Manuscrit auteur, publié dans "N/P"

sic_0001003, version 1 - 15 Jun 2004

Document et texte
Permanence et transformations

Roger T. Pédauque
Document de travail soumis à la discussion[1]
Version du 15-06-2004

1. Le contexte de la réflexion

2. Il est apparu rapidement au sein du Réseau thématique pluridisciplinaire « Document et contenus : création, indexation, navigation »[2] du département STIC du CNRS qu'une réflexion globale et approfondie sur la notion de document était nécessaire. Un premier texte collectif a été rédigé, signé d'un pseudonyme représentant l'ensemble du collectif : Roger T. Pédauque[3]. Son objectif était de repérer les principales problématiques qui se posaient autour du document, confronté au développement explosif du numérique. Ce premier balisage a fait ressortir la richesse des travaux en cours et l'importance du chemin déjà parcouru, mais il a aussi pointé bien des questions encore en suspens. Il a été depuis à l'origine de débats et publications.
3. Pour avancer dans la démarche collective du réseau, il paraît utile d'approfondir le travail théorique. Il ne s'agit pas de revenir sur des questions déjà clairement exposées dans le premier texte de Roger T. Pédauque, mais en s'appuyant sur ce dernier et sur bien d'autres contributions sur le même objet de poser des questions plus théoriques. L'ambition, à terme, est de construire une « théorie du document » afin de fournir aux différents chercheurs engagés sur cet objet un cadre d'analyse transdisciplinaire. Nous verrons, le moment venu, si elle est couronnée de succès.
4. Ce document de travail propose à la discussion une réflexion qui relève des deux premières entrées pédauquienne (forme-texte). Il vise à comprendre plus finement les relations entre les choix d'ingénierie documentaire et leurs présupposés et conséquences sur les objets (ou services) qu'ils manipulent. Il s'agit de proposition de pistes théoriques qui doivent être débattues, amendées, complétées.. selon la méthode pédauquienne qui a fait la preuve de son efficacité (voir en fin de texte).
5. La difficulté pour les acteurs contemporains qui s'emparent de cette question (analystes, producteurs, usagers), c'est qu'ils se la posent dans des conditions historiques particulières. Ils ne peuvent abstraire la conceptualisation de la situation dans laquelle elle est aujourd'hui élaborée. C'est pourquoi, en même temps que nous nous efforçons de conceptualiser ces questions, nous ne devons pas croire, illusoirement, que nous pourrions les aborder de façon totalement objective et neutre. En effet, les termes même de notre analyse ont le caractère d'héritages historiques, en même temps que d'outils conceptuels. Et nos interventions techniques et sociales sont des prises de parti dans cette histoire.
6. Nous héritons ici de deux traditions. L'une, très ancienne, imprègne nos vies sociales et personnelles : la culture de l'imprimé. L'autre, récente mais puissante, transforme nombre de nos objets et habitudes quotidiennes : l'informatisation de la société. L'une et l'autre conditionnent nos façons de penser et de raisonner. Nous ne pouvons d'autant moins nous en extraire que l'objet même de notre réflexion, le document, est en même temps un des outils qui la porte !
7. A l'automne 2004 un autre travail privilégiera les croisements avec la 3e entrée pédauquienne : le Medium. Afin de ne pas entretenir d'ambiguïtés, nous avons préféré retenir le mot « texte » pour désigner l'objet de notre analyse, à laquelle on aurait pu reprocher la sous-estimation de la dimension sociale si l'on s'était arrêté à celui de « document »

8. Repenser le texte et la lecture

9. Dans ce contexte, le texte, au lieu de circuler sous la forme d'un objet relativement stable, associant un support matériel et une configuration formelle – ce qui conditionne l'idée même de document dans son sens classique – devient une réalité en permanence recomposée, une sorte *d'événement réitéré.* Ce qui se conserve n'est plus le texte donné à lire mais l'ensemble des modèles formels qui en représentent la

---

[1] Le présent document est le résultat d'une première réflexion collective menée par Bruno Bachimont, Jean Charlet, Jean-Yves Dion-Dury, Yves Jeanneret, Jean-Michel Salaün et Christine Vanoirbeek. Le travail visait à synthétiser des pistes de recherche, au-delà des idées des seuls contributeurs, il met donc en perspective un ensemble d'idées émises par la communauté scientifique.

[2] Dit RTP-DOC http://rtp-doc.enssib.fr.

[3] Pour la démarche collective http://rtp-doc.enssib.fr/pedauque/index.html. Pour le texte :
Document : forme, signe et médium, les re-formulations du numérique.
Roger T. Pédauque. Article. 08 juillet 2003. Working paper.
<http://archivesic.ccsd.cnrs.fr/sic_00000511.html>

sic_00001003, version 1 - 15 Jun 2004

structure. Paradoxalement, le texte devient insaisissable, en même temps qu'il reste incontournable. En effet, quelles que soient les formes de communication, un texte se trouve toujours, *in extremis*, reconfiguré. Mais il l'est par un mixte d'interventions : les actes des scripteurs (toujours d'une certaine façon pluriels, si on intègre l'écriture des outils) ; les opérations du système ; le geste de lecture pour *actualiser un nombre limité des relations rendues possibles par le modèle sous-jacent du texte*.

10. L'objet texte ici désigné n'est pas un objet purement sémiotique, ce n'est pas seulement un ensemble de codes ; c'est un objet fabriqué, artisanalement et/ou industriellement (dans le régime des médias informatisés par un mixte des deux), doté de propriétés matérielles, inscrit dans un certain type d'échange. L'importance de cet objet matériel dans la communication n'est pas nouvelle. Elle a déjà été soulignée par les historiens de la lecture et les « sociologues des textes ». Elle prend seulement aujourd'hui un relief nouveau, compte tenu du déplacement des propriétés de ces objets-textes.
11. Le texte – manuscrit, imprimé ou informatisé – *n'a jamais été un objet purement linguistique*, même dans le cas particulier où il est constitué essentiellement de *mots*. En effet, les occurrences des énoncés linguistiques n'adviennent qu'au sein de constructions inter-sémiotiques plus complexes, soit dans une sémiotique du corps (la voix), soit dans une sémiotique de l'image complexe (l'écriture, le cadre, la page, la typographie, etc.).
12. Lorsqu'on relit les théoriciens du texte (sémioticiens, historiens du livre, historiens des techniques) on voit que celui-ci se situe à la croisée de trois composantes toutes indispensables :
    a. la création de formes repères qui permettent l'anticipation et la reconnaissance d'organisations de la pensée (dimension soulignée par des auteurs comme Chartier ou Illitch) ;
    b. l'exercice d'une association et d'une dissémination des interprétations qui redistribue en permanence les textes (dimension particulièrement soulignée par Barthes et Derrida) ;
    c. et le processus de la réécriture, dans lequel le texte apparaît comme un objet toujours singulier mais toujours mis en série (dimension particulièrement soulignée par Bakhtine et les historiens de la culture lettrée).
13. Revenir à ces questions – qui montrent que, si les objets étaient naguère plus stables, le texte lui-même a toujours été une réalité difficile à cerner – permet de préciser la question qui se pose aujourd'hui. Y a-t-il encore du texte, et par là du document ? Cela dépend de la définition qu'on retient : oui, s'il s'agit de ce lien nécessaire et- toujours reconstruit entre les trois entités décrites ci-dessus (objets, signes et pratiques), mais non s'il s'agit de l'objet identifiable et autonome qui le portait.
14. La numérisation des contenus et l'informatisation de leur exploitation a pour effet de modifier profondément les conditions sous lesquelles les contenus sont constitués et exploités, de leur création à leur consultation en passant par leur circulation, conservation et documentation. En démultipliant les possibilités de manipulation, en découplant la forme consultée de la ressource enregistrée, la numérisation et l'informatisation déconstruisent l'unité documentaire et éclate la cohésion de la lecture. Autrement dit, ces innovations technologiques ouvrent de nouveaux espaces de possibilités où l'on ne retrouve pas nécessairement les pratiques anciennes, pratiques qu'il faut par conséquent, repenser, reconfigurer et déplacer dans ces espaces.
15. A cet égard, l'évolution du texte affecte nécessairement une conception de la culture. Le texte livresque et stable est porteur d'une figure de l'œuvre, même s'il tolère largement une transformation et une circulation. L'éclatement des ressources numériques engage sans doute un processus de remise en cause de la notion d'œuvre, sans pour autant dissiper la question de l'auteur, car toute textualisation est exercice d'un pouvoir et mise en place d'un ordre.
16. Un bonne façon de se reposer la question est peut-être de partir de l'activité de lecture ou de reconstruction des textes numériques. Une analogie avec le cinéma nous parait ici tout à fait éclairante, en ce qu'elle introduit une dimension un peu oubliée, la temporalité :
17. Le spectateur n'accède au contenu qu'à travers une consultation temporelle, où le flux de sa conscience se synchronise avec le flux des images et des sons ;
18. Ces images et ces sons sont construits et assemblés à partir de fragments dispersés, de ressources éparpillées sur les bandes magnétiques ou supports numériques.
19. Le spectateur reconstruit le sens de ce qu'il voit en en effectuant la synthèse temporelle au cours de sa consultation, au fur et à mesure que cette consultation progresse. En regardant le film, le spectateur se raconte une histoire, l'histoire qu'il est en train de voir, à partir de la succession temporelle d'images et de sons qui lui est proposée.
20. La lecture est analogue au phénomène cinématographique : lire, c'est consulter dans une succession temporelle les éléments dispersées sur des supports. Lire et constituer du sens, c'est vivre cette succession temporelle comme un film, le film de la conscience.
21. Finalement, il ne pourrait s'agir alors que de proposer pour l'instrumentation numérique des documents d'avoir une cinématographie des contenus, où les interactions doivent s'intégrer et se synthétiser dans un montage dynamique pour la conscience. Il serait par conséquent essentiel de comprendre comment la

sic_00001003, version 1 - 15 Jun 2004

dispersion spatiale des éléments constituant les ressources enregistrées d'un contenu peut donner lieu à une synthèse temporelle du sens, où le lecteur se saisit des possibilités d'interaction sur le contenu pour déployer dans son vécu temporel des actions d'interprétation et de construction du sens.

22. **La boite à outils : algorithmes et écrit**

23. Une autre façon de poser les mêmes questions est d'ouvrir directement la boîte à outils et de s'interroger sur leur nature, choix et fonctionnalité. Les évolutions récentes dans le traitement des documents sont en grande partie déterminées par la spécification algorithmique de leurs transformations, et par leur exécution via des moyen informatiques. Par ailleurs, il apparaît que, même dans un contexte de plus en plus ouvert au multimédia, l'écrit est l'outil très largement dominant pour la manipulation des textes documentaire dans un contexte numérique. Ainsi, il nous paraît important d'interroger les opportunités et les choix réalisés au sein de ces deux technologies, et leur combinaison, dans le champ qui nous intéresse ici.
24. Compte tenu des éléments rappelés dans la partie précédente, nous voudrions souligner d'abord l'importance de ce que l'on pourrait appeler une « transformation documentaire numérique » (nous dirons simplement « *transformation* » dans la suite). Il ne s'agit pas simplement d'un « programme ». Ainsi le simple comptage de mots n'est pas une *transformation*, tandis que la réalisation d'un index en est une. Cette dernière fait passer d'un document à un autre (enrichi) sans modifier les propriétés formelles qui l'ont défini comme « document ».
25. Toute *transformation* peut se décrire par les propriétés du document d'origine. La première de ces propriétés est la conformité au format, qui permet de restituer le document électronique dans sa forme physique perceptible, qu'elle soit image sur papier ou sur écran. Sans cette conformité « formelle », un logiciel de traitement de texte ne peut, par exemple, interpréter la suite d'octets, forme numérique élémentaire du document, afin de reconstruire les pages et d'organiser la présentation du texte et des images. Le document est alors dépouillé de sa sémantique de présentation, puisque même si l'information est intégralement présente dans le segment d'octets, ce dernier ne peut pas être interprété par le programme d'édition, et par conséquent devient inexploitable pour le lecteur humain.
26. L'exemple du calcul d'un index met en évidence l'importance des propriétés d'invariance de plus haut niveau (plus abstraites) que les invariances de format : ainsi les entrées constituant un index (des « expressions » donc) doivent être présentes dans le texte du document. Plus précisément, les références de pages associées aux entrées de l'index doivent exactement correspondre aux pages dans lesquelles on y trouvera leurs occurrences. Cet exemple montre que ce qui semble le plus essentiel à la transformation documentaire est en premier lieu concerné par l'invariance, ce qui pourrait sembler paradoxal au premier abord. Toutefois, l'invariance ne suffit pas a décrire complètement les propriétés d'une *transformation*.
27. Toute *transformation* peut aussi se décrire par les propriétés des documents produits. Dans l'exemple précédant, la construction de l'index doit aussi respecter des règles de présentation qui facilitent son utilisation : tabulation, alignement, et autres conventions typographiques.
28. Ces trois points, propriétés du document d'origine, invariance et propriété du document produit, sont les sommets d'un « triangle » qui détermine le cadre de la problématique transformationnelle. Son pivot est dans la spécification concrète des opérations de transformation.
29. Dans les architectures de traitement de documents, les transformations jouent un rôle croissant, et tendent à se spécialiser et à se combiner en réseaux de composants qui enchaînent les traitements afin d'obtenir les résultats escomptés. Les facteurs de cette évolution sont économiques (diminution des coûts et des durées de développement) mais aussi techniques (amélioration de la fiabilité globale des architectures, en se reposant au maximum sur des transformations modulaires et bien caractérisées). Dans de tels réseaux, il devient primordial de savoir comment les connaissances disponibles sur les documents sont propagées au fil des transformations. Dans le cas contraire, de l'incertitude est accumulée, entraînant des dysfonctionnements et des surcoûts, tant du point de vue des performances à l'exécution, que du point de vue des développements logiciels. C'est ce problème de propagation « compositionnelle » des propriétés formelles du document qui nous parait fondamental.
30. Si une vision « moderne » du document chez les informaticiens tend à l'assimiler à une structure de données, sa finalité en fait un objet bien plus riche. Un document pourrait donc être considéré de manière abstraite comme une structure de donnée muni d'une fonction d'interprétation permettant sa restitution sensible et organisée selon les règles coutumières qui régissent la composition typographiques et stylistique. Cette « fonction d'interprétation » peut se montrer en adéquation avec le modèle « triangulaire » si elle joue un rôle par exemple en tant qu'invariant dans les chaînes de transformation.

31. Cette dernière fonction n'est pas sans rapport avec l'utilisation massive de l'écriture comme technologie de gestion documentaire au travers du numérique. Le Web, par exemple, même s'il est ouvert au multimédia, reste un média dominé par l'écrit. Il y a sans doute des raisons techniques (bande passante et mémoires

sic_00001003, version 1 - 15 Jun 2004

limitées), mais notre hypothèse est que cette pré-éminence est due surtout à la performance technologique de l'écriture : à la fois signe graphique, icones repérables et descriptibles, combinaison de symboles abstraits, discrets, soumis à des règles morphosynthaxiques et traitables statistiquement, signifiants et traduisibles.

32. Cette performance a sans doute déjà été à l'origine de la multiplication des documents écrits dans l'univers du papier. Déjà se sont préfigurés dans cet univers des éléments d'organisation des structures : titres de chapitres, index, etc. De même tout un appareillage s'est construit pour « documenter » par l'écrit des objets divers, que l'on appelait autrefois « informations secondaires », aujourd'hui « métadonnées ». Cet appareillage a pour objectif une modélisation du monde des documents et il a pris une dimension nouvelle avec l'informatique et maintenant le Web (balisage, ontologies etc..).
33. Deux caractéristiques nous paraissent essentielles à souligner et travailler :
    a. l'écriture permet une hybridation du temps (par la linéarité, la transcription de la langue) et de l'espace (par le graphisme, la mise en page) ; autrement dit dans sa rencontre avec le numérique, elle a déjà en partie permis de développer des systèmes répondant à la problématique développée dans la partie précédente. Sans remonter très loin dans le temps, l'exemple le plus frappant est peut-être celui des DVD qui autorise, grâce à des entrées écrites, un « feuilletage » et une « documentation des films.
    b. l'écriture permet un calcul « bricolé » grâce à son caractère discret (qui autorise les comptages et traitement statistiques), son suivi de règle morphosynthaxique (qui autorise des traitements logiques) ou encore son caractère iconographique (qui permet des repérages de forme et topographique). Mais il ne s'agit que d'un bricolage (au sens fort) dans la mesure où l'interprétation des signes donne du jeu aux constructions formelles. Néanmoins, c'est peut-être en grande partie ces caractéristiques qui autorisent les *transformations* présentées ci dessus.
34. De ce fait, même pour la gestion des images, l'écriture est utilisée à grande échelle. Mieux, le fait que le numérique mette sur un même signal représentations iconiques et graphiques, permet à l'écriture d'être utilisée pour casser certaines contraintes de lecture des images animées ou nom (légende, feuilletage, chapitrage des DVD etc..).
35. On peut penser que nous allons assister dans ce domaine à des évolutions profondes qui seront sans doute différentes ou hybrides selon les types d'écriture. Certains symptômes, comme l'évolution de l'écriture des messages courts, le préfigurent peut-être.

36. Fragmentations et normalisations

37. Ainsi d'une part, il paraît nécessaire de repenser le texte et la lecture, d'autre part les outils qui aujourd'hui participent à la transformation des documents puisent à la fois dans la logique la plus formelle et dans l'efficacité structurelle du langage. Il se produit alors un décalage entre la transformation des documents et leur identité sémantique. Ce décalage, s'il n'était pas sinon maîtrisé, au moins repéré et compris, pourrait conduire, et nous y sommes peut être déjà, à des effets de sens dangereux pour la communication humaine.
38. Pour mieux mesurer le phénomène, il faut pointer les deux mouvements fondamentaux de la transformation du document dans son passage au numérique : le repérage de sa structure et la prise en compte de sa sémantique (au sens du Web sémantique)et donc les tentative de modélisation et de manipulation de l'une et l'autre..
39. L'approche "document structuré", issue du monde éditorial et tout particulièrement de celui de l'édition de documents techniques à grande échelle, a privilégié un point de vue sur le document, en termes de représentation abstraite mais formelle et, par conséquent, appréhendable par des programmes.
40. Si l'objectif de publication a été largement prépondérant et conduit à considérer la structure logique "éditorale" d'un document indépendamment de sa structure physique sur un support de restitution, il est très clair que le potentiel offert par les technologies ne se limite pas à ce seul point de vue. Deux problèmes majeurs sont alors à résoudre :
    a. Comment, de manière formelle, prendre en compte et maintenir la co-existence de, potentiellement, plusieurs structures logiques sur un même document ?
    b. Comment assurer la stabilité du contenu d'un document dont la responsabilité n'incombe plus seulement à un ou plusieurs auteurs et peut être altérée par l'intervention d'applications informatiques ?

41. Étape suivante, le Web sémantique (WS) est une opportunité extraordinaire de normaliser enfin un certain nombre de langages informatiques importants pour ce qu'on appelle l'interopérabilité des systèmes. On peut considérer que cette opportunité est en passe de se réaliser et que le WS a des conséquences sur toute l'informatique :

sic_00001003, version 1 - 15 Jun 2004

a. Les langages de description textuelle. SGML était normalisé mais assez lourd à mettre en œuvre. Maintenant, XML est normalisé et devient même l'esperanto de biens des applications informatiques. Cette normalisation vient avec l'existence d'éditeurs, de *parseurs,* de transformateurs, etc. et de toute une galaxie de langages de traitement et de transformation de ce même XML.
b. Les normes d'échange. EDIFACT existait bien mais sa mise en œuvre était difficile. Le WS a créé l'opportunité de développer beaucoup plus facilement des applications de toutes sortes et a appelé le renouvellement d'une telle norme avec de nombreuses avancées dans les services Web et les normes d'interopérabilité bâties sur XML comme SOAP ou des initiatives de consortiums comme ebXML.
c. Les langages dits de « représentation des connaissances » de l'intelligence artificielle n'étaient que des tentatives jamais abouties en terme de normalisation. L'empilement des langages du WS (le « cake » de Tim Berners Lee) a abouti à ce jour à normaliser la description des ontologies, et un certain nombre d'inférences que l'on peut faire, avec, à travers un langage comme OWL.

42. Le WS et l'organisme de propositions qu'est le W3C, n'est pas le seul auteur des normalisations dont on parle ici. Mais il est bien à l'origine de l'appel d'air créé autour de cette question de normes et standards. Et l'apparition, à ce rythme, de normes et leur respect est une révolution en soi.
43. Peut-il alors encore exister des applications qui ne respecteraient pas les normes du WS ? Il semble que l'on peut d'abord se poser la question au regard des expériences de l'intelligence artificielle.
44. En intelligence artificielle et plus particulièrement en ingénierie des connaissances, on a mis en avant, à l'occasion d'un projet qui a marqué son époque – KADS et CommonKADS –, la nécessité de décrire les inférences et les tâches des systèmes à base de connaissances (SBC) dans des « bibliothèques de tâches » réutilisables. Si l'objectif était compréhensible, les bibliothèques ont été peu utilisées. Il était difficile de trouver des situations qui se moulaient bien dans ces « patrons » de raisonnement prédéfinis. Avec le WS, le problème se pose à nouveau : si la prédominance des normes qu'il propose – et *in fine* impose – se confirme au rythme actuel, on ne décrira plus que des systèmes fondés sur des ontologies avec des raisonnements de subsomption validés dans le langage formel des *logiques de description*.
45. Déjà sur le Web un texte n'existe sous format numérisé que pratiquement structuré en XML, respectant un schéma XML ou, plus anciennement, une *Définition de type de document* (DTD). Ce respect est valable pour des textes scientifiques comme plus littéraires (*text Encoding Initiative*). En balayant les assises techniques et économiques de la diffusion papier, le Web est-il en train de proposer une nouvelle direction de travail : *Il n'existe ou n'existerait de document diffusé que sous forme numérique et encodé en XML* ?
46. Restons dans cette hypothèse et revenons au WS et à ses normes : un document sera repérable sur la toile *sémantisée* grâce à des balises qui lui donneront un sens ; balises choisies dans un référentiel partagé, une *ontologie.* Le WS serait alors en train de contraindre encore plus l'hypothèse précédente et la question qui va avec : *il n'existerait de document diffusé que sous forme numérique et dont le contenu serait repéré par des ontologies* ?
47. Si l'on rapproche ces remarques de celles qui introduisent la réflexion de ce document, nous pouvons noter qu'un tel futur ignore l'intertextualité, qu'il a du mal aujourd'hui à prendre en compte la temporalité, deux dimensions essentielles des textes dans leur mise en document. Sans être pessimiste, admettons le caractère d'urgence de cette réflexion. Les ontologies, pour utiles qu'elles soient, sont des objets formels (dans le cadre du WS) et leur compréhension par des êtres humains passe par la prise en compte d'une dimension linguistique, manifestement ignorée, si ce n'est des concepteurs de ce WS au moins de nombre d'utilisateurs (nous parlons ici des chercheurs).
48. Inversement, on peut remarquer que les normes autour du WS ont ravivé le désir de formalisation de l'intelligence artificielle et les rêves – sans connotation péjorative – de « raisonnement artificiel » en même temps qu'elles permettent de réfléchir sur les annotations et les gloses des documents – des textes et du contexte, laissant l'intepretation à l'être humain – et leur mise en œuvre informatique. Le WS contiendrait-il le poison et l'antidote ?

Si vous souhaitez participer à la rédaction de ce texte, il vous suffit d'envoyer vos remarques, en précisant le cas échéant le numéro du ou des paragraphes auxquels elles réfèrent, à pedauque@enssib.fr. Votre message sera envoyé au forum Pédauque qui regroupe tous les contributeurs et vous serez inscrit à ce forum. Si vous êtes déjà abonné au forum, il vous suffit d'envoyer le message à la liste selon la procédure habituelle.
Ce forum est consacré uniquement à la discussion scientifique sur la notion de document. Seuls les messages s'inscrivant dans cette perspective seront pris en compte. Les simples demandes d'inscription, ou les annonces sans contribution seront rejetées. Le forum sera clos au 15 septembre 2004.