gestion ; nous y reviendrons ci-dessous. Le problème est,

---

<hi> (mis en évidence) : distingue un mot ou une expression comme graphiquement distincte du texte environnant, sans en donner la raison ;

néanmoins, d'une autre nature quand l'on veut traiter à la fois des allographes et des abréviations. En effet, s'il faut, d'une part, indiquer la nature graphique des signes visibles (les lettres en leurs divers allographes et les signes d'abréviations) et, d'autre part, indiquer comment résoudre l'abréviation, ce sont deux fonctionnalités indépendantes et à traiter séparément, car correspondant à des facteurs distincts de l'histoire des écritures.

En limitant le nombre d'allographes étudiés aux formes qui ont perduré via la longue tradition typographique ou dans d'autres alphabets (a, *ɑ* ; d, ð ; e, ę ; i, j ; s, ſ, *ʃ*, ß ; u, v, w), et, en particulier aux formes de **d** et **s**, qui ont déjà suscité des études paléographiques, mais sans tenir compte ni des autres lettres (r, demi-r, r-rond), d'une part, et, d'autre part, en négligeant d'autres phénomènes paléographiques tels que la fusion des oves contraposés, les ligatures, et les élisions et superposition (Zamponi 1988), l'on pouvait espérer traiter les différents phénomènes graphiques par un seul outil, en créant des entités distinctes pour les abréviations selon les allographes qui les signalent, en distinguant par exemple, les abréviations « dnſ » et « ðn*ʃ* ». La question vaut, du reste, également pour les abréviations qui sont réalisées sous des formes différentes, mais restent, dans l'analyse paléographique, une seule et même abréviation (par exemple, « que » avec **q** et semi-colon réalisé sous forme de **z** plongeant).

Avec des entités correspondant à la théorie paléographique, le nombre de celles-ci s'est, nous l'avons dit considérablement accru. Du point de vue théorique de la modélisation d'information comme du point de vue pratique, il est déraisonnable de conserver des entités différentes, en dédoublant ou triplant les abréviations pour tenir compte des allographes. Une autre voie doit être trouvée. La première serait un retour à une écriture sans entité, où les allographes pourraient être insérés directement au fil du document XML. Néanmoins, l'utilisation d'entités correspond parfaitement à l'opération intellectuelle d'identification de signes et de proposition d'une résolution en tenant compte du contexte historique et linguistique. Sans doute la solution résidera dans une approche en stand-off qui reste à spécifier, pour concorder avec un autre objectif majeur du projet ORIFLAMMS : établir une ontologie des formes.

3.2. Gagner en granularité grâce à l'alignement du texte et de l'image : vers l'ontologie des formes

Pour dépasser les distinctions binaires, il est nécessaire d'aller au-delà d'une analyse graphique d'après une transcription qui est, par la nature des choses, déjà une interprétation, une réduction à une chaîne de caractères de ce qui est une réalité graphique en deux (voire trois) dimensions, et d'élaborer aussi finement que possible, ce que nous appelons « ontologie des formes ».

Le mot « ontologie » est ici entendu dans son acception des sciences informatiques, c'est-à-dire comme une description formelle de la réalité et des relations entre concepts, répartissant les objets et les concepts (« individus ») en différentes classes subordonnées ou associées les unes aux autres, et définies par leurs attributs et leurs relations. En appliquant cette approche aux écritures médiévales et aux signes qui les composent, nous considérerons que chaque réalisation graphique d'un signe (par exemple les différents « a » qui composent un exemplaire précis d'un texte) est un élément qui doit être attribué à une « classe » ; chacune de ces « classes » correspond à un « allographe ». Celles-ci pourront, les cas échéant, être réunies dans des classes de niveau supérieur.

Cette ontologie sera formalisée à l'aide d'une grammaire spécifique, le langage OWL (Web Ontology Language, http://www.w3.org/TR/2012/REC-owl2-primer-

20121211/). L'intérêt d'une formalisation extérieure aux documents de transcription et d'encodage est double. Tout d'abord une telle ontologie formalisée, avec ses critères de distinction et de rapprochement, contribuera aux efforts actuels de la communauté paléographique pour créer des critères objectifs d'expertise. Ensuite, en permettant l'emploi des technologies du web sémantique et des regroupements variés, elle offrira la possibilité de tester des hypothèses multiples sur la réalité graphique et de s'interroger sur la perception par les médiévaux des formes qu'ils employaient.

Pour éviter un raisonnement tautologique et une interprétation des formes à l'aune de nos connaissances préalables sur les écritures employées, nous considérons qu'il faut éviter de sélectionner des « spécimens » dans des échantillons d'écriture, mais, au contraire, extraire l'ensemble des signes écrits pour, ensuite, proposer une classification formelle objective, et étudier particulièrement les formes hybrides et intermédiaires.

Aussi s'avère-t-il indispensable d'indiquer les coordonnées sur l'image d'origine des caractères transcrits. C'est en effet ainsi que l'on pourra associer chaque lettre – ou chaque allographe – avec ses instanciations, c'est-à-dire ses réalisations concrètes et chaque fois uniques, et de permettre une analyse fondée sur la forme pure, et non catégorisée *a priori*.

À cette fin, les équipes du projet ORIFLAMMS – en particulier le LIRIS – développe un outil d'alignement pour associer l'image et le texte, apte à enrichir les transcriptions de métadonnées de structure et de coordonnées graphiques, par extraction des données structurelles de l'image, encodage de l'image et de la transcription associée, mise en relation de l'image et de la transcription.

En effet, plutôt que de recommencer le travail à neuf, l'idée qui a prévalu à la conception du projet ORIFLAMMS est d'exploiter les corpus adéquats déjà existants, dont certains étaient déjà munis de transcriptions allographétiques (Charrette, Graal, Fontenay). L'objectif est de disposer d'un système qui aligne les transcriptions actuelles et les images et fournisse une interface pour corriger et enrichir les transcriptions, à partir d'éditions séparant texte et image (fig. 1). À l'heure actuelle, le logiciel développé permet déjà d'aligner un texte transcrit avec indication de changement de colonnes et de lignes. Une première interface de visualisation s'ouvre à l'utilisateur, où les mots alignés sont superposés à l'image du texte ; par défaut ils sont sur fond jaune, pour indiquer qu'ils n'ont pas été validés par intervention humaine (fig. 2). Dans une interface de validation, les multiples occurrences d'un même mot sont visualisées de façon tabulaire, de sorte que l'œil humain identifie aisément les erreurs d'alignement ; l'opération de validation consiste à rejeter les alignements fautifs pour en permettre la correction manuelle (fig. 3). Dans l'interface de visualisation, les mots rejetés sont marqués en rouge ; ceux validés, en vert. La correction de l'alignement est réalisée dans l'interface de visualisation : en effet, les limites extrêmes des mots alignés peuvent être déplacées par glisser-déposer dans l'interface de visualisation (fig. 2). La double interface, de validation et de correction, cherche la plus grande ergonomie en tirant parti des caractéristiques cognitives humaines qui rendent les individus à même de percevoir facilement l'altérité dans un ensemble homogène. La validation dans une interface tabulaire permet de n'avoir pas à relire tout le texte lors de la correction manuelle, et de se concentrer, au contraire, sur les alignements fautifs dans l'interface de correction.

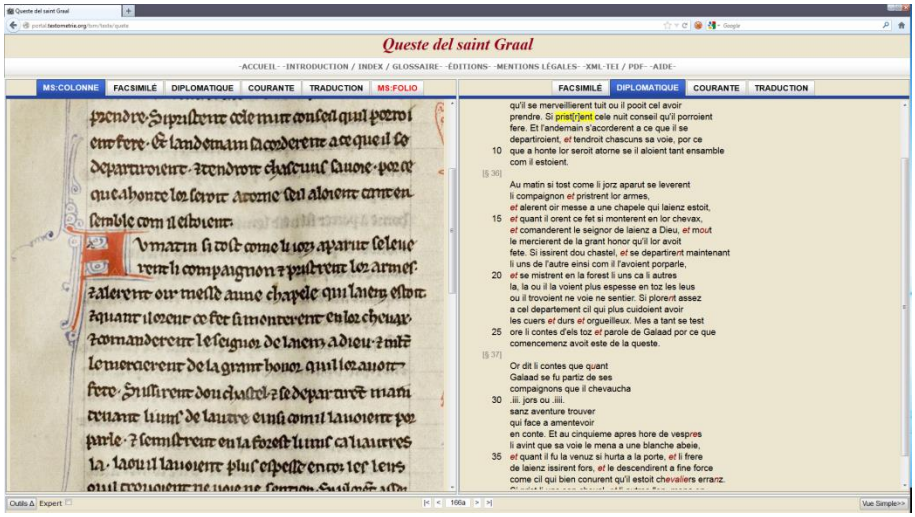

Figure 1 : Interface actuelle de l'édition électronique de la Queste du Graal (éd. C. Marchello-Nizia, collab. A. Lavrentev)

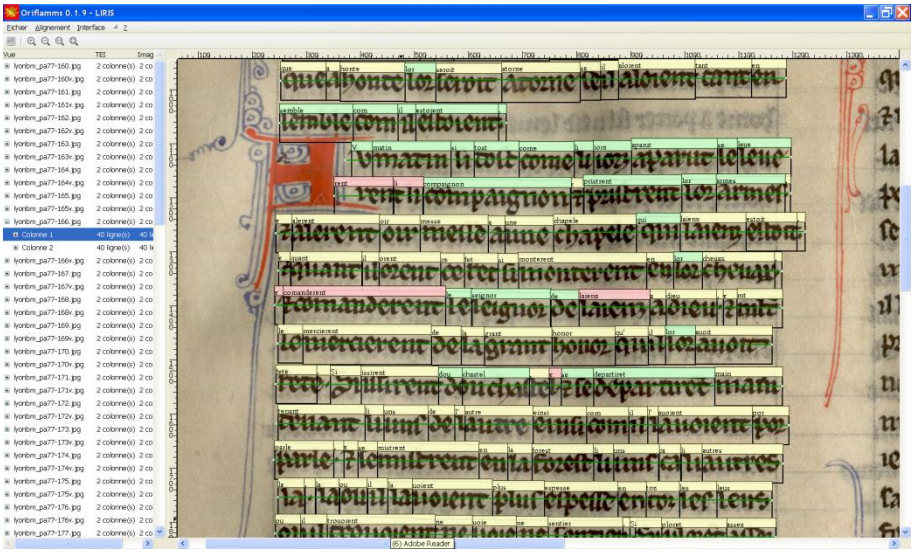

Figure 2 : Interface de visualisation et de correction de l'alignement texte-image (logiciel développé par Y. Leydier et le LIRIS dans le cadre du projet ORIFLAMMS)

Figure 3 : Interface de validation de l'alignement texte-image (logiciel développé par Y. Leydier et le LIRIS dans le cadre du projet ORIFLAMMS)

À cette étape l'alignement corrigé permet déjà de mettre à disposition un terrain d'entraînement pour la reconnaissance de mots (*wordspotting*). Dans une étape ultérieure, nous nous proposons de descendre au niveau de la lettre. Accéder à ce niveau de granularité plus fin permettra de mieux modéliser chaque signe graphique et ouvrira la voie à une spécification plus claire des allographes et à leur hiérarchisation, autrement dit, à l'établissement d'une ontologie des signes. La distinction entre allographes doit en effet combiner l'analyse humaine, historique et graphique des *ductus*, et la vision par ordinateur. L'ontologie des signes doit, à son tour, non seulement améliorer les algorithmes d'alignement et de reconnaissance par la meilleure connaissance des formes, mais aussi alimenter la recherche en paléographie et en épigraphie. En effet, une meilleure qualification des allographes et des éléments de styles augmentée des possibilités d'analyse d'image permettra enfin d'aborder de façon renouvelée la question des ambigüités et parallèles graphiques et de leur rôle dans les modes de signification d'une écriture. Ainsi il est bien connu des historiens de l'écriture qu'à certaines époques les lettres **o** et **r**, **e** et **c**, **g** et **s** partagent certaines de leurs réalisations ; que d'autres s'opposent par des symétries (**p** et **q**, **p** et **b**, **b** et **d**, partiellement **a** et **e**), des rotations (**p** et **d**, **N** et **Z**), parfois augmentées de positions divergentes par rapport à la ligne d'écriture. Ces ambiguïtés qui fondent aujourd'hui la substance de nombreux logos d'entreprises, influencèrent aussi l'histoire évolutive des écritures selon les assimilations qui eurent lieu successivement. Comprendre les interactions entre les formes et les signes, selon leur nature graphique et leur fréquence au sein du système graphique constituera une base pour ce que Marc Smith appelle la « graphonomie historique » et pour de futurs développeemnts en reconnaissance de l'écriture manuscrite.

4. Conclusion

Dans le cadre d'ORIFLAMMS, les sept partenaires institutionnels, publics et privés, travaillent concomitamment à une spécification commune des normes

d'encodage et à l'élaboration d'une ontologie des signes graphiques médiévaux. L'encodage des abréviations doit rendre compte de l'opération intellectuelle de restitution tout en autorisant l'enregistrement des caractéristiques graphiques et une interprétation linguistique adéquate. Les spécifications actuelles et les formats doivent être affinés. L'ontologie des formes, établie sur la base de l'alignement des textes et images dans le cadre du projet de recherche, servira de base, tant graphique que descriptive, à l'analyse des systèmes graphiques des écritures médiévales et de leur variabilité selon les temps, les lieux, les langues et les modes d'écriture. Ce n'est que l'analyse *a posteriori* des données qui permettra de contribuer au débat sur la valeur « contrastive » des différentes formes, donc leur signification au sein du système graphique, et leur valeur comme caractère. Les conclusions de cette analyse nourriront aussi les logiciels de reconnaissance et les modèles heuristiques des SHS pour redéfinir ce qu'est une écriture, les zones de stabilité et de forte variabilité des systèmes graphiques médiévaux.

Bibliographie non numérotée et références

Agence Nationale de la Recherche. 2013. « Projet ANR. Corpus, données et outils de la recherche en sciences humaines et sociales (Corpus) 2012 : projet ORIFLAMMS ». *ANR - Agence Nationale de la Recherche*. http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2%5BCODE%5D=ANR-12-CORP-0010.

Burghart, Marjorie, et UMR 5648 - Histoire, Archéologie, Littératures des Mondes Chrétiens et Musulmans Médiévaux. 2011. « Album interactif de paléographie médiévale / Interactive Album of Mediaeval Palaeography ».http://ciham.ish-lyon.cnrs.fr/paleographie/.

Ciula, Arianna. 2004. Modelli digitali di scrittura carolina. *Gazette du livre médiéval*, nº 45, p. 27‑38.

———. 2005. Digital palaeography: using the digital representation of medieval script to support palaeographic analysis. *Digital Medievalist*, nº 1, http://www.digitalmedievalist.org/journal/1.1/ciula/.

Cloppet, Florence, Hani Daher, Véronique Églin, Hubert Emptoz, Matthieu Exbrayat, Guillaume Joutel, Frank Lebourgeois, et al. 2011. New Tools for Exploring, Analysing and Categorising Medieval Scripts. *Digital Medievalist*, nº 7), http://www.digitalmedievalist.org/journal/7/cloppet/.

Daher, Hani, Véronique Églin, Stéphane Brès, et Nicole Vincent. 2011. « Étude de la dynami que des écritures médiévales. Analyse et classification des formes écrites », *Gazette du livre médiéval*, nºs 56-57, p. 21‑41.

École nationale des chartes. 2001. *Conseils pour l'édition de textes médiévaux (Fascicule I, Conseils généraux. Fascicule II, Actes et documents d'archives. Fascicule III, Textes littéraires)*. *Orientations et méthode*s. Éd. du CTHS/École des chartes, Paris.

———. 2007. Dossiers documentaires. *THELEME : Techniques pour l'Historien En Ligne, Etudes, Manuels, Exercices*, http://theleme.enc.sorbonne.fr/dossiers/.

Ginther, James, et Abigail Firey. 2012. T-PEN. Transcription for paleographical and editorial notation. *T-PEN*. http://t-pen.org/TPEN/.

Hassner, Tal, Malte Rehbein, Peter A. Stokes, et Lior Wolf. 2013. Computation and Palaeography: Potentials and Limits. *Dagstuhl Manifestos*, vol. 2, p. 14‑35.http://dx.doi.org/doi:10.4230/DagMan.2.1.14.

ICARUS – International Centre for Archival Research. 2011. *Monasterium.Net*. http://www.mom-ca.uni-koeln.de/mom/home.

Joutel, Guillaume. 2011. Analyse d'écritures médiévales basée sur la décomposition en curvelets. *Gazette du livre médiéval*, nos 56-57, p. 58-71.

Lebourgeois, Frank, et Ikram Moalla. 2011. Caractérisation des écritures médiévales par des méthodes statistiques basées sur les cooccurrences. *Gazette du livre médiéval*, nos 56-57, p. 72‑100.

Mazziotta, Nicolas. 2008. Traiter les abréviations du français médiéval. Théorie de l'écriture et pratiques d'encodage. *Corpus*, no 7. http://corpus.revues.org/index1517.html.

Muzerelle, Denis. 2011. À la recherche d'algorithmes experts en écritures médiévales. *Gazette du livre médiéval*, nos 56-57, p. 5‑20.

Oeser, Wolfgang. 1971. « Das "a" als Grundlage für Schriftvarianten in der gotischen Buchschrift ». *Scriptorium*, vol. 25, no 1, p. 25‑45.

———. 1994. Beobachtungen zur Strukturierung und Variantenbildung der Textura. Ein Beitrag zur Paläographie des Hoch- und Spätmittelalters. *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde*, vol. 40, p. 359‑439.

———. 2001. Beobachtungen zur Differenzierung in der gotischen Buchschrift. Das Phänomen des Semiquadratus. *Archiv für Diplomatik, Schriftgeschichte, Siegel- und Wappenkunde*, vol. 47-48, p. 223‑283.

Siddiqi, Imran, Florence Cloppet, et Nicole Vincent. 2011. Writing property descriptors. A proposal for typological groupings. *Gazette du livre médiéval*, nos 56-57, p. 42‑57.

Stutzmann, Dominique. sous presse. L'écriture, réalité esthétique ? Ordre et régularité chez les Cisterciens de Fontenay. *Bollettino dei Classici dell'Accademia Nazionale dei Lincei. Supplemento*.

———. 2009. *Écrire à Fontenay. Esprit cistercien et pratiques de l'écrit en Bourgogne (XIIe-XIIIe siècles)*. Thèse en histoire médiévale ; Université Paris 1 Panthéon-Sorbonne.

———. 2010. Paléographie statistique pour décrire, identifier, dater... Normaliser pour coopérer et aller plus loin ? *Kodikologie und Paläographie im digitalen Zeitalter 2 - Codicology and Palaeography in the Digital Age 2*, (*Schriften des Instituts für Dokumentologie und Editorik* , no 3). Norderstedt, p. 247‑277.

———. 2013a. Diplomatique et paléographie numériques : exploiter le profil scribal collectif pour dater et attribuer les chartes. *« Digital diplomatics » or: The computer as a tool for the diplomatist?*(*Archiv für Diplomatik. Beiheft*, vol. 14).Böhlau, Köln, sous presse.

———. 2013b. Système graphique et normes sociales : pour une analyse électronique des écritures médiévales. *Medieval Autograph Manuscripts. Proceedings of the XVIIth Colloquium of the Comité International de Paléographie Latine, held in Ljubljana, 7-10 September 2010* (dir. Nataša Golob) (*Bibliologia*, vol. 36). Brepols, Turnhout, p. 429-434.

———. 2013c. Paléographie latine et vernaculaire (livres et documents). *Annuaire de l'École pratique des hautes études (EPHE), Section des sciences historiques et philologiques* , vol. 144, p. 115‑128.

———. 2014. Variability as key factor for understanding medieval scripts: the ORIFLAMMS project. *Digital Palaeography*(dir. Peter A. Stokes, Malte Rehbein, et Stewart J. Brookes). Ashgate, Farnham, sous presse.

Zamponi, Stefano. 1988. Elisione e sovrapposizione nella littera textualis. *Scrittura e civiltà,* vol. 12, p. 135‑176.

Biographie

Dominique Stutzmann est chercheur en histoire de l'écriture médiévale. Il est actuellement responsable de la section de paléographie latine au sein de l'Institut de Recherche et d'Histoire des Textes (CNRS, UPR 841). Il dirige le projet ANR Oriflamms, ainsi que le projet « Saint-Omer » en collaboration avec la Bibliothèque d'Agglomération de Saint-Omer et l'École nationale des chartes. Archiviste paléographe (2002) et docteur en histoire (2009), de formation, il a aussi été chargé de conférences à l'École pratique des hautes études de 2007 à 2013. Il est actuellement membre du bureau exécutif de Digital Medievalist et du comité de direction de la revue *Scriptorium.*