Les documents anciens

coordonnateurs

Jacques André
Marie-Anne Chabin

Document numérique
Volume 3 • n° 1-2 – juin 1999

Hermes

Les documents anciens

Sommaire

Volume 3 – n° spécial 1-2/1999

• Éditorial – Numériser des documents anciens, et après ?
Jacques ANDRÉ, Marie-Anne CHABIN 7

• La numérisation dans les archives de France — Catherine DHÉRENT 13

• La numérisation des manuscrits médiévaux à l'Institut de recherche et d'histoire des textes – Élisabeth LALOU 29

• Archim : une banque d'images numériques pour le public du Centre historique des Archives nationales — Florence CLAVAUD 39

• Réflexions sur la représentation des documents anciens : le projet Philectre
Gusnard de VENTADERT 57

• An Electronic Edition of Don Quixote for Humanities Scholars
Sueh-Cheng HU, Richard FURUTA, Eduardo URBINA 75

• Pour un système de philologie numérique — Andrea BOZZI 93

• Conception d'un poste d'édition et de lecture d'hypermédias littéraires
Éric LECOLINET, Laurent ROBERT 103

• Analyse historique de sources manuscrites : application de TEI à un corpus de lettres de rémission du XVIe siècle — Jean-Daniel FEKETE, Nicole DUFOURNAUD 117

• Représentation et exploitation de métadonnées complexes : le cas des documents anciens — François ROLE 135

• Vers un standard européen de description des manuscrits : le projet Master
Lou BURNARD, Peter ROBINSON 151

ÉDITORIAL

Numériser des documents anciens : et après ?

L'idée d'un numéro spécial de Document numérique *sur les documents anciens s'imposait. En effet, les projets de numérisation de livres ou d'archives des siècles passés se multiplient. Il s'agit aussi bien de mettre des documents rares à disposition du public sur cédéroms ou sur les réseaux, que de structurer des corpus afin non plus seulement de les éditer mais de les rendre disponibles et manipulables par chacun selon son propre travail et d'approfondir des études comparatives. Ces réalisations font naître une problématique spécifique de l'utilisation de l'informatique pour le traitement des documents d'autrefois.*

Commençons par définir un certain nombre de termes.

Numériser. *Pour beaucoup, il s'agit d'un synonyme de scanner. Pour nous, il s'agit bien plus que l'obtention de* bitmaps *ou d'images, que ce soit en format tif, gif, ou autre. Cela veut dire d'abord un environnement (références, base de données/corpus, etc.), mais aussi un accès aux informations (indexation, transcriptions, etc., avec tous les pointeurs d'une vision à l'autre). Scanner un livre du XVIIIe siècle n'a d'intérêt pour un chercheur que s'il a accès non seulement aux images du texte, mais aussi au texte lui-même.*

Hypertextes. *On a trop tendance à réduire la notion d'hypertexte à la seule gestion de liens de références : je clique ici et je fais apparaître tel texte ou image, ce que finalement on savait faire, ergonomie mise à part, du temps de Diderot. L'hypertexte permet en effet d'autres liens, de typer ces liens et d'en faire des réseaux sémantiques, cognitifs, de s'approprier l'information.*

Document ancien. *D'après le* Petit Robert, *est « ancien » ce « qui existe depuis longtemps, qui date d'une époque bien antérieure » ; l'adjectif ancien recouvre donc les deux notions de durée et d'antériorité. Celles-ci ne sont d'ailleurs pas incompatibles.*
Il faut donc entendre par « documents anciens » à la fois les documents produits au cours des siècles passés et les documents plus récents que d'autres dans une relation de comparaison entre un document donné et une version plus « ancienne » de ce

même document. L'expression « documents anciens » vise aussi bien les documents d'archives, créés dans le cours d'une activité définie (par exemple des registres de baptêmes, des actes notariés, des manuscrits d'écrivains mais aussi les cadastres napoléoniens) et qui en général n'existent qu'en un seul exemplaire, que les livres (y compris les partitions musicales, les recueils de cartes ou gravures, etc.), créés pour être diffusés et dont il existe ou a existé plusieurs exemplaires.
Il n'y a pas vraiment de terminus ad quem *de la notion de document ancien. Elle se déplace au rythme de l'évolution des techniques d'écritures, un nouvel outil repoussant vers le monde des anciens les documents créés à l'aide des outils de la génération précédente : techniques d'imprimerie face à la plume des moines copistes, photographie face aux estampes, enregistrement numérique face à l'enregistrement analogique. « Ancien » doit donc être entendu dans un sens beaucoup plus large que « médiéval ». Par ailleurs, les documents anciens comportent aussi bien du texte que de l'image, voire, demain, de l'image animée et du son. De la même façon, les documents anciens sont aussi bien manuscrits qu'imprimés, ou issus d'autres techniques plus récentes mais appelées à vieillir.*
La technique d'écriture la plus ancienne reste la main de l'homme. C'est aussi celle qui, par opposition à la rigueur mécanique d'une machine, présente le plus de fantaisie, de nuance, de particularismes. C'est celle qui, vis-à-vis de l'environnement numérique actuel, soulève le plus d'interrogations. Ce n'est donc pas un hasard si la majorité des articles rassemblés ici traitent de la numérisation de documents manuscrits, expression qui aurait aussi bien pu être le titre de ce numéro spécial.

Manuscrit. *Il convient d'attirer l'attention du lecteur sur l'ambivalence de ce mot. D'une part, il renvoie à un texte ou à des dessins tracés à la main avec un crayon ou un pinceau sur un parchemin ou un papier, donc y compris les documents calligraphiés précédant l'invention et l'utilisation de l'imprimerie ; d'autre part, il désigne l'état préparatoire d'un document qui sera mis au propre avant d'être imprimé. La notion d'antériorité par rapport à l'état final d'un texte est donc à bien distinguer d'une facture définitive due à la main de l'homme.*

Image. *Pour conclure cette parenthèse terminologique, il faut signaler que le mot « image » revêt lui aussi deux acceptions distinctes sous la plume des différents auteurs de ce numéro. À côté du sens traditionnel d'illustration, de dessin ou d'enluminure, le mot image est utilisé, dans le contexte numérique, pour désigner une page ou une partie de manuscrit numérisée en mode image, et comportant éventuellement texte et illustrations.*

Le propos de ce numéro est donc l'utilité et l'utilisation de l'informatique pour les documents du passé. D'une manière générale, la technologie numérique appliquée au domaine documentaire se répartit en quatre secteurs : la création de nouveaux documents, le stockage, la description ou structuration des documents et la diffusion. En ce qui concerne plus précisément les documents anciens ou manuscrits au sens défini ci-dessus, la numérisation présente deux autres domaines d'application : d'un côté, la structuration des documents numérisés pour l'étude analytique et surtout l'étude de corpus ou dossiers ; de l'autre côté, la numérisation de documents historiques ou ar-

tistiques dans le cadre de la diffusion du savoir et des connaissances. Autrement dit, le document numérisé est envisagé tantôt comme document final du travail (diffusion des connaissances), tantôt comme document intermédiaire (étude). Numériser des documents anciens n'est pas une fin en soi, ce n'est que la première étape de l'important : le processus de recherche et non le produit de ce processus.

Le numérique permet à un public plus large d'accéder au contenu de documents uniques, précieux et fragiles qui forment l'essence du patrimoine écrit national. Certes, les albums de photographies et les éditions de textes jouaient déjà ce rôle avec le support papier, mais la maniabilité des documents numériques et l'existence des réseaux permettent de décupler le nombre de documents mis à disposition, à un coût bien moindre, ce qui est particulièrement important pour des publications pointues qui souvent frustraient le chercheur par leur incomplétude et le lecteur cultivé qui se perdait dans l'ésotérisme du langage des chercheurs. Dans ce contexte de diffusion, les deux objectifs principaux sont la stabilité des standards techniques car la démarche s'inscrit dans le temps, et la qualité de la restitution de l'image du document. La qualité de la restitution est d'autant plus importante que le numérique est de plus en plus envisagé comme support de substitution pour les documents qui se conservent mal (« numériser pour sauver » commencent à dire certains archivistes).

La technologie numérique appliquée à l'étude des documents anciens est tout autre. La durabilité des standards techniques et la qualité de l'image des documents ne sont pas négligeables mais la spécificité tient aux difficultés de structuration des documents, à la dualité texte/image des documents traités, à la définition des différents éléments d'analyse, à la formalisation des éléments de description et des métadonnées, à la navigation entre les différentes versions étudiées parallèlement.

L'étude des documents manuscrits ou la comparaison entre manuscrit et imprimé, ou encore la mise en perspective d'un texte (manuscrit ou imprimé) et de sa transcription nécessitent une décomposition minutieuse de chaque paragraphe, de chaque ligne, chaque mot, voire de chaque signe. Le travail d'analyse du chercheur sur un document ou un ensemble de documents anciens se traduit par des annotations complexes d'ordre technique, linguistique et historique, auxquelles il faut ajouter des appréciations personnelles. Les manuscrits littéraires modernes, créés au gré de la liberté du scripteur, cumulant parfois ratures et mauvaise qualité du papier, représentent un défi particulier pour les langages de structuration des documents. Un dossier informatisé est alors formé de corpus mais aussi de tout le matériel cognitif du chercheur.

Les articles de ce numéro spécial ont tous été envoyés spontanément en réponse à un appel à publication. Aucun n'a été sollicité. Ce qui explique que nous ne prétendons pas montrer l'état de l'art dans ce domaine, mais plutôt des visions partielles : tel organisme n'est pas présent, telle réalisation n'est pas décrite.

Une vingtaine d'articles ont été soumis ; le comité de rédaction n'en a retenu qu'une dizaine, éliminant des propositions d'articles qui se limitaient à la présentation de projets non encore concrétisés, dans la mesure où nous souhaitions mettre

l'accent sur des réalisations, achevées ou en cours, porteuses de solutions ou de questions argumentées. Le sujet de la numérisation des documents anciens est loin d'être clos et on peut penser que d'autres publications suivront.

Les articles présentés s'articulent autour de trois thèmes.

1. La numérisation des grandes collections nationales, avec notamment les articles de Catherine Dhérent et de Florence Clavaud sur la numérisation aux Archives nationales de France et celui d'Élisabeth Lalou sur la numérisation des manuscrits médiévaux à l'Institut de recherche et d'histoire des textes.

2. La présentation de systèmes (ou postes de travail) de philologie numérique, avec la présentation des projets Philectre (article collectif signé Gusnard De Ventadert et celui d'Éric Lecolinet et Laurent Robert), Bambi (article d'Andrea Bozzi) et d'un système américain actuellement dédié à l'étude de Don Quichotte *(article de Sueh-Cheng Hu, Rick Furuta et Eduardo Urbina).*

3. Des expériences de structuration de corpus de manuscrits : application de la TEI à des lettres de rémission (article de Jean-Daniel Fekete et Nicole Dufournaud) ou exploitation de métadonnées (article de François Role) et la présentation du projet européen Master de description des manuscrits (article de Lou Burnard et Peter Robinson).

Nous avions envisagé de joindre à ce numéro un état des industries liées à la numérisation des documents, dans la mesure où diverses entreprises ou sociétés de service se spécialisent sur la scannérisation de documents anciens, la production de cédéroms, de banques d'images ou de logiciels de traitement des documents numériques. Nous remercions notamment les services d'archives et leurs prestataires qui ont répondu à notre enquête. Toutefois, l'opération s'est révélée un peu prématurée et, plutôt que de présenter un état lacunaire ou partiel, nous avons préféré attendre d'avoir recueilli des informations plus complètes. Ce recensement pourra intervenir, dans un deuxième temps, dans un cadre plus ciblé.

La confrontation de la dizaine d'articles de ce numéro spécial fait apparaître deux axes de réflexion à approfondir quant aux travaux actuellement menés sur la numérisation des documents anciens.

Droits de reproduction et d'exploitation. *La technologie numérique remet en cause les repères classiques d'évaluation des droits d'auteur et des droits d'exploitation et de diffusion. La rapidité et la relative facilité des opérations de numérisation des documents et de manipulation des documents numériques provoquent de nouvelles habitudes : le « copillage » numérique prend le relais du « photocopillage ». La transversalité des réseaux rend très difficile le contrôle de la diffusion et de l'exploitation qui est faite des documents numérisés. Les projets de numérisation mêlent facilement l'exploitation patrimoniale ou scientifique à l'exploitation commerciale. Il est vrai que cela n'est pas propre aux documents anciens. Cependant, la numérisation ou la mise en ligne de documents anciens soulèvent des problèmes spécifiques. Par exemple, l'article de J.-D. Fekete et N. Dufournaud met en évidence le cas de docu-*

ments de collection publique, dépourvus de droits d'auteur, qui sont librement accessibles au chercheur mais le détenteur des fonds n'autorise pas la recopie numérique par le chercheur. Il n'est pas question de permettre n'importe quelle exploitation sans l'aval du propriétaire ou sans participation de celui-ci au bénéfice, quel qu'il soit, de l'opération. Les chercheurs ont de tout temps obligation morale de citer leurs sources mais, là, les prolongements et les utilisations induits par la numérisation posent des questions en matière de contrôle d'accès à l'information publique, voire sur le plan financier. Pourtant, on ne peut envisager non plus que la recherche reste longtemps bloquée par un vide juridique ou même réglementaire.

En ce qui concerne les documents littéraires, il apparaît que les possibilités de la technologie numérique appellent une redéfinition, un nouveau cadrage des droits d'utilisation. Il n'est pas concevable que le progrès technologique soit freiné par un environnement juridique inadapté.

Réutilisation des systèmes élaborés. *Sur le plan technique, la question de la réutilisation possible des systèmes élaborés dans le cadre d'un projet défini apparaît déterminante dans l'évaluation des projets et des expérimentations. Le système et la technique qui supportent la méthode d'un chercheur sur un corpus peuvent-ils être réutilisés par un autre chercheur pour ce même corpus ou pour un autre ? Comment résoudre les problèmes de formation des chercheurs à l'outil de traitement des documents ? Les travaux réalisés sur un corpus dans le cadre d'une problématique donnée peuvent-ils être réutilisés pour une autre problématique sur le même corpus ? Sont-ils d'ailleurs conçus pour être réutilisés ?*

Toutes ces questions trouveront peu à peu leurs réponses, au fur et à mesure des réalisations, des bilans et des normalisations, mais il est sans doute nécessaire que les équipes de projet les intègrent à leur réflexion dès le début de leurs travaux, afin de favoriser l'émergence de solutions plus générales.

Jacques ANDRÉ (Irisa/Inria-Rennes)
et Marie-Anne CHABIN
Coordonnateurs de ce numéro