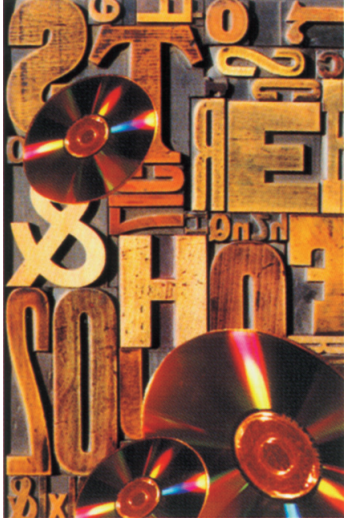


septembre-décembre 2013

Document numérique

RSTI série DN • Volume 16 – n° 3/2013

Gestion informatisée des écritures anciennes

sous la direction de

Christine Bénévent
Rémi Jimenes
Guillaume Sarah

Hermes Lavoisier

Document numérique

Sommaire Volume 16 – n° 3/2013

GESTION INFORMATISÉE DES ÉCRITURES ANCIENNES

INTRODUCTION

Le présent numéro de la revue *Document numérique* réuni sept articles issus d'un colloque organisé au Centre d'Études supérieures de la Renaissance (Tours) les 21 et 22 mai 2013. Cet événement trouve son origine dans la rencontre, un an auparavant, de deux équipes de recherche travaillant sur des sources de nature et de périodes différentes. À l'Institut de Recherche sur les Archéomatériaux (Centre Ernest Babelon, Orléans), Guillaume Sarah se penchait sur la question de la transcription des légendes monétaires et réfléchissait aux possibilités de réalisation d'une police pour l'épigraphie du haut Moyen Âge. À Tours, l'équipe des Bibliothèques Virtuelles Humanistes était confrontée au problème de la transcription des caractères dits « spéciaux », nombreux dans les imprimés de la Renaissance. Un appel à projet de la Maison des Sciences de l'Homme Val de Loire rendit possible la rencontre de nos deux équipes. Une première réunion de travail nous permit de constater les similitudes des problèmes, tant épistémologiques que techniques, auxquels nous étions respectivement confrontés. L'idée d'élargir le cadre de ce dialogue s'imposa dès lors à nous comme une évidence : il s'agissait de réunir des spécialistes issus de divers horizons professionnels, travaillant sur des sources et des périodes historiques variées, afin de dresser un état des problèmes posés et des perspectives ouvertes pour la *gestion informatisée des écritures anciennes*. Les problématiques étant largement partagées, nous avons jugé utile d'ouvrir au public ces séances de travail. Au vu de l'affluence, restée constante deux jours durant, ce colloque répondait manifestement à une attente.

Consacrer un colloque à l'épineuse question de la « gestion informatisée des écritures anciennes », c'était postuler que les caractéristiques graphiques des documents anciens méritent d'être informatisées. La validité de ce présupposé n'est pourtant pas universellement reconnue. Le stimulant article de Peter Scharf qui ouvre ce volume le montre bien. Désignant explicitement les défauts inhérents aux systèmes actuels de codage de textes sanskrits, l'auteur y défend une position radicale : il suggère de substituer à la dimension graphique de l'écriture un codage strictement phonétique de la langue. Même si la position défendue par Peter Scharf diffère de celle qu'adoptent les autres contributeurs de ce volume, sa grande vertu est de désigner en termes explicites les difficultés rencontrées dans la transcription numérique des textes anciens. Mais le problème n'est pas toujours posé de façon aussi nette. L'accès aux sources anciennes étant souvent difficile, les historiens des

textes se sont longtemps contentés d'éditions critiques qui reproduisaient une œuvre en normalisant sa graphie et sa mise en page, ou en tentant de reproduire en « quasi-facsimilé » l'aspect des documents originaux. L'étude des écritures anciennes a dès lors pu être considérée comme une simple « science auxiliaire », apanage d'un petit nombre d'experts sollicités en cas de besoin, mais dont on croyait la plupart du temps pouvoir se passer. Cette situation est en passe de se modifier. Le signalement et la numérisation massive des documents anciens facilitent désormais l'accès aux sources ou à leurs reproductions photographiques. Les formats numériques d'encodage offrent par ailleurs de nouvelles possibilités d'enregistrement et de représentation de l'information – même si les technologies numériques n'offrent pas encore de solution entièrement satisfaisante pour ce faire. Les historiens des textes se trouvent ainsi confrontés plus souvent qu'auparavant à la réalité graphique des sources originales et envisagent d'en rendre compte dans leurs travaux de transcription.

Mais à quoi bon enregistrer les informations graphiques ? Les objections ne manquent pas, exprimées sur différents modes. *Philosophique* : « Pourquoi sanctuariser les formes graphiques, quand le texte doit d'abord être porteur de sens ? ». *Corporatiste* : « Laissons les questions paléographiques aux spécialistes ; la plupart de nos collègues n'utilisent les textes que pour y chercher des informations sémantiques. ». *Pratique* : « Cette profusion de caractères spéciaux risque de rebuter le public et de perturber nos outils d'interrogations ! ». *Indépendant* : « De quel droit prétendez-vous imposer aux chercheurs vos propres manières de faire ? ». *Raisonnable* : « N'êtes-vous pas en train de céder à une sorte de fétichisme graphique ? ». Tel le Cyrano de Rostand, celui qui prétend reproduire fidèlement une source ancienne risque de passer à la fois pour un doux rêveur et pour une âme trop fièrement campée dans ses principes.

Il en va pourtant de l'économie de la recherche. Le document écrit est un *tout*. Son contenu intellectuel est manifesté par une réalité matérielle qui contribue à lui conférer sa valeur historique – c'est-à-dire sa capacité à représenter « un moment déterminé de l'évolution dans un domaine quelconque de l'activité humaine » (Aloys Riegl). Un même document est dès lors susceptible d'être soumis à des analyses sémantiques, stylistiques, linguistiques, paléographiques ou archéologiques, et chaque spécialiste, muni de ses propres grilles d'analyse, y découvrira des ressources différentes. Naturellement, l'opération de transcription ne permet pas de respecter l'intégrité de la source originale, dans la mesure où elle implique la perte des informations relatives à la disposition précise des éléments, à la texture du support, à la couleur des encres, etc. Mais une transcription peut néanmoins rendre compte avec fidélité de certaines informations graphiques, comme la présence dans le texte d'une ligature, d'un signe abréviatif, d'une variante allographétique... Une transcription idéale pourrait donc servir de base à des analyses non seulement littéraires, lexicales, syntaxiques, mais aussi à des études orthographiques ou paléographiques. Elle serait à ce titre *définitive*. Le texte ainsi établi constituerait une base solide à laquelle chacun pourrait apporter, moyennant quelques modestes transformations, les enrichissements nécessaires à ses propres

travaux (balisage sémantique, encodage de structures grammaticales, lemmatisation, regroupement des entités allographétiques, etc.).

Pour que de telles transcriptions puissent être produites, il importe de disposer des moyens techniques de les réaliser. Or les technologies numériques ne permettent pas encore de rendre compte fidèlement des écritures anciennes. C'est un triple problème qui se pose à nous :

1. la relative méconnaissance des anciens systèmes d'écritures (dimension proprement *paléographique*) ;

2. la question des modalités d'enregistrement des données graphiques (le *codage* numérique de l'information) ; et

3. la question de son rendu à l'écran ou à l'impression (le problème des *fontes*).

Inventaire et classement des signes d'écritures

Transcrire un document historique implique d'identifier chaque signe composant le texte original. L'analyse paléographique précède et conditionne donc l'édition des textes anciens. Cette analyse est certes affaire de spécialistes, l'expertise paléographique relevant de compétences singulières et irremplaçables. Pour autant, la nécessité d'établir une typologie cohérente et exhaustive des signes d'écriture se fait désormais sentir de façon pressante à tous les chercheurs impliqués dans la transcription de documents anciens. Loin de constituer un simple mode d'accès aux sources historiques, l'informatisation des textes constitue un enjeu décisif pour la recherche : elle est amenée à nourrir les travaux paléographiques les plus fondamentaux.

Une telle perspective implique d'inventorier les signes d'écriture utilisés dans les documents historiques. Mais ce travail d'inventaire est par essence infini. L'idée d'une liste close s'oppose en effet au principe d'ouverture sur lequel reposent certains systèmes graphiques. Dimitri Meeks rappelle que l'écriture hiéroglyphique égyptienne constitue un système ouvert et irréductible, qui ne connaît « pas de limite logique » ; aucun inventaire de hiéroglyphes ne pourra jamais prétendre à l'exhaustivité. Même les systèmes d'écriture en apparence fermés (tel le système alphabétique) peuvent difficilement être réduits à une simple liste de signes, dès lors que le facteur humain (la liberté du scribe) entre en jeu : Dominique Stutzmann explique avec raison que « la réalité graphique [...] n'est pas une réalité discrète, c'est au contraire un continuum évolutif » ; le chercheur peut dès lors être confronté à une infinité de « variantes graphiques » qui rendent vaine toute prétention à l'exhaustivité des travaux d'inventaire. Stutzmann n'hésite donc pas à relativiser la valeur des listes de signes produites par les chercheurs : « Les répertoires de "glyphes" [...] n'aident guère l'analyse, car ils sont à la fois trop riches et trop peu structurés ». La problématique est donc moins celle de l'*inventaire* proprement dit que celle du *classement*, et donc de la description, des signes d'écriture. Il s'agit de rattacher chaque signe à une classe prédéterminée ou, si l'on préfère, d'opérer une distinction claire entre le signe (théorique) et ses différentes manifestations graphiques.

Cette opération est complexe et impose une réflexion préalable sur les cadres de classement des signes graphiques. La plupart des contributions de ce volume s'accordent ainsi à rejeter l'idée d'une partition binaire entre « caractère » (signe linguistique) et « glyphe » (manifestation graphique). Concernant l'écriture hiéroglyphique égyptienne, Dimitri Meeks d'une part, Stéphane Polis et Serge Rosmorduc d'autre part, proposent des méthodes de classement relativement proches, malgré des divergences dans le vocabulaire employé : ces deux contributions introduisent un niveau intermédiaire (appelé *type* par Meeks, et *classe* par Polis et Rosmorduc) entre le signe linguistique (*famille* de signes, ou *graphème*) et ses manifestations graphiques (*glyphe* ou *forme*). Pour la transcription des imprimés anciens, l'article de Jacques André et Rémi Jimenes introduit le concept de *typème*, « chaînon manquant entre caractère et glyphe », étroitement lié à la dimension technique de la typographie au plomb. Pour l'analyse des graphies manuscrites médiévales, Dominique Stutzmann suggère quant à lui de « dépasser les distinctions binaires » et de « bâtir une ontologie des formes » à *n* niveaux, dans laquelle chaque signe se rattacherait à une classe, les différentes classes pouvant être unifiées à des niveaux supérieurs.

Pour se repérer dans le « *mare magnum* des variantes graphiques », Odd Einar Haugen propose un très utile questionnaire, applicable à tous types d'écritures, et destiné à déterminer si une « variante graphique » mérite d'être considérée comme une entité spécifique. Cette grille d'analyse tient en quatre questions simples : le signe considéré est-il susceptible de refléter une particularité phonétique ? Ce signe peut-il aider à localiser dans le temps ou dans l'espace l'origine d'un document ? Son utilisation est-elle soumise à des règles de position relative dans le texte ? Enfin, peut-il être caractérisé par une forme clairement identifiable ? Ce questionnaire, éclairant de simplicité, ne manquera pas de s'avérer utile pour les travaux à venir.

Enregistrement de l'information graphique

Une fois identifiées les spécificités graphiques pertinentes des documents anciens, leur intégration aux fichiers numériques demeure problématique. Peter Scharf remarque ainsi que les codages typographiques sont inadaptés au traitement des langues anciennes, précisément parce qu'ils ont été élaborés dans un environnement dominé par les langues modernes d'Europe occidentale. Même le plus complet des codages typographiques (Unicode) ne couvre pas les besoins des spécialistes : si la plupart des systèmes d'écritures y sont représentés, nombre de caractères usités par les anciens scribes n'y ont pas encore été intégrés. Meeks fait ainsi remarquer qu'Unicode ne valide que 1200 caractères hiéroglyphiques égyptiens, « les plus usités de l'époque classique », alors qu'il faudrait disposer de plus de 10 000 caractères pour pouvoir travailler sur l'ensemble des textes hiéroglyphiques. Unicode ainsi conçu « laisse au bord du chemin, non seulement la plupart des doctorants, mais aussi les philologues, épigraphistes, lexicographes, éditeurs de textes pour lesquels une palette beaucoup plus large est indispensable ». Même pour les systèmes graphiques moins anciens et moins riches, Unicode est loin d'être satisfaisant. Jacques André et Rémi Jimenes rappellent que la typographie

numérique ne permet pas de reproduire fidèlement la typographie au plomb et déplorent les « incohérences » d'Unicode, qui ne reconnaît pas l'existence des ligatures latines mais officialise pourtant, en vertu d'un « principe de convertibilité », certaines d'entre elles.

Face aux lacunes du standard Unicode, divers partis techniques peuvent être adoptés. On peut, par exemple, tenter d'en compléter la grille en soumettant à Unicode des propositions d'ajouts et en utilisant des numéros spécifiques de la *Private Use Area* (zone des caractères à usage privé) pour le cas des caractères refusés par le consortium. Cette démarche a été initiée dès le début des années 2000 par les médiévistes du groupe MUFI, dont les travaux sont ici représentés par son président, Odd Einar Haugen.

Les verrous techniques liés aux spécificités d'Unicode peuvent être contournés par le recours à des systèmes alternatifs d'enregistrement de l'information ou à des codages spécifiques, propres à une langue ou à un type de source. La contribution de Stéphane Polis et Serge Rosmorduc montre ainsi quel travail préparatoire suppose la révision du *Manuel de Codage* de l'égyptien ancien publié en 1988, qui permet d'associer à chaque signe des opérateurs de positionnement. Dimitri Meeks reste, quant à lui, partisan de l'utilisation d'une version révisée et largement complétée d'Unicode. Si ces solutions diffèrent du point de vue technique, les deux contributions égyptologiques s'accordent sur la nécessité de coder les signes au niveau intermédiaire (« type » ou « classe »). Telle est aussi la solution préconisée par Jacques André et Rémi Jimenes, qui proposent la mise en œuvre d'un codage « typémique » univoque, fondé sur une version d'Unicode à la fois augmentée (ajout des caractères manquants) et amputée (réduction des glyphes polysémiques à un seul et unique code). On peut également enregistrer les spécificités graphiques par un balisage extérieur au texte proprement dit, en joignant à la transcription une couche d'informations descriptives. Pour ce faire, la souplesse du langage XML peut être mise à profit : on lira ainsi avec intérêt la présentation du projet Oriflamms par Dominique Stutzmann, qui propose de recourir au XML-TEI pour définir des entités décrivant les abréviations tout en leur associant la valeur de l'abréviation résolue, mêlant ainsi étroitement informations graphiques et sémantiques.

La question des polices

Une fois l'information graphique enregistrée, se pose enfin le problème de son rendu à l'écran ou à l'impression. Les chercheurs se sont longtemps contentés d'une typographie imitative, et en la matière tous les bricolages étaient permis : mélange de fontes, dessins de polices « complémentaires » (sans souci de codage rationnel des glyphes), utilisation d'un caractère « ressemblant » pour un autre (le signe mathématique intégrale ∫ en guise de s-long ſ). De telles pratiques pouvaient paraître légitimes, faute de mieux, dans un contexte éditorial dominé par la publication papier ; elles ne peuvent plus être considérées comme satisfaisantes dès lors que les textes circulent sous la forme de fichiers numériques susceptibles d'être échangés et partagés.

Le problème posé par les polices est considérable. S'il est nécessaire de disposer préalablement d'un codage rationnel des caractères dits spéciaux, il importe également de réaliser, pour chaque système d'écriture, des fontes comportant autant de glyphes que nécessaire, faute de quoi les codages mis au point par les chercheurs ne pourront jamais être utilisés efficacement. Ces fontes doivent assurer le compromis entre des critères d'exigence parfois contradictoires : recherche de lisibilité pour le lecteur moderne, fidélité aux modèles historiques, soucis de neutralité vis-à-vis de la variété des formes originales. Pour être utilisées, ces fontes doivent également être associées à des outils spécifiques facilitant la saisie à l'intérieur des logiciels de composition et de traitement de textes : pilotes de clavier, tables de caractères, raccourcis claviers appelant des macros d'insertion, etc. Ce problème des fontes n'est ici abordé frontalement que par Florence Codine, qui présente un projet de « Police pour les Inscriptions Monétaires » dont on suivra avec intérêt les progrès. Remarquons toutefois que plusieurs des contributeurs de ce volume sont impliqués, à des degrés divers, dans la réalisation de fontes spécifiques à leurs travaux.

Le programme du colloque « GIÉcA » a pu surprendre, tant il bousculait les frontières linguistiques et disciplinaires. Rares sont en effet les occasions de dialogue entre des spécialistes de l'écriture hiéroglyphique, du sanskrit, de la paléographie latine, des langues nordiques ou du copte. Rares sont aussi les moments de rencontre entre bibliothécaires, informaticiens, dessinateurs de caractères et universitaires Il ne nous était malheureusement pas possible de reproduire en un seul volume l'intégralité des communications prononcées lors des deux journées « GIÉcA », et les sept contributions que nous publions ici ne fournissent qu'un aperçu de la richesse et de la variété des discussions[1]. La qualité des débats a largement confirmé la validité de notre postulat initial : la variété des langues et des supports d'écritures (épigraphiques, numismatiques, manuscrits, imprimés) ne doit pas faire oublier le caractère commun des problèmes théoriques et techniques rencontrés.

1. Aux sept contributions publiées ici s'ajoutaient : Marc Smith (École nationale des Chartes) : « La typographie face aux écritures anciennes, entre reproduction et transcodage » ; Andreas Stötzner (dessinateur de caractères) : « On the identification, encoding and design of special characters » ; Andreas Stötzner et Odd Einar Haugen (MUFI) : « The MUFI project : history, development and results » ; Elisa Pallottini (Univ. de Rome) : « De l'écriture à l'édition numérique des textes épigraphiques » ; Marion Lamé (Anhima), Francisco Soler et Victoria Luzon (Université de Grenade) : « De la transcription graphique à la reconstitution diplomatique » ; Jonathan Perez (dessinateur de caractères) : « Une police de caractères copte scripte pour l'édition scientifique » ; Julia Joffre (dessinatrice de caractères) : « Création d'une police adaptée à l'étude des écritures gothiques », Frédéric Rayar et Jean-Yves Ramel (Université de Tours), « Les logiciels Agora et Retro : des outils pour l'extraction et l'indexation des caractères anciens ». En périphérie des sessions scientifiques proprement dites, Thomas Huot-Marchand a également donné une présentation de l'*Atelier National de Recherche Typographique* (Nancy).

Ces problèmes ne trouveront certes pas leur résolution dans les pages qui suivent. Mais nous espérons que les pistes de réflexion et les propositions techniques évoquées par les contributeurs de ce numéro spécial de *Document numérique* s'avéreront utiles pour rassembler des corpus de transcriptions numériques préservant au maximum l'information graphique.

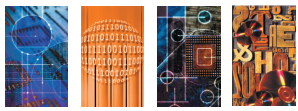
RÉMI JIMENES
Centre d'études supérieures de la Renaissance, Tours

Remerciements

Nous adressons nos plus vifs remerciements aux membres du comité scientifique qui ont assuré le choix des intervenants et la relecture des contributions. Nous tenons également à remercier l'ensemble des personnels du CESR, et plus particulièrement l'équipe des Bibliothèques Virtuelles Humanistes, largement mise à contribution, dont le soutien moral et matériel a été décisif pour le bon déroulement des deux journées « GIÉcA ».

Christine Bénévent, Rémi Jimenes, Guillaume Sarah

Revue des Sciences et Technologies de l'Information



série Document numérique • Volume 16 – n° 3/2013

Depuis deux décennies, les projets impliquant le traitement informatisé de corpus de documents historiques se multiplient. L'exploitation des données et métadonnées ainsi produites constitue un enjeu important dont témoigne le développement d'outils et de formats d'indexation toujours plus performants.

Mais, en amont des problèmes posés par l'indexation des données, un constat s'impose aux chercheurs : celui de l'inadéquation de la typographie numérique avec les formes graphiques présentes dans les documents anciens. Plusieurs domaines de recherche (tels l'étude des pratiques orthographiques, l'indexation des légendes monétaires, la transcription de documents épigraphiques, manuscrits ou d'imprimés…) nécessitent l'utilisation de caractères actuellement absents des codages numériques. Ce simple constat masque en réalité un triple problème :

1. la relative méconnaissance des anciens systèmes d'écritures (dimension proprement *paléographique*) ;
2. la question des modalités d'enregistrement des données graphiques (le codage numérique de l'information) ;
3. la question de son rendu à l'écran ou à l'impression (le problème des fontes).

Tels sont les thèmes abordés dans les sept articles de *Gestion informatisée des écritures anciennes*.

dn.revuesonline.com

9 782746 246430