Document : forme, signe et médium, les re-formulations du numérique

Roger T. Pédauque

Document : forme, signe et médium, les re-formulations du numérique

Roger T. Pédauque, STIC-CNRS
Contacts pedauque@enssib.fr

Version 3, 8 -07-2003

Résumé

Ce texte présente un travail collectif de réflexion en cours au sein du réseau thématique pluridisciplinaire 33 du département STIC du CNRS. Il se propose de préciser la notion de document dans son passage au numérique à partir de recherches qui privilégient plutôt la forme (comme un objet matériel ou immatériel), le signe (comme un porteur de sens) ou le médium (comme un vecteur de communication).
Ces entrées montrent chacune que des transformations radicales sont en cours. Leur superposition souligne l'importance de la pluridisciplinarité pour une analyse lucide et complète de la notion et de son évolution.

Contexte

Très peu d'articles scientifiques proposent une définition du document, encore moins la discutent. Le document se repère directement comme objet d'analyse dans quelques rares communautés scientifiques : les chercheurs des sciences de l'information (ou *Information Science*), issus des travaux concernant les techniques documentaires largement renouvelés par le traitement informatique, et des chercheurs s'intéressant à la numérisation et aux problèmes d'indexation-catégorisation qui ont souvent élargi leurs réflexions à la gestion électronique de documents ou encore ceux développant des outils d'édition électronique (*Electronic Publishing*). Par ailleurs, parce qu'il est un outil indispensable à la construction et l'avancée de la discipline, on le discute en histoire et tout particulièrement en archéologie, ou en géographie spécialement concernant les cartes, ou encore en droit pour les textes et articles de lois, règlements ou circulaires, mais sous un angle instrumental et rarement directement.
Nombre de dictionnaires, de répertoires de normes, d'encyclopédies présentent des définitions, qui relèvent plus de la désignation ou de la description que d'une réflexion approfondie sur la notion.
Depuis le latin *documentum* qui donne au mot des racines professorale (*docere* = enseigner), jusqu'à sa marginalisation par l'emploi plus récent, plus fréquent mais guère plus précis, du terme "information", il semble que la notion s'appuie communément sur deux fonctions : la preuve (la bien nommée "pièce à conviction" des juristes ou l'élément d'un dossier) et le renseignement (la représentation du monde ou le témoignage). L'archivistique contemporaine, par exemple, reconnaît ces deux fonctions en admettant pour le document une "valeur d'évidence" (de l'activité) qui a un sens un peu plus large que la preuve juridique, et une "valeur d'information" qui correspond au terme renseignement ci-dessus.
Un très grand nombre d'autres travaux de recherche utilisent pour désigner des objets comparables un vocabulaire différent, parfois rigoureusement défini, souvent aussi sujet à des interprétations diverses. Ainsi les chercheurs en informatique depuis l'étude des réseaux, en base de données, en fouille de textes, en recherche d'informations, en traitement automatique de la

langue, jusqu'à l'ingénierie des connaissances ou encore les linguistes de corpus, les sémiologues, les psychologues de l'apprentissage, les sociologues de la culture ou de l'organisation, les économistes des médias ou de l'information, les juristes de la propriété intellectuelle et, d'une façon générale "les humanités" usent de vocables divers comme information, donnée, ressource, fichier, écrit, texte, image, papier, article, oeuvre, livre, journal, feuille, page, etc. qui, bien entendu, ne sont pas synonymes, qui ont chaque fois une justification dans le contexte particulier de la recherche concernée, mais qui, tous, ont un rapport (la plupart du temps non assumé) avec la notion de document.

Enfin, les documents sont omniprésents dans notre vie courante (notamment au plan administratif et même dans l'activité scientifique). Ainsi la notion est intuitive pour chacun d'entre nous sans que nous ressentions le besoin de la préciser.

Ce flou fait aujourd'hui problème. En effet le numérique bouscule profondément la notion de document sans que l'on puisse clairement en mesurer les effets et les conséquences faute d'en avoir au préalable cerné les contours. Du papier, support le plus courant, au numérique, ces transformations se repèrent facilement, par exemple par l'aspect matériel, le traitement cognitif, la perception ou encore l'usage. Cette remise en cause, même si elle fut annoncée par les textes de quelques pionniers et préparée par la convergence de plus en plus manifeste entre l'écrit et l'audiovisuel est toute récente, encore chaotique et sans doute sans retour. Il est probable que les nombreux chercheurs qui abordent ces questions sous de multiples facettes gagneraient à une vue d'ensemble leur permettant de se positionner plus lucidement.

Le contraste entre la relative stabilité qui prévalait jusqu'à présent et la vitesse et radicalité des bouleversements d'aujourd'hui explique sans doute le retard de l'analyse. On n'avait pas besoin de s'interroger, sinon comme historien, sur un objet trop courant pour ne pas être évident, et aujourd'hui on n'a pas vraiment eu le temps de prendre du recul.

Le document a été construit comme un objet, dont la concrétisation la plus banale est la feuille de papier, au cours d'un processus séculaire où se sont entrelacés outils, savoirs et statuts. Depuis quelques dizaines d'années avec le numérique, nous sommes entrés dans une phase nouvelle dont certaines caractéristiques sont en filiation directe avec la période précédente, tandis que d'autres marquent au contraire un changement radical et peut-être l'émergence d'une notion différente reprenant tout ou partie de l'utilité sociale de ce que nous appelions "document". La manifestation la plus évidente du changement est donc la perte de la stabilité du document comme objet matériel et sa transformation en un processus construit à la demande, qui ébranle parfois la confiance que l'on mettait en lui.

Le questionnement entre rupture et continuité ne se pose pas seulement sur l'objet. Les méthodes d'analyses ou les épistémologies sont, elles aussi, en évolution rapide.

Une réflexion pluridisciplinaire

Notre sentiment est que ces difficultés ne peuvent être levées que par une réflexion résolument pluridisciplinaire. Nous avons été conforté dans cette opinion par le département STIC du CNRS qui a lancé un réseau thématique pluridisciplinaire, baptisé "Document et contenus : création, indexation, navigation" (http://rtp-doc.enssib.fr) qui regroupe une centaine de chercheurs. Pour certaines des disciplines auxquelles ils appartiennent, le document n'est pas une notion centrale et les chercheurs n'en ont qu'une appréhension partielle. L'objectif du réseau est donc de déplacer cette attention oblique pour faire du document un objet principal de recherche, au moins pour un temps, en croisant les apports partiels des uns et des autres.

Il n'est pas sûr qu'il existe, entre les disciplines ou même à l'intérieur de chacune d'elles, un consensus sur les questions que nous abordons. Notre objectif n'est pas d'uniformiser, de définir une ligne, un courant ou une école de pensée, mais d'éclaircir et de préciser les notions afin de lever les malentendus, d'ouvrir les perspectives et de repérer les éventuels désaccords. Notre conviction est que le dialogue entre les disciplines ne pourra être fructueux que si l'on a réussi à dégager les notions essentielles, pour les discuter ou pour s'en servir comme point d'appui.
Cette tentative n'est pas sans péril. D'une part, le contresens, ou simplement la superficialité, est possible. D'autre part les bases différentes des disciplines ou des courants peuvent être contradictoires. Outre les difficultés conceptuelles, l'objectif peut aussi rencontrer des obstacles plus terre à terre. Chaque spécialité développe naturellement sa propre culture et son propre vocabulaire, pour de bonnes (rigueur) ou de mauvaises (protection) raisons. Les mêmes mots prennent parfois des sens différents selon les communautés, souvent même ils sont étrangers aux extérieurs. Nous sommes contraints dans un texte transversal d'utiliser un vocabulaire commun, dans tous les sens du terme, au risque de caricaturer.
Concrètement, ce texte est le résultat d'un travail collectif au sein du réseau. Compte tenu de la méthode employée pour sa rédaction et des nombreuses contributions qui lui ont donné corps, nous avons choisi de ne noter aucune citation ou référence bibliographique directe. L'inverse conduirait en effet à biaiser la dynamique collective en induisant une concurrence entre auteurs ou écoles de pensée. Néanmoins, une bibliographie est proposée sur le site du rtp-doc.

Propositions

Pour organiser nos propositions, nous utiliserons une analogie avec la distinction en linguistique entre syntaxe, sémantique et pragmatique. Sans entrer dans une discussion sur la validité de cette analogie, ni même sur la légitimité de cette tripartition en sciences du langage, nous constatons qu'elle permet un classement assez simple des recherches en cours et des courants qui les sous-tendent. Nous distinguerons :

- Le document comme forme ; sous cette rubrique, nous rangerons les approches qui analysent le document comme un objet, matériel ou immatériel et qui en étudient la structure pour mieux l'analyser, l'utiliser ou le manipuler.
- Le document comme signe ; pour ces chercheurs le document est perçu avant tout comme porteur de sens et doté d'une intentionnalité ; ainsi le document est indissociable du sujet en contexte qui le construit ou le reconstruit et lui donne sens ; en même temps, il est pris dans un système documentaire ou un système de connaissances.
- Le document comme médium ; cette dimension enfin pose la question du statut du document dans les relations sociales ; le document est une trace, construite ou retrouvée, d'une communication qui s'est affranchie de l'espace et du temps ; en même temps, il est un élément de systèmes identitaires et un vecteur de pouvoir.

L'analogie avec la linguistique reste cavalière. On pourrait arguer que la première catégorie est à mettre plus précisément en relation avec la morphosyntaxe, que la seconde comprend à la fois sémantique et pragmatique. Mais comparaison n'est pas raison, il nous suffit qu'elle soit efficiente sur notre terrain.
Chaque catégorie doit être vue comme une dominante et non comme une dimension exclusive. Ainsi, par exemple, les chercheurs qui abordent le document par la première entrée ne négligent pas nécessairement les deux autres, néanmoins leur analyse et raisonnement privilégient le premier éclairage, les deux autres restant des compléments ou des contraintes extérieures. Plus

précisément, le terme "d'entrée" serait peut être le plus approprié, chacune de ces entrées est en effet une façon d'aborder l'objet de la recherche, le document, à partir de laquelle les autres dimensions seront retrouvées à l'occasion de développements, de contraintes, d'obstacles ou de limites apparues dans le raisonnement premier. Mais la tendance de chacune des approches est sans doute aussi de trop relativiser les autres.
Nous aborderons chaque catégorie selon le même schéma :

- tout d'abord, un repérage des principales disciplines, savoir-faire ou spécialités qui privilégient ce point de vue. L'objectif n'est pas de discuter leur validité ou leur scientificité, mais de passer en revue la diversité des travaux représentatifs de l'orientation sans jugement d'importance, ni de valeur.
- puis, nous proposerons une interprétation de l'évolution des points de vue dans le passage du document traditionnel au document numérique.
- nous construirons progressivement une définition du document à partir de chaque entrée.
- nous pointerons quelques questions qui restent en suspens à l'intérieur de cette catégorie, au-delà de l'approfondissement des recherches particulières en cours.

Chaque fois nous essaierons de dégager l'essentiel, sans trop nous arrêter aux nuances, exceptions et cas particuliers. L'objectif est de souligner les fondamentaux, pas d'être exhaustif. Concernant la définition, une méthode consisterait à rechercher systématiquement les cas ne lui correspondant pas et ainsi construire une définition universelle. Cette modalité ne nous parait pas très opératoire. Notre objectif n'est pas de répondre à tout mais de construire une définition générique, quitte à repérer des exceptions qui représentent soit des cas très particuliers, soit des situations intermédiaires ou transitoires quand ils ne relèvent pas tout simplement d'une analyse incomplète ou erronée.
Nous proposerons en conclusion une synthèse des trois entrées visant à souligner les éléments de continuité et ceux de rupture face à la période précédente.

Document comme forme

Le terme de "forme" peut être contesté car il prête à confusion, nous l'utilisons car nous n'en avons pas trouvé de meilleur. Il faut l'entendre ici à la fois dans le sens de "contour", voire de "figure" autrement dit le document est appréhendé comme un objet, ou une inscription sur un objet, dont on repère les frontières, et en référence au "formalisme", c'est-à-dire que cet objet ou cette inscription répond à des règles qui le constituent.
Ici, le document est donc vu comme un objet de communication régit par des règles de mise en formes plus ou moins explicites, qui matérialisent un contrat de lecture entre un producteur et un lecteur. Le document est principalement étudié sous l'angle de ce protocole implicite de communication quel que soit son contenu textuel ou non textuel précis.

Spécialités concernées

Il faut souligner dès le départ la place particulière prise par l'écriture, technique dont l'apprentissage largement partagé place depuis son apparition le document dans une situation sociale fondamentale.
Les savoir-faire, professionnels ou non, qui privilégient ce point de vue sont nombreux, parfois très anciens comme la calligraphie, la typographie, ou bien, pour d'autres formes de la représentation, les techniciens de la musique, de la vidéo ou du cinéma et aussi la

bibliothéconomie dont le cœur est le catalogage, le classement et la gestion des documents, ou encore la diplomatique archivistique.
Il est logique que les informaticiens qui partent d'objets matériels pour les numériser, ceux de l'image donc, aient de fortes parentés avec ces premières spécialités. Ils s'intéresseront vite à la structure interne des documents avec les systèmes automatiques de reconnaissance des formes, en tout premier la reconnaissance automatique des caractères, puis de l'écriture manuscrite ou encore de la forme des mises en page ou des images. Dans leur domaine, ils retrouveront les problèmes de formats, d'échange, de stockage, de description, d'adressage, de conservation, ou de traitement des grands nombres. Il s'agit ici de la lecture automatique ou analyse de documents. Les chercheurs tentent de décoder l'objet en explicitant/exploitant le protocole de communication sous-jacent (le contrat de lecture).
De la même façon, tous ceux qui s'intéressent aux caractères typographiques, aux mises en page, aux formats éditoriaux, à la normalisation internationale de ces domaines, au traitement de texte, ceux qui construisent des systèmes numériques de vidéo déclinent en les renouvelant profondément de vieux savoir-faire.
D'autres spécialités informatiques ont opté pour ce premier point de vue. La conception de systèmes de gestion électronique de documents, comme le nom l'indique, part bien de l'idée que le document préexiste comme objet repérable, même s'il peut être virtuel. Si cette fois, le point de départ n'est plus un objet concret, mais un fichier informatique, bien des problèmes posés relèvent des mêmes interrogations fondamentales.
Enfin, un changement brutal d'échelle est apparu avec l'invention et le succès explosif du World Wide Web. Autour de cette toile, une intense activité de recherche, conception, négociation, normalisation, développement s'est déployée, notamment mais pas seulement, au sein du consortium W3C. Même si ces chercheurs emploient peu le mot de "document" et lui préfère celui de "ressource" qui englobe bien d'autres objets, bien des questions posées par les concepteurs du Web, dans sa version actuelle (c'est-à-dire avant le "Web sémantique"), relèvent aussi largement de cette première approche : il s'agit de savoir comment relier, à une échelle planétaire, des ressources entre elles et donc de définir des normes et des systèmes applicables sur toutes les machines et de donner une adresse repérable aux dites ressources. Parmi celles-ci, beaucoup ont les caractéristiques du document tel qu'on l'entend dans cette première dimension : fichier HTML ou XML, images, enregistrements sonores ou vidéo, flots multimédia, etc..

Evolution

Une première définition du document pourrait être représentée par l'équation : *Document traditionnel = support + inscription.* Dans un premier temps en effet, l'accent est mis sur un support manipulable (au sens premier) sur lequel est fixé une trace interprétable selon sa forme par un oeil, une oreille humaine, ou par le toucher pour la lecture en braille et, pourquoi pas demain, d'autres sens, avec ou sans prothèse. Cette trace représente le contenu, matérialisé par une inscription.
Le support traditionnel dominant (mais pas exclusif) est le papier, la trace l'écriture, manuscrite ou imprimée. La feuille écrite, élément de base, peut être enrichie par les mises en page et le paratexte et allongée par les reliures, renvois etc. conférant à ce document une très grande plasticité et complexité. Le "codex" (le livre avec des pages reliées) est sans doute la forme la plus élaborée du document traditionnel. Sa qualité peut se mesurer à la robustesse de "ses spécificités", pratiquement inchangées depuis plus d'un millénaire !

Au prix d'un important effort social (l'école qui permet l'acquisition du protocole de lecture), ce document est directement perceptible, c'est-à-dire sans outil intermédiaire de forte technicité (sinon pour certains des lunettes), par une part plus ou moins importante de la population d'une société donnée : celle qui a appris à lire.
Quand, dans l'histoire, cette notion (*support* + *inscription*) s'est étendue à d'autres formes de représentation, comme la musique enregistrée, le cinéma puis l'audiovisuel, le support n'a pas gardé sa faculté d'appropriation directe. Alors même que la représentation se rapprochait de la perception humaine immédiate (et donc nécessitait un apprentissage moins lourd pour être décryptée), le dispositif de lecture s'est sophistiqué. Il est indispensable de disposer de machines pour écouter un disque (gravé), projeter un film (inscrit sur une pellicule), ou une bande vidéo (enregistrée). L'objet est toujours nécessaire à la lecture, mais il n'est plus suffisant.
Mieux, la radiodiffusion puis la télévision ont permis de séparer le décryptage de la transmission du signal. Ainsi l'auditeur ou le téléspectateur écoute ou regarde un "programme" dont la transmission lui échappe. Il n'est pas maître de son moment de lecture, sauf ponctuellement par l'enregistrement sur bande magnétique. D'une certaine façon, la diffusion hertzienne en entrant dans son domicile l'a dépossédé d'une partie de l'autonomie spatio-temporelle qu'il avait gagnée en manipulant des objets gravés ou enregistrés.
Ainsi, l'audiovisuel a ouvert la voie à une évolution de l'utilisation des supports, mais la mutation essentielle pour nous est le passage de l'inscription d'un signal analogique à un signal numérique, avec toutes les facilités de traitement informatique qui l'accompagnent. Celle-ci a des conséquences radicales pour l'ensemble des documents, écrits, images et audiovisuels. Les mutations se repèrent sur les dispositifs d'écriture-lecture et sur les documents eux-mêmes.
Concernant les dispositifs, on observe tout d'abord un étonnant chassé-croisé entre l'écrit et l'audiovisuel. Le premier intègre un dispositif familier au second. Il n'est plus possible de lire sans machine. Même si la production de l'imprimé passe par un fort appareillage technique, la lecture du papier est, nous l'avons dit, directe ou presque. Disques optiques ou magnétiques, bandes enregistrées, machines pour traiter et restituer le signal ainsi que connections par réseau sont les outils indispensables à acquérir individuellement pour une lecture électronique, quitte à revenir à l'état précédent grâce à son imprimante. L'auditeur ou le téléspectateur peut, de son côté, retrouver dans le Web la maîtrise du lancement et de l'arrêt d'un flux temporel perdue dans le réseau de radiodiffusion qui transmet un flot de programmes ininterrompable, simplement modulable jusqu'ici par l'enregistrement du magnétoscope.
La seconde conséquence sur les dispositifs est l'entrelacement des supports et des signaux. La notion de support se complexifie et devient ambiguë. Est-ce le fichier, l'outil matériel qui l'héberge ou encore la surface de l'écran sur lequel il s'affiche ? Lorsqu'il traverse le réseau, le document est copié par fragments dans les routeurs pour une durée très brève, mais surtout il peut être stocké en totalité dans des caches pour une durée variable. Les mêmes « supports » d'ailleurs peuvent accueillir indifféremment n'importe quel type de représentation pourvue qu'elle soit numérique et même, sous réserve de compatibilité de format, les représentations elles-même peuvent se confondre : on pourra "lire", étroitement imbriqué, du texte, de l'image, de l'audio ou de l'image animée.
Alors que l'imprimerie avait privilégié le support matériel, du fait de la lourdeur technologique inhérente à toute activité de production de document, la publication électronique a rendu possible la production à la demande d'un document (indifféremment sur support écran ou sur papier). De ce fait le support a perdu sa position privilégiée au profit de la publication électronique. On peut, à ce propos, se souvenir qu'une des évolutions importantes de la publication assistée par

ordinateur a été l'avènement du *wysiwyg* (*What you see is what you get*) qui a rendu possible la visualisation à l'écran comme sur le papier.
Enfin, depuis le développement conjoint de la micro-informatique et des télécommunications, les machines elles-même se multiplient et s'autonomisent, entre ordinateurs portables, tablettes, téléphones, et outils intégrés de diverses factures recherchant la meilleure adéquation entre les comportement des lecteurs et leurs besoins, génériques ou spécifiques. Il n'est pas indifférent pour le devenir des documents que la diffusion du téléphone portable ait été beaucoup plus rapide et plus étendue que celle du micro-ordinateur.
Ainsi la notion de support a perdu sa limpidité première. Mais dans notre équation (*support* + *inscription*) les conséquences du numérique sur le second terme, l'inscription, sont tout aussi radicales. L'inscription relève du codage, une opération familière des informaticiens. Ils ont alors cherché à isoler les éléments logiques qui constituent cette dimension du document, pour les modéliser, automatiser les opérations et réagencer les différents éléments ainsi perfectionnés.
On peut faire sur ce volet un rapprochement avec la notion de programme telle qu'elle est souvent présentée en informatique : *programme = logiciel+ données*. Un document ne serait qu'un cas particulier d'un programme informatique dont la partie *logiciel* représenterait la "structure" et la partie *données* le "contenu", et l'équation deviendrait : *document numérique = structure + données*. En cohérence avec cette première entrée sur la forme, les chercheurs n'interviennent pas sur le contenu et étudient au contraire de près la structure qui, par définition, est modélisable et représente en quelque sorte, indépendamment du support, le "contrat de lecture", passé entre le producteur du document et ses lecteurs potentiels.
La structure varie énormément selon le type de document. Certains sont peu ou pas structurés, par exemple certaines oeuvres d'art ou textes spontanés où forme et fond sont indissociables, d'autres, au contraire, suivent des règles formelles rigides. La structure est aussi différente selon le type de médias. L'audiovisuel introduit, par exemple, une dimension temporelle peu présente dans l'écrit. Néanmoins, l'analyse a permis de repérer et d'isoler plusieurs niveaux de structuration dans le cas le plus général. Ceux-ci ont été construits à partir de deux courants de recherche, les uns partent de l'analogique pour aller vers le numérique, les autres font le cheminement inverse. Avant de revenir sur la notion de structure, il est préférable de comprendre la logique de leur raisonnement.
Le premier courant s'est donné comme tâche de traduire en un format numérique les documents traditionnels afin qu'ils puissent profiter des performances de l'outil. Autrement dit, il faudra faire passer un document traditionnel d'une équation à l'autre : *support* + *inscription* vers *structure* + *données*. Cette opération peut se faire à partir du document premier que l'on numérise. On cherche précisément à dématérialiser le document en s'appuyant sur une démarche de traitement d'images et de reconnaissance de formes. On peut aussi raisonner simplement à partir de la représentation d'un document en reconstruisant directement, sans l'aide d'un ancien support, l'équivalent visuel de tout ou partie de sa représentation. Notons que l'opération n'est pas socialement triviale, elle doit pouvoir s'effectuer dans les deux sens, notamment pour des raisons juridiques, nous y reviendrons dans la troisième entrée. C'est dans ce premier courant que l'on trouve les traiteurs d'image ; comme leur nom l'indique, leur recherche consiste à reconstruire l'image, donc la représentation formelle, d'un document.
Le principe est la reconnaissance des formes. Pour être reconnue, une forme doit donc avoir été préalablement connue. Plus le document d'origine s'appuie sur des structures génériques, plus il sera facilement transposable. Ainsi la complexité s'accroît quand on passe des caractères typographiques au graphique, aux schémas, puis à l'image et enfin aux objets en trois dimensions. Même si l'objectif est de reproduire une perception similaire ou homologue à celle de l'objet d'origine, il s'agit bien néanmoins d'une traduction nouvelle qui pourra occulter des éléments

signifiants ou au contraire en faire découvrir ou redécouvrir de nouveaux, selon les choix technologiques effectués et l'utilisation ultérieure des fichiers. (..)
D'autres chercheurs partent directement de l'équation finale (*document numérique = structure + données*) ; autrement dit, ils font le chemin inverse. A partir du cœur du raisonnement informatique, l'élaboration d'algorithme, ils reconstruisent les documents, remontant pas à pas leur logique ou structure interne pour déboucher sur une représentation lisible à l'écran. Ce second courant est issu de l'utilisation courante du texte dans les langages informatiques en intégrant progressivement le souci formel (*wysiwyg*), il a débouché d'abord sur la bureautique, il a rencontré les chercheurs développant des outils pour la publication électronique, et enfin il a été confronté à la nécessité de pouvoir échanger les documents à grande échelle. Il a alors véritablement explosé avec la révolution du Web.
Ces informaticiens ont raisonné par couches, pour isoler et traiter séparément les éléments de la structure du document. Ils ont ainsi découvert ou redécouvert les différents niveaux logiques de cette dernière, le niveau le plus bas étant celui du texte ou du signal analogique, que l'on a cherché à unifier sous forme d'unicode, MPEG etc.
La notion de balisage du texte, de son côté, est attachée à celle de la structure du document depuis la transformation des photocomposeuses électromécaniques en photocomposeuses numériques. Progressivement elle a instauré deux principes : les balises décrivent la structure plutôt que les caractéristiques physiques du document et sont compréhensibles aussi bien par un programme que par un interprète humain.
Sans refaire l'histoire de tout ce cheminement, disons que, du point de vue qui nous occupe ici, le Web peut donc être décrit comme une infinité de documents reliés entre eux. Son architecture est bâtie sur trois piliers : des ressources identifiées par un schéma d'adressage universel (identification), qui sont représentées par un ensemble non-exclusif de schémas (représentation) et échangés selon des protocoles standardisés (interaction). Cette architecture suppose que l'on puisse accéder à des documents de partout, sur n'importe quel type de matériel et selon la spécificité des groupes d'utilisateurs.
Les deux courants, celui de la reconnaissance de documents traditionnels et celui visant la construction directe de documents numériques, ne sont pas indépendants et, partant de points différents pour atteindre le même but, ils convergent. Ils ont permis notamment de souligner deux niveaux fondamentaux de structuration des documents : la structure logique (la construction d'un document en parties et sous-parties articulées entre elles), et la représentation formelle de la présentation, les "styles" au sens informatique (pour le texte, par exemple, les choix typographiques). Pour notre propos, la révolution fondamentale est peut-être l'uniformisation progressive du format (au sens informatique) du document qui permet justement un traitement simple de ces deux niveaux.
Un document devrait être lisible sur tout type d'ordinateur et déchiffrable par des applications variées. La tendance est à la fragmentation : les formats "propriétaires" envahissent le marché et condamnent les formats "universels" et "gratuits" à rester le luxe de spécialistes. D'autre part, les formats "non universels" entraînent des situations d'illisibilité : un programme ne peut pas lire un fichier, une application ne peut pas ouvrir un document, une page Web ne s'affiche pas correctement sur un écran. Ajoutons à cela que le format doit pouvoir transcrire l'alphabet : le "format" serait scriptible en plusieurs langues. La normalisation est donc essentielle.
Il est probable que le succès de plus en plus manifeste de la norme XML, et de ses nombreux dérivés particuliers, marque une nouvelle étape sinon un aboutissement de ces mouvements. En effet, cette norme, issue à la fois de l'informatisation des techniques éditoriales (SGML) et de la sophistication des premiers balisages du Web (HTML), intègre dans un même fichier structure et

contenu par un balisage normalisé du texte, permettant de retrouver et de dépasser très largement la plasticité et complexité des feuilles reliées que nous avions soulignées au démarrage de cette partie et dont on avait perdu quelques fonctionnalités en route. Mais elle introduit aussi des questions inédites en renouvelant les termes de l'ancien contrat de lecture où le lien entre représentation perçue et structure logique était pérennisé par le support. Avec une approche à la XML, on capture la structure et le contenu ; la forme peut en être dérivée de différentes façons. Elle n'est pas représentée intrinsèquement. On peut dire qu'on délaisse la forme comme dimension primordiale du document. Mais d'autre part, beaucoup de travaux se mènent sur les différents moyens de représenter et produire la forme d'un document électronique, en particulier pour les documents XML.
Ainsi, nous pourrions bien avoir une nouvelle transformation de notre équation : *document numérique = structure+ données* deviendrait *document XML = données structurées + mise en forme* dont la seconde partie (le "style") serait largement modulable. Dans le monde XML, la forme est définie séparément de la structure des données, par l'intermédiaire d'une feuille de style (XSL).
Une évolution possible, mais non certaine, serait que les documents ainsi « rédigés » rejoignent des bases de données, centralisées ou distribuées, et que l'ensemble des fichiers s'apparente de plus en plus à un ou plusieurs vastes jeux de « legos » où des briques de différentes tailles, formes et usages seraient agencées selon des configurations très variées. Un dernier pas serait ainsi en train de se franchir : un document n'aurait de forme à proprement parler qu'à deux moments : celui de sa conception par son auteur qui devra le visualiser ou l'entendre, pour s'assurer qu'il correspond à ses choix (et encore ce n'est pas obligatoire si sa production relève du processus) et celui de sa re-construction par un lecteur. Il est peu probable que le document sera toujours identique dans l'un et l'autre cas. Une autre façon de concevoir cette évolution serait de considérer que le document est maintenant la base de données elle-même dont les différentes sorties ne seraient qu'une interprétation partielle de la richesse. Une communauté de chercheurs réfléchit à cette question, dans le contexte du Web sémantique, en termes de « documents virtuels personnalisables ».
Cette évolution pose le problème de la gestion des temporalités diverses d'un et de plusieurs documents et de son écriture, de son enrichissement ou de sa ré-écriture par des intervenants variés. Déjà la gestion des versions successives d'un même document est délicate aussi bien pour les personnes, les organisations ou encore à l'échelle du Web. Il s'agit d'inventer les procédures permettant de rattacher un texte à un auteur (ou à un collectif d´auteurs), tout en permettant à chacun de s´approprier - de se ré-approprier - tout ou partie de documents produits par d´autres ou par eux-mêmes afin de limiter la prolifération «bruyante» des versions différentes d´une même information sur le réseau et d´identifier la nature et les origines de ces modifications dans l´optique d´une gestion cohérente de l´ensemble des documents électroniques actuellement disponibles, indépendamment de leur format, de leur statut et en dehors de toute institution centralisée.
Nous percevons très clairement les prémisses de cette dernière étape ; nous entrevoyons aussi bien des problèmes qu'elle soulève ; il est beaucoup plus aventureux d'en prévoir l'évolution et donc les conséquences, sinon pour dire qu'elles seront à coup sûr très importantes et durables.

Définition 1

L'observation de cette première dimension nous conduit à formuler une définition du document numérique, à ce stade incomplète mais représentative d'un important mouvement en cours. Cette

définition doit prendre en compte la marginalisation du support, le rôle fondamental pris à l'inverse par l'articulation entre la structure logique et les styles pour redéfinir le contrat de lecture, compris ici comme le contrat de lisibilité.
Un document numérique est un ensemble de données organisées selon une structure stable associée à des règles de mise en forme permettant une lisibilité partagée entre son concepteur et ses lecteurs.
Cette définition est sans doute trop longue pour être facilement mémorisable. Rappelons donc la transformation de l'équation :
Document traditionnel = support + inscription en *Document numérique = structures + données.*
Et suggérons l'évolution en cours dont l'issue reste encore incertaine :
Document numérique = structures + données deviendrait *document XML = données structurées + mise en forme.* En rappelant que*, stricto sensu*, la norme XML ne définit pas de mise en forme, celle-ci est définie par XSL.

Questions

Cette première approche, par la forme, laisse plusieurs questions en suspens. Nous insisterons ici sur celles qui se rapportent à la relation entre le monde perceptible et l'organisation numérique du nouveau monde documentaire.
Une première série de questions relèvent de l'affichage des documents. Alors même que la "bibliographie matérielle" a étudié de très près l'objet "livre" sous tous ses aspects, il semble que le passage au numérique ait surtout approfondi la question de la structure à partir de son entrée logique et à des fins de traitement. Ainsi, dans cette première dimension, les chercheurs considèrent volontiers que, puisque la structure est intégrée au fichier, n'importe quel affichage est alors possible, et donc les problèmes de perception relèveraient d'une autre problématique. Cette conception, poussée à la caricature, supposerait que structure et contenu sont indépendants, ce qui est pour le moins discutable. Il y a du sens dans la forme, et certains étudient, par exemple, depuis longtemps l'importance cognitive des cheminements possibles dans les liens hypertextes. Néanmoins, il reste un important travail à mener sur la lecture électronique pour mieux comprendre les mécanismes d'interdépendance entre les deux termes de l'équation.
Il est à noter que cette séparation détruit les bases de la diplomatique archivistique dont l'un des buts est d'authentifier le fond du document par l'analyse de sa forme, il en découle que l'authentification (la validation) doit (devra) être assurée par d'autres moyens, de type technique (filigrane électronique) ou organisationnel (tiers archiveurs certifiés). Pourrait -on imaginer que les exigences de forme - et entre autre la forme authentique - soient imposées et/ou validées par l'existence d'une feuille de style "signée" ou "filigranée" ?
Ces questions sont d'autant plus délicates qu'un même document peut se lire couramment sur différents appareils de lecture. Ainsi doit-on raisonner comme si les terminaux n'avaient pas d'influence sur la perception ? Il suffit pour se convaincre du contraire de comparer les écrans d'un micro-ordinateur, d'une tablette, d'un agenda électronique ou d'un téléphone portable. On retrouve alors le support de lecture dont on avait cru pouvoir s'affranchir. Ces questions font l'objet de débats nourris, notamment au sein du consortium W3C (*device independence*).
Les progrès réalisés sur la mise en écran, suite notamment aux travaux du laboratoire Parc de Xerox puis au perfectionnement des outils bureautiques, s'en tiennent à une organisation visuelle qui en reste, pour le document, à une mise en page et une mise en dossier. Quelques graphistes proposent d'intéressantes compositions. Mais ces efforts paraissent peu liés aux recherches précédentes. De même les livres électroniques ou les espoirs mis dans l'encre électronique n'ont

pas encore débouché sur des applications très probantes dans la vie courante, même si la possibilité de reconstruire un codex électronique fait rêver. Certains pour qui la représentation spatiale est fondamentale, comme les géographes ou les architectes, ont déjà mené une réflexion poussée sur ces questions, mais ils sont l'exception et non le cas général. Une nouvelle fois c'est peut-être l'audiovisuel qui ouvre des voies prometteuses avec la réalité augmentée où sont intégrés des aspects analogiques et des reconstructions numériques.
Une deuxième série de questions concerne la pérennité des documents numériques. Ces questions sont souvent débattues. D'un côté, elles ne diffèrent pas des très anciens problèmes d'archivage et de conservation, simplement transposés sur d'autres techniques. D'un autre côté, des problèmes radicalement nouveaux sont posés : les fichiers XML, pour peu qu'ils soient régulièrement rafraîchis et conservés dans de bonnes conditions, sont en théorie inaltérables puisqu'ils contiennent la totalité de leurs informations sous forme numérique. Ainsi certains considèrent que, sous peu, les problèmes de pérennité seront résolus. Mais à l'inverse, ces fichiers ne représentent pas, loin s'en faut, la (ou les) forme(s) sous lesquelles sont lus les documents. Ainsi une mémoire complète de ces documents supposerait de conserver la totalité des matériels et systèmes de lecture successifs qui permettent d'y accéder. Là encore, un gros travail théorique et pratique reste à mener.
Enfin, sans prétendre épuiser les problèmes, soulignons une troisième série de questions. Le document traditionnel est un objet matériel manipulable. Cet objet s'efface dans le numérique, jusqu'à, dans l'avancée ultime, n'être même plus qu'une sorte de puzzle dont les morceaux sont agencés à la demande du lecteur. Néanmoins, le lecteur a toujours accès au document à partir d'une machine, le terminal où ce dernier s'affiche. Ira-t-on vers une version extrême de cette idée où le document ne serait plus que l'avatar moderne d'une ardoise magique où s'afficheraient à la demande des éléments signifiants multimédias, seulement contraints par une logique de sens et de besoins spécifiés ? Ou aura-t-on une restructuration de documents types répondant à des besoins ou situations particulières dont l'éventuelle dynamique sera confinée à des plages strictement définies ? Et ne peut-on faire l'hypothèse que la stabilité visuelle du papier, la maniabilité et la co-existence des feuilles jouent un rôle important dans la cognition ? Alors ne faudrait-il pas encourager les efforts en direction d'un "codex électronique" ? Quel impact les nouveaux régimes de lecture vont-ils avoir sur nos régimes de savoir ? Qu'en est-il la responsabilité, juridique ou simplement morale, de l'auteur (individuel ou collectif) ? Ou plus directement en rapport avec notre entrée sur la forme : l'élaboration d'un document peut-elle se détacher de sa forme perceptible et donc, est-il simplement concevable d'envisager une rupture formelle entre l'élaboration par l'auteur (qui est aussi le premier lecteur) et la proposition faite aux lecteurs ? Le succès des formats de "fac-similé" (PDF), est souvent analysé comme une résistance momentanée au changement. Ne s'agit-il pas plutôt d'une indispensable stabilité perceptive.
On pourrait résumer ces questions par une seule : en effaçant le support, n'a-t-on pas trop délaissé justement la forme ?

Document comme signe

Comme pour l'entrée précédente, les termes du titre de cette partie ne doivent pas être lus dans une acception trop académique. Depuis fort longtemps, le signe fait l'objet de très nombreux travaux savants. Même si certains d'entre eux nous servent, nous ne cherchons pas ici à discuter le concept. Notre objectif est simplement de regrouper et présenter les recherches qui prennent le document d'abord comme un objet signifiant.

L'entrée qui nous intéresse ici est celle du traitement du contenu. Si la forme est parfois prise en compte, elle ne l'est que comme porteuse de sens.

Spécialités concernées

Cette catégorie concerne des disciplines sensiblement différentes de la précédente, certaines se présentant comme une avancée historique par rapport à celle-là, comme si, passant de la forme au signe, on se rapprochait du cœur d'un problème.
Ainsi, du côté des savoir-faire professionnels, on passe de la bibliothéconomie à la documentation puis aux professionnels de l'information, qui, plutôt que gérer des objets, fournissent des réponses aux questions du lecteur. Ou encore la GED devient le Knowledge Management (KM) qui, au-delà d'un système de gestion de stocks de fichiers, permet de repérer directement les connaissances utiles pour une organisation. Et surtout, le Web gagne un adjectif en se qualifiant de "Web sémantique", voulant ainsi signifier qu'une meilleure utilisation des capacités des machines interconnectées pourrait autoriser un traitement du contenu des fichiers en ligne en vue de l'organisation de services plus proches des demandes cognitives des internautes.
D'un point de vue académique, cette catégorie réunit d'abord ceux qui travaillent sur le texte, la parole, l'image donc des linguistes ou des sémioticiens toutes écoles confondues aussi bien ceux qui font de l'analyse de discours, de la linguistique de corpus, de la sémantique que ceux qui construisent des outils de traitement automatique de la langue en vue de traduction ou de recherche automatiques d'informations. Un rapprochement est en cours avec une seconde catégorie d'informaticiens issus plutôt des travaux sur l'intelligence artificielle, qui, à partir d'une tentative de modéliser le raisonnement, cherchent à construire des outils capables, eux aussi, de répondre aux questions sur la base d'une recherche dans des fichiers. Dans le même temps, on passe de la notion d'information à celle de connaissance qui a l'avantage sur la première d'intégrer le raisonnement. Une discipline nouvelle baptisée "ingénierie des connaissances" est ainsi en émergence.
Très vite, il est apparu qu'un travail d'information sur l'information était utile, parfois indispensable. Du catalogage à l'indexation, puis des thésaurus aux ontologies, les "métadonnées" sont devenues un outil et un objet de recherche essentiel.
Ici, comme pour l'entrée précédente, l'explosion du Web a modifié la donne en changeant l'échelle des ressources disponibles. Ainsi, la tentative de construction d'un Web sémantique lancée par les architectes du Web traditionnel a rencontré de façon heureuse les chercheurs de cette dimension.

Evolution

La définition du document traditionnel selon cette dimension pourrait être symbolisée par l'équation : *Document = inscription + sens*.
Ici, le support est accessoire, y compris pour le document traditionnel du moment que l'inscription est préservée. L'important est le contenu, matérialisé par l'inscription, qui est porteur de sens. Le sens, lui-même se construit par rapport au contexte de production et de diffusion du document qui va conditionner l'interprétation du contenu.
Trois idées forces nous semblent fonder cette dimension, dans un triangle classique en sémantique. La première concerne la création des documents, la seconde leur interprétation et la troisième les signes qui les constituent.
"Penser c'est classer", en réalisant des documents nous isolons et rangeons des discours pour nous aider à penser le monde. La mise en document est une façon de construire, ou de traduire, notre

compréhension sociale. Ainsi la notion de genre textuel et celle de collection sont fondamentales. En effet, les documents se regroupent par grandes catégories dont les différents items ont une homologie et une relation entre eux. Cette opération se réalise à la fois en amont (mise en document) et en aval (mise en collection). Le classement varie selon les situations et les époques. Il peut être très formalisé, comme simplement implicite. Il peut faire référence à des actions très précises et organisées (papiers d'identité, formulaires, contrats, etc.) ou à de simples attentions, impressions, sensations (médias, fictions, etc.). Il marque notre représentation sociale et nos lectures du monde. Il passe nécessairement par un système qui permettra de placer le document dans un ensemble et de l'y retrouver, une indexation au sens strict ou figuré, et donc des systèmes de classification concrets ou abstraits.

Le second point important est l'interprétation. Quels liens le document suggère-t-il ou instaure-t-il et sous quelle forme ? Un document n'a de sens que s'il est lu ou interprété par un lecteur. Cette interprétation dépend largement du contexte dans lequel elle est pratiquée. Un même document pourra prendre des sens différents, voir opposés, selon l'époque et la situation sociale ou individuelle de l'interprétant. D'une certaine façon, ce dernier re-crée le document chaque fois qu'il l'isole et en prend connaissance. Le lecteur doit ici être entendu dans un sens général, il peut s'agir aussi bien d'une personne physique que d'un groupe de personnes dans des espaces et des temps différents.. peut-être même d'une machine.

Ainsi un document, pour la dimension qui nous occupe ici, est pris dans une double relation, relation au monde documentaire (classement) et relation au monde naturel (interprétation). Ces relations se réalisent grâce à un "horizon d'attente", un ensemble de signes familiers qui construit le contrat de lecture entre le lecteur et le document lui permettant d'en décrypter sans difficulté le sens, car il sera placé d'emblée dans son contexte d'interprétation. Les éditeurs, par leur intervention sur le texte, sa mise en forme, et aussi par leur action commerciale, sont les premiers artisans de cette construction pour les documents publiés. La notion donc de "contrat de lecture", dont nous avions souligné l'importance dans l'entrée précédente par la forme, prend une épaisseur supplémentaire puisqu'il est aussi indispensable à la compréhension du document.

La troisième idée force concerne les signes eux-mêmes. Tout objet est potentiellement un signe et pourrait-être un "document". Une discussion, désormais classique, a montré par exemple qu'une antilope dans un zoo (donc dans un système social de classement) était un document. Mais la très grande majorité des documents sont construits à partir du langage, écrit majoritairement ou parlé. Le zoo, lui-même, est construit autour d'un discours, l'antilope est en quelque sorte "documentée". On peut faire la même remarque au sujet des documents audiovisuels qui sont toujours accompagnés de "légendes de lecture" sous la forme d'un grand nombre de textes dès le moment de leur fabrication jusqu'à celui de leur exploitation. La structure de la langue écrite depuis la lettre d'alphabet dans les langues indo-européennes jusqu'au discours organise donc la plupart des documents. Ceux-ci sont fait de morceaux discrets, plus ou moins isolables et réagencables, analysables, soumis à des règles de syntaxe, de mise en discours et de style. Cette utilisation de la langue naturelle confère au document une très grande plasticité.

L'explosion documentaire, c'est-à-dire la brutale augmentation du nombre de documents manifeste dès la fin du 19ème siècle et sans rémission depuis, a conduit à l'invention de ce qu'on a appelé les "langages documentaires" (références bibliographiques, index, thésaurus, résumés etc.), organisés de façon associative ou hiérarchique, qui sont directement issus de la triade précédente : il était possible en effet de construire à partir des textes des documents (ou des images, ou des objets eux-même) un langage artificiel ou formel permettant de les classer pour les retrouver à la demande. De longue date, les archivistes ont également collecté des métadonnées sur les documents et leurs producteurs, dans le cadre de la description archivistique,

qui présuppose la notion de contexte du document comme le préalable essentiel de son exploitation future.
La construction de ces "langages" pose de nombreux problèmes. Elle suppose d'abord une normalisation, un certain nombre de règles communes sur lesquelles les différents protagonistes font consensus. Le consensus n'est pas suffisant, il faut y ajouter la motivation. Chaque personne qui participe à l'effort commun doit en récolter un avantage clair, faute de quoi il est peu probable que la construction collective sera effective. Ces langages enfin balancent continuellement entre l'universel et le contingent. L'hésitation est souvent mal comprise. Il ne s'agit pas d'une faiblesse conceptuelle ou d'une incapacité à choisir. C'est, au contraire, une dynamique fondatrice du mouvement documentaire, reposant sur la triade présentée en introduction de cette entrée : des signes pris dans une dialectique entre général qui classe et particulier qui réfère (interprétation).
L'insistance, justifiée ou non, des documentalistes à se différencier des bibliothécaires par le service qu'ils rendent : la recherche d'information est aussi révélatrice d'une conception de l'information et de son détachement du support. Les premiers s'attacheraient à analyser le contenu des documents pour présenter directement à leurs usagers les réponses qu'ils attendent et non simplement le ou les documents qui les contiendraient éventuellement. Ainsi, les documentalistes participent à l'interprétation des documents disponibles en reconstruisant, en quelque sorte, pour le lecteur un document, ou un dossier documentaire, adapté à son besoin.
Les "sciences de l'information" sont issues de ce mouvement. Le terme d'"information" à mi-chemin entre "donnée" et "connaissance", reste mal défini. Il faudrait sans doute plutôt parler d'unités documentaires. Les sciences de l'information s'attachent à comprendre comment des unités s'emboîtent (une idée scientifique est exposée dans un article, publié dans un numéro de revues, diffusé sous un titre, récolté dans une collection, etc.), se distribuent selon des lois statistiques d'une grande régularité, à perfectionner les langages documentaires et, surtout, à analyser finement le processus de recherche d'informations entre un usager ou un lecteur et un système d'accès.
Dans un premier temps, le numérique a été utilisé par les documentalistes simplement comme un outil performant de classement des items des langages documentaires sous forme de bases de données bibliographiques. Rapidement le traitement informatique de la langue naturelle puis la production et la gestion directe de documents électroniques, le succès du Web et enfin la modélisation du raisonnement ont changé la donne.
Le traitement automatique de la langue dépasse largement la problématique documentaire. Néanmoins à l'inverse, le traitement du document est nécessairement concerné par les avancées et les difficultés des outils de traitement de la langue dès lors que l'on s'attaque au texte intégral. Soit pour de l'indexation automatique, soit pour des résumés, soit encore pour des systèmes de question-réponse, les informaticiens et les linguistes ont réuni leurs compétences en utilisant des outils statistiques et morphosyntaxiques. A leur manière, ils ont suivi un chemin parallèle à celui des documentalistes, reconstruisant à l'aide de filtres et de calculs, sinon un langage au sens informatique, du moins du texte censé représenter d'une façon structurée le contenu du document et autorisant ainsi un traitement automatique par les machines. Les résultats furent, dans un premier temps, moins probants que ne le pensaient les zélateurs. Les meilleurs outils ont dû intégrer une part de travail humain, se présentant plus comme des outils d'aide que comme des outils automatiques.
Néanmoins, leur efficacité est, pour l'internaute sinon pour l'initié, spectaculaire dans leur application au Web sous forme de moteurs. Il est frappant de retrouver dans ces derniers, peut être parce qu'ils s'adressent maintenant aux très grands nombres (de documents, comme d'internautes) les vieilles questions bibliothéconomiques déjà retravaillées par les sciences de

l'information : lois bibliométriques (Zipf), collections (copies caches), indexation et mots clés (métadonnées), citations (liens), prêts (hits).. évidement largement renouvelées par la puissance de calcul, utilisant les apports du traitement automatique de la langue, mais résultant souvent plus d'un bricolage empirique de méthodes que d'une analyse scientifique trop rigoureuse.
Comme dans l'approche précédente, les informaticiens ont cherché à isoler les éléments logiques pour les modéliser. Mais ici, ils se sont attaqués directement au contenu. Ainsi, comme précédemment, nous pourrions représenter la transformation par celle de l'équation première : *Document = inscription + sens* devient avec le numérique *Document numérique = texte informé + connaissances*. Le remplacement d'*inscription* par *texte informé* voudrait signifier que le texte (pris au sens large, y compris audiovisuel) a été soumis ou pourrait être soumis à un traitement permettant d'en repérer les unités d'information. Le remplacement de *sens* par *connaissances* voudrait introduire la notion de personnalisation pour un lecteur ou un usager donné.
L'arrivée annoncée du Web sémantique peut être comprise à la fois dans la continuité de ces résultats et comme, sinon une rupture, du moins un saut méthodologique. Pour la première interprétation, on notera, par exemple, la structure toujours plus formalisée des documents (XML) et l'insistance sur l'indexation (RDF). Il s'agit de ce point de vue de construire une bibliothèque distribuée multimédia à l'échelle du réseau des réseaux, intégrant des outils de recherche plus performants. L'ambition est aussi plus large. L'objectif est de passer d'un Web constitué simplement d'un ensemble de fichiers reliés entre eux à un réseau utilisant pleinement les capacités de calcul des machines reliées notamment concernant les traitements sémantiques des textes. A cette fin l'utilisation des "métadonnées" que l'on peut modéliser et combiner est essentielle. Ainsi, à leur manière, les promoteurs du Web sémantique construisent des sortes de langages documentaires, qu'ils ont baptisé "ontologies".
La rencontre des promoteurs du Web sémantique avec les chercheurs de l'ingénierie des connaissances, dont l'objectif est la modélisation du raisonnement, était alors inévitable. Ceux-ci réfléchissent, depuis les années quatre-vingt-dix à la façon de rendre compte des raisonnements contenus dans des documents. Ils intègrent, notamment, la question des statuts des documents, de la modélisation des raisonnements et surtout celle des ontologies. Les ontologies ont été définies comme des représentations d'un domaine, accentuant la dissociation (parfois provisoire) entre le raisonnement heuristique et la description des concepts manipulés par ces heuristiques. Cette dissociation assumée était aussi un moyen de simplifier la modélisation de deux types de connaissances, considérées comme indépendantes dans un premier temps. Les ontologies se focalisent donc sur l'essence d'un domaine (comme la médecine, ou une spécialité de la médecine par exemple), sur son vocabulaire et, au-delà, sur le sens dont il est porteur. Ce sens a deux facettes, celui compris par l'être humain et c'est la sémantique interprétative et celui « compris » par la machine et c'est la sémantique formelle de l'ontologie. Les ontologies peuvent être vues comme une structuration plus riche que les thésaurus ou les lexiques utilisés jusqu'ici car elles introduisent d'une part une dimension sémantique (le réseau conceptuel) et d'autre part, dans certains cas, une dimension lexicale qui améliore les accès aux documents. Mais une des principales richesses des ontologies est, justement, leur côté formel qui va permettre leur usage par un programme informatique là où un thésaurus est en échec. Ce formel est obtenu en décontextualisant les concepts inclus dans l'ontologie. D'où la nécessité, pour des raisons de compréhensibilité, de maintenance, de relier cette ontologie à la dimension lexicale dont elle est issue, aux textes.
Ainsi, comme pour l'entrée précédente mais d'une façon sans doute moins avancée, nous sommes peut être au seuil d'une nouvelle étape pour le document numérique grâce aux apports du Web sémantique. Nous pourrions représenter cette étape par la transformation de l'équation précédente

: *Document numérique = texte informé + connaissances* deviendrait *Document WS = texte informé + ontologies.*
Néanmoins l'augmentation du nombre de documents accessibles dans une forme ne comprenant pas de métadonnée est beaucoup plus importante que celle de documents "indexés". Pire, la concurrence sur le Web conduit à des stratégies opportunistes d'indexation, visant à tromper les moteurs de recherche. Ainsi, il est vraisemblable que, au moins dans un premier temps, on trouve deux dynamiques parallèles. D'un côté, pour les communautés auto-régulées qui ont intérêt à développer une recherche documentaire performante (experts, entreprise, médias, etc.), des "langages métiers" seront appliqués aux documents le plus en amont possible de leur fabrication, vraisemblablement d'une façon manuelle-assistée. D'un autre côté, des méta-langages automatiques plus légers, éventuellement adaptés aux comportements de recherche de grandes catégories, continueront à se perfectionner pour les outils ouverts largement aux internautes.
Quand on observe l'évolution et les avancées des recherches dans cette dimension, il ressort un aspect cyclique : les changements de support, d'échelle ou d'outil obligent à reposer des questions anciennes. La construction controversée d'un langage parallèle émerge à chaque étape. Ainsi les tenants de l'étape précédente ont l'impression que les nouveaux venus redécouvrent des problèmes anciens tandis que ces derniers pensent que le saut réalisé oblige à remettre à plat l'ensemble des problèmes. Il n'est pas vraiment étonnant que la construction d'un tel langage soit cyclique. Chaque changement de support ou d'échelle nécessite de reconstruire sa structure. Il faut prendre maintenant en compte outre la masse de données à représenter, leur aspect multiculturel et multilingue. En même temps, ses fondements ne sont pas pour autant remis en cause, ils sont (ou devraient être) simplement mieux connus et plus solides.

Définition 2

Selon cette deuxième dimension, nous pourrions donc présenter une nouvelle définition du document, sans prétendre toujours qu'elle englobe la totalité de la notion. Cette définition doit prendre en compte la capacité de traitement du contenu en vue soit de la recherche d'information, soit tout simplement du repérage du document. Elle relève du deuxième volet du contrat de lecture que nous avons repéré, celui de l'intelligibilité :
Un document numérique est un texte dont les éléments sont potentiellement analysable par un système de connaissance en vue de son exploitation par un lecteur compétent.
Une nouvelle fois, la définition est un peu laborieuse. Rappelons donc nos transformations d'équations plus schématiques, mais plus mémorisables :
Document = inscription + sens devient avec le numérique *Document numérique = texte informé + connaissances* qui pourrait déboucher avec le Web sémantique sur : *Document WS = texte informé + ontologies.*

Questions

En se rapprochant de la communication humaine, les chercheurs de cette entrée ont largement augmenté la complexité des problèmes à traiter. Bien des questions restent encore non résolues. Du côté des langues, par exemple, on peut s'interroger sur l'application des outils aux langues dont la structure et l'écriture sont en rupture avec les langues indo-européennes. Par ailleurs, la frontière entre l'automatisation et le travail intellectuel humain reste mouvante.
Mais pour notre propos, notons surtout que, pour les chercheurs qui privilégient cette entrée, le document parait souvent une notion secondaire, seul le texte, le contenu, compte vraiment. Pourtant le contenu n'a de valeur, nous l'avons vu en introduction de cette approche, que par

rapport à un contexte. Le document n'est-il pas justement une des constructions de ce contexte en positionnant les informations qu'il contient par rapport à celles contenues dans d'autres documents et en permettant au lecteur d'avoir une indication de la valeur du contenu par le statut du document ? Autrement dit en portant trop exclusivement son attention sur le traitement du texte, ne sous-estime-t-on pas la valeur sémantique de la mise en document ? Une meilleure prise en compte de la mise en forme matérielle est au cœur de nouveaux projets de recherche. Le plus immédiat est de bénéficier des balisages structurels (et à l'avenir sémantique) pour moduler les analyses de textes, l'identification de connaissances ou l'annotation. Une étude plus fine irait vers l'intégration d'éléments de mise en forme matérielle comme les polices, les casses, les indentations ou énumérations, etc. Des collaborations entre spécialistes des documents et de modélisation des connaissances sont alors indispensables.

Il se pose alors une série de questions à partir de la triade notée en introduction de cette entrée :

Est-il possible de traiter du sens d'un document sans relation proche ou distante à l'ensemble auquel il se réfère (collection, catégorie, renvois, bibliographie, etc.) ? Autrement dit, comment, au-delà d'un travail de laboratoire sur des corpus fermés, intégrer l'analyse du document comme "tête de réseau", génératrice de structures travaillant le sens ? C'est toute la question, renouvelée par le numérique, de l'enrichissement en liens des documents entre eux, dans les nouvelles situations d'hypertextualisation ou de construction motivée de collections.

Peut-on valider une information, autrement que par l'authenticité du document qui la contient ? Le problème de la confiance est en ce moment un sujet étudié par les chercheurs informaticiens (entre autres en représentation des connaissances) intéressés par le Web Sémantique et par le consortium W3C. Ils cherchent une solution technique (rajouter une couche formelle, interprétable par des agents logiciels) qui permette à des utilisateurs/lecteurs d'un site de renforcer ou diminuer la crédibilité, la confiance que l'on peut accorder aux informations qu'il contient. Cette vue technique du problème rejoint l'idée que sur le Web, ce sont les internautes qui valident l'information et rendent populaire ou non un site. Le projet pourrait être pertinent au sein d'une communauté spécialisée et qui possède ses conventions, il atteint très vite des limites à l'échelle du Web. Le problème étant plus complexe qu'un "vote" qui plébiscite un site et lui accorde du crédit. Des approches plus sociologiques s'imposent.

Lorsqu'il s'agit d'analyser le contenu d'un document pour en tirer des modèles de connaissances pour des usages particuliers, la validité et la pertinence du document sont mises à mal par d'autres sources de connaissances que sont les experts d'un domaine ou les utilisateurs. Suivant les objectifs poursuivis, les méthodes mêmes d'extraction de connaissances à partir de textes accordent un poids égal sinon plus fort à la formulation orale des connaissances. La valeur ajoutée par l'authenticité, la certification ou la reconnaissance du texte peut être donc mise de côté dans certains cas.

Dans quelle mesure peut-on isoler un élément signifiant dans un ensemble qui a une unité de sens, le document comme un tout ? Cette unité n'a-t-elle pas souvent un poids décisif dans la signification des éléments qui la compose ? Comment prendre en compte la signification globale, l'unité sémantique, d'un document en n'appréhendant que ses parties ? Les questions posées redoublent largement celles qui se posent autour des textes en sémantique. Le passage de texte à document mériterait sans doute une analyse plus fine.

Ces questions ont sans doute des réponses différentes selon les types de documents auxquels on les applique. Mais faute, pour le moment, d'une avancée réelle sur la typologie, elles restent, nous semble-t-il, largement ouvertes.

Document comme médium

Renouvelons pour la dernière fois notre précaution de vocabulaire : le terme "médium" doit être pris ici dans un sens large. Il regroupe toutes les approches qui analysent le document comme un phénomène social, un élément tangible d'une communication entre des personnes humaines.
Cette entrée relève donc d'une analyse de la communication, une communication particulière où le document est compris comme le vecteur d'un message entre des personnes. Nous pouvons affirmer ainsi qu'il s'agit d'une troisième dimension du contrat de lecture, celle de la sociabilité.

Spécialités concernées

Notons d'abord que le champ social concerné peut être partagé en deux : d'un côté, les organisations qui usent des documents pour leur régulation interne et pour atteindre les objectifs qu'elles se sont fixés et, d'un autre côté, les sociétés ou collectivités ouvertes dans lesquelles les documents circulent
Sans doute tous les chercheurs précédents pourraient être inclus dans cette catégorie, puisque tous s'intéressent à une activité sociale, néanmoins nous classerons ici ceux dont l'entrée est d'abord sociale avant d'être instrumentale.
Du côté des savoir-faire traditionnels aussi, les métiers déjà cités se retrouvent ici, mais nous insisterons sur les archivistes, dont la mission première est de garder trace de l'activité humaine en conservant des documents produits au fil de l'eau par celle-ci, et les éditeurs, dont le métier est de favoriser la construction et de rendre public des documents intéressants une collectivité.
Les disciplines des sciences humaines et sociales qui s'intéressent aux échanges sont potentiellement concernées par cette dimension. De fait, des sociologues, des économistes, des juristes, des historiens quelques psychologues, un nombre certain de philosophes, et bien entendu des chercheurs en sciences de la communication, en sciences politiques et en sciences de la gestion s'intéressent, directement ou indirectement, aux documents à partir de leur entrée disciplinaire.
Le numérique a renouvelé l'intérêt de nombre de chercheurs de ces disciplines tant sur le phénomène global que sur des situations particulières. Ainsi, sans qu'il soit toujours assumé, il y a un rapport entre la réflexion sur le document et l'intérêt nouveau pour les communautés d'intérêt, les collaboratifs, le travail en réseau, la mémoire et le patrimoine, la propriété intellectuelle, etc.
Mais dans cette entrée le décalage entre les informaticiens et les autres chercheurs est plus grand. Bien peu des spécialistes en sciences sociales et humaines ont une connaissance réelle de l'informatique. Inversement, les informaticiens ont une compréhension souvent rapide des problématiques sociales. Ce fossé conduit parfois à des enthousiasmes fascinés ou à l'inverse à des rejets radicaux, de la part aussi bien des tenants des sciences sociales et humaines que de ceux des sciences de l'ingénieur.

Evolution

Un document donne un statut à une information, à un signe matérialisé. Il est porté par un groupe social qui le suscite, le diffuse, le sauvegarde et l'utilise. Comme nous l'avons suggéré en introduction de ce texte, c'est une preuve qui fait foi d'un état des choses et c'est une annonce qui prévient d'un évènement. C'est aussi un discours dont la signature le rattache à un auteur. C'est un témoignage, sans nécessairement que cet objectif ait été voulu au moment de sa conception. C'est une pièce de dossier.

Pour être en harmonie avec les entrées précédentes, proposons la troisième et dernière définition sous forme d'une équation : *Document = inscription + légitimité*. Il nous semble que cette équation permet de représenter le processus social de mise en document. Le statut de document s'acquérrait sous deux conditions : l'inscription doit dépasser la communication intime (entre quelques personnes privées) pour devenir légitime *et* la légitimité doit s'affranchir de l'éphémère (dépasser le moment de son énonciation) et donc être enregistrée, inscrite. Ces conditions impliquent que si tout signe peut être un document, un signe particulier (même répondant aux deux dimensions précédemment traitées) ne l'est pas nécessairement. Par exemple, un journal intime n'est pas un document, sauf si quelqu'un prend l'initiative de le rendre public ou au moins de le communiquer au-delà du cercle restreint des familiers de son auteur. Ou encore, une émission de radio ou télévision en direct n'est pas un document, sauf si quelqu'un l'enregistre pour une utilisation sociale future.
Cette position ne fait pas l'unanimité complète des contributeurs de ce texte. Pour certains, la valeur d'un document pourrait en effet pré-exister à son partage ou son enregistrement.
Le statut de document n'est pas acquis pour l'éternité, il se donne et il peut aussi se perdre dans l'oubli collectif définitivement ou encore se retrouver si quelqu'un redécouvre et re-légitimise un document disparu de la conscience collective, mais non détruit.
Pourtant l'équation ne rend pas compte de la fonction sociale des documents. Les documents sont utilisés pour la régulation des sociétés humaines en assurant une communication et une pérennisation de la norme et des connaissances nécessaires à leur survie ou leur continuité. D'une certaine façon, nous pourrions dire que le contrat de lecture, dont nous avons repéré deux dimensions correspondant aux deux entrées précédentes : la lisibilité et la compréhension, prend ici sa troisième dimension : la sociabilité, l'appropriation par laquelle le lecteur en prenant connaissance d'un document marque sa participation à une société humaine ou, inversement, l'inscription sur un artefact d'une représentation du monde naturel et son insertion dans un patrimoine collectif.
Un document n'est pas nécessairement publié. Bien des documents, parce qu'ils règlent par exemple des questions privées (dossier médical, transaction entre des personnes), ou parce qu'ils contiennent des secrets non-divulgables, ne sont consultables que par un nombre très limité de personnes. Néanmoins, ils ont un caractère social dans le sens où ils sont rédigés selon des règles établies qui fondent leur légitimité, qu'ils sont utilisés dans des relations formelles, et que, en cas de disfonctionnement, ils auront valeur de référence. Inversement, la publication, large ou limitée, constitue un moyen simple de légitimation. En effet, une fois rendu public, c'est-à-dire consultable par un nombre conséquent de personnes, un texte fait partie du patrimoine commun. Il ne peut plus être modifié sans difficulté, sa valeur est appréciée collectivement.
La multiplication des documents est donc liée à l'évolution des sociétés, par deux dynamiques, externe et interne, qui se confortent l'une l'autre : celle de l'usage social des documents tout d'abord et celle de leur économie propre ensuite.
L'organisation politique et sociale s'appuie sur la production et l'échange de documents. Les religions et leurs clercs, les Etats et les administrations, les organisations productives et le commerce, la société civile dans leurs différentes composantes, leur évolution historique, leurs géographies et cultures propres, leurs fonctions changeantes se sont servies et se servent encore largement des documents pour leur régulation interne comme pour l'affirmation concurrentielle de leur identité et position. Ainsi, citons les principales sources de l'activité documentaire dans les pays occidentaux sans prétendre à l'exhaustivité :

- En France, le passage de l'Ancien régime à la République, puis celui de l'Etat-gendarme à l'Etat-providence, et enfin aujourd'hui l'intégration de l'Etat à des

ensembles plus vastes, comme l'Europe, ou la mondialisation, n'a pas été sans conséquence sur la production des documents, leur rôle et leur nombre. Il suffit, par comparaison, d'évoquer l'importance du document dans l'histoire parallèle de l'administration de la Chine pour percevoir combien celui-ci est à la fois fondamental et pourtant bien spécifique à chaque civilisation.

- L'industrialisation avec tous les savoirs et normalisations techniques, organisationnels, transactionnels et comptables qui l'ont accompagnée, a "produit" un nombre considérable de documents. C'est peut-être le facteur principal de l'explosion documentaire notée plus haut.
- Les progrès scientifiques et ceux de l'éducation ont considérablement augmenté le nombre des producteurs et des consommateurs de documents, pour le fonctionnement interne de la science et, plus encore, pour la popularisation des innombrables savoirs partiels qui l'accompagnent.
- Les échanges commerciaux ou non commerciaux, qui ont explosé avec le développement des transports, celui des télécommunications, et l'ouverture des frontières utilisent nombre de documents pour se "fluidifier" (matérialisation des transactions, notes techniques accompagnant produits et services, informations commerciales, etc.).
- Le développement du temps libre, l'allongement de la durée de vie, l'accroissement de "l'espace public" sont encore des facteurs essentiels du développement de la culture et d'un de ses principaux vecteurs : le document.

Cette dynamique a fait l'objet d'éclairages particuliers, mais les quelques tentatives d'appréhensions générales, peut être parce qu'elles restent des travaux solitaires, nous semblent des essais plus spéculatifs que démonstratifs.

La seconde dynamique qui permet de fonder le document comme médium est celle de son économie interne, qui se construit à partir de l'évolution des technologies qui le constitue (évolution développée dans les deux entrées précédentes) et, d'autre part, par les modalités de la mise en document. Celles-ci supposent, en effet, un travail dont il faut bien trouver les moyens de la réalisation. La mise en document peut être analysée comme un acte de communication ordinaire avec d'un côté un (ou plusieurs) expéditeurs et, de l'autre, un (ou plusieurs) destinataires. Des métiers se sont spécialisés sur tel ou tel moment du processus ou tel ou tel domaine d'application. Des systèmes se sont construits et formalisés pour répondre à la régularité de la production. Des entrepreneurs, petits ou gros, se sont lancés dans l'aventure ou des organisations l'ont prise en charge. Ces dispositifs ont un coût de mise en place, d'entretien et une inertie.

Deux courants de recherche principaux se sont attachés à étudier l'économie de cette mise en document. Le premier s'intéresse à la communication organisationnelle et étudie plutôt les documents dans un processus de travail ; le second analyse la communication des médias et s'intéresse au processus de publication.

La communication organisationnelle étudie le document immergé dans des pratiques et des situations professionnelles caractérisées, et donc contraintes, par des systèmes de règles. Elle le regarde alors à plusieurs niveaux : tout d'abord, comme un écrit identifié dans un contexte, formalisé par des règles d'écriture, de circulation, d'usage, inscrivant une intention liée à l'action, et gardant trace des négociations socio-techniques menées autour de lui. Cela amène, notamment, à étudier les processus de fabrication et de gestion des documents, activités qui ne sont plus seulement le fait d'acteurs spécialistes, mais qui se redistribuent dans l'ensemble de l'organisation. Elle le regarde ensuite comme un élément structurant l'organisation, en tant que

support de coordination. Enfin, elle le voit comme un moyen utilisé par les acteurs individuels ou collectifs dans leurs différentes stratégies. Sur le plan méthodologique, le document est pris comme un « observable » permettant d'étudier à la fois les relations entre les acteurs, (il est un médiateur), les modes de régulation (il est un outil de management) et les recompositions organisationnelles (il en est l'un des révélateurs).

Les premières avancées de l'analyse ne touchent qu'une partie de l'activité documentaire et sont encore loin de formaliser les contours d'une économie générale des documents. Il reste bien des incertitudes à lever par exemple sur les relations entre les systèmes documentaires et les systèmes d'organisation, l'évolution des différents métiers de la médiation ou encore l'économie des bibliothèques ou des archives. Chez les archivistes en particulier, les discussions sont nourries autour des pratiques de *Records management* et du *Business reingineering*, qui sont en voie de normalisation,. La doctrine actuelle (mais elle est loin d'être assimilée par les producteurs documentaires institutionnels) veut que les missions (objectifs institutionnels) engendrent des processus (fonctions organisationnelles), qui engendre des procédures (méthodes d'action formalisées), qui engendre des documents (ou des transactions).

Du côté des médias, l'économie de plusieurs secteurs est bien connue parce qu'elle a fait l'objet d'analyses particulières. Citons la communication scientifique avec le rôle de la publication des articles, la révision par les pairs, les citations, les pré-publications. Ou encore les médias grand public avec la gradation entre une édition, artisanale fondée sur la vente individuelle d'objets, la dialectique entre le fonds et le best-seller, les réseaux de distribution, et la radio-diffusion, plus industrielle organisée sur la captation à son domicile de l'attention du destinataire vendue à des annonceurs intéressés.

D'autres thèmes ont été approfondis, compte tenu des enjeux économiques qu'ils représentent, comme le droit de la propriété intellectuelle. Dans ce domaine, les traditions différentes entre le droit d'auteur latin et le copyright anglo-saxon permettent de distinguer quelques propriétés du document. Le premier privilégie l'attachement de l'auteur à son oeuvre tandis que le second met en avant la notion de publication, donnant la propriété intellectuelle à celui qui en prend l'initiative. D'une certaine façon, nous pourrions dire du droit d'auteur est un droit de l'œuvre, tandis que le copyright est un droit du document.

Notons un dernier point concernant l'économie des documents, de première importance pour notre propos : plus l'existence d'un document est connue, plus il sera lu et plus il sera lu, plus son existence sera connue. Grâce aux relations entre les lecteurs et celles entre les documents, il peut se développer un phénomène de résonance, qui prend différentes formes et différents noms selon les secteurs ou les spécialités. Les professionnels du marketing ou les stratèges des médias l'utilisent régulièrement en construisant des notoriétés qu'ils revendent sur d'autres supports. Dans la communication scientifique, le facteur d'impact, basé sur le comptage de citations dans les articles, relève du même processus (et conduit aux même dérives..). Cette qualité est peut être l'explication de bien des caractéristiques de la distribution des documents : best-sellers pour l'édition, prime-time pour la radio-télévision, modes diverses, concentration et éclatement ou encore régularité quasi-parfaite des lois de la bibliométrie quand de grandes quantités de documents sont accessibles de manière égale par un grand nombre de demandeurs.

Le numérique se traduit par des mouvements contradictoires dont l'interprétation n'est pas aisée. Le premier constat pourrait être celui de l'effacement, difficilement mesurable, car il s'est effectué en ordre dispersé, d'un nombre important de documents qui, tenus sous une forme traditionnelle, rendaient compte des procédures. La diffusion des outils informatiques s'est traduite assez souvent par la dissociation de fonctions assumées jusqu'alors par un type unique de document. Tel est le cas des registres de l'état civil, qui continuent d'être tenus sur papier pour

des raisons juridiques tout en étant numérisés pour les besoins de la consultation. De plus en plus souvent, coïncidant avec la disparition de cadres intermédiaires, le remplacement est total : formulaires, tableaux fiches, pilotages, modes d'emploi qui faisaient les beaux jours de la bureaucratie publique ou privée sont remplacés par des bases de données et de l'échange de données informatisées. Ce mouvement qui fut baptisé il y a quelques années "informatisation de la société" risque de s'accélérer encore avec les développements soulignés dans les dimensions précédentes.

Mais, dans le monde des organisations, il se constate conjointement une montée d'une mise en écrit et en documents, amplifiée de façon exponentielle par la démarche qualité. Le document y est porteur de normes sociales et organisationnelles, ce qui le positionne autant comme un support d'action que comme une mémoire des relations. Le fait de mettre des données ou procédures en banques de données n'efface pas leur valeur prescriptive, bien au contraire. Les Intranets, par exemple, allouent aux documents un statut (de référence, d'outil, entre autres) en y associant des règles d'identification et de circulation tout en modélisant et anticipant les usages possibles. ils amplifient la visibilité des décisions et des activités, en les rendant largement accessibles.

Dans cette perspective, moins que jamais, le document ne ferait sens isolément, mais serait constitué par l'enregistrement informatique de transactions préalablement définies. L'affichage des informations, éphémère et nécessairement dépendant de technologies évolutives, ne constituerait pas à lui seul le document ; il devrait être validé par des procédures certifiées. Nombre de documents peuvent s'apparenter à la transcription de procédures ou d'une de leurs étapes : ainsi, ces documents ne pourraient finalement se comprendre qu'en fonction du mode de traduction informatique dont les procédures qu'ils transcrivent ont fait l'objet. Si l'on ajoute les progrès de la signature électronique, bien des transactions pourraient à l'avenir se réaliser sans les formalismes adoptés pour le papier.

Cette mutation accroît considérablement les possibilités de contrôle par les facilités de croisement des informations. Dans le domaine social, la France s'est doté d'une protection légale, mais fragile compte tenu du développement des opérations numériques, par la loi "Informatique et liberté". Dans le domaine économique, certains analystes y ont vu l'émergence d'une nouvelle économie dont l'étude des déboires dépasse largement le cadre de cet article. Retenons, pour notre propos, l'idée d'un changement radical des structures socio-économiques. Comme l'ère industrielle a été marquée par interchangeabilité des parties, la société de l'information serait caractérisée par la possibilité de ré-utiliser l'information.

Ainsi, du côté de la communication organisationelle, nous aurions repéré une première évolution de notre équation première *Document = inscription + légitimité* en *Document numérique = texte + procédure.* Mais cette dernière équation ne rend pas compte d'un autre très important mouvement en cours du côté des médias à partir de l'avènement du Web.

Le Web a projeté brutalement le numérique à l'échelle de la société toute entière. Pour comprendre le succès qu'il a rencontré, mesuré par sa diffusion explosive dans les populations et selon les types d'activité, il faut revenir à l'esprit qui fonde son architecture. L'organisation du Web, est conforme aux orientations des concepteurs de l'Internet imaginé comme un réseau de communication de plusieurs à plusieurs où chaque pôle, grand ou petit, devait disposer des mêmes outils et être à la fois producteur et consommateur. Le Web suppose une conception sociale, ou plutôt de la communication sociale, proche de la "République des sciences" ou du mouvement des logiciels libres. Dans une telle société, chaque personne est acteur et responsable devant la communauté de ses actes. Si l'on traduit cela dans notre domaine, nous dirons que chacun est capable de lire ou d'écrire des documents qui concernent la vie collective et chacun

aura à coeur de ne rendre public que des documents qui enrichissent la collectivité. Les pionniers géniaux du Web, articulation de l'Internet et de l'hypermédia, ont construit un système à leur image, ou plutôt à l'image de la communauté informationnelle à laquelle ils appartiennent.
Cette idée est très présente dans nombre de discours et d'initiatives du domaine à commencer par ceux du consortium W3C. L'industrie du contenant (industrie des logiciels et des télécommunications), non sans débats, batailles et compromis, est très attentive à ces développements qui confortent ses positions puisqu'ils favorisent l'augmentation du trafic et des traitements au détriment d'une industrie du contenu.
Mais, tout le monde ne peut parler à tout le monde, ce serait une cacophonie, il faut des représentants. Jusqu'à présent cette difficulté était résolue par un ou des systèmes de filtres qui permettent de sélectionner les auteurs pertinents et de configurer des documents représentatifs et utiles. Ces systèmes ont un coût qui ne saurait se diluer dans le fonctionnement communautaire, puisque l'égalité des acteurs a disparu. Quelques uns seulement écrivent au nom des autres et des médiateurs professionnels organisent l'ensemble du système de publication et d'accès. Le système éditorial est un avatar de cette organisation, compromis entre intérêts privés et publics.
Il y a donc un malentendu, volontaire ou innocent, au départ entre les systèmes conçus par les pionniers de l'Internet et confortés par les industriels du contenant et la réalité ordinaire de la communication sociale. Ce malentendu est néanmoins d'une grande fertilité, car il permet aux collectivités dont la communication est bridée par le système traditionnel de trouver un espace pour échanger. Il donne aussi à nombre d'institutions, à commencer par celles d'intérêt général, un outil simple pour communiquer avec la population. Ainsi, le Web est un vaste bazar où l'on trouve une multitude de documents consultables gratuitement pour le lecteur et reliés entre eux.
Certains considèrent qu'il ne s'agit que d'une organisation provisoire, illustrant la jeunesse du média. Cette analyse ne prend peut-être pas assez la mesure de la rupture que produit le Web. Il est possible aussi que le filtrage ou la sélection ne se fasse plus dans le Web *a priori* comme dans les médias traditionnels, mais *a posteriori* selon un système de "percolation" où les documents les plus pertinents seraient progressivement repérés et mis en valeur, par le nombre de liens et par le jeu des moteurs. L'important alors serait la toile elle-même qui, par son mouvement continuel (les liens qui se font et se défont, les moteurs qui tournent, les pages qui apparaissent et disparaissent), autoriserait un repérage des documents. L'implication d'un nombre conséquent d'internautes, jusqu'ici à l'écart du petit monde fermé des médias, et le succès manifeste du média dans les pratiques donnent consistance à cette hypothèse en procurant une dimension et une rapidité inédite à la dynamique ordinaire de la légitimité par la notoriété.
En suivant un raisonnement parallèle, mais dans une perspective plus littéraire, plusieurs chercheurs ou essayistes ont vu dans l'avènement du Web, et plus précisément dans les techniques d'hypertextualité ou d'hypermédia, un effacement des documents. Ainsi la triade classique auteur-oeuvre-lecteur, à l'origine de la construction du document littéraire, pourrait laisser place à un processus interactif où les liens entre les pages accessibles joueraient un rôle plus important que le texte tel qu'il était auparavant construit par l'auteur. Néanmoins, même si d'intéressantes expériences d'écriture hypertextuelle ont été et sont toujours conduites avec des conséquences sémantiques et cognitives non négligeables, il apparaît que le développement explosif du Web a conduit au contraire à une multiplication exponentielle de documents mis en ligne. Les liens entre les pages paraissent se structurer progressivement pour construire de nouvelles normes de paratexte, renforçant au contraire l'aspect documentaire du Web.
Nous pourrions résumer notre développement sur le Web par une transformation de l'équation initiale deviendrait *document Web = publication + accès repéré.* La publication seule ne ferait plus la légitimité, il faudrait lui adjoindre la notoriété par le repérage de l'accès.

En cohérence avec notre raisonnement, les médias traditionnels n'ont pas pu construire sur le Web de modèles économiques viables. Seuls quelques secteurs, qui avaient déjà des affinités avec les réseaux ont trouvé des modalités de financement : l'information financière et l'information scientifique. Il est possible aussi que la musique à la suite du mouvement d'échanges entre internautes soit en train de redéfinir son mode de distribution et de valorisation.
Inversement, le numérique a permis de renforcer les "anciens" médias, édition et audiovisuel, en leur autorisant d'intéressants gains de productivité en favorisant les synergies et les diversifications. Ainsi, à partir d'une même base de données par exemple, un journal peu décliner les actualités à la fois sur des sorties papier, sur le Web, à la radio, par SMS, par audiotel, etc. Ou encore chaque média peut faire valoir ses propres domaines d'excellence (notoriété pour la télévision ou la radio, interactivité pour le Web et le téléphone, appropriation pour l'édition) et par la résonance notée plus haut conduire à des profitabilités inédites. Ces changements récents restent à évaluer. Ils conduisent aussi à des investissements lourds dont les retours sont à moyen terme alors même que l'avenir reste incertain. Ainsi, après engouement et tâtonnement, le Web n'apparaît que comme un média supplémentaire dont il faut bien comprendre les qualités propres pour l'articuler avec les médias existants.
Les développements annoncés du Web sémantique laissent présager d'autres développements, notamment dans les relations entre document et service. Mais cela reste encore de la prospective difficile à prendre en compte dans cette entrée à dimension sociale.

Définition 3

Dans cette perspective, nous avons donc repéré des mouvements forts, parfois divergents, souvent chaotiques. Il n'est pas simple à ce stade de proposer une définition reflétant clairement cette troisième entrée. C'est pourquoi nous nous en tiendrons à une définition très générale :
Un document numérique est la trace de relations sociales reconstruite par les dispositifs informatique
En rappelant les équations que nous avons construites et transformées : *document = inscription + légitimité* devient *document numérique = texte + procédure* et *document Web = publication + accès repéré.*
Malgré la difficulté de construction d'une définition, soulignons l'importance du troisième volet du contrat de lecture repéré, celui de la sociabilité.

Questions

La première série de questions concerne la notion d'archive dont la base est l'enregistrement et la conservation de documents. Le rôle de l'archive est de garder la mémoire d'une activité humaine. Un rôle nouveau, plus actif, émerge pour les archives avec le numérique. Archives ouvertes, récupération des programmes audio-visuels à la source ou des émissions de télévision diffusées, archivage du Web, bien des activités inédites se développent, tandis que les pratiques de l'archivistique se renouvellent, la mise en place du *records management* dans les organisations apparaissant comme une condition pour un bon archivage électronique. On y sent aussi bien des questions encore sans réponse définitive sur un rôle différent à assumer : hésitation entre le témoignage d'une action passée et l'enregistrement d'une action en cours ; confusion entre l'archivage et la publication ; simple enregistrement ou préparation d'une utilisation à venir. Plus encore, comment garder la mémoire d'un mouvement continu de renouvellement de pages reliées entre elles ?

La seconde série de questions concerne la notion d'attention (types de perceptions et d'intentions), sans laquelle un document ne saurait avoir de lecteur. L'attention humaine est limitée, par le temps disponible, par la fatigue du lecteur ou par les compétences techniques ou intellectuelles dont il dispose. Cette problématique est bien connue des radiodiffuseurs.
L'internaute étant nécessairement actif, il ne saurait être "capté" comme peut l'être l'auditeur de la radio ou le téléspectateur. Autrement dit le Web marie la liberté de choix de l'édition avec l'accessibilité de la radiodiffusion, ou élargit les services de bibliothèque à toute la planète pour la collection et le domicile pour la consultation. Ainsi les lois bibliométriques et les effets de résonance risquent de jouer à une échelle inédite par secteurs : l'attention se concentrant très fortement sur un nombre réduit de documents et se dispersant sur un très grand nombre. Ces phénomènes et leurs conséquences sont encore bien peu étudiés
Par ailleurs, l'intention des promoteurs du Web est de mettre à disposition de façon égale pour l'ensemble de la planète les sites ou les documents. Mais la diffusion des innovations est très inégalitaire. Le Web et le numérique n'échappent pas à la règle. Pire, il semble que ce soit le média le moins bien partagé entre les pays et entre les populations à l'intérieur de chaque pays.
La troisième série de questions concerne l'oubli du financement du contenu. La loi du moindre effort appliquée à l'accessibilité du Web fait que l'internaute préfèrera éviter tous les obstacles et barrières à sa navigation plutôt que les affronter. Ainsi, il contournera toute demande de financement directe. Dans la même dynamique, un mouvement militant fait de la gratuité une sorte de qualité naturelle du Web ou une accessibilité au savoir et à la culture libérée des contraintes commerciales.
Opportunisme et politique se conjuguent pour que, progressivement, l'économie du contenu sur le Web se configure comme un marché institutionnel, du "B2B". Est-on vraiment sûr que cette structure de financement garantit à moyen terme la diversité et la pluralité des documents mis en ligne ? Est-on sûr simplement qu'elle contient les ressources suffisantes, secteur par secteur, pour alimenter la production et gestion de documents ?
L'objectif de traduire l'ensemble des documents existants d'un support traditionnel à un état numérique est hors de portée. Le développement explosif de l'informatique oblige néanmoins à envisager des traitements de très grande ampleur, sauf à s'accommoder d'une amnésie radicale de notre culture documentaire. Il faudra donc dans un avenir proche être en mesure de faire des choix raisonnés (que numériser en priorité ?) et de construire des outils capables de traiter à un coût raisonnable de grandes masses de documents.

Conclusion

Les trois entrées abordées ont fait ressortir plusieurs thématiques fondatrices du document, confortées ou contestées par sa version numérique. Reste à envisager une synthèse proposant un éclairage général, qui englobe les trois points de vue, un peu comme la palette entière du peintre peut se construire à partir de trois couleurs primaires. 0u, dit d'une façon plus académique, est-il possible d'envisager une théorie du document à partir de laquelle nous pourrions mieux mesurer les conséquences actuelles et à venir du numérique ?
Remarquons d'abord sans surprise que nous avons repéré dans chacune des entrées des étapes dans l'histoire de la numérisation des documents que nous pouvons maintenant mettre en parallèle. Le document traditionnel repose sur un support, un texte et une légitimité. Une première phase de numérisation, celle dans laquelle nous sommes sans doute encore, a fait ressortir ses structures internes, l'importance des métadonnées pour le traitement et la difficulté

de la validation. Une seconde phase, sans doute commencée mais dont l'aboutissement reste incertain, insiste sur le format XML qui intègre la structure mais pas la forme, s'appuie sur les ontologies pour retrouver et reconstruire les textes, et met en avant l'accès personnalisé. Il y a dans cette évolution générale un sens dont il faudrait mieux comprendre l'orientation, les conséquences et les limites.

Soulignons ensuite que l'opposition entre papier et numérique est vaine. Quasiment tous les documents d'aujourd'hui ont été à un moment de leur vie sous un format numérique, et ceux qui échappent à cette règle risquent de tomber dans l'oubli. Inversement, de très nombreux documents numériques sont à un moment ou un autre imprimés, sur une imprimante individuelle ou dans une imprimerie professionnelle. Ainsi, l'important est bien de mieux cerner la notion de document en général, dont le numérique est à la fois un révélateur et un facteur d'évolution.

Notons enfin que dans chaque entrée nous avons insisté sur l'idée de contrat de lecture traduit par la lisibilité dans la première, par la compréhension dans la seconde et par la sociabilité pour la troisième. Il est probable que ce contrat à trois facettes présente, dans toutes les nuances que nous avons exposées, la réalité de la notion de document. Un document ne serait finalement qu'un contrat entre des hommes dont les qualités anthropologiques (lisibilité-perception), intellectuelle (compréhension-assimilation) et sociales (sociabilité-intégration) fonderaient une part de leur humanité, de leur capacité à vivre ensemble. Dans cette perspective, le numérique n'est qu'une modalité de multiplication et d'évolution de ces contrats. Mais l'importance qu'il a prise, sa performance, et la rapidité de sa diffusion rendent d'autant plus nécessaire une fine et juste analyse. Nous avons clairement montré, en particulier dans les séries de questions qu'aucune des entrées n'était indépendante. Il serait vain de tenter de les séparer, au contraire la notion ne s'éclaire réellement que dans leur superposition. Mais nous avons aussi constaté que chacune était prise dans un mouvement de recherches pluridisciplinaires qui avait son autonomie et dont la spécialisation implique des expertises trop pointues pour être totalement partagées.

Ce texte est un appel pour approfondir aussi bien chaque approche que leur croisement.

Roger T. Pédauque
CNRS - STIC
08-07-2003