Le document comme contenant, contenu et médium. Les reformulations du numérique

Alain Nossereau

La réflexion engagée par le département STIC du CNRS, dans le cadre du RPT 33, sur le thème des impacts du numérique sur la notion de document a conduit à la production d'un texte de référence (Le document : forme signe et médium, les re-formulations du numérique). A l'occasion de la rénovation des sections de STT (Sciences et Technologies Tertiaires), qui deviennent STG (Sciences et Technologies de la Gestion), dont les nouveaux programmes de communication sont orientés vers la production et l'enrichissement de documents numériques, ainsi que la gestion documentaire, l'Enseignement secondaire, et notamment les professeurs d'Economie Gestion en charge de ce futur enseignement, s'invitent à cette réflexion. Voici un document de travail, conçu et rédigé par l'un d'entre eux, sur la base du texte de Roger T. Pédauque.

La notion de document nous est tellement familière, tellement intuitive que nous ne ressentons pas le besoin de la préciser. En effet, chacun dans notre domaine professionnel ou privé, émettons et consultons quotidiennement un grand nombre de documents et la banalité de l'acte n'incline pas à la réflexion. Les vocables pour désigner l'objet document ne manquent pas et les variations sur le thème sont nombreuses : information, donnée, ressource, fichier, écrit, texte, image, papier, article, œuvre, livre, journal, feuille, page… Évidemment, chacun de ces termes se réfère à un contexte particulier ; mais ils se rattachent tous, à des degrés divers, à la notion de document, du latin « *documentum* » (du verbe « *docere* » : enseigner).

Il est traditionnel d'attribuer au document, notamment écrit, une double fonction ou valeur :
- de preuve (à la façon d'une « pièce à conviction » produite lors d'un procès) ;
- de renseignement (ou d'information).

Depuis une dizaine d'années, les progrès des TIC et l'irruption du numérique bousculent profondément la notion de document elle-même, sans que l'on puisse clairement en mesurer les conséquences au niveau de ses fonctions, faute d'en avoir, au préalable, cerné les contours.

Le contraste est grand entre la très longue stabilité du concept et la brutalité de ses évolutions récentes. Il n'y a pas de différence de nature entre une tablette sumérienne (4000 ans avant J.C.) et une banale lettre manuscrite. Support, inscription, texte, structure, mise en forme, message et communication : les composantes de base du document se retrouvent. En revanche, quand sont apparues les premières machines à écrire à mémoire, dont le relais a été rapidement pris par les ordinateurs lorsque ceux-ci ont investi massivement le domaine du traitement et la production de textes, un saut méthodologique a été franchi. Le document a pu exister et être conservé sur un autre support que le papier, dans une mémoire de masse. Deux conséquences majeures sont alors apparues :
- l'imbrication, la continuité entre le support traditionnel et l'inscription étaient rompues ;
- la lecture et l'écriture sur ce nouveau support exigeaient l'interface d'une machine.

Le phénomène a été doublement caractérisé :
- on a parlé de *numérisation* des documents, pour indiquer que l'inscription originelle en langue naturelle avait été codée et stockée en langage d'ordinateur (binaire), inaccessible à l'homme en l'état ;
- on a parlé de *dématérialisation* pour indiquer que cette inscription originelle cédait le pas à son image numérique et que, d'une certaine façon, la copie devenait l'original.

L'effacement progressif du support traditionnel conduit à une conception plus abstraite du document, dont le sens réside dans ses données et surtout dans sa structure logique.

Si l'on considère que tout document convoque nécessairement les mécanismes du langage, que se soit au moment de son écriture ou de sa lecture, il ne semble pas illégitime d'étudier les documents sous l'angle de la linguistique. Cette investigation scientifique du langage humain, à travers la diversité des langues[1], nous offre ses trois niveaux traditionnels d'analyse : la syntaxe[2], la sémantique[3] et la pragmatique[4].

Nous les reprendrons à notre compte pour articuler notre propos, avec les précautions d'usage, car comparaison n'est pas raison… Cependant, cette analogie nous fournit une tripartition relativement opérationnelle pour couvrir notre sujet.

<table>
<tr><td rowspan="3">SUPPORT</td><td>SYNTAXE</td><td>SÉMANTIQUE</td><td>PRAGMATIQUE</td></tr>
<tr><td>Inscription</td><td>Sens</td><td>Trace</td></tr>
<tr><td>Signes</td><td>Texte</td><td>Communication</td></tr>
<tr><td colspan="2">CONTENANT</td><td>CONTENU</td><td>MÉDIUM</td></tr>
<tr><td colspan="2">Document vu comme un contenant</td><td></td><td></td></tr>
<tr><td></td><td colspan="2">Document vu comme un contenu</td><td></td></tr>
<tr><td></td><td></td><td colspan="2">Document vu comme un médium</td></tr>
</table>

Organiquement, tout document est à la fois *contenant*, *contenu* et *médium*. Au cours du cycle de vie[5] du document, l'importance respective de chacune de ces fonctions est modulée selon les parties prenantes : producteurs, dépositaires, lecteurs au sens large.

I – Document vu comme contenant

Cette approche considère le document comme un objet matériel et en étudie la structure. Il est constitué d'un *support* manipulable sur lequel est fixée une *inscription* interprétable selon sa forme. L'inscription matérialise le *contenu* du document.
Les règles de forme, auxquelles obéissent les signes qui constituent l'inscription, appartiennent au « contrat de lecture », sorte protocole de communication implicite entre un producteur et un lecteur, au niveau de la lisibilité et de la perception du document.

A) Évolution

a) Un document traditionnel (non numérisé) peut être résumé par l'équation suivante :

DOCUMENT TRADITIONNEL = SUPPORT + INSCRIPTION

b) La numérisation du document va conduire en premier lieu à l'effacement du support :

- la notion de support devient ambiguë : s'agit-il du fichier, de l'unité logique, du disque physique, du « cache », ou bien encore de l'écran où le document est affiché ? ;
- pour les informaticiens, l'inscription relève du codage : démarche qui conduit à isoler les éléments logiques du document pour les modéliser, dans l'optique d'un traitement automatisé ultérieur. En définitive, le document ne serait qu'un cas particulier, un avatar d'une application informatique, dans laquelle :

APPLICATION INFORMATIQUE = INSTRUCTIONS + DONNÉES[6]

En symétrie, on obtiendrait :

DOCUMENT NUMÉRIQUE = STRUCTURE + DONNÉES

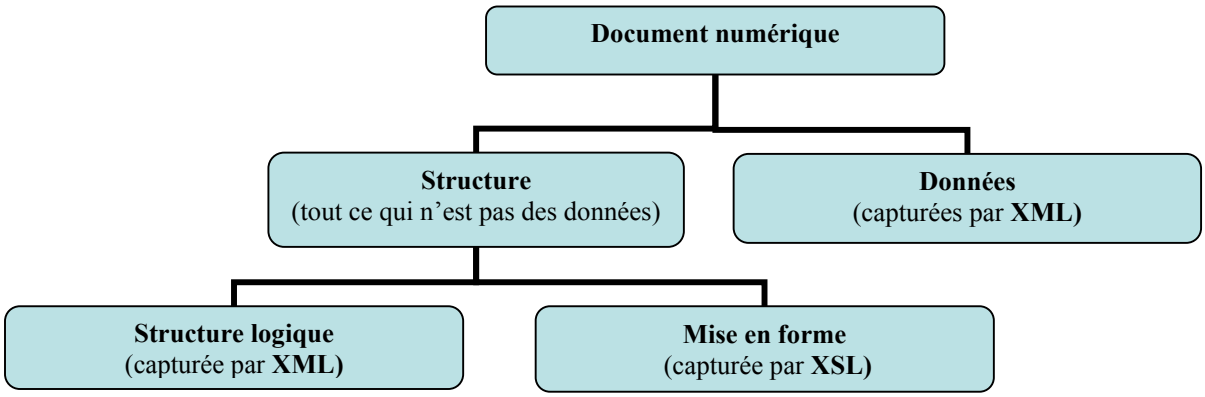
- la « structure » ainsi isolée comprend en fait tout ce qui ne constitue pas expressément des données, notamment :
  - la structure logique du document, en général décrite sous forme arborescente ;

  - le format binaire de sauvegarde, qui peut être « fermé » (propriétaire) ou « ouvert »[7] ;
  - le format de restitution, qui comprend les instructions de mise en forme.
- cette partition induit un nouveau contrat de lecture, puisque, nous l'avons vu, elle rend la structure du document modélisable, donc susceptible d'être traitée, puis « lue » par une machine ;
- le passage de l'analogique au numérique (« l'acquisition » du document) suppose évidemment la possibilité de restitution, à l'inverse, du numérique vers l'analogique. On obtiendra, dans ce cas, une production similaire, homologue, ou encore « homothétique » au document d'origine, mais non *identique* ; car il s'agira d'une traduction nouvelle, qui pourra occulter certains éléments signifiants (ou, au contraire, en faire découvrir de nouveaux…) ;
- La dématérialisation des documents a contraint les chercheurs à repérer et à isoler plusieurs niveaux de structuration, et notamment à opérer une distinction radicale entre la structure logique du document (son articulation en parties et sous parties), et sa représentation formelle (les « styles » : par exemple, pour les textes, les choix typographiques) ;
- la multiplication des formats « propriétaires » ont conduit à des situations d'illisibilité, difficilement tolérables pour l'utilisateur. En réaction, on assiste actuellement, dans le monde documentaire, à une uniformisation progressive autour de la norme XML.

c) Le format XML regroupe la structure logique et les données :

- XML capture uniquement la structure logique du document et ses données, qu'il intègre dans le même fichier. XML est l'héritier du SGML (langage utilisé dans l'informatisation des techniques éditoriales) et du HTML (langage courant utilisé sur le Web), tous deux étant des langages dits « à balises » ;
- Les balises utilisées en XML ont une portée strictement logique, structurelle (comme en SGML), c'est-à-dire qu'elles servent à repérer les différentes parties et sous parties du document. Ainsi le contenu d'un document au format XML est structuré grâce à des balises en forme arborescente. A l'opposé, la majorité des balises utilisées en HTML ne concerne que des attributs de mise en forme (polices, attributs, etc.) ;

On peut représenter cet état de fait de la façon suivante :



d) En lui-même, le format XML n'intègre aucune instruction de présentation :

- la présentation physique d'un document au format XML est définie séparément de la structure des données, dans des « feuilles de styles », à l'aide du XSL[8] ;
- ainsi, la forme peut être « dérivée » de différentes façons, mais n'est pas représentée intrinsèquement ;
- XML marginalise la forme et la délaisse comme dimension primordiale du document ;

- XML bouscule l'ancien contrat de lecture, où le lien entre la représentation perçue et la structure logique était pérennisée par le support.

A ce stade, on obtient :

DOCUMENT XML = DONNÉES STRUCTURÉES + MISE EN FORME

e) Prospective des documents XML :

- les documents ainsi rédigés en XML pourraient rejoindre des bases de données, centralisées ou distribuées, pour constituer, en quelque sorte, de vastes jeux de « Lego »®, à l'aide de « briques » de formes et d'usages très variés, mises à la disposition de tous ;
- un document n'aurait de forme à proprement parler qu'à deux moments : celui de sa conception par son auteur et celui de sa « reconstruction » par un lecteur. Comme nous l'avons vu, il est peu probable que le document soit identique dans l'un et l'autre cas ;
- une autre façon de concevoir cette évolution serait de considérer que le document constitue dorénavant la base de données elle-même, dont les différentes sorties (dérivations) ne seraient que des interprétations partielles de la richesse.

B) Définition proposée :

« Un document numérique est un ensemble de données organisé autour d'une structure stable, associé à des règles de mise en forme permettant une lisibilité partagée entre son concepteur et ses lecteurs ».

C) Questions posées :

- Il y a du sens dans la forme : l'indentation des paragraphes, les polices de caractères, les attributs de rehaussement, la ponctuations sont des éléments qui participent à l'intelligence d'un document (la forme est au service du fond) ;
- La forme joue un rôle indiscutable dans la cognition : l'apprentissage met en jeu la « mémoire visuelle ». De même, les chercheurs ont noté l'importance cognitive des cheminements possibles dans un ensemble de liens hypertextes. D'ailleurs, un travail sur la « lecture électronique » doit être mené pour mieux comprendre l'interdépendance entre fond et forme ; si l'on estime que la stabilité visuelle du papier, sa maniabilité joue un rôle important dans la cognition, ne faudrait-il pas encourager les efforts déployés en direction des « codex électroniques » (*e-books*) ?
- Quel serait l'impact de ces nouveaux régimes de lecture sur nos régimes de savoir ?
- Il devient impossible d'authentifier le fond d'un document par l'analyse de sa forme. L'authentification devra être assurée par d'autres moyens (filigrane électronique, tiers archiveurs certifiés) ;
- Les différents terminaux de restitution n'ayant pas les mêmes performances, la perception du document dépendra du terminal utilisé. Ainsi, faudra-t-il conserver, pour un document XML donné, la totalité des matériels et des systèmes de lecture qui permettent (ou qui ont permis) d'y accéder ?
- La gestion des temporalités successives d'un document, de son écriture par son auteur à son enrichissement, ou son remaniement par des intervenants variés devront être assurées. Or, la gestion des différentes versions (*versioning*) d'un document est déjà délicate au niveau d'un seul individu ; elle se complexifie au niveau des organisations, et devient carrément inextricable au niveau du Web…
- Il y aura nécessité d'inventer les procédures qui permettront à la fois d'attribuer avec certitude un document à un auteur (ou à un groupe d'auteurs), et à chacun de s'en approprier tout ou partie, tout en limitant la prolifération « bruyante » d'une même ressource…

En résumé, l'élaboration d'un document peut-elle se détacher de sa forme perceptible ? Est-il concevable d'envisager une rupture formelle entre l'élaboration par l'auteur et la proposition faites aux lecteurs. On s'interroge notamment sur le succès des formats électroniques de « fac-similé », comme le .PDF d'Adobe. En un mot, en effaçant le support, n'a-t-on pas trop délaissé la forme ?

II – Document vu comme un contenu (texte au sens large)

Dans cette seconde approche, le document est pris comme un *objet signifiant*. Si la *forme* est prise en compte, elle ne l'est que comme porteuse de sens. Le *support* est également accessoire, y compris pour le document traditionnel, du moment que le contenu, matérialisé par l'inscription, est préservé. Le *sens* d'un document se construit aussi par rapport au *contexte* de production et de diffusion et conditionne son interprétation. La construction du sens sera parachevée par l'inclusion du document dans un système social de *classement*. Le sens quitte le territoire de *l'information* pour gagner celui de la *connaissance*[9]. Le contrat de lecture est appréhendé ici sous l'angle de l'assimilation, de l'appropriation du sens du document.

A) Évolution

Au départ, on admet que :

DOCUMENT TRADITIONNEL = INSCRIPTION + SENS

a) Du texte brut au texte « informé »

Pour reprendre la célèbre théorie de Claude Shannon dans laquelle le « bruit » d'un système de communication est représenté par l'entropie, c'est-à-dire l'état d'ignorance, d'incertitude dans laquelle se trouve placé le récepteur (pour nous, le lecteur de document) par rapport à l'état du système en question ; son inverse (la « néguentropie ») constituerait l'information proprement dite. Dans sa « Théorie mathématique de la communication », l'ambition de Shannon était uniquement quantitative : il s'agissait, pour lui, d'optimiser le processus afin de réduire le bruit et d'augmenter corrélativement la quantité d'informations transmises. Il n'était pas tenu compte du sens du message : la quantité d'informations était assimilée à la quantité de signes émis et correctement reçus.

Le propos est plutôt de bâtir une méthode qui permette de repérer et d'évaluer la « quantité de sens » contenue dans un texte :

- Jean Pierre Balpe[10], de l'Université de Paris VIII, a développé une théorie de la « valence informationnelle » dans le cadre des hypertextes. Si l'on admet qu'un document hypertexte constitue une « unité d'information », sa valence informationnelle serait fonction du nombre de concepts auxquels il fait référence. Plus l'unité d'information contient de concepts donnant accès à d'autres unités d'information, plus la valence informationnelle de l'unité de départ est importante ;
- Mais, lorsque l'on s'attache à l'évaluation de la véritable substance informationnelle d'un texte, il faut tenir compte des redondances et des éléments discursifs considérés comme sans valeur. En clair, un concept peut être évoqué dans un texte, sans qu'il soit fait un apport informationnel particulier (cas de la simple citation). Les outils automatiques[11] dans ce domaine sont donc dotés de fonctions d'évaluation plus ou moins sophistiquées : du simple calcul de fréquence lexicale d'apparition à l'indentification de fragments de phrases porteurs de sens (dits fragments de phrases indicateurs), balisés par ce que l'on appelle des « marqueurs linguistiques d'extraction » (MLE)[12] ;
- Ainsi, le texte d'un document est dit « informé » lorsqu'il a subi un traitement (manuel, mais aujourd'hui de plus en plus automatisé) qui a permis d'extraire les

différents concepts qu'il contient, et de répertorier les différentes unités d'information auxquelles il renvoie. Les sciences de l'information s'attachent à comprendre comment ces unités se hiérarchisent et s'emboîtent. Ce traitement donne lieu à la production d'importants paratextes, comme par exemple les méta données[13] du *Dublin Core*[14] (titre, domaine, auteur, réseau de mots-clés, résumé, note de synthèse, etc.). La valeur informationnelle du document ainsi décrit est rendue accessible et évaluable sans accéder à son contenu.

b) la construction du sens : classement et contextualisation

« Penser, c'est classer ». En réalisant des documents, nous isolons et nous rangeons des discours pour nous aider à penser le monde et à traduire notre compréhension sociale. Cette construction se réalise en amont par la « mise en document » et en aval, par la « mise en collection ». Ces deux opérations sont porteuses d'une intentionnalité sociale forte[15].

La mise en collection de documents suppose l'adoption d'un ou plusieurs systèmes de classement, qui peuvent varier énormément en fonction des situations ou des époques (parfois très formalisés, ils peuvent être également implicites). Un système de classement permet de ranger un document dans un ensemble, mais également de l'y retrouver. Il repose nécessairement sur un système d'indexation[16].

L'inclusion d'un document dans un système de classement va permettre son interprétation. En effet, un document ne fait sens que s'il est lu par un lecteur[17] (individu, groupe, machine). En un mot, le système de classement fournit au lecteur le contexte d'interprétation du document, élément essentiel de son exploitation. Il permet notamment de repérer les liens que ce dernier instaure ou suggère avec les autres unités documentaires.

Puisque le lecteur re-crée en quelque sorte le document chaque fois qu'il l'isole et en prend connaissance, le contrat de lecture exige qu'il soit placé dans son contexte d'interprétation.

c) l'irruption du numérique dans les systèmes documentaires

Dès le début du 19ème siècle, avec ce qu'il est convenu d'appeler l'explosion documentaire, due à l'industrialisation, se sont développés des langages documentaires. L'élaboration de ces langages a soulevé de nombreux problèmes, dont celui de la normalisation des vocabulaires, en fonction des domaines concernés.

- Historiquement, ce sont les bibliothécaires qui ont utilisé les premiers l'outil informatique avec la constitution de bases de données bibliographiques, dont les traitements étaient limités aux méta données attachées aux documents répertoriés ;
- Progressivement, le traitement des textes documentaires a pu être automatisé, avec le développement parallèle des sciences de l'information, soit au niveau de l'indexation, soit au niveau des systèmes de mot-clés et d'abstracts, soit enfin au niveau des interfaces d'interrogation, pour lesquelles linguistes et informaticiens ont conjugué leurs compétences ;
- Les plus grandes difficultés sont apparues avec les outils de traitement automatique de la langue, dès lors que l'on s'attaquait à l'acquisition et au traitement du texte intégral. Les meilleurs outils ont du intégrer une part importante de travail humain et se présentent plus comme des outils d'aide, que comme des outils strictement autonomes.

d) Le développement des moteurs de recherche sur le Web[18]

Même si les résultats obtenus par ces outils sont de nature à satisfaire la majorité des internautes, l'indexation (y compris l'indexation « plein texte ») des documents collectés

s'effectue toujours en aveugle, sur des entités purement lexicales (*tokens*) et non sur des unités sémantiques (les concepts), excluant ainsi tout traitement sémantique. En conséquence, ces systèmes souffrent de nombreuses limitations, dont :

- ils sont incapables de gérer correctement les formes fléchies des substantifs, des adjectifs, des verbes[19];
- ils sont aveugles aux phénomènes d'homonymie, de polysémie et de synonymie ;
- tous les termes présents dans un document ont le même statut. En l'absence d'information sur la sémantique de tel ou tel mot, la recherche proposée aux utilisateurs ne peut être que purement lexicale, ce qui génère inévitablement un taux de bruit important (pourcentage de documents non pertinents trouvés en réponse à une requête).

Deux voies s'ouvrent pour améliorer les possibilités de recherche d'information sur le Web : la « lemmatisation[20] », issue de l'ingénierie linguistique ; la seconde consiste à structurer le Web pour rendre explicites les relations sémantiques entre les unités informationnelles que contiennent les documents. C'est dans cette seconde approche que s'inscrivent XML et RDF[21] (*Resource Definition Framework*). Ainsi, à terme, se constituerait un véritable « Web sémantique ».

e) La gestion du contenu au niveau d'Internet : le Web sémantique (WS)

L'annonce du Web sémantique peut être resituée dans la continuité des efforts de structuration des documents déployés dans le cadre d'XML, débouchant sur une formalisation de plus en plus poussée de la structure des documents, que ce soit avec les DTD (*Document Type Definition*)[22], et plus récemment, depuis une recommandation de mai 2001 du W3C[23], les schémas XML (*XML Schema*)[24]. L'avènement du Web sémantique induira certainement une avancée dans la modélisation des connaissances, car des outils comme RDF permettent de représenter les connaissances grâce à des schémas appropriés, en offrant à la fois des modèles et un formalisme adapté.

Avec ces outils, on entre de plain pied dans le domaine du traitement du sens des contenus : l'objectif est de passer d'un Web actuellement constitué d'un ensemble de fichiers reliés entre eux à un réseau utilisant pleinement les capacités de calculs[25] des machines interconnectées, pour un véritable traitement sémantique des documents. A leur manière, les promoteurs du Web sémantique construisent des sortes de langages documentaires perfectionnés, qu'ils ont baptisé « ontologies »[26]. Leur rencontre avec les chercheurs de l'ingénierie des connaissances, qui travaillent sur la modélisation du raisonnement, était inévitable. Les ontologies se focalisent sur l'essence d'un domaine (médecine, par exemple), sur son vocabulaire et, au-delà, sur le sens dont il est porteur avec la constitution d'un réseau conceptuel. Le sens, ainsi formalisé, présente toujours deux facettes : celui compris par l'être humain (sémantique interprétative) et celui « compris » par les machines (sémantique formelle). Les ontologies peuvent être vues comme une structuration plus riche que les lexiques ou les thésaurus utilisés jusqu'ici, car leur sémantique formelle va permettre leur utilisation dans le cadre d'applications informatiques, là où les thésaurus étaient en échec.

Ainsi, on obtient :

DOCUMENT WS = TEXTE INFORMÉ + ONTOLOGIES

B) Définition proposée :

« Un document numérique est un texte dont les éléments sont potentiellement analysables par un système de connaissance, en vue de son exploitation par un lecteur compétent ».

C) Questions posées :

- Quid de l'application de ces outils aux langues dont la structure et l'écriture s'écartent des schémas indo-européens ?
- Ces nouveaux outils, pour séduisants qu'ils sont, n'établissent pas une frontière nette entre ce qui relève du travail intellectuel humain et ce qui relève de la machine ;
- Dans cette approche, le document apparaît comme secondaire, seuls le texte et son contenu comptent vraiment. Pourtant, la mise en document, comme nous le verrons dans la troisième partie, est l'un des éléments essentiels de la construction du contexte. La valeur sémantique de la mise en document n'est-elle pas sous-estimée ?
- Dans de nombreux cas, il est impossible de valider une information autrement que par l'authenticité du document qui la contient ;
- Comment réagissent les systèmes automatiques d'extraction de connaissances lorsqu'ils sont confrontés à des sources d'informations contradictoires ? Comment résolvent-ils le problème de l'accréditation des sources ?
- Dans quelle mesure peut-on isoler un élément signifiant de l'ensemble qui le contient et qui constitue une unité de sens ? Car un document forme un tout…

III– Document vu comme un médium

Le terme *médium* doit être pris au sens large. Il regroupe les approches qui analysent le document comme un *phénomène social*, comme *vecteur de message* entre individus, ou groupes, qui participe de la troisième dimension du contrat de lecture : celle de la sociabilité.

A) Évolution

Un document donne un statut, confère une légitimité à une information, car il est porté par le groupe social qui le suscite. Il constitue simultanément une *preuve* qui fait foi d'un état des choses ; une *annonce* qui prévient d'un événement ; un *discours* qui se rattache à un auteur ; un *témoignage*, une pièce de dossier.

Ainsi :

DOCUMENT TRADITIONNEL = TEXTE + LEGITIMITÉ

a) La mise en document est un processus social :

- la diffusion d'un texte doit dépasser la communication intime (la sphère privée) pour accéder à la légitimité. Ainsi, un journal intime n'est pas un document, sauf si quelqu'un prend l'initiative de le rendre public ;
- la présence même de texte dans un document suppose nécessairement une opération préalable d'enregistrement, lui ayant permis de s'affranchir de l'éphémère. Ainsi, une émission de télévision ou de radio en direct n'est pas un document, sauf si quelqu'un l'enregistre pour une utilisation sociale future ;
- le statut de document n'est pas acquis pour l'éternité, il se donne, mais peut se perdre dans l'oubli collectif, et se retrouver à un moment de l'Histoire. C'est le cas de la « découverte », donc de la relégitimisation d'un document disparu de la mémoire collective, mais non détruit.

b) Les fonctions sociales des documents :

Ils pérennisent les normes des connaissances nécessaires à leur développement et à leur survie. Certains documents, dont l'utilité sociale est capitale, ne sont cependant pas publiés et font l'objet d'une diffusion restreinte pour des raisons de confidentialité : dossiers médicaux, secrets scientifiques, diplomatiques ou militaires.

c) La multiplication des documents obéit à deux dynamiques régulatrices :

- dynamique externe (qui assure la fonction sociale des documents). Les organisations politiques et sociales s'appuient sur la production de documents (exemples : religions et clercs, états et administrations, entreprises). La mise en document peut être analysée comme un acte d'administration, avec, d'un côté, un ou plusieurs émetteurs détenteurs de pouvoir, et, de l'autre, les destinataires qui leur sont assujettis. Ainsi, en assurant une communication contrôlée, les documents participent à la régulation des sociétés humaines. La promulgation des lois dans sphère juridique en est l'exemple le plus évident. Dans la sphère commerciale, les documents de communication externe émis par les entreprises et autres acteurs économiques permettent à ces derniers d'affirmer leur identité concurrentielle. Cette dynamique externe amène les chercheurs à s'interroger sur la nature de la communication propre aux médias[27], ainsi qu'aux processus de publication, comme par exemple la traditionnelle gradation des étapes de la communication scientifique (écritures d'articles, révisions par les pairs, citations, pré-publications, etc.) ;
- dynamique interne (qui répond à l'économie propre des organisations, à leur besoin de régulation interne). La communication organisationnelle étudie les documents comme étapes de processus de travail, immergées dans des pratiques et des situations professionnelles caractérisées. Elle est nécessairement amenée à explorer et à critiquer les processus de fabrication de documents, en voie de normalisation grâce à des outils comme le BPM[28] (*Business Process Management*), le BPR[29] (*Business Process Reengineering*), ou encore le RM[30] (*Records Management*), …

d) L'impact du numérique dans ce domaine donne lieu à des mouvements contradictoires :

- numérisation (et donc effacement) d'un nombre important de documents papier qui rendent compte de procédures, comme les registres d'Etat Civil pour l'Administration ; formulaires, tableaux, fiches et modes d'emploi dans le domaine des entreprises, qui sont maintenant stockés dans des bases de données et dont la transmission est assurée par EDI (échange de données informatisé). Le fait de basculer des documents, ou des procédures, en BDD n'affaiblit nullement leur valeur prescriptive : les Intranets, par exemple, allouent aux documents numériques qu'ils manipulent un statut de référence et d'outil, en y associant des règles d'identification et de circulation. Ce type d'applications amplifie la visibilité des activités, des décisions en les rendant accessibles, sinon transparentes ;
- montée massive d'une mise en écrit, et donc d'une mise en document, entraînée notamment par la démarche qualité : édition de manuels qualité, de référentiels[31].

Schématiquement, on obtient :

DOCUMENT NUMÉRIQUE = TEXTE + PROCÉDURES

e) En ce qui concerne l'impact du Web sur les documents :

- plusieurs théories rendent compte de cet impact. La plus productive semble être celle du filtrage à posteriori des sources : les documents les plus pertinents seraient progressivement repérés et mis en valeur par le nombre de liens les concernant, notamment ceux des moteurs de recherche. Il s'agirait d'un véritable système de « percolation », accélérant la dynamique classique de légitimation par la notoriété : plus l'existence d'un document est connue, plus il sera lu, et plus il sera lu, plus son existence sera connue ;
- certains chercheurs ont vu dans l'avènement du Web un effacement, une disparition du document derrière la « richesse » et la « complexité »[32] que peuvent engendrer les

techniques d'hypertextualité et d'hypermédia. D'intéressantes expériences d'écriture hypertextuelle ont mis en évidence des apports sémantiques et cognitifs non négligeables ;
- à l'inverse, d'autres chercheurs ont remarqué que le développement explosif du Web a conduit à une multiplication exponentielle du nombre des documents. Les liens entre pages paraissent se structurer pour construire de nouvelles normes de paratexte, renforçant l'aspect documentaire du Web ;
- en ce qui concerne le Web sémantique, ses conséquences supposées appartiennent encore à la prospective…

Ainsi :

DOCUMENT WEB = PUBLICATION + ASPECT REPÉRÉ

B) Définition proposée :

« Un document numérique est une trace de relations sociales reconstruite par des dispositifs informatiques ».

C) Questions posées :
- Le droit de propriété intellectuelle de tradition latine, qui privilégie l'attachement de l'auteur à son œuvre évoluera (et évolue déjà) vers le *copyright* anglo-saxon, qui met en avant la notion de publication, conférant la propriété intellectuelle à celui qui en prend l'initiative. Sous cet aspect, le droit d'auteur est un droit de l'œuvre, alors que le *copyright* s'avère être plutôt un droit du document ;
- La généralisation d'outils comme le *Records Managment* entretient l'ambiguïté entre archivage et publication : leur objet est-il de témoigner d'actions passées ou bien, au contraire, d'enregistrer des actions en cours ?
- En ce qui concerne la pratique des lecteurs-internautes, les lois bibliométriques jouent à plein : comme nous l'avons vu, leur attention se concentre fortement sur un nombre réduit de documents, et au contraire, se disperse sur un très grand nombre, à la manière d'une loi des « 20/80 »… Cette tendance peut représenter à terme un facteur d'appauvrissement du contenu, du fait des risques de résonance[33] qu'elle entraîne ;
- On peut noter un oubli de taille : celui du financement du contenu. Pendant combien de temps le Web pourra-t-il reposer sur l'idéologie du savoir et de la culture gratuite ? Il est clair que, sans financement explicite, la richesse du contenu du Web est condamnée à décliner…

Conclusion

Le document traditionnel repose sur un support, un texte et une légitimité.

Le document a été construit comme un objet, dont la concrétisation la plus banale était la feuille de papier. Le numérique a fait émergé une notion différente de ce que nous appelons document. Principalement la perte de stabilité en tant qu'objet matériel et sa transposition en un processus construit à la demande ont ébranlé la confiance qu'on pouvait mettre en lui.

Une première phase de numérisation a fait ressortir l'importance de la structure interne du document, des métas donnés et des difficultés de validation.

La seconde phase, dominée par la norme XML, intègre la structure et les données, mais pas la forme. Elle s'appuie sur des ontologies pour retrouver et reconstruire des textes, mettant en avant l'accès personnalisé.

Cependant, l'opposition papier/numérique est vaine, car tous les documents actuels ont été convertis au format numérique, ou le seront un jour. Ceux qui échapperaient à cette règle risquent fort de tomber dans l'oubli. Inversement, de très nombreux documents numériques,

seront, à un moment ou à un autre, imprimés. Ainsi, le numérique est à la fois un révélateur et un facteur d'évolution de la notion de document.
La notion de contrat de lecture a également largement évolué, mais s'articule toujours autour de trois concepts : lisibilité, compréhension et sociabilité. Ainsi, un document ne serait, en définitive, qu'un contrat entre des hommes dont les qualités *anthropologiques* (lisibilité-perception), *intellectuelles* (compréhension-assimilation) et *sociales* (sociabilité-intégration) fonderaient un part de leur humanité, de leur capacité à vivre ensemble.

--Cet article a été rédigé par Alain Nossereau, professeur d'Economie Gestion au Lycée Beaussier de La Seyne sur Mer, formateur en communication et gestion de l'information Académie de Nice.

Mots clés
Dématérialisation, document, classement, contexte, contrat de lecture, méta donnée, méta langage, moteur de recherche, numérisation, ontologie, support, thésaurus, web sémantique, XML, XSL.

Pour aller plus loin
XML.
http://www.educnet.education.fr/dossier/xml/default.htm

Méta-donnée.
http://www.educnet.education.fr/dossier/metadata/default.htm

Hypermédia et l'apprentissage.
http://www.educnet.education.fr/dossier/hypermedia/default.htm

Manuel numérique.
http://www.educnet.education.fr/dossier/hypermedia/default.htm

Livre numérique, livre électronique.
http://www.educnet.education.fr/dossier/eformation/default.htm

E-formation.
http://www.educnet.education.fr/dossier/eformation/default.htm

Séminaire « Numérique et manuels scolaires & universitaires » Abbaye de Fontevraud - 29 et 30 septembre 2004.
http://www.educnet.education.fr/documentation/manuel/alternatifs.htm

Bibliothèques et livres numériques.
Jean-Michel Salaün. Article, Lectures (Bruxelles) (123):34-38. 01 novembre 2001. Publié ou en cours de publication.
http://archivesic.ccsd.cnrs.fr/sic_00000062.html

Chronique inachevée d'une réflexion collective sur le document.
Jean-Michel Salaün. Article, Communication et Langages (140). 01 juin 2004. Publié ou en cours de publication.
http://archivesic.ccsd.cnrs.fr/sic_00001028.html

Documents et numérique.
Jean-Michel Salaün. Article, Contribution au rapport du Conseil d'Analyse Economique sur la Société de l'information. 01 décembre 2003. *Working paper*.
http://archivesic.ccsd.cnrs.fr/sic_00000831.html

Introduction : un dialogue pluridisciplinaire pour penser le « document numérique ».
Jean-Michel Salaün et Jean Charlet. Article, Revue i3. 4(1):7-18. 05 juillet 2004. Publié ou en cours de publication.
http://archivesic.ccsd.cnrs.fr/sic_00001012.html

Le document numérique: un objet fédérateur en sciences de l'information.
Ghislaine Chartron, Brigitte Guyot, Thierry Lafouge, Sylvie Lainé-Cruzel, Genevieve Lallich-Boidin, Marie-France Peyrelong et Jean-Michel Salaün. Article, Documentaliste Sciences de l'Information. 39(6):298-305. 03 décembre 2002. Publié ou en cours de publication.
http://archivesic.ccsd.cnrs.fr/sic_00000417.html

Le document numérique comme « lego »® ou La dialectique peut-elle casser des briques ?
Dominique Cotte et Marie Després-Lonnet. Article, Revue I3. 4(1):159-172. 05 juillet 2004. Publié ou en cours de publication.
http://archivesic.ccsd.cnrs.fr/sic_00001021.html

Réinterroger les structures documentaires : de la numérisation à l'informatisation.
Stéphane Crozat et Bruno Bachimont. Article, Revue I3. 4(1):59-74. 05 juillet 2004. Publié ou en cours de publication.
http://archivesic.ccsd.cnrs.fr/sic_00001015.html

Documents, ressources, données : les avatars de l'information numérique.
Sylvie Lainé-Cruzel. Article, Revue I3. 4(1):105-120. 05 juillet 2004. Publié ou en cours de publication.
http://archivesic.ccsd.cnrs.fr/sic_00001018.html

Effets de la numérisation et de la mise en réseau sur le concept de document.
Sylvie Leleu-Merviel. Article, Revue i3. 4(1):121-140. 05 juillet 2004. Publié ou en cours de publication.
http://archivesic.ccsd.cnrs.fr/sic_00001019.html

Réflexions sur la modélisation des documents.
Veronika Lux-Pogodalla et Jean-Yves Vion-Dury. Article, Revue i3. 4(1):19-40. 05 juillet 2004. Publié ou en cours de publication.
http://archivesic.ccsd.cnrs.fr/sic_00001013.html

---

[1] 2796 langues parlées, environ 350 langues écrites.

[2] Partie de la grammaire qui étudie les règles régissant les relations entre les mots ou les syntagmes (groupes de mots qui forment une unité fonctionnelle) à l'intérieur d'une phrase. (Source : Dictionnaire Hachette).

[3] Etude du langage du point de vue du sens. (Source : Dictionnaire Hachette).

[4] Usages et conventions implicites dans une langue, qui permettent de l'interpréter. Ex : « Ils vont encore nous augmenter les impôts ». « Ils » désigne forcément le gouvernement, car seul le gouvernement a le pouvoir d'augmenter ou de baisser les impôts… (Source : Jacques Moeschler - Université de Genève)

[5] Les étapes constitutives du cycle de vie d'un document sont respectivement : la création, la modification, la consultation, la transmission, la conservation, l'archivage ou la destruction.

[6] Respectant, en cela, la structure définie par le mathématicien John Von Neumann, dans laquelle le programme et les données sont stockés sur le même support : la mémoire centrale de l'ordinateur.

[7] Voir l'article du WikiPédia (L'Encyclopédie libre) : http://fr.wikipedia.org/wiki/Format_ouvert

[8] (*eXtended Stylesheet Language*). Standard du *World Wide Consortium* basé sur XML, permettant de construire des feuilles de style. L'association d'une feuille de style XSL à un document XML normalise la présentation des données. (Source : http://www.laboratoire-microsoft.org/def/3217/)

[9] Ceci amène à s'interroger sur la nature du schéma cognitif classique : données->informations->connaissances. Il s'agit probablement d'un « continuum » :

Donnée : … --- …
Information : S O S
Connaissance : « *En cas d'alerte : déclencher les secours* ». La connaissance permet de produire de nouvelles données, informations et connaissances : inférence
(Source : Olivier Corby, INRIA)

[10] Voir à ce sujet Y. Claeyssen, intervenant à l'Université de Lille 3 : http://home.nordnet.fr/~yclaeyssen/d13.html

[11] Agents dits intelligents, de différentes catégories : les "fouineurs", les "récupérateurs", les "synthétiseurs", les "résumeurs", les "filtreurs" (exemple : pour les messageries électroniques), les "veilleurs" (surveillance des pages HTML et des sites). (Source : http://www.decisionnel.net/veille/index.htm).

[12] Chaque phrase se voit attribuer une pondération particulière définie comme la somme des poids des marqueurs linguistiques qu'elle contient. (Voir en ce qui concerne les logiciels « résumeurs », comme le module de synthèse automatique de MS Word : http://www.ofil.refer.org/tribune/n3536/EvaluationResumeurs.htm)

[13] Une méta donnée est littéralement une donnée sur une donnée. (Source : Educnet http://www.educnet.education.fr/dossier/metadata/quoi1.htm)

[14] La norme de méta données du *Dublin Core* propose un ensemble d'éléments pour décrire une grande variété de ressources en réseau. Elle comprend 15 champs standard dont la sémantique a été établie par un consensus international de professionnels provenant de diverses disciplines telles que la bibliothéconomie, l'informatique, le balisage de textes, la communauté muséologique et d'autres domaines connexes… Grâce à cette norme les moteurs de recherche du Web seront en mesure de « comprendre » les différentes sémantiques et de les utiliser de façons différenciées. Ils feront la différence, par exemple, entre un mot-clé auteur et un mot-clé purement descriptif, et ne confondront plus ainsi un article écrit par "Pierre Dupont" avec un article qui parle, entre autres, de "Pierre Dupont". (Source : http://www.mutu-xml.org/xml-base/shared/KEY-DUBLINCORE.html)

[15] Selon un exemple classique, une antilope en liberté dans la savane ne constitue pas un document. Mais le même animal, capturé et exposé dans un zoo, dans le carré des animaux africains par exemple, entre dans un système social de classement et se trouve ainsi « documenté ».

[16] Littéralement : établissement d'un lien (pointeur, en informatique) entre une « étiquette » (identifiant) et un objet. Ce lien est réciproque : la création de l'étiquette permet de ranger l'objet, la consultation de l'étiquette permet de le retrouver.

[17] Un texte n'accède au statut de document que s'il a fait l'objet d'une diffusion minimum, hors de la sphère privée : un journal intime n'est pas un document, comme on le verra par la suite.

[18] Cet alinéa est largement inspiré de l'ouvrage d'Alain Michard (INRIA) : « XML, langage et applications ». Éditions Eyrolles. 2001. (p. 313 et suivantes)

[19] Cette carence entraîne des bruits importants en cas de recherche avec troncature : « cloche* » donnera « cloche » et « cloches », mais aussi « clocher » et « clocheton ».

[20] Dans un dictionnaire, action de donner à un ensemble de mots une adresse lexicale unique (canonique) : le lemme, qui permet de les réunir et éventuellement de les trier. Exemple : l'infinitif pour un verbe.

[21] Métalangage spécifié par le W3C qui consiste à coder la sémantique de documents web de manière à obtenir une meilleure pertinence des résultats fournis par les moteurs de recherche. RDF est rédigé suivant la syntaxe de XML. (Source : http://www.laboratoire-microsoft.org/def/3048/)

[22] Une DTD est, comme son nom l'indique, une définition d'une structure type de document, prédéfinie et généralement stockée dans un fichier extérieur. Tout nouveau document XML, faisant référence explicite à une DTD, en respectera obligatoirement la structure, et sera déclaré « valide » en terminologie XML. Ainsi, l'ensemble des documents ayant adopté la même DTD hériteront de la même structure et formeront une « famille » homogène, même si, évidemment, ils varient par leurs données.

[23] Le *World Wide Web Consortium*, dirigé par Tim Berners-Lee, également créateur de l'HTML.

[24] L'utilisation des DTD ne posait pas de problèmes dans le monde documentaire. En revanche, lorsque la norme XML s'est développée et que des échanges de données entre applications informatiques se sont directement réalisés en format XML, la technique des DTD s'est révélée insuffisante : performante pour décrire précisément des structure arborescentes, elle ne permet pas de définir le type des éléments manipulés par ces structures. Par exemple une date y sera toujours enregistrée comme une chaîne de caractères, dont la forme pourra considérablement varier, en fonction de ce qu'aura saisi l'opérateur : Exemple « 31 mai 2005 » ou « 31/05/2005 » ou encore « 31.05.05 ». L'application informatique réceptrice sera incapable d'interpréter cette chaîne comme une date et d'y appliquer le traitement prévu. Une erreur sera générée, mettant en péril le déroulement complet de l'application. Les schémas XML résolvent ces problèmes, car ils permettent de définir des types d'éléments. Par exemple, ici, le schéma XML précisera que cette chaîne est bien une date au format « JJ/MM/AAAA ».

[25] De fait, la navigation sur Internet par elle-même n'exige que peu de ressources machine…

[26] Voir, dans l'excellent site de Karl Dubost, non dépourvu d'humour, la hiérarchie suggérée entre les différents moyens d'indexation : le vocabulaire contrôlé, la taxonomie, le thésaurus et l'ontologie : http://www.la-grange.net/2004/03/19.html et http://www.la-grange.net/2003/04/28.html

[27] Telle qu'on pu l'étudier Dominique Wolton, directeur de recherche au CNRS et Pierre Bourdieu, philosophe, décédé en 2002. Une polémique a opposé ces deux auteurs sur ce sujet.

[28] Terme générique désignant à la fois la modélisation et la traduction, à l'aide d'un "moteur", des processus modélisés dans la réalité de l'entreprise. S'y ajoutent des outils (indicateurs, tableaux de bord…) de suivi de l'exécution de ces processus et d'évaluation de leurs performances.
(Source : http://www.alaide.com/dico.php?q=BPM&ix=3455)

[29] Démarche de remise en question et de redéfinition en profondeur des processus d'une organisation en vue de la restructurer pour la rendre plus efficace tout en réduisant les coûts. Cette réorganisation des méthodes de travail constitue souvent la première phase d'un projet d'informatisation: on commence par rationaliser une activité de l'entreprise (la prise en compte d'une commande d'un client) afin de bien cerner tous les cas de figure et de pouvoir déclencher des actions adéquates de manière automatique et sans ambiguïté. (Source : Journal du Net)

[30] Le *Records Management* gère les « records », c'est-à-dire, selon des cultures d'entreprise et d'administration : les dossiers vivants, les archives courantes et intermédiaires, les documents internes de référence.
Il prend en compte les exigences légales de conservation, la responsabilité et la traçabilité, les besoins d'information, l'efficacité des systèmes et des procédures, le "rapport qualité/prix" de l'information archivée. Le Records management se situe au carrefour de plusieurs compétences : documentation, archivistique, contrôle qualité, nouvelles technologies et droit. (Source : http://www.archive17.fr/RM.html)

[31] Référentiel : Liste d'une série d'actes, de performances observables détaillant un ensemble de capacités : référentiel de formation, ou de compétences : référentiel de métier. (Source : Norme AFNOR).

[32] Voir « Hypertexte et complexité » de Jean Clément, Maître de Conférence au département hypermédia de l'Université de Paris VIII : http://hypermedia.univ-paris8.fr/jean/articles/clement.pdf

[33] Phénomène qui peut affecter un système physique dynamique comme une balançoire, dont les oscillations s'amplifient, même sous l'effet d'impulsions modérées, lorsqu'elles sont appliquées au bon moment. Sens figuré, ici : la notoriété acquise sur Internet peut présenter un caractère artificiel.