Introduction : un dialogue pluridisciplinaire pour penser le « document numérique »

Jean-Michel Salaün*, Jean Charlet**

*ENSSIB
17,21 Boulevard du 11 novembre 1918
69623 Villeurbanne Cedex
salaun@enssib.fr
**STIM/AP-HP & INSERM ERM 202,
jc@biomath.jussieu.fr

Résumé

Un travail collectif de réflexion pluridisciplinaire sur la notion de document a servi d'appui à l'appel à contributions pour ce numéro thématique consacré au document numérique. Il est résumé dans cette introduction, à partir de ses trois entrées sur la forme, le texte et le médium Les onze articles retenus sont présentés succinctement. Ils proposent de développer différentes modélisations, de repérer les transformations entre document traditionnel et document numérique et illustrent les problèmes posés par le changement en cours sur plusieurs terrains. La dynamique en cours a donc montré déjà des résultats conséquents et doit être poursuivie.

Mots-clés : *document, document numérique, numérisation, pluridisciplinarité.*

Abstract

A multidisciplinary research work on Electronic Document was used in the call for this issue of the journal It is summarized in this introduction, namely its three approaches on the form, the text and the medium. The eleven selected articles are briefly presented. They propose to develop different modeling, to point changing between traditional and electronic document and to show the change in progress on different fields. This work in progress have already showed real results and consequently must be continued.

Key-words: *Document, Electronic Document, Digitalisation, Multidisciplinary Research*

1 INTRODUCTION

Curieusement, le « document », omniprésent dans notre quotidien, n'a été que peu soumis à la réflexion critique des chercheurs qui lui préfère souvent le terme d'« information », un peu comme si le contenu prévalait sur le contenant. Peut-être l'objet-document est-il trop familier, trivial, la notion est-elle trop intuitive. Ils n'étaient ou ne paraissaient pas problématiques, dans tous les sens du terme. Pourtant aujourd'hui le numérique bouscule profondément le document, devenu électronique. Et les chercheurs français sont nombreux à se pencher sur les aspects numériques du document, directement ou indirectement. Mais ils sont dispersés géographiquement, éclatés dans leurs approches et ils n'échangent régulièrement qu'entre tenants d'une spécialité pointue. Pire, même si les uns et les autres le déplorent souvent, les relations entre les tenants des sciences de l'ingénieur et ceux des sciences humaines et sociales restent plutôt rares, à la notable exception près de la linguistique et dans une moindre mesure de l'archéologie.

À rebours de l'évolution de la réflexion de ces cinquante dernières années, notre hypothèse est que le concept de document a été trop vite délaissé et qu'il est, dans de nombreux cas, plus approprié pour comprendre et décrire les situations que celui d'information. Les nations de civilisation ancienne, celles de l'Europe notamment, ont un besoin urgent d'outils conceptuels d'une part parce qu'elles ont du mal à évaluer les décisions à prendre concernant leur patrimoine documentaire, d'autre part parce que l'omniprésence du document dans la gestion sociale, à l'échelle de pays entiers comme pour de tous petits groupes, pèse sur les changements en cours.

Ainsi, sur l'initiative du département Sciences et Technologies de l'information et de la communication (STIC) du CNRS, et particulièrement de Catherine Garbay, un réseau de chercheurs thématique et pluridisciplinaire a été organisé en 2002 sur le document numérique[1]. Le réseau a initié une réflexion collective et transversale fondamentale. La rédaction collective d'un document (évidemment) de

[1] http://rtp-doc.enssib.fr

travail, signé d'un pseudonyme Roger T. Pédauque[2], a permis un premier repérage analytique des travaux en cours et des questions en suspens. Ce document a servi de base à l'appel à contribution pour ce numéro de la revue I3. Nous en donnons ci-dessous un bref résumé, puis nous présenterons les articles de ce numéro pour conclure enfin sur quelques perspectives à venir.

2 DES AVANCEES

Une façon simple de repérer les travaux des chercheurs autour du document numérique est de les regrouper en trois catégories par rapport à leur angle principal d'approche : le signe ou la forme, le texte ou le contenu, le médium ou la relation.

Dans la première entrée sur le signe ou la forme, trois communautés de chercheurs se croisent sans toujours bien se connaître : les traiteurs d'image, les éditeurs numériques et la communauté du Web. Des travaux du PARC à Palo-Alto au consortium W3C le chemin parcouru en quelques années est impressionnant. Même si les chercheurs sont partis de problématiques différentes, ils convergent nettement aujourd'hui dans la compréhension des structures logiques du document et dans les interrogations sur les formes perceptibles. Le travail principal concerne les formats, au sens informatique, c'est-à-dire des outils de mise en forme. D'une certaine façon, le numérique a déplacé la question du support du document, qui en assurait la stabilité en fixant l'inscription, vers la problématique de sa structure. L'inscription transformée en signal se déplace d'un support à l'autre et la stabilité du document réside alors dans celle de l'organisation du signal. La popularisation de la norme XML marque une étape mais elle laisse encore largement ouverte la problématique de la perception et de la lecture. En effet, en séparant de façon radicale la structure logique d'un texte de sa représentation visuelle, elle autorise des traitements formels différents pour un même contenu à une échelle inédite.

La seconde entrée sur le texte (il faut comprendre ce terme dans un sens générique sans référence à un mode d'expression particulier, comme synonyme de contenu) fait se rencontrer les chercheurs en linguistique, en sciences de l'information et en ingénierie des connaissances. Le problème

[2]Document : forme, signe et médium, les re-formulations du numérique. Roger T. Pédauque. Article. 08 juillet 2003. Working paper
http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/05/11/index_fr.html

principal est de construire des modèles permettant de traiter le contenu des documents, afin de les retrouver, et éventuellement les réagencer pour en produire de nouveaux, adaptés à la demande du lecteur. Il s'agit de produire du sens pour le lecteur en dépassant la confusion de l'accumulation des informations, à partir notamment d'une modélisation de données sur les documents, les métadonnées. Depuis les bases de données bibliographiques jusqu'aux « moteurs » d'aujourd'hui, ici aussi les développements et perfectionnements ont été très rapides. Au-delà de la performance des outils, la question posée est celle de la construction des savoirs dans la relation entre les métalangages et les documents eux-mêmes. Celle-ci est posée aussi bien dans la relation entre un lecteur et un texte que dans celle de sa « navigation » entre plusieurs textes. Manipuler des ontologies, c'est-à-dire entre autres articuler des concepts, n'est pas sans effet sur l'organisation de nos connaissances, particulière à une communauté ou générique pour une société.

La troisième entrée sur le médium comprend deux terrains sensiblement différents : d'une part la communication organisationnelle, d'autre part les médias de masse. On y trouve principalement des chercheurs en télécommunication ; c'est ici aussi que les sciences sociales et les sciences de la communication sont majoritaires. Le document y fonde son statut dans une diffusion dépassant l'intime et l'éphémère, en s'articulant avec d'autres dans des collections. C'est-à-dire qu'un document n'existe que dans une collectivité élargie et une durée pour alimenter un patrimoine de connaissances partagées. Deux pratiques sociales très « normées » se modifient au travers du Web. D'une part, la frontière entre la communication inter-personnelle et la communication « flottante » ou publique, entre correspondance privée et publicité, se déplace et tous les codes sociaux et modalités organisationnelles qui s'y rapportent sont alors ébranlés. D'autre part, illustrées par le préfixe « hyper- » (hypertexte, hypermédia), les relations entre documents se distribuent différemment entre partage et liens, entre mise en ligne et mise en réseau. Suite à ces deux déplacements, la notion de publication qui s'était stabilisée au cours des siècles, se transforme brutalement et celle de bibliothèque, dont les contours étaient limités par la taille des collections et l'implantation physique et institutionnelle, explose en bibliothèque numérique, centralisée ou distribuée et largement accessible.

3 DES INTERROGATIONS

Si l'on croise ces entrées, on constate qu'une évolution possible, explorée par bien des chercheurs en pointe sur ce champ mais non certaine, serait que les documents structurés, informés et contextualisés, rejoignent des bases de données, centralisées ou distribuées, ou des hypertextes, toujours en construction, et que l'ensemble des fichiers s'apparente de plus en plus à un ou plusieurs vastes jeux où des briques de différentes tailles, formes et usages seraient agencées selon des configurations très variées. Selon cette hypothèse dont la validité est sans doute variable selon la spécificité des contextes, un document n'aurait alors d'existence à proprement parler qu'à deux moments : éventuellement celui de sa conception par son auteur qui devra le visualiser ou l'entendre, pour s'assurer qu'il correspond à ses choix et surtout celui de sa re-construction par un lecteur.

Nous percevons très clairement les prémisses de cette évolution, par exemple dans la multiplication de sites Web dynamiques, où les pages sont construites à la volée à chaque requête de l'internaute ou encore dans la page actualité de Google réactualisée toutes les quinze minutes par un moteur balayant 500 sites d'informations journalistiques[3]. On assiste même à des sortes de mise en abîme où des lecteurs mettent en ligne leurs impressions dans des chroniques souvent reliées entre elles, les blogs. Nous entrevoyons aussi bien des problèmes qu'elle soulève, par exemple, celui de la validation et de la hiérarchisation ou celui de la responsabilité éditoriale. Il est beaucoup plus aventureux d'en prévoir les avancées, les résistances ou déviances et donc les conséquences, sinon pour dire qu'elles seront à coup sûr importantes et durables.

Il se construirait alors sous nos yeux un nouveau compromis entre une multiplicité d'acteurs pour ré-inventer des documents ou des artefacts de substitution. Dans ce processus, le numérique joue un rôle majeur mais il n'est sûrement pas le seul phénomène en cause. En l'absence des outils pour une analyse pertinente, le jeu des acteurs et les multiples micro-négociations se réalisent « à l'aveugle », privilégiant les ajustements de court terme et les intérêts immédiats, conduisant à de nouveaux circuits/chemins de connaissance, de nouveaux processus de catégorisation, de nouveaux arts de disposer (dispositio) et de mémoire…. Le risque est le manque de cohérence et les malentendus multiples. Ainsi en est-il, pour ne prendre que deux exemples, des discours et initiatives contradictoires et confus sur le droit d'auteur dans le numérique ou

[3] http://news.google.fr/

encore des multiples et envahissants parasitages de la publication numérique.

4 UN APPROFONDISSEMENT NECESSAIRE

L'objet de ce numéro de revue est donc de poursuivre la réflexion engagée par des contributions, cette fois individuelles et signées, au débat lancé. Nous revenons ci-dessous sur chaque article retenu en pointant les avancées qu'il propose et aussi soulignant quelques éléments qui nous paraissent mériter encore débat.

Veronika Lux-Pogodalla et Jean-Yves Vion-Dury proposent un classement analytique des outils de l'édition électronique et surtout des logiques qui les sous-tendent en montrant leur détachement progressif du médium physique. Leur raisonnement préfigure, ou voudrait préfigurer, une modélisation générale de la notion de document opérationnelle pour agencer les développements informatiques, passés ou à venir, autour des outils de production du document. On comprend facilement toute l'importance de leur réflexion, d'abord pour son apport à une théorie générale du document à partir d'une entrée d'ingéniérie et aussi tout l'intérêt qu'il y aurait à la relire à partir d'une analyse plus sociale.

Dans le prolongement de l'article précédent, Olivier Beaudoux propose d'associer les outils d'interaction avec l'utilisateur au modèle même du document en spécifiant un modèle générique, le modèle DPI (Document, présentation, instrument). Ce modèle est instancié par l'utilisateur sur chaque document et lui permet de choisir les composants de son espace de travail en fonction des tâches qu'il veut accomplir. La « philosophie » affirmée d'un tel outil étant de masquer les applications au bénéfice de composants articulés entre eux. Rejoignant la volonté modélisatrice de l'article précédent, ce travail pose la question de la possibilité et la finalité de la modélisation : jusqu'où doit-on et peut-on aller dans la modélisation des documents, leur opérationalité et leur interaction avec l'utilisateur ? Peut-on se donner les primitives de modélisation de toutes les interactions possible ou, sinon, doit-on garder une espace de liberté – contextuel, non modélisable – correspondant à des interactions non anticipées ?

Stéphane Crozat et Bruno Bachimont s'inscrivent eux aussi dans cette tradition modélisatrice. Ils proposent de séparer pour mieux les faire apparaître dans une organisation de l'écran trois niveaux de structure, proches des trois entrées de Pédauque. La proposition a fait l'objet de développements pour des documents pédagogiques. Le besoin d'outils

adaptés aux nouvelles modalités d'écriture et de publication est clair et les auteurs font une proposition stimulante qui rejoint leurs réflexions sur la relation fond/forme (*cf.* plus bas). Mais pour être vraiment convaincants, il leur reste à confronter le modèle à une pratique continue en situation réelle. De plus, la problématique sous-jacente est celle des modèles éditoriaux et de leurs filières de production. On peut se demander s'il ne serait pas utile de les analyser plus finement, quitte ensuite à y renoncer en totalité ou en partie.

Dans le dernier article de cette première série, Nathalie Aussenac-Gilles et Anne Condamines tirent les leçons d'un travail pluridisciplinaire qu'elles ont piloté sur les ressources terminologiques, en pointant les problèmes posés par le traitement automatique de la langue dans les problématiques documentaires. Les relations entre le document et la langue passent par les textes et les outils terminologiques qu'on leur associe pour pouvoir les manipuler via une machine ou par une intervention humaine. Il parait clair que les avancées viendront du dialogue entre l'ingénierie des connaissances, les terminologues et les sciences de l'information. Toute la difficulté provient de la contradiction entre la nécessité de normalisation et la contrainte de l'évolutivité des connaissances en contexte. Une perspective pourrait être la notion de genre textuel qui introduit une généricité sans enfermer pour autant les documents dans une trop grande rigidité classificatoire.

Bruno Bachimont et Stéphane Crozat font en quelque sorte la transition entre les articles précédents dont le point de départ était le numérique et la calculabilité, donc un objectif de modélisation directement issu de l'opportunité ouverte par les performances de l'outil informatique, et les suivants qui, à l'inverse, tentent d'abord de définir le document traditionnel, pour mieux mesurer ensuite les changements apportés par le numérique. Ainsi les auteurs de cet article de transition s'interrogent sur les relations entre le fond et la forme d'un document à partir d'une théorisation de l'acte de lecture/écriture. Et, ils s'appuient sur les capacités calculatoires d'une ingénierie documentaire pour suggérer la possibilité d'une résolution du paradoxe entre la pluralité des représentations possibles d'un document et l'unicité du contenu de référence. Quoiqu'il en soit de la faisabilité d'une solution, le problème est ici clairement posé et il va rebondir dans tous les articles qui suivent.

Pour Sylvie Lainé-Cruzel, le document traditionnel relève de l'évidence, c'est-à-dire que sa caractéristique principale est d'être vu comme un objet signifiant et fini. Il est donc préférable, si l'on suit cette

auteure, de réserver le terme « document » aux fichiers stabilisés pour utiliser celui de « ressources » pour ce qu'on appelle encore, à tort selon elle, « document électronique » et qui ne relève que d'un service ponctuel et éphémère d'information. Cette proposition a le mérite de la clarté et de la simplicité. Elle rencontre les préoccupations des bibliothécaires appelés à repérer dans le foisonnement d'objets informationnels ceux qui méritent d'être mis en collection, ou encore les préoccupations des juristes ou de toute profession pour laquelle la « preuve » et la stabilité de l'élément qui la supporte est fondamentale. Néanmoins, ne peut-on prétendre *a contrario* qu'un document est, et a toujours été (même si nous sommes aveuglés par le papier imprimé), un objet en construction et qu'une preuve est dépendante d'un régime de vérité et donc toujours reconstruite ?

À partir d'une proposition de départ proche de la précédente, Sylvie Leleu-Merviel considère le document d'abord comme une image qui fait sens, répondant à différentes fonctions sociales qu'elle énumère. Elle montre alors comment il est possible, de son point de vue, d'analyser le document en différentes couches, les unes relevant d'aspects techniques, d'autres d'une dimension sémiotique. Logiquement, elle propose *in fine* d'avancer sur une représentation de la topologie du Web pour marquer sa métrique et les relations entre les différents fragments signifiants. L'accent mis ici sur la représentation topologique dans un univers hypertextuel est fondamental. Le découpage analytique proposé peut être confronté avec ceux des autres articles. Il n'y a pas consensus sur cette question et il reste sans doute là un travail de réflexion pluridisciplinaire.

Marie-Anne Chabin s'intéresse à une définition du document en construction. À la différence des deux auteures précédentes mais sans réelle contradiction, le document n'est pas considéré ici prioritairement comme un objet ou une image, mais comme le résultat d'un processus, ou plus précisément de deux processus : celui de l'auteur, individuel ou collectif, et celui du lecteur. Logiquement elle propose de séparer les documents en deux grandes catégories : les documents-trace et les documents-source. Ainsi, il ne reste plus qu'à transposer l'analyse dans un environnement numérique et pointer les difficultés qui pourraient survenir dans la fixation de l'information et la maintenance des objets, pour les corriger. L'intérêt de la proposition est de bien faire ressortir la dualité du processus qui conduit au document. Mais on peut ici rétorquer que si cette dualité est constitutive du document, son statut social lui permet de la dépasser en lui procurant une valeur de régulation au-delà de la simple relation à l'auteur ou au lecteur.

Les trois derniers articles illustrent sur des terrains particuliers bien des questions posées dans les précédents.

Dominique Cotte et Marie Després-Lonnet contestent l'idée que l'on puisse reconstruire des documents à partir de l'agencement d'éléments disjoints. Prenant l'exemple de la presse grand-public, ils montrent combien le travail éditorial suppose de multiples et fins ajustements pour construire un journal. Dans ce contexte, une trop grande automatisation conduit à des ruptures signifiantes nécessitant un travail humain pour rendre unité et cohérence au journal et inclure les articles dans un ensemble. L'importance de la médiation humaine pour la signification d'un document est ainsi, exemples à l'appui, soulignée avec force. Mais on peut aussi se demander si le raisonnement n'est pas circulaire : un journal-papier étant construit quotidiennement comme une unité, il résiste naturellement au découpage. Mais cette forme est particulière, sa caractéristique est-elle transposable ? Mieux, n'assiste-t-on pas aujourd'hui sur ce même terrain à l'arrivée en masse de quotidiens gratuits, construits sans trop de soucis d'unité éditoriale ? Sans parler, en repartant du premier exemple des auteurs, des journaux électroniques construits à partir de sites qui mettent explicitement leur contenu à disposition à travers un mécanisme de « syndication », le RSS[4], et revendiquent donc de servir de briques de contenu à d'autres sites.

Prenant un point de départ inversé, Dominique Boullier et Franck Ghitalla rendent compte d'un travail d'observation d'internautes naviguant sur le Web. Ils montrent combien ces derniers y ont perdu leurs repères habituels de lecture de documents faute d'une structuration claire de l'espace de l'écran. Ainsi des micro-stratégies ou tâtonnements, cohérents avec les « styles cognitifs » de chacun, bien éloignés d'une facture éditoriale unifiée leur permettent de récupérer les éléments signifiants qui les intéressent. Les auteurs repèrent ainsi quatre styles qui pourraient représenter quatre orientations éditoriales en cours dans l'interaction entre le succès d'une offre donnée et un type de comportement d'internautes.

Sandra Bringay, Catherine Barry et Jean Charlet, enfin, en partant des propositions de Roger T. Pédauque pour analyser le dossier médical des patients comme un document, intègrent l'une et l'autre approche. Ils

[4] Le RSS pour *Really Simple Syndication* est un format qui permet d'indexer le contenu d'un site et de le mettre instantanément à disposition d'autres sites : c'est la *syndication des contenus* .

montrent notamment que celui-ci peut être analysé comme une construction collective intégrant la structure proposée par le concepteur du document, les connaissances du ou des rédacteurs et l'apport des lecteurs par des annotations. L'intérêt de l'approche est notamment de prendre en compte frontalement la multi-structuralité, l'hypertextualité et l'évolutivité d'un document en contraignant la pluralité des documents au sein d'un même métier, soigner. Néanmoins, on peut se demander comment ces propositions pourraient être appliquées dans un environnement plus ouvert et hétérogène.

5 POURSUIVRE LE PROCESSUS

L'ensemble de ces contributions amène, nous semble-t-il, des avancées significatives. Les propositions se croisent largement, se recoupent et se contredisent parfois. Il serait présomptueux d'en présenter une lecture unifiée. Au contraire, il nous semble qu'elles suggèrent plusieurs pistes qui méritent approfondissement et confrontation et qu'il reste donc encore du chemin à parcourir dans l'interdisciplinarité dans le sens d'une « théorie du document » pour pouvoir rendre compte des phénomènes auxquels nous assistons et construire des outils adaptés à la situation.

Ainsi pour ne prendre que quelques exemples des thèmes qu'une lecture transversale pourraient faire ressortir, sans souci ici de hiérarchie ni d'exhaustivité, remarquons que :

- Tous les auteurs s'accordent à souligner l'importance de la mise à plat de la structure du document qu'autorise ou que favorise le numérique en vue de la manipulation de sa forme. Mais les analyses divergent largement sur les niveaux ou même les définitions des structures. Ainsi l'insistance de Pédauque sur la normalisation XML ne débouche pas, au moins chez nos auteurs, sur une vision unifiée de la structure d'un document. Il y a là un vaste chantier pour un travail pluridisciplinaire.
- Il reste manifestement chez beaucoup une ambiguïté sur le statut social du document qui est plutôt vu comme un outil de mémoire externe, individuel ou collectif, basé sur l'inscription que comme un objet social, fondé sur un statut. Cette conception autorise tous les développements des outils d'écriture ou de représentation, mais interdit d'en penser lucidement les conséquences sociales. Il reste ici à ouvrir un vrai dialogue entre sciences de l'ingénieur et historiens, sociologues, économistes ou juristes. Ainsi, comment passe-t-on d'un texte à un document et vice-versa ? Peut-on retrouver dans le

numérique la permanence que le document institue ? En interconnectant les réseaux, n'a-t-on pas favorisé une confusion entre inscription et publication ?

Enfin un tel numéro est forcément lacunaire, tributaire des contributions reçues. Certaines communautés s'intéressant pourtant de près au document ne s'y sont pas exprimées, comme par exemple celle des traiteurs d'image ou celle de la recherche (*retrieval*) documentaire. Ainsi, il faut rappeler que ce numéro ne constitue qu'une étape, que nous espérons constructive, d'un processus de longue haleine qui doit être poursuivi.