Le texte en jeu Permanence et transformations du document

Roger T. Pédauque

Le texte en jeu

Permanence et transformations du document

Roger T. Pédauque
Contacts pedauque@enssib.fr

Version 4, 07-04-2005

Résumé

Ce document de travail, réalisé collectivement dans le cadre du RTP-DOC, attire l'attention de la communauté scientifique sur les risques qu'il y aurait à ne pas préciser la notion de texte et sa relation avec celle de document, afin de mieux adapter les développements du numérique à une communication humaine.

La brutalité des évolutions actuelles s'appuie sur des implicites, séparant structure et contenu et rapprochant les activités de lecture et d'écriture. Trois modélisations informatiques, successives et distinctes, sous-tendent les développements récents des outils autour du document codé en XML : les DTD, les Schémas et le Web sémantique. Ces modélisations négligent des réflexions multidisciplinaires indispensables :

1. Les développements informatiques utilisent des disciplines comme la linguistique qui n'ont pas analysé clairement la notion de texte. Pour imaginer une structure indépendante d'un contenu, il faut rendre compte des multiples dimensions, intellectuelles et matérielles, de la textualité.
2. La différence entre texte et document, déjà intégrée par exemple dans le raisonnement juridique, peut se décliner dans les dispositifs techniques par des Schémas en repérant les invariants pour préserver le statut documentaire dans la transformation du document d'un état à un autre. Néanmoins, la notion d'invariance montre ses limites dans la traduction.
3. Enfin la tentative de dépassement par le Web sémantique peut conduire à un appauvrissement ou une confusion de l'ordre des savoirs humains si une analyse correcte n'est pas faite du rôle des ontologies, refusant le « nominalisme ».

Un retour sur les notions fondamentales devrait faciliter un développement des outils allié à une meilleure maîtrise par les communautés humaines des conséquences de leur utilisation. En particulier, il paraîtrait utile de préciser ou modifier la notion de texte dans un environnement documentaire de plus en plus multimédia et de mieux analyser le rôle de la médiation, humaine ou automatisée sans la confondre avec une « écriture automatique ».

Sommaire

Le texte en jeu
Permanence et transformations du document

Roger T. Pédauque, STIC-CNRS[1]
Contacts pedauque@enssib.fr

Version 4, 07-04-2005

Avant-propos

Il est apparu rapidement au sein du Réseau thématique pluridisciplinaire « Document et contenus : création, indexation, navigation »[2] des départements STIC et SHS du CNRS qu'une réflexion globale et approfondie sur la notion de document était nécessaire[3].

Pour contribuer à ce mouvement maintenant largement alimenté par des contributions et initiatives nombreuses, nous proposons ici un approfondissement qui relève prioritairement de la deuxième entrée (texte), parmi les trois repérées dans un précédent document du même auteur. 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. Ce faisant, nous ne pouvons négliger la troisième entrée (médium) puisque les présupposés s'inspirent d'une communication imaginée et les conséquences portent sur une communication réalisée, ni la première (forme) puisque les choix définissent des formes qui configurent objets et services. Néanmoins, il nous semble possible de déduire du balisage réalisé dans le premier texte de Roger une méthode d'analyse qui confère à chaque entrée une autonomie relative sans pour autant abandonner les éclairages indispensables des deux autres. Les recherches ne peuvent s'abstraire d'une analyse triangulaire en prenant nécessairement en compte l'ensemble des entrées, néanmoins chacune nécessite des expertises précises qui relèvent d'une logique propre, car irréductible aux autres entrées. Ainsi pour produire un savoir spécialisé utile, il est obligatoire de mettre l'accent sur une entrée, mais les autres entrées devront lui fournir des données extérieures indispensables à sa construction. Un prochain travail collectif privilégiera l'entrée « médium ».

La méthode d'écriture collective de ce document induit un mode de lecture. Les cinq sections proposent cinq éclairages sur la même problématique. Elles ont été élaborées et discutées de concert, mais ont fait en réalité l'objet de débats distincts. Aussi, on peut soit lire le document dans sa continuité, en suivant le fil proposé dans la première section, soit s'attacher plus exclusivement à l'une ou l'autre des sections, chacune développant un argumentaire propre.

---

[1] Le présent document vise à synthétiser des pistes de recherche autour du document numérique. Au-delà des idées des seuls contributeurs, il met en perspective un ensemble d'idées émises par la communauté scientifique.
Responsables de la synthèse : Bruno Bachimont, Jean Charlet, Yves Jeanneret, Jean-Michel Salaün, Monique Slodzian, Christine Vanoirbeek et Jean-Yves Vion-Dury.
Contributeurs : Olivier Beaudoux, Abdel Belaid, Dominique Boullier, Evelyne Broudoux, Marie-Anne Chabin, Christophe Choisy, Anne Condamines, Dominique Cotte, Gabriel Gallezot, Eric Guichard, Brigitte Guyot, Laurence Likforman-Sulem, Niels W. Lund, Yannick Maignien, Claude Poissenot, François Rastier, Christian Rossi, Eric Thivant, Christian Vandendorpe, Michel Volle, Manuel Zacklad.

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

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

1 Aux origines des choix d'ingénierie documentaire

1.1 Héritages historiques et violence des déplacements contemporains

Prégnance historique

Pour prendre en compte toute la complexité des phénomènes, nous devons les placer dans la perspective d'une histoire des constructions techniques et symboliques afin de relier l'invention de dispositifs matériels d'information et de communication avec la mise en place de formes et d'usages du texte. En effet, les questions discutées ici tiennent à une évolution fort ancienne, qui unit le mouvement général des innovations techniques à la mise en ordre des savoirs, la mise en forme des expressions, la mise en place des relations. Ce type d'invention couplée était déjà en jeu avec les tablettes sumériennes, avec le livre, avec la presse, avec l'émission de radio et de télévision. Le problème que nous étudions n'est pas d'une nature différente et nous y rencontrons les réflexions des historiens du livre et de ceux de la technique.

Des objets comme la page ou, à d'autres niveaux, le titre ou le genre prennent des formes nouvelles, mais ces formes ne peuvent être saisies que dans le fil d'une histoire longue. Les modes de production du texte et les gestes qui sont associés à sa manipulation relèvent de processus industriels et d'usages renouvelés, mais qui recyclent beaucoup de façons de faire plus anciennes. C'est pourquoi l'analyse des mutations actuelles du texte ne doit pas ignorer celles des transformations de l'imprimerie, de l'institution des modèles éditoriaux, du développement des pratiques documentaires, de l'invention des formes graphiques, des usages sociaux de l'imprimé, entre beaucoup d'autres transformations plus anciennes.

Pourtant, cette filiation de longue durée ne doit pas nous dédouaner d'une analyse des caractéristiques inédites des outils contemporains. A l'occasion de l'écriture de ce texte, une discussion s'est développée sur la position de l'informatique, comme science ou technique, et son inscription par et dans le social. Celle-ci déborde largement le sujet précis de ce document et ne saurait être reproduite. Mais son émergence et la vivacité des débats qui l'ont accompagnée montrent combien la position de l'informatique déroute et combien l'absence d'une analyse partagée handicape l'avancée de notre réflexion.

Nous avons montré dans le premier texte de Roger[4] la nécessité de repérer les différentes origines et les conséquences de l'application de l'informatique sur les documents. Nous verrons dans celui-ci combien la relation ambiguë entre la calculabilité et l'organisation de la langue, la polysémie du terme « langage » (machine ou humain) ou encore les différentes manières de modéliser en informatique, pèsent sur les choix des acteurs et les orientations proposées.

La difficulté pour les contemporains qui s'emparent de cette question (analystes, producteurs, usagers), est bien qu'ils se la posent dans des conditions historiques particulières. Or, si beaucoup de questions se conservent, les conditions pour les poser se déplacent, parce que le rapport entre dispositif technique, organisation des informations et lisibilité du texte ne se définit plus dans les mêmes termes. 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

[4] *Opus cite*, *Cf.* note 2

marqués par l'avancée de l'informatique. Et nos interventions, scientifiques, techniques et sociales, sont des prises de parti dans cette histoire.

Rupture technico-économique

Il y a, au tournant de ce millénaire, une violence dans la rapidité de diffusion et de transformation des outils de lecture-écriture numérique, dans la performance des choix techniques, dans l'importance symbolique, pour les individus, pour les organisations comme pour les sociétés, des objets en cause – rien de moins que notre patrimoine cognitif et culturel ! – qui donne au changement une teneur totalitaire. Les modes de représentation des textes, dont l'évolution lente et continue s'étalait sur des siècles permettant une appropriation progressive, sont maintenant discutés jour et nuit dans des forums internationaux très actifs, puis appliqués par les industriels dans une course dont le gagnant pense, à tort ou à raison, tirer le meilleur profit en maîtrisant la technologie dominante parce que la première proposée. Ainsi les outils paraissent s'imposer sans contestation possible.

Sans doute, l'acceptation d'un standard et la configuration d'une norme ne vont pas de soi. Le consensus se construit dans des jeux de rôles aux résultats incertains et les situations diffèrent selon, par exemple, que l'on raisonne globalement ou sur des applications et des communautés spécialisés. Il faudrait analyser plus finement le processus que nous ne saurions le faire ici. Néanmoins, la rapidité et l'étendue des échanges tendent à accélérer les changements et favoriser les conformismes. Se singulariser par ses outils ou simplement refuser de suivre le mouvement général de ses pairs, nécessite une maîtrise forte de la technicité, réservée à un nombre très réduit d'acteurs qui disposent du savoir-faire et du temps suffisants. Sans ces qualités, le risque est grand de se marginaliser en freinant la communication.

Dès lors, la brutalité du changement rend difficile le recul, l'analyse critique fine et nuancée. Il faut trancher vite, réagir. C'est l'heure des prises de positions fortes et militantes – pour ou contre – moins celle d'une déconstruction exigeante et positive des orientations choisies. Par un effet de perspective, les changements paraissent pluriels et multiples, ils sont simplement plus rapides, plus radicaux et plus étendus.

La finesse d'analyse est bien présente, mais elle est le plus souvent réservée aux expertises très pointues, par exemple dans les spécifications techniques des « standards » telles qu'elles sont discutées au sein du consortium W3C. En comparaison, les présupposés généraux qui sous-tendent les choix globaux, tels qu'ils sont énoncés par exemple dans le même groupe, paraissent généreux, mais presque naïfs ou simplement tautologiques[5].

De même, les entreprises du secteur, très récentes et pourtant parfois déjà très prospères et en position de quasi-monopole, recherchent d'abord la domination du marché. Celles qui s'imposent écrasent leurs concurrentes en utilisant avec une grande habileté les caractéristiques économiques particulières du domaine (verrouillage, externalités).

Cette situation paradoxale, entre complexité des choix techniques particuliers ou astuce des choix stratégiques et simplicité des orientations générales, donne le sentiment erroné d'un destin technico-économique, heureux ou malheureux, que l'on doit subir et dont les termes alternatifs se réduisent souvent dans les débats à un choix binaire entre des militants des

---

[5] « (..) le Consortium du World Wide Web (W3C) crée des standards pour le Web. Sa mission est de **mener le Web à son potentiel maximal**, tout en développant des technologies (spécifications, lignes directrices, logiciel et outils) qui favorisent l'échange d'information, le commerce, l'inspiration, le libre arbitre, et la compréhension collective. »

Extrait de « Le W3C en 7 points » : http://www.w3.org/Consortium/Points/w3c7.fr.htm

sources ouvertes et des défenseurs des intérêts commerciaux. Les développeurs, eux-mêmes, considèrent qu'ils manipulent un outil neutre dont les orientations sont définies par les utilisateurs, soit en amont comme donneurs d'ordre, soit en aval dans des applications diverses parfois très éloignées de leurs projets initiaux.

En réalité, des choix fondamentaux sont réalisés par consensus d'un groupe d'experts à la suite de discussions longues et très techniques, puis des mises en pratique sont engagées par les groupes industriels les plus puissants ou les collectifs les plus pointus qui ont les moyens à la fois de participer aux discussions et d'investir (en temps ou en argent) en recherche-développement. L'organisation du Web favorise ce genre de dynamique, à la fois transparente (en théorie, tout le monde a accès à la discussion et à ses résultats) et opaque (en pratique, seuls quelques-uns peuvent en comprendre les éléments et en maîtriser les enjeux).

De multiples décisions sont donc prises sur des standards, sur des modes d'écriture et de lecture qui découlent de modèles implicites, rarement exposés car ne posant pas problème pour ceux qui développent les outils, et dont les conséquences sont pourtant lourdes sur nos modes de représentation des textes.

1.2 Modèles « implicites »

Fond, forme et grammatisation

Le document est souvent abordé à travers les notions de *fond*[6] et de *forme*, qui permettent de distinguer une forme sous laquelle un document se présente du fond ou contenu (les deux termes seront synonymes dans cette section) qu'on lui prête. Cette distinction est, comme on le sait, fragile. Elle repose en effet sur des pratiques et des conventions qui permettent d'opposer trois niveaux distincts pour les documents inscrits sur un support :

- La *forme*, c'est-à-dire la forme matérielle sous laquelle se présente un document et dont les propriétés perceptives et son ancrage dans différentes traditions sémiotiques, culturelles ou sociales conditionnent l'interprétation. Dans le cas du texte imprimé par exemple, c'est le papier noirci que l'on a sous les yeux.

- Le *contenu*, c'est-à-dire ce qui constitue les « caractéristiques essentielles » du document, qui permettent de dire qu'il s'agit du même contenu quand on considère ses différentes présentations possibles. Ces caractéristiques sont dégagées par la tradition, historique, sociale ou culturelle. Dans le cas du texte, la tradition occidentale nous a légué le fait que le *contenu* se caractérise comme le codage alphabétique du discours et sa structuration en chapitres, sections et paragraphes. Les autres attributs de mise en forme participent de la *forme*, et n'ajouteraient que des attributs inessentiels au *contenu*. Autrement dit, l'intégrité de l'œuvre se caractérise par le *contenu*, et non par la *forme*.

- La *signification*, c'est-à-dire l'ensemble des interprétations explicitant ce qui est compris ou retenu d'un document. C'est alors l'ensemble des documents et actions possibles constituant le réseau d'interprétation du *contenu*. La signification passe par un réseau de *formes* et la dynamique qui permet de sans cesse l'enrichir. Nous ne la développerons pas ici.

Seule la *forme* existe concrètement. En effet, le *contenu* est un objet abstrait, qui n'existe pas en tant que tel. Ce serait plutôt un « invariant » dégagé des multiples présentations possibles considérées comme reflétant le même contenu.

[6] Par convention et afin de faciliter la lecture, nous mettons un mot en italique, pour signifier qu'il est employé comme un concept et non dans le sens commun.

Le *contenu* ne serait donc pas physiquement constitué : quand il l'est, c'est qu'on a retenu une présentation particulière comme étant canoniquement reliée au contenu d'un document. Cette forme canonique matérialise un contenu en lui donnant une forme physiquement manipulable et perceptible. Un contenu étant par nature abstrait, la forme canonique qui le matérialise n'est jamais fidèle ni pure vis-à-vis de ce dernier : elle dépend de règles conventionnelles qui peuvent se reconfigurer au gré des mutations techniques, historiques et culturelles.

Par exemple et dans nos sociétés, le manuscrit d'un auteur constitue une forme canonique matérialisant le contenu de son oeuvre. La représentation canonique découle du choix des lettres, de la ponctuation, et de la structuration, en faisant abstraction des graphies particulières, de la couleur de l'encre, etc. Dans le contexte technologique contemporain, la forme canonique s'apparente de plus en plus à un texte numérique codé en Unicode[7] et structuré avec un balisage XML. Les formes canoniques présentant un contenu ne sont donc pas uniques : on considère aujourd'hui que le texte ASCII est le même contenu que son équivalent manuscrit, ce sont des formes particulières d'un même invariant, qui se déclinera ensuite dans des présentations de formes diverses par la publication.

Il s'agit d'un exemple de ce que certains ont appelé une « *grammatisation* »[8] qui matérialise la substance d'une expression en éléments matériels discrets et manipulables. Mais en matérialisant un contenu, on effectue des choix qui comportent une part d'arbitraire, renvoyant à plus tard une éventuelle meilleure représentation. Il est impossible, en effet, d'avoir une expression exacte : tout support de l'expression la modifie par les suppléments incontrôlés qu'il y ajoute. Ainsi toute expression est inadéquate à ce qui est exprimé car, grammatisée, elle hérite d'une autonomie liée aux possibilités manipulatoires des unités de *grammatisation*, qui déplacent son sens autant qu'elles le constituent. En même temps, l'expression y gagnera une formalisation qui permettra de lui faire subir toutes sortes d'opération.

Il n'y a pas lieu de se lamenter ou de se réjouir de cette influence de la mise en forme, ni de tenter d'y échapper par un illusoire appel à une neutralité de la technique, qui masquerait le problème au lieu de le résoudre. En revanche, il incombe à une théorie du document de thématiser cette influence.

XML et ses trois modèles documentaires

On trouvera un cas d'école en considérant les *grammatisations* informatiques des documents par les normes proposées en ingénierie documentaire, notamment à travers les travaux du W3C.

Les techniques documentaires fondées sur XML objectivent et réifient la tradition du document papier. Il est d'usage de considérer qu'un document structuré en XML[9] recèle le *fond* – l'équivalent du manuscrit avant sa mise en forme typographique –, et que sa publication à travers par exemple des feuilles de style en est la *forme*. Mais les différentes représentations possibles d'un document en XML renvoient à des *grammatisations* différentes. Selon qu'elle adopte le formalisme des DTDs (Définition de Type de Document),

---

[7] http://www.unicode.org

[8] Auroux Sylvain, *La révolution technologique de la grammatisation*, Mardaga, 1992.
Derrida Jacques, *L'écriture et la différence*, Seuil, 1968, *Marges de la philosophie*, Seuil 1972, *De la grammatologie,* Seuil, 1967.

[9] Pour une présentation simple de XML et de ses dérivés : W3C, XML en 10 points
http://www.w3.org/XML/1999/XML-in-10-points.fr.html

des Schémas[10], ou bien encore des représentations logiques du Web sémantique, la codification XML renvoie à autant de considérations distinctes sur le *contenu* :

- Les DTDs, sont des grammaires qui organisent et décrivent la structure d'un document. Leur objectif est de permettre la production et la description des documents, sans contraindre particulièrement ce qui est écrit ou formulé. Un document obéit à des règles de construction relevant d'une tradition ancestrale du monde de l'édition. Les possibilités offertes par le numérique ont progressivement convergé vers SGML (*Standard Generalized Markup Language*), issu du domaine particulier de l'édition technique dont les textes sont par nature très structurés afin d'en faciliter l'utilisation. Celui-ci a notamment permis le développement d'outils facilitant la vérification de la cohérence d'un document – via le mécanisme de DTD – et la déclinaison de présentations variées – via l'association de feuilles de styles – à un même *contenu*, ici pratiquement synonyme de « texte ». Cette approche est notamment privilégiée par les entreprises qui se sont développées à partir de la bureautique.
- Dans le contexte des Schémas, la préoccupation est double : il s'agit, non seulement structurer le document, mais aussi de contrôler la nature de ce qui est écrit : qu'un numéro de téléphone soit bien une suite de huit chiffres, par exemple. Dans cette optique, le *contenu* est considéré comme un ensemble d'informations, le Schéma en contraint les modalités de présentation par la prise en compte d'un certain nombre d'invariances. L'enjeu est alors de pouvoir échanger des informations entre des applications ou des bases de données. Les Schémas XML sont donc issus d'un autre monde que la précédente *grammatisation* des DTDs : celui des bases de données. S'éloignant de la fonction première de SGML, la représentation et le traitement de documents, XML s'est alors imposé comme un langage universel et a été adopté par la majorité des applications informatiques. En conséquence, ce langage est désormais majoritairement utilisé pour représenter des informations factuelles et plus seulement pour supporter un discours construit d'un document. Cette approche est plus favorable aux entreprises qui ont une implantation ancienne dans l'informatique fondamentale et en particulier dans les bases de données.
- Enfin, les représentations formalisées du Web sémantique accompagnent les documents d'expressions formelles pour s'y substituer en fait. Le *contenu* du document n'est plus seulement une information dont on contrôle l'expression pour l'échange entre applications informatiques comme dans les Schémas, mais une représentation formelle de connaissances dont on contrôle la sémantique (formelle) pour la déclinaison et le calcul par des agents produisant des services. La volonté d'élargir le traitement des informations publiées sur le Web, et de mobiliser des connaissances non seulement valables pour une transaction précise entre bases de données mais utiles pour un ensemble de traitements, a conduit à la modélisation des connaissances d'un domaine (les ontologies) pour paramétrer l'exploitation des contenus. On trouve ici aussi les firmes les plus jeunes et les plus ambitieuses qui ont construit leur croissance autour du développement du réseau. Nous reviendrons plus précisément dans la section 4 sur l'importance et l'ambiguïté des mouvements autour du Web sémantique et sur la problématique des ontologies.

[10] Par convention, nous mettons ici une majuscule, pour différencier le terme de son acception ordinaire de « représentation graphique ».

Ces approches sont souvent présentées comme une évolution et une progression naturelles de l'instrumentation documentaire, la sophistication informatique croissante devant souligner davantage un approfondissement de la notion de *contenu* qu'un changement. Dans cette optique, le troisième modèle marquerait une rupture avec les précédents, puisque ici la sémantique est mise à l'œuvre, pour, en quelques sortes, doubler le *contenu* par un langage parallèle plus approprié aux traitements informatiques.

Mais il s'agit aussi, et peut-être surtout, de conceptions différentes sur ce qui fait le *contenu* d'un document. Les arguments précédents suggèrent qu'il s'agit d'une concurrence entre communautés de pratique et entre orientations stratégiques de firmes, où les règles conventionnelles dégagées par l'usage ou les choix conceptuels ou encore les avantages concurrentiels commerciaux sont radicalement distincts. L'instrumentation technique n'est donc pas neutre et implique une conception du *contenu* et de la *signification*. Si cela n'est pas nouveau, il ne faut pas l'ignorer ici et savoir effectuer les choix raisonnés entre les différents paradigmes technologiques proposés.

On voit, sur cet exemple, le jeu de la *grammatisation*. A partir d'un langage désormais véhiculaire, c'est-à-dire utilisée comme un simple instrument syntaxique, des conceptions distinctes du *contenu* se déploient en différentes couches. La syntaxe XML, premier élément de *grammatisation*, initialement élaborée pour exprimer et matérialiser le contenu d'un document, a en fait démultiplié et différencié les points de vue sur le document. Au lieu de fixer un *contenu*, la syntaxe XML en a différé l'explicitation en permettant une différenciation inédite.

La *grammatisation* est donc bien une manipulation de l'expression qui la détourne de son intention initiale. Ce n'est pas un défaut, mais une vertu : sans ce décalage nous ne pourrions sans doute produire un sens nouveau. Mais, pour exploiter ces possibilités, il est nécessaire de quitter toute conception neutraliste de la technique, pour thématiser aussi explicitement que possible les déplacements et transformations effectuées. En considérant les techniques du génie documentaire comme des couches sédimentées d'instrumentations successives et de *grammatisations* différentes, on aborde une généalogie des documents numériques et il devient possible d'élaborer une philologie numérique, c'est-à-dire une discipline capable d'établir les textes en fonction de leurs mises en forme et formalisations successives.

Il reste à interroger les choix effectués et questions posées par chacun de ces modèles. De manière rapide, on peut s'interroger sur le rapport entre le « modèle DTD » et les textes, dont il semble faire l'impasse (section 2) ; également, on peut se demander ce qu'impliquent, quant aux documents, les notions de typage et d'information dans le « modèle Schémas » (section 3) ; enfin, le « modèle ontologies », impliqué par la conception formelle des documents, présuppose un accord sur une vision du monde, acquis seulement dans certaines communautés très structurées ou sous certaines conditions (section 4). La figure 1 ci-dessous rend compte de cette gradation.

On trouvera en annexe un exemple illustratif de l'application de ces trois modèles à un extrait du dossier médical d'un patient.

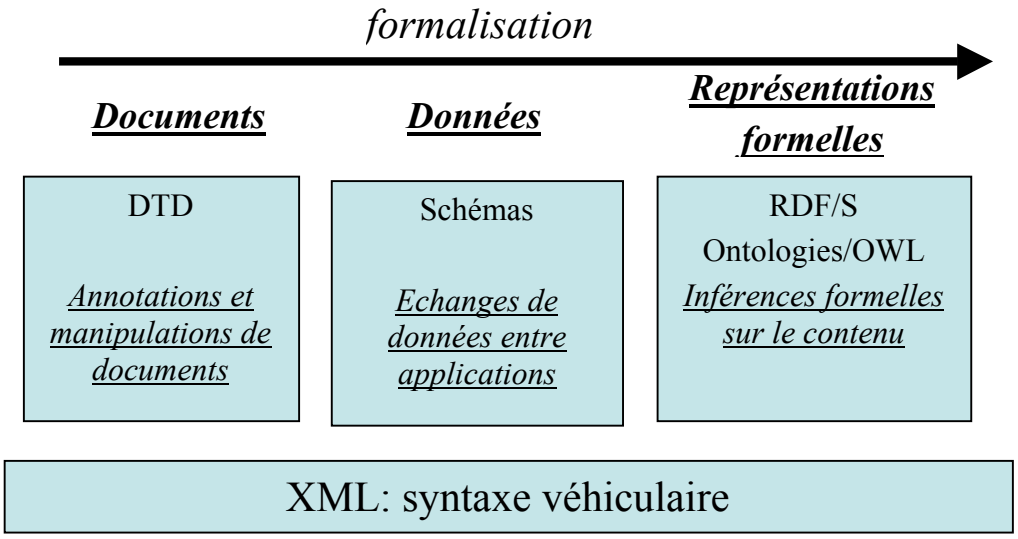

FIG. 1 : *XML et la grammatisation des documents. On distingue 3 postures : les documents, les données, les connaissances. Les deux premières sont proches dans la mesure où les Schémas prolongent les DTDs en précisant le type des contenus et des structures. Contrôlant des structures d'arbre, elles spécifient une organisation de contenu. L'approche « connaissance » reformule le contenu en laissant ambigu le fait de savoir si la formalisation annote ou remplace le contenu. Mais on peut voir que ces trois approches procèdent d'une grammatisation accrue et d'une formalisation plus importante à chaque fois puisque l'on contrôle toujours davantage le contenu, au point de le remplacer par sa formalisation.*

2 L'incertain concept de *texte*

Le premier modèle décrit ci-dessus, celui des DTDs issu des traditions de l'imprimerie, fait implicitement l'hypothèse que le maintien du *contenu* assimilé au *texte* est suffisant pour nous permettre de nous orienter dans les documents, grâce à l'efficacité nouvelle des outils informatique. Mais cette hypothèse repose sur le postulat que le *texte* lui-même est un invariant simple à identifier. Nous verrons dans cette section que cette hypothèse est hasardeuse.

Au moment où l'objet documentaire devient de plus en plus difficile à cerner, la construction textuelle deviendrait donc déterminante pour identifier les structures cognitives, symboliques et sociales définissant un espace documentaire. Là où le volume, la qualité du papier, l'origine distinguent, par exemple, facilement dans l'édition traditionnelle une brochure d'un manuel, seule l'organisation des signes différencie deux sites portés par le même espace-écran. En rendant incertain un objet autrefois facile à identifier, les changements en cours supposent d'autres points de repère, mais il n'est pas sûr que nous ayons encore les outils suffisants pour les construire.

Nous ne pouvons, notamment, trouver dans la linguistique une définition claire du *texte*. De plus, celui-ci déborde largement la simple transcription de la langue pour s'enrichir de toutes sortes de signes, statiques et dynamiques, pour lesquels le contour d'un contenu reste flou. Enfin, sa mise en situation montre des usages forts différents, dont certains directement en relation avec l'action, qui induisent des configurations différentes, élargies ou éphémères.

2.1 Le *texte*, point aveugle de la linguistique

Le *texte* étant à l'évidence une dimension du langage, on s'attendrait à ce que les textes concrets constituent l'objet empirique essentiel de la linguistique. Ce n'est pas le cas. Le texte n'a jamais vraiment constitué une unité minimale d'étude pour la linguistique, en partie parce que celle-ci souhaitait se démarquer de la littérature, jugée non scientifique.

La logique de la phrase et du mot

On constate que la dimension d'analyse privilégiée des linguistes est la phrase. Celle-ci est en effet à la fois un artefact culturel (l'exemple des grammairiens et des linguistes, la citation des écrivains, etc.) et une problématique logico-grammaticale. Concernée avant tout par les questions de référence et de vérité, la tradition logico-grammaticale privilégie le mot (lieu de la référence) et la proposition (lieu de la vérité). La phrase présente l'avantage d'être hiérarchisée (subordinations, rections internes), segmentable en unités inférieures (de la proposition au morphème) et achevée (bornage de la ponctuation). L'impératif du codage informatique a accentué la domination de la phrase comme unité de description.

Il existe en linguistique des théories du texte, mais elles sont minoritaires et les modèles linguistiques dominants relèvent de l'extension au texte de propriétés du syntagme et de la phrase, selon un principe d'analyse compositionnelle qui tend à mettre « hors jeu » la singularité du texte comme objet empirique et la relation globale qu'il entretient avec le sens. Cette tendance est aujourd'hui renforcée par le fait que les techniques d'automatisation de l'analyse qui ont été privilégiées font du lexique et de l'organisation intra-phrastique un objet crucial. Mais cette tendance tend à masquer le texte en tant qu'unité concrète d'analyse et, traduite en opérations de reformulation automatique, elle efface l'hétérogénéité des objets textuels au bénéfice de modèles logiques qui, pour avoir été plus ou moins banalisés comme la forme usuellement considérée du langage, n'en sont pas moins, en eux-mêmes, des interprétations instrumentalisées d'une sorte particulière.

Quant à la sémantique du mot, du fait qu'elle repose sur un modèle logiciste qui fait de lui le lieu de la référence, elle n'est pas applicable au texte. En effet, si dans le cas du mot la position mentaliste consistant à relier les mots au monde par l'intermédiaire du concept continue de prévaloir dans les sémantiques classiques, elle ne résiste pas au texte. Pour attribuer au mot une signification constante, il faut présupposer que celui-ci reflète une réalité permanente et des concepts stables. Or la polysémie et la non-prédictibilité du sens des unités qui le composent sont la règle du texte. On notera, par exemple, qu'une langue comme le français offre cinq temps morphologiques pour exprimer la seule relation temporelle du passé. Dans les textes, l'interaction des divers marqueurs syntaxiques et lexicaux qui régit la temporalité est fondamentalement polysémique. Ainsi l'ensemble des phénomènes qui caractérisent un texte – de sa prosodie à son fonctionnement lexical – ressortit à une sémantique du global : les propriétés du tout ne peuvent être expliquées à partir des propriétés des parties. Les textes n'étant pas composés d'unités correspondant à des signifiants stables et isolables, les critères dévolus à la sémiotique du signe ne sont simplement pas pertinents pour la sémantique textuelle.

Il y a néanmoins des linguistiques qui ont mis le texte au centre de leurs préoccupations. Il s'agit de théories qui s'attachent à étudier les structures argumentatives des textes, les interactions et actes du langage. Elles relèvent de théories communicationnelles et, en cela, d'approches fonctionnalistes caractérisant les textes par des critères externes (but, interactions typiques, etc.). Elles décrivent des normes de production textuelle dans des espaces sociolinguistiques identifiés. Leur intérêt est d'introduire la problématique des genres et discours. La contrepartie de ces classifications pragmatiques est l'hétérogénéité (on trouve en

vrac la signalisation routière, l'article de presse, la conversation, le poème) et une grande confusion entre les notions de genre et discours. Leur limite épistémologique tient au fait qu'elles conçoivent les acteurs et les fonctions narratives comme des instances cognitives extrinsèques au texte, à la langue et à sa culture.

En cela elles ne s'émancipent pas des préjugés logico-grammaticaux : de la segmentation du texte en unités propositionnelles à la hiérarchisation des thèmes et de leurs propriétés, on reconstitue dans le processus d'analyse textuelle un modèle de connaissance imprégné de la tradition grammaticale. Par exemple, l'analyse du discours descriptif est classiquement comprise comme une opération de construction de la référence par analogie avec la fonction de désignation du mot. Cette équivalence entre attribution d'une signification lexicale et discursive se traduit par la représentation du texte en structure arborescente. Cela revient à réduire le discours à une équivalence statique sans attention aux transformations du thème au sein du texte et à faire de la linguistique textuelle un système de classement (discours juridique, esthétique, etc.) Or la tradition rhétorique nous a appris à aborder le texte comme le lieu-même de la complexité : le discours y est analysé comme déploiement de séquences différentes (descriptif *vs* narratif *vs* argumentatif), croisement et reprises de thèmes, jeu de contrastes lexicaux (antonymie, registre soutenu/familier, etc.)

Fonder une sémiotique du texte

L'enjeu actuel est de dépasser l'héritage référentialiste pour fonder une sémiotique du texte tirant profit à la fois des possibilités d'analyse de la linguistique de corpus et de la problématique de l'hypertextualité et des techniques multimédias. Il s'agit de fonder le texte comme unité d'analyse intrinsèque avec ses corrélats méthodologiques d'intra- et d'intertextualité. A la différence de la phrase où les parties du discours ont des signifiants isolables, le texte ne peut être segmenté en unités sémantiques directement identifiables. L'intratextuel ne se réduit pas évidemment à quelques mots clés qui en subsumeraient la thématique. Le sens d'un texte résulte de connexions de traits sémantiques qui se manifestent à différents paliers du texte (phonèmes, morphèmes, lexies, prosodie…). Ces traits relatifs à la fois au *fond* et à la *forme* sont identifiés par l'interprétation dans la construction du sens, la *signification*.

Tout texte relève en même temps de l'intertextuel, c'est-à-dire des rapports entre textes qui se sont construits à l'intérieur d'une culture à un moment donné de son histoire. Ainsi, tout texte appartient à un genre dans une culture donnée. Le genre codifie la production et l'interprétation du texte, il relève du social et non du système fonctionnel de la langue (par exemple, la recette de cuisine dont la textualité varie de culture à culture). L'analyse des textes implique donc la comparaison de textes de même genre, principe trop peu pratiqué dans la linguistique de corpus.

Le genre d'un texte numérisé peut être appréhendé par des informations externes : mise en page (choix de police, variété de couleurs, tableaux, colonnes, éléments multimédia, etc.) et mode de navigation (liens). On note ainsi que certains types de discours, par exemple les textes scientifiques, actualisent volontiers différents systèmes sémiotiques (diagrammes, illustration, etc.) qui produisent un effet d'objectivation de l'information, induisant un trait de catégorisation générique.

Ces formes sémiotiques, en partie non linguistiques sont interprétables par le corps du texte ou par un appareil de légendes ou de notes. Leur signifiance est assurée par leur liaison au textuel. Par exemple, une carte d'état major sans noms de lieux ni légendes n'est pas interprétable. Les courbes de niveaux sont des symboles conformes à un codage prédéfini par un méta-texte qui peut se situer hors du champ du document.. En effet, la relation

fondamentale d'interprétance assurée par le linguistique vis-à-vis des autres systèmes sémiotiques dans le document numérique, est elle-même d'ordre sémiotique et la langue est le seul système qui puisse se catégoriser et s'interpréter lui-même.

Il est probable que le document numérique donne une dimension nouvelle à la notion d'intertextualité. Celle-ci se prolonge dans l'inter-sémioticité. L'inter-sémiocité du document numérique qui intègre texte, écrit, image et son ne change pas la nature du *texte*, elle en accroît en revanche la complexité. Les rapports d'inclusion réciproque constituent un objet de recherche nouveau qui devrait mieux spécifier la relation entre sémantique et sémiotique.

2.2 Le *texte*, objet sémiotique complexe

Une sémiotique multidimensionnelle

Nous sommes donc devant une difficulté sérieuse puisque la linguistique, dont on pouvait attendre un éclairage, n'est pas en mesure, du moins aujourd'hui, de nous fournir une définition satisfaisante du concept de *texte*. Cette difficulté est encore accrue par le fait que celui-ci excède sa seule dimension linguistique. En fait, 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.)

Et, bien entendu, cette sémiotique multidimensionelle est l'essence même des documents multimédias dont on comprend maintenant que la séparation *forme*/*fond* relevée dans la section précédente est encore plus délicate. En effet, pour les documents non textuels, comme les objets sonores, vidéos, photographiques, comment caractériser leur *contenu*, c'est-à-dire l'invariant qui serait commun entre plusieurs objets considérés présenter la même chose ? Les règles associées à la caractérisation de cet invariant sont floues ; transcoder n'altère pas le contenu, sauf si la qualité est trop altérée ; en quoi passer un film de définition standard (SD) en haute définition (HD) concerne-t-il le contenu (on créée pourtant de l'information) ? En quoi retraiter (changer la couleur, le cadrage – par exemple la télévision diffuse en 4/3 des images tournées généralement en 16/9) une image permet de créer une nouvelle image, ou une nouvelle présentation de la même image ? Le droit d'auteur est attentif à ces questions, et y recherche des critères pour démarquer le piratage, la malfaçon de la création.

Cette réalité inter-sémiotique du *texte* n'est donc pas nouvelle, mais elle prend un relief particulièrement accru avec la numérisation. Pour l'expliciter, on pourrait par exemple adopter la définition suivante du *texte* : *un contenu symbolique aux formes sémiotiques variées, à la combinatoire infinie, dont l'agencement est susceptible de faire émerger une pluralité de significations s'agrégeant singulièrement à chaque lecture*. L'enjeu de l'analyse est en effet de reconnaître les différentes dimensions du *texte* dans les médias informatisés sans nécessairement les définir à partir des modèles qui nous sont familiers : modèle langagier, littéraire, livresque, etc., afin de mieux comprendre et éventuellement d'intervenir dans le débat sur les modèles XML repérés précédemment.

En tant qu'objet sémiotique, le *texte* peut être considéré :

- dans sa généralité comme *une configuration de signes de toute nature* ;
- dans une acception limitée au seul matériau langagier, abstraction faite des autres types de signes, comme *une suite de mots* ;

- dans une acception plus matérielle, comme *l'inscription sur un support déterminé* de l'une ou l'autre de ces constructions, avec les modes d'accès correspondants (imprimé, enregistrement, etc.) ;
- dans une approche technique, *comme une suite discrète de symboles* (par exemple une séquence alphabétique), éventuellement accompagnée de balises, qui pourra être traitée automatiquement.

Par exemple, le portail d'un musée constitue en lui-même un texte complexe : *a*) le linguiste y distingue le « texte » des pages, des « illustrations » qui l'accompagnent ; *b*) du point de vue ergonomique, il est différent d'être placé devant ce texte écrit, mais activable, plutôt que devant une brochure imprimée ou face à une personne qui nous ferait visiter les lieux ; *c*) enfin, du point de vue de l'informatique, certains fichiers avec lesquels le site a été confectionné (et seulement certains d'entre eux) sont des « fichiers texte », c'est-à-dire qu'ils contiennent une information portée par une chaîne de caractères. Plus généralement, un tel objet définit des rapports communicationnels de nature sociale : comment l'institution se met-elle en scène ? Quels sont les codes qu'elle utilise ?

Un « texte » sans consensus

Le jeu entre les diverses acceptions possibles de la notion de *texte* ne manque pas de susciter la discussion. On peut contester son emploi dans une perspective inter-sémiotique, telle qu'on l'utilise ici. Il garde, en effet, de sa source langagière, et plus particulièrement de la forme de l'écriture alphabétique à laquelle on le rattache le plus souvent, une teneur particulière qui surdétermine le raisonnement. Ainsi, il rend mal compte de documents ou parties de document non-écrits ou de documents immédiatement orientés vers l'action. Le terme de « texte » peut, en fonction de l'univers auquel il est spontanément associé, nous pousser à négliger systématiquement les qualités des différents médias ou substances de l'expression (la différence réelle, et non simplement superficielle, entre l'écriture, le son et les images) et à sous-estimer l'impact qu'ils ont sur la globalité d'un document.

D'autre part, la tension entre deux définitions du texte, en tant que formation matérielle existante et en tant que phénomène interprétatif, voire pour certains, en tant qu'ensemble d'invariants du sens d'une forme matérielle à une autre montre que le débat sur la *grammatisation* soulevé à la section 1 n'est pas encore épuisé.

Trois questions épistémologiques majeures restent ouvertes : 1) le statut donné dans nos analyses à la matérialité des formes, et notamment si une forme d'invariant sémantique peut subsumer cette matérialité ; 2) le rapport établi entre les formes signifiantes et les formes logiques ; 3) la place, déterminante ou limitée, faite à la réalité du réseau dans les phénomènes de transformations documentaires et textuelles, depuis l'idée que le réseau incarnerait une actualisation de l'archive telle qu'elle est définie par Foucault[11] jusqu'au souci affirmé au contraire de maintenir une distinction entre l'ordre des objets techniques et textuels et celui de la culture en tant qu'interprétation produite par les hommes.

On peut dire, pour ponctuer provisoirement ce débat, qu'interroger l'évolution du texte avec le numérique a permis de cerner des questions essentielles, mais que le concept de *texte*, central dans ce travail, reste essentiellement un objet controversé et à discuter dans l'avenir. Des termes comme « production sémiotique » ou « configuration » ont d'ailleurs été proposés pour éviter les difficultés ici décrites, tout en conservant la teneur de la discussion en cours.

[11]Foucault Michel, *L'archéologie du savoir*, Gallimard 1969.

2.3 La production textuelle

Le *texte* n'est pas seulement un objet sémiotique, c'est un objet fabriqué, artisanalement ou industriellement, 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.

Le texte en action

Poser la question du *texte* dans le régime technique des médias informatisés et des documents numériques, c'est réfléchir aux relations qui s'établissent entre la matérialité de ces objets textuels et les types d'investissements nécessaires pour faire de ces objets, non de simples choses, mais des objets organisés interprétables dans une société donnée. Investissement qui se situe nécessairement à la croisée de trois composantes toutes indispensables :

- la création de *formes repères* qui permettent l'anticipation et la reconnaissance d'organisations de la pensée, les cadres et les genres du texte, c'est-à-dire de modèles d'organisation qui, par-delà la singularité de chaque texte, rendent une forme textuelle générale reconnaissable ;
- l'exercice d'une association et d'une *dissémination des interprétations* qui redistribue en permanence les textes, par le travail de la reprise, de la réécriture, de l'« archive », auquel l'inscription informatique des textes donne une extension et une visibilité particulières ;
- le *processus de la réécriture ou de la réappropriation*, dans lequel le texte apparaît comme un objet toujours singulier mais toujours mis en série.

Le texte n'est pas un objet ponctuel, mais un ensemble associant une réalité matérielle (l'objet texte), des formes qui l'organisent (la textualité) et des moyens culturels pour le qualifier (les pratiques interprétatives).

Concrètement, pour définir les objets textuels du réseau, nous devons penser autant au bloc-notes ou au mode d'emploi qu'à l'œuvre littéraire ; pour évoquer le contexte de leur usage, nous devons intégrer l'écriture et la lecture, mais aussi l'annotation, la transformation, la duplication ; pour les relier à des enjeux intellectuels, nous devons imaginer des situations d'échange, de co-construction, de démantèlement de ces textes autant que des productions d'auteur.

L'interprétation d'un texte passe d'ailleurs par des procédés de réécriture qui en font un nouveau texte : travail qui, associant des hommes et des « machines à texte », produit de nouvelles formes, qui vont du texte alphabétique à diverses formes d'organisation visuelle, de listes, cartes, tableaux, etc. Ces opérations, qui inscrivent dans l'écriture la trace de lectures, contribuent à produire et stabiliser des représentations du sens des textes sur lesquels elles opèrent. C'est pourquoi ces objets intermédiaires sont essentiels pour comprendre la notion de *texte*.

D'autre part, comme il a été indiqué plus haut, nous devons mettre en relation ces dimensions du *texte*, sans pour autant les réduire aux objets qui ont pu, à une certaine époque, dans une culture dont nous sommes imprégnés, la culture livresque, les rassembler. Pour se repérer dans les mouvements en cours, l'analyse doit à la fois porter ses efforts sur la définition de la spécificité des objets plus ou moins inconnus auxquels nous sommes confrontés aujourd'hui, et sur une indispensable prise de repère à partir de ce que nous sommes capables de concevoir comme un *texte* – et qui reste, nécessairement, adhérent à l'ordre contingent des

textes que nous connaissons. La difficulté est accrue du fait que bien des phénomènes présentés comme nouveaux ont eu leur précédent dans l'histoire longue des documents : versionning *vs* palimpsestes, annotations des scribes et des copistes, etc.

Les trois dimensions précédemment définies (un objet matériel, l'organisation observable des signes sur cet objet et les modèles culturels qui le font reconnaître comme *texte*) restent essentielles. Le processus de l'échange continue de reposer sur l'émergence et le partage des formes symboliques, qui permettent de donner un caractère signifiant à des objets matériels. Mais les conditions de cette émergence et de ce partage tendent à changer.

Beaucoup de différences de détail distinguent les textes imprimés des textes numériques, mais la différence qui nous semble la plus essentielle a été soulignée dans la section précédente : la dissociation entre une forme logique et abstraite et une forme concrète et perceptible. Dans ces conditions, le texte ne circule plus sous la forme d'un objet relativement stable, associant un support matériel et une configuration formelle (ce qui conditionnait l'idée même de document). Il devient une réalité en permanence recomposée, qu'on peut apparenter à un *événement réitéré.* Ce qui se conserve n'est plus le texte donné à lire mais l'ensemble des modèles formels qui en représentent la 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.*

Les conséquences du design

D'autre part, le texte numérique n'est pas seulement un texte qui aurait été numérisé, c'est un texte régi par des programmes qui conditionnent les gestes d'écriture, de lecture, de manipulation. Les programmes écrivent de plus en plus à l'avance, non seulement les contenus textuels, mais les pratiques qui leur sont associées. C'est ce qui donne son importance à l'intervention de la réflexion critique en amont des processus de design.

De ce point de vue, toutes les modélisations sur lesquelles repose le texte informatisé sont porteuses de sens et génératrices d'effets de pouvoir. Elles sont la traduction du lien entre les imaginaires (de la communication, de l'information, du sens) qui habitent les concepteurs de systèmes et des procédures par lesquelles ces systèmes sont conçus – y compris les contraintes et objectifs liés à la dimension économique de cette production.

La numérisation des contenus et l'informatisation de leur exploitation a pour effet de modifier profondément les conditions dans 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 éclatent 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.

Soulignons deux points : d'une part, dans cette structure la question de l'intertextualité (la façon d'établir des liens entre les textes, liens matériels et liens de l'interprétation) se pose dans des termes nouveaux, cette intertextualité pouvant être appareillée et formalisée autrement qu'elle ne l'était dans l'espace du livre et de la note ; d'autre part, le texte informatisé se prête particulièrement bien au morcellement, à la manipulation et à la

réécriture, produisant, dans certains cas, des textes fragmentaires, procéduraux, très liés à l'action, qui sont fort éloignés du discours continu auquel nous a habitués le livre et qui ne doivent pas être considérés seulement comme des objets à interpréter (ce qu'ils demeurent toujours) mais aussi comme des objets à manipuler.

Espace de l'écran, temps du signal

Sans prétendre épuiser le sujet, nous voudrions conclure cette section par une proposition pour avancer dans une meilleure maîtrise du texte numérique en reprenant plusieurs pistes énoncées par la prise en compte d'une dimension moins analysée et pourtant, selon nous, essentielle : la temporalité.

Le numérique associe deux dimensions, autrefois réservées à deux médias distincts : l'inscription sur un support (autrefois prioritairement le papier, aujourd'hui l'écran) passe par une dimension de persistance spatiale ; le flux du signal (autrefois prioritairement audio-vidéo, aujourd'hui multimédia y compris écrit), suit une dimension temporelle. Par cette facilité, le texte numérique acquiert une fluidité plus grande, sans pour autant perdre ses qualités de persistance visuelle. Ainsi, un nombre considérable d'utilisations de l'écrit ou de l'image et du son, souvent déjà existantes mais bridées par les frontières spatiales ou temporelles, sont libérées et nous les voyons exploser sous nos yeux. Les exemples sont déjà multiples, multiformes et de nature très diverse, citons sans souci d'ordre ni d'exhaustivité : le calcul en direct des pages Web et leur actualisation continue, l'insertion de séquences vidéos dans du texte, la réalisation de schémas mouvants à partir de formules aux variables changeantes, les modélisations 3D en temps réel multipliant les points de vue, les innombrables déclinaisons des jeux-vidéos mélangeant les modes d'expression, le développement très rapide de Weblogs de plus en plus multimédias, réactifs et reliés, le chapitrage ou les annotations sonores des films dans les DVDs, les échanges sur téléphones de 3^{e} génération, les systèmes mélangeant documents et téléprésence, etc.

Reste que de nouveaux problèmes de perception se posent. L'écrit permet à la parole (entre autres) de persister. L'interaction permet alors le passage de cet axe du temps (flux) à un axe spatial (persistant) par le biais du geste qui, lui, s'exécute selon ces deux axes. Nous sommes habitués à cette situation, issue d'une évolution de plusieurs millénaires. Mais à partir du moment où la trace devient éphémère, juste l'activation lumineuse d'un écran par un signal, elle retrouve une dimension temporelle immédiate, créant une situation inédite à l'échelle de l'humanité. Tout le problème est de comprendre comment nous pouvons nous repérer dans ces changements. Comment et à quel prix pouvons-nous nous adapter à ce nouvel environnement symbolique ? Ou, pour parler comme les chercheurs en psychologie écologique, quelle « affordance[12] » faut-il construire pour que nous ne perdions pas nos repères cognitifs ?

---

[12]Définition :
Ensemble des aspects psychologiquement pertinents et significatifs de l'environnement d'un être vivant.
Note(s) :
Selon James Jerome Gibson (1979).
Les affordances sont des propriétés réelles des objets qui peuvent avoir une valeur utile pour leur observateur. Elles portent sur ce que l'on perçoit en fonction de ce sur quoi on peut agir. Ainsi, nous percevons qu'un petit objet est préhensible, alors qu'un grand ne l'est pas. Les affordances sont déterminées conjointement par les caractères physiques d'un objet et par les capacités sensorielles, motrices et mentales d'un être vivant. Pour un même objet, elles diffèrent d'une espèce à l'autre. Ainsi, un caillou peut être perçu comme un presse-papiers, l'élément d'un jardin de rocaille ou un marteau (Office québécois de la langue française).

Suggérons qu'une façon de se reposer la question est peut-être de partir de l'activité de lecture-écriture ou de reconstruction des textes numériques en combinant l'héritage de l'inscription sur papier et celui de l'audiovisuel.

Les images et les sons du cinéma ont été construits et assemblés à partir de fragments dispersés, de ressources éparpillées sur les bandes magnétiques ou supports numériques. Le « texte » ainsi construit a sa temporalité propre, souvent éloignée de la temporalité naturelle des événements qu'il relate, y compris par exemple dans les fragments sonores les plus élémentaires. Le spectateur, pour sa part, n'accède au contenu qu'à travers une consultation temporelle, qu'il ne maîtrise pas mais où le flux de sa conscience se synchronise avec le flux des images et des sons. En regardant le film ou un document audiovisuel, le spectateur se raconte une histoire ou déroule un raisonnement, celle ou celui qu'il est en train de voir et entendre, à partir de la succession temporelle contrainte d'images et de sons qui lui est proposée. Le spectateur reconstruit le sens de ce qu'il voit en en effectuant une synthèse temporelle au cours de la projection, au fur et à mesure que celle-ci progresse.

Dans la lecture sur papier, au contraire, le lecteur est libre d'explorer la surface de la page à sa guise, en choisissant sa propre temporalité, en faisant des retours en arrière, en sélectionnant les passages qui l'intéressent, en se rendant, grâce à l'index, à toutes les occurrences d'un même mot, etc. L'histoire du livre, notamment par le passage du rouleau au codex, a contribué à détacher le texte du fil continu et temporel de la parole pour en faire un objet de plus en plus « tabulaire ». C'est ainsi qu'il faut aussi comprendre la dimension non-narrative d'une quantité croissante de textes.

Pour analyser correctement la nouveauté d'un texte mu par un signal et néanmoins inscrit sur un écran, il faut peut-être marier les deux approches. Au-delà du nécessaire travail sur l'organisation spatiale des pages, il faudrait ajouter pour l'instrumentation numérique des documents 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 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.

Cette proposition n'est pas contradictoire avec la pratique de plus en plus courante du « zapping », ou le saut rapide d'un texte à un autre. Elle suggère au contraire de prendre au sérieux l'action de consultation de l'internaute et d'en repérer le sens pour que les affordances du document numérique, construites par les ingénieurs, permettent au spectateur-lecteur-auteur-acteur de construire son propre texte signifiant à partir des opportunités qui lui sont ouvertes.

3 Invariance et transformation

On pourrait considérer le problème de la représentation du *texte* comme, sinon résolu, du moins renvoyé à plus tard, si, grâce au deuxième modèle indiqué en section 2, celui des Schémas, il suffisait de définir les invariants des modèles de documents pour pouvoir les reconstruire à partir d'un *contenu* géré par des bases de données. Ce n'est pourtant pas réellement possible aujourd'hui, si l'on désire garantir la fiabilité du résultat, car les processus de déconstruction/reconstruction, en tant qu'éléments dynamiques, échappent à une caractérisation aussi rigoureuse que les Schémas, éléments descriptifs statiques des modèles.

Jusqu'à une période récente, les termes de « texte » et de « document » étaient pratiquement synonymes. Le texte étant fixé sur un support qui lui procurait stabilité, le statut

documentaire relevait du juridisme ou de l'institution, mais ne concernait pas (plus précisément, ne concernait plus) les ingénieurs et la technique. Mais, nous l'avons largement développé dans le premier texte de Roger[13], le support et l'inscription ne sont plus aujourd'hui inséparable, bien au contraire.

Pour autant, la stabilité documentaire est essentielle à bon nombre d'usages et pour la lecture nous avons besoin de retrouver des repères familiers, c'est-à-dire des éléments perceptibles stables. Aussi, il est possible de raisonner par analogie en considérant qu'un document est défini par les éléments qui lui procurent une stabilité. Autrement dit, le statut documentaire serait spécifié par un certain nombre d'invariants qui régissent sa cohérence au sein des différentes formes qu'il peut revêtir.

Cette cohérence documentaire est plus que jamais centrale, parce que menacée, comme le montre la montée des préoccupations juridiques sur l'intégrité des textes. Elle devrait donc être préservée par la transformation informatisée alors même que, de part sa nature algorithmique, elle tend à introduire de nombreux aléas et limitations propres à toute activité partiellement ou complètement automatisée. Autrement dit, les systèmes techniques ne doivent pas altérer la cohérence documentaire.

Nous proposerons donc dans cette section une réflexion sur la transformation documentaire informatisée, en prenant soin de ne pas confondre transformation et traduction. Dès que la transformation concerne la dimension linguistique du texte lui-même, où l'objectif est d'en préserver la signification, les transformations automatisées trouvent leurs limites, ne serait-ce que parce qu'on ne sait pas repérer d'invariants fiables et universels.

Le but d'une telle réflexion est de montrer que la cohérence documentaire n'est pas seulement un enjeu important et sous-estimé, mais aussi qu'elle requière la conception de modèles dynamiques de transformations, symétriques des modèles statiques de structuration et contenu que sont les Schémas.

3.1 La nécessaire intégrité juridique

Les questions juridiques posées par le numérique sont une illustration éclairante des enjeux de la transformation des documents. Trop souvent la dimension juridique du document est réduite aux aspects économiques de la propriété intellectuelle. En réalité, c'est bien le statut de document qui est en cause et, plus précisément, la relation et la différence entre *texte* et *document* dans le passage d'un format ou d'un support à un autre. Prenons rapidement deux exemples :

- La communication, en particulier la communication institutionnelle, d'un organisme à un autre ou d'un particulier à un organisme, s'appuie sur de nombreux documents. Autant le numérique peut faciliter les échanges, autant il est indispensable d'assurer une préservation mutuellement acceptée d'un contenu des documents échangés pour éviter tout malentendu qui pourrait avoir d'importantes conséquences sociales. Bien souvent ici le terme d' « informations » est considéré comme synonyme de *contenu.* Un document devra préserver les informations, ses différentes présentations étant homologues pourvu que cette règle soit respectée. Le Québec par exemple, afin d'assurer la sécurité juridique des communications, s'est doté d'une loi qui définit une notion de document indépendamment des supports[14] et précise les règles de transfert, conservation, consultation et transmission.

[13] *Opus cite*, *cf.* note 2.

[14] Extrait de la Loi du Québec concernant le cadre juridique des technologies de l'information. Adoptée le 21 juin 2001 :

- Le droit moral d'un auteur s'appuie sur l'intégrité de son oeuvre qui, très souvent, se matérialise par un document dont la forme canonique doit être préservée. Ceci n'implique pas nécessairement des restrictions de diffusion. Sous l'impulsion du juriste L. Lessig par exemple, une licence s'inspirant des pratiques en cours dans les communautés du logiciel libre a été conçue pour s'appliquer à la diffusion gratuite de textes sur le Web, baptisée : « *Creative Commons*[15] ». Mais le respect de la création suppose toujours le respect de sa représentation.

L'une et l'autre initiatives, qui visent à favoriser les échanges numériques, insistent sur l'intégrité du *document*, c'est-à-dire l'importance et le maintien du statut de document pour un texte donné. Ce statut est essentiel pour que le texte ait valeur de preuve (loi du Québec) ou qu'il puisse être considéré comme l'œuvre d'un individu (*Creative Commons*).

Sans que ce soit explicitement précisé dans l'un et l'autre cas, le raisonnement tenu suppose que l'on puisse définir un certain nombre d'invariants qui donne à un objet déterminé un statut documentaire, quelles que soient les transformations qu'il subisse.

Mais on voit bien qu'il rend délicat la prise en compte de « documents » évolutifs et collectifs, qui restent encore souvent des textes flottants au statut incertain. Il y a là un défi conceptuel important souligné par la mise en place des réseaux et outils numériques, qui ne peut se résoudre, de notre point de vue, que par une conjointe élucidation de la notion de document et de sa prise en compte dans des outils adaptés.

3.2 Pour une élucidation de la transformation documentaire

La notion d'« invariant »

La photocopie est un exemple simple de transformation documentaire automatisée. En tant qu'opérateur de transformation, la propriété d'invariance qu'il doit satisfaire est la préservation de l'information graphique, de façon à pouvoir garantir la lecture correcte des pages reproduites. Peu de personnes se soucient de savoir au-delà de combien de photocopies en cascade cette information est perdue. Et pourtant, cette caractéristique d'invariance peut revêtir une grande importance dans certains contextes, interdisant par exemple l'utilisation de procédés xérographiques pour des systèmes d'archivage à très long terme, ou simplement en limitant le nombre de chaînons dans la transmission de copies en cascade.

Le problème central des opérateurs transformationnels qui se développent massivement aujourd'hui reste fondamentalement le même mais se décline de manière plus complexe et plus subtile. En effet, d'une part, les formes documentaires se sont multipliées et, d'autre part, elles ont évoluées qualitativement vers une mixité et une abstraction croissante, comme on l'a dit dans les sections précédentes. Par ailleurs, les opérations de transformation ont suivi une évolution encore plus forte, suivant la progression générale de l'informatique et la plasticité des traitements qu'elle autorise. Lorsqu'un programme réorganise au vol une page

---

« LA NOTION DE DOCUMENT
3. Un document est constitué d'information portée par un support. L'information y est délimitée et structurée, de façon tangible ou logique selon le support qui la porte, et elle est intelligible sous forme de mots, de sons ou d'images. L'information peut être rendue au moyen de tout mode d'écriture, y compris d'un système de symboles transcriptibles sous l'une de ces formes ou en un autre système de symboles.
Pour l'application de la présente loi, est assimilée au document toute banque de données dont les éléments structurants permettent la création de documents par la délimitation et la structuration de l'information qui y est inscrite. »
http://www2.publicationsduquebec.gouv.qc.ca/home.php#

[15] http://fr.creativecommons.org/

Internet pour l'adapter à un lecteur dont on a modélisé le profil, quels sont les invariants souhaitables pour caractériser cette transformation? Doit-on « filtrer » le contenu et conserver la présentation ? Doit-on modifier le style des sections et des tables, doit-on calculer et ajouter un glossaire personnalisé ou une liste de références et d'hyperliens adaptés aux connaissances du lecteur ? Est-on seulement sûr que le résultat sera correctement visualisable ?

A vrai dire, la question n'est pas aujourd'hui posée en ces termes par les développeurs, et ses enjeux ne sont pas encore universellement reconnus. Plus encore, aucun programme ne peut garantir une réelle fiabilité dans la réalisation des processus transformationnels de ce type, et cela se voit peu dans le cas de notre exemple car les butineurs ont été conçus pour offrir une tolérance maximale aux erreurs.

Il est probable que cette situation ne durera pas éternellement, ne serait-ce que dans la mesure ou des documents à la finalité de plus en plus riche et socialement « importants » sont en jeu. Mais la question ne saurait trouver les réponses du seul côté de l'informatique. Il y a là un enjeu important pour une recherche transdisciplinaire.

Les trois niveaux de validité

Ces aspects, directement liés à la virtualisation des formes documentaires, appellent une réflexion théorique quant à la qualité des opérateurs transformationnels. En tout premier lieu, la notion de propriété formelle et opérationnelle est nécessaire pour qualifier la nature exacte d'un document en tant qu'ensemble d'informations organisées. Un premier pas est déjà franchi dans cette direction, avec le concept de validation automatisée d'un document, hérité de SGML et provenant des travaux antérieurs sur les grammaires formelles. Dans ce modèle, fondamental dans le paysage esquissé par le consortium W3C, un document possède plusieurs niveaux de validité formelle croissante[16] :

1. la correction formelle (« *well-formedness* ») correspond aux propriétés lexicales et syntaxiques minimales qu'une succession de caractères doit vérifier pour être considérée comme un document manipulable en machine.
2. Le second niveau, la validité formelle (« *schema conformance* ») définit la conformité à une grammaire (appelée le schéma du document) qui décrit les constituants logiques internes au document et les règles régissant leur organisation.
3. Le troisième niveau, la validité sémantique, est plus flou, mais vise à établir et vérifier des propriétés du contenu documentaire, tels que l'utilisation de dates ou d'acronymes correctement formés, ou encore la cohérence des références intra-documentaires, par exemple.

En fait si, ces propriétés sont statiques – c'est-à-dire vérifiables sans transformer le document –, elles n'en sont pas moins justifiées par des besoins transformationnels. Ainsi, la correction formelle autorise la transformation d'un segment d'octet en arbre (analyse structurelle), la conformité à un Schéma simplifie les transformations ultérieures de l'arbre en éliminant les incertitudes sur la localisation et la nature des informations. Par exemple, pour un document conforme au Schéma « rapport technique », on peut garantir qu'un résumé suit nécessairement la liste des auteurs, qui, elle-même, est non vide et suit nécessairement le titre.

---

[16] *Extensible Markup Language (XML) 1.0 (Third Edition)* ***W3C Recommendation*** *4th February 2004, François Yergeau, Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler*
Le vocabulaire est ici un vocabulaire d'informaticiens : les termes « syntaxes », « grammaire » ou « sémantique » n'ont de sens que dans une perspective opérationnelle de construction ou de manipulation de langages informatiques.

Il s'agit seulement d'un premier pas, car en l'état actuel de la technologie, peu de langages transformationnels sont capables de garantir l'invariance du niveau 1 (le résultat ne sera pas nécessairement bien formé) et aucun ne préserve le niveau 2, hormis quelques prototypes peinant à sortir des laboratoires. Ainsi, une transformation se proposant de transcrire un formulaire XML (conforme au Schéma standard XForm) en un document SVG susceptible d'être visualisé avec une grande qualité graphique pourra totalement dysfonctionner dans des cas licites mais non prévus par les développeurs. Plus grave encore que la violation des corrections ou validité formelles, il se pourrait que des informations sensibles disparaissent du résultat final (ou au contraire, apparaissent alors qu'elles ne le devraient pas !).

L'exemple du calcul d'un index met en évidence l'importance des propriétés d'invariance de niveau 3 (plus sémantique). Les entrées constituant un index (des « expressions » donc) doivent être présentes dans le texte du document. Plus précisément, les références de pages associées, ou les liens correspondants, aux entrées de l'index doivent exactement correspondre aux pages, ou emplacements, dans lesquelles on y trouvera leurs occurrences. L'invariant caractéristique ici est l'association d'une expression avec sa page, ou son ancre, couple qui doit être préservé dans toute entrée de l'index afin de garantir son efficience.

Pour un modèle de transformation spécifique au document

Le problème de fiabilité des applications informatiques n'est pas nouveau mais il prend une connotation très particulière dans le domaine documentaire. Pourtant le défi technique est abordé aujourd'hui au travers de la problématique des langages de programmation généralistes, ce qui place la difficulté à un fort haut niveau puisque les programmeurs ne bénéficient d'aucune facilité liée aux particularités de l'objet document. Ce dernier est manipulé comme une donnée parmi d'autres, mais particulièrement complexe. Dans le cas de langages spécialisés, il s'agit d'intégrer le document au sein des modèles de données sous-jacents au langage et d'aborder la préservation des propriétés du document comme on traite la vérification des types de données. Cette approche s'avère difficile et les prototypes de recherche se montrent difficilement transférables aux acteurs industriels.

Une autre voie, *a priori* moins ambitieuse, serait de prendre en compte les spécificités du document au sein d'un modèle de transformation général, probablement moins expressif qu'un langage de programmation mais proposant de manipuler les invariants comme des entités de premier ordre, c'est-à-dire syntaxiques, et offrant une combinatoire bien maîtrisée. Par exemple, un invariant peut être défini comme une clause logique qui doit être vérifiée avant et après la transformation. Cette clause, si elle est nommée, peut être réutilisée, ou composée avec d'autres.

L'avantage résiderait dans la mise en avant de l'invariance comme une préoccupation centrale lors de la conception des opérateurs transformationnels. Cette approche se justifierait assez bien car, 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 essentiel.

La question fondamentale qui se cache derrière l'approche compositionnelle est de savoir s'il existe un jeu de transformations élémentaires susceptible d'être bien caractérisé par rapport aux propriétés des documents, et pouvant se combiner de manière efficace pour être en mesure d'offrir une puissance expressive réaliste. De plus, ce modèle doit rester suffisamment abstrait pour s'affranchir des détails d'implantation et rester au niveau de généralité caractéristique d'une bonne « théorie », en ce sens qu'elle doit permettre de modéliser, explorer et comprendre le comportement des architectures de transformation sur des documents potentiellement complexes, sans recourir à des hypothèses d'exécution trop précises.

Enfin, il faut ajouter ici que les transformations documentaires les plus riches font immanquablement appel à des processus interactifs dans lesquels des intervenants humains coopèrent avec des procédés automatisés. De telles étapes de transformations, de part leur importance, devraient pouvoir se modéliser dans le cadre d'une réflexion sur l'invariance documentaire ; en ce cas, la finesse du niveau de description des interactions sera probablement la question centrale à aborder. Dans le cas où l'intégration des dimensions interactives se montrerait trop ardue, il conviendrait de ne pas complexifier excessivement le modèle théorique pour tenter de le généraliser, mais plutôt de restreindre son applicabilité aux transformations automatiques, et de définir un autre cadre spécifique, plus favorable aux transformations interactives. Cette seconde option n'interdirait en rien d'envisager une compatibilité des modèles.

La réponse à la question ne saurait être confiée aux seuls ingénieurs ou chercheurs en informatique. Ses termes ne sont pas sans rapport avec les problèmes posés précédemment dans la Section 2.3 dans une perspective de sciences humaines et sociales. La rencontre des deux points de vue pourrait déboucher sur des perspectives in-envisagées jusqu'à présent.

Dès à présent, il nous semble qu'en ce qui concerne la « profondeur » des invariants, c'est-à-dire de leur rapport à l'axe *forme*/*contenu*/*signification* du document tel que présenté à la section 1.2, une position modeste focalisée sur la forme « grammatisée » et le contenu « opératoire » tel que décrit par des Schémas paraît être la plus réaliste. En effet, la problématique générale de la transformation des textes, en rapport avec des invariants de « signification », dépasse très largement celle de l'invariance documentaire, comme la section suivante le montre.

3.3 La nécessaire « trahison » du traducteur

Contrairement à l'homologie que nous avions trouvée entre une problématique juridique et celle, technique, de la transformation documentaire, la subtilité du passage d'une langue à une autre interdit tout rapprochement avec le problème précédent. Pourtant, l'arrière-plan de la problématique de la traduction est devenu en quelque dix ans celui de l'accès électronique à des collections étendues de textes et à leur traduction. Il est désormais possible d'étudier sur de larges échantillons des phénomènes linguistiques récurrents dans le passage d'une langue à l'autre. Toutefois, les études sur corpus privilégiant aujourd'hui l'aspect quantitatif (textes mesurés en kilo-octets) sur le qualitatif (critères de caractérisation et de différenciation des textes tels que le genre, le discours, le registre, etc.), la complexité des opérations mises en œuvre dans la traduction échappe souvent aux analystes qui imaginent que des segments équivalents de textes sources et cibles affichant une représentativité jugée suffisante en corpus ont automatiquement statut de ressources traductionnelles réutilisables.

Une approche de la traduction qui se résume à la fourniture de ressources extensives (lexicales ou phrastiques) pour les mémoires de traduction est insuffisante, même si l'on s'en tient à des domaines relativement restreints, à moins bien entendu qu'il ne s'agisse de textes en langage contrôlé au lexique fortement normalisé, à l'instar du système canadien METEO.

Le point aveugle des approches technologiques de la traduction concerne la compréhension de ce qu'est un texte et de la manière dynamique dont se construit son sens. Le traducteur professionnel résout les problèmes locaux en fonction d'une appréhension globale du texte, en combinant les différents niveaux du texte. Il s'ensuit que le texte traduit est toujours une transformation de l'original. Contrairement aux idées reçues, la *trahison du traducteur* résulte d'une fidélité captive, d'un assujettissement aux mots de l'original. Le traducteur ne raisonne pas en termes de traduction littérale mais d'équivalence, le processus de traduction étant délinéarisé puisqu'il se fait par va-et-vient entre les différents niveaux du texte original et par comparaison avec d'autres textes. Tout texte – y compris scientifique ou technique – s'inscrivant dans l'espace culturel d'une société donnée se constitue à partir d'autres textes (intertexte) et comporte une part d'interprétation. La part de création que suppose toute traduction passe par des prises de décision qui ne relèvent pas d'une axiologie explicite. On peut imaginer que les possibilités d'interactivité offertes par le Web seront mises à profit pour élaborer des systèmes d'évaluation collective (notation et pondération) pour améliorer « l'art de traduire ».

4 La tentative de dépassement du Web sémantique

Le troisième et dernier modèle implicite de construction des documents numériques tente d'intégrer les connaissances. Il relève du mouvement du Web sémantique, initialisé et popularisé par Tim Berners-Lee, et s'appuie notamment sur ce qu'il est convenu aujourd'hui d'appeler dans la communauté de l'ingénierie des connaissances des « ontologies ».

Il s'agit cette fois clairement d'entrer dans le *contenu* pour y appliquer un raisonnement. Plus précisément, il s'agit de construire un métalangage, fondé sur les ontologies, représentant de façon formelle le *contenu* des documents qui pourra donc servir de base à des modélisations informatiques. Bien des travaux se mènent dans ce sens. Les résultats sont significatifs pour des communautés très structurées.

Mais l'ambition est plus grande, peut-être démesurée. Pour conjurer les risques de formatage de la pensée, ou simplement d'incohérence, qu'elle contient, il peut être utile de la comparer à d'autres tentatives passées de simplification radicale du vocabulaire.

4.1 Relativité et formalisme des ontologies

Les ontologies sont apparues au début des années 90 dans la communauté de l'ingénierie des connaissances, dans le cadre des démarches d'acquisition des connaissances pour les systèmes à base de connaissances (SBC). Faisant suite aux systèmes experts qui séparaient une base de connaissances « déclarative » et un moteur d'inférence « procédural », les SBC proposaient alors de spécifier, d'un côté, des connaissances du domaine modélisé et, de l'autre, des connaissances de raisonnement décrivant les règles heuristiques d'utilisation de ces connaissances du domaine. L'idée de cette séparation modulaire était de construire mieux et plus rapidement des SBC en réutilisant le plus possible des composants génériques, que ce soit au niveau du raisonnement ou des connaissances du domaine. Ces dernières précisant tout ce qui a trait au domaine.

Dans ce contexte, les chercheurs ont proposé de fonder ces connaissances sur la spécification d'une « ontologie », ensemble structuré par différentes relations, principalement la subsomption[17], des objets du domaine.

Les ontologies sont développées dans un contexte informatique – que ce soit celui de l'Ingénierie des connaissances, de la gestion et des systèmes d'information ou du Web sémantique – où le but final est de spécifier un artefact informatique et non de réaliser une représentation du monde. Dans le même temps, l'objectif de cet artefact est bien une gestion plus performante, pour le couple homme-machine, de connaissances qui sont une représentation du monde. Cette tension entre un système technique et une représentation fait à la fois l'ambiguïté et l'intérêt de la tentative. C'est une manifestation de la *grammatisation* repérée en Section 1.4. Pour en atténuer les contradictions potentielles, il convient d'insister sur les caractères relatif et formel des ontologies.

Ontologies et point de vue sur le monde

La communauté de préoccupation avec la proposition de modélisation du monde de Aristote qui a créé le concept – mais pas le nom – explique l'utilisation du mot « ontologie ». Cet emprunt a souvent été reproché aux informaticiens de l'Ingénierie des connaissances, d'autant que le caractère universaliste des ontologies de l'informatique est des plus discutés.

La science a toujours eu pour premier but de repérer et classifier les objets du monde pour les comprendre, comprendre leur fonctionnement et leur genèse. Les classifications ainsi construites sont des taxinomies. Elles comportent la classification elle-même et les critères qui la gouvernent. Les motivations des classifications ont évolué dans le temps.

Prenons l'exemple des taxinomies en sciences naturelles pour éclairer cette question. En botanique au début du XVI^{e} siècle, les critères de classification reposaient sur les effets supposés bénéfiques pour l'homme. Durant ce siècle et le suivant, ces classifications évolueront vers des critères liés aux organes de reproduction puis des critères floraux distinguant les végétaux supérieurs et inférieurs. Plus près de nous, au XX^{e}siècle, les classifications effectuées jusque-là en suivant le système de Linné ont été remises en cause (250 ans plus tard quand même) : les théories de Darwin sont apparues un siècle après Linné et les espèces sont alors considérées comme des entités provisoires. À l'opposé d'une conception statique du monde, les biologistes pensent que les espèces représentent des entités mouvantes qui évoluent dans le temps.

Prenons l'exemple des mammifères dans le groupe des vertébrés. « Allaiter ses petits grâce à des glandes lactéales » et « avoir un tégument à poils » est, pour les 3500 espèces qui constituent à ce jour la classe des mammifères, toujours vrai pensons-nous. Ces deux caractères seraient l'essence même des mammifères et donc définitoires. Pourtant, au regard d'une taxinomie phylogénétique et en essayant donc de reconstituer l'histoire paléontologique des mammifères, les choses se compliquent avec la découverte de fossiles reptiles présentant des caractères prémammaliens. C'est ainsi que l'origine reptilienne des mammifères a été affirmée et qu'il a fallu rechercher, du côté du squelette, de nouveaux caractères de différenciation entre les reptiles et les mammifères au cours de l'évolution. On a là un exemple illustratif de la remise en cause de la taxinomie des vertébrés au XX^{e} siècle Ainsi décider du définitoire et du contingent n'est pas toujours évident même dans ce domaine

[17] La subsomption est, par définition, l'action de penser un objet comme appartenant à un ensemble (Le Grand Robert). Elle est un des modes privilégiés pour modéliser – et penser – l'organisation du vivant. Elle est retenue de la même manière pour fonder et organiser les modèles des artefacts informatiques.

pourtant ancien. En fait, les critères de classification dépendent des buts poursuivis et n'ont rien d'immuables.

Rapporté au niveau des modélisations dans le cadre des artefacts informatique, des expérimentations – p. ex. en médecine ou dans des domaines techniques – montrent que la construction d'une ontologie suppose des choix, des engagements ontologiques qui sont justifiés par l'application à développer et le point de vue sur le monde qu'elle réifie. Par exemple, dans le cas d'un SBC en médecine, il a pu être nécessaire de conserver et modéliser deux points de vue sur les maladies, les maladies comme processus physiopathologique et les maladies comme état pathologique. Là où la souplesse de la langue permet d'utiliser un seul mot et de s'y retrouver grâce au contexte d'énonciation, la conceptualisation nécessite la prise en compte de deux points de vue explicitement et logiquement différenciés.

Ainsi le but premier des ontologies – et toujours poursuivi par certains –, proposer une modélisation générique et universelle du monde, nous parait discutable :.

Formalisme et définition

Une ontologie en informatique doit donc servir dans un artefact informatique. De plus, dans des applications complexes, elle peut être la base de systèmes d'inférences. Elle doit, pour cela, être définie formellement au sens où le langage dans lequel elle est décrite doit être muni d'une sémantique formelle qui décrit les manipulations autorisées. Dans le contexte du Web sémantique, ce langage est OWL (*Ontology Web Language*) et il respecte le formalisme des « logiques de description ».

Il est alors possible de proposer une définition complète de ce qu'est une ontologie :

> **Ontologie**[18] : *Une ontologie implique ou comprend une certaine vue du monde par rapport à un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts – e.g. entités, attributs, processus –, leurs définitions et leurs interrelations. On appelle cela une conceptualisation.*
>
> *[...]*
> *Une ontologie peut prendre différentes formes mais elle inclura nécessairement un vocabulaire*[19] *de termes et une spécification de leur signification.*

Cette définition permet de préciser deux contraintes qui s'imposent successivement au concepteur d'ontologies : une ontologie est bien une conceptualisation avec toutes les difficultés que cela implique et doit être par la suite utilisée dans un artefact informatique dont on veut spécifier le comportement ; l'ontologie devra également être une théorie logique pour laquelle on précisera le vocabulaire manipulé.

Ainsi, après avoir mis en avant le caractère subjectif – c.-à-d. lié à un point de vue – et donc sujet à interprétation d'une ontologie et le caractère formel, indispensable, soulignons, une nouvelle fois, la tension que subit cet artefact : objet de consensus pour les humains qui l'interprètent dans le contexte de leur activité et objet formel permettant son exploitation par un ordinateur, il doit permettre de relier le contenu exploitable par la machine à sa signification pour les humains. Les deux domaines n'étant pas soumis aux mêmes logiques, leur articulation est nécessairement en tension permanente.

---

[18] Cette définition est tirée des travaux de T. R. Gruber qui a été le premier à parler d'ontologies en informatique (« A translation approach to portable ontology specifications », in *Knowledge Acquisition*, 5, 1993, p. 199-220).

[19] « Vocabulaire » est utilisé ici tel qu'il apparaît dans le texte. Il doit être compris dans un sens logique et être vu comme le vocabulaire des expressions manipulées par une théorie logique.

4.2 Le « Cake » de T. Berners-Lee

Le Web sémantique est vu par son concepteur – Tim Berners-Lee – comme une extension du Web qui le transformera en un vaste espace d'échange de ressources entre êtres humains et machines permettant une exploitation, qualitativement supérieure, de grands volumes d'informations et de services variés. Les utilisateurs seraient alors déchargés d'une bonne partie de leurs tâches de recherche, de construction et de combinaison des résultats, grâce aux capacités accrues des machines à accéder aux *contenus* des ressources et à effectuer des raisonnements sur ceux-ci. Ceci nécessite une représentation sémantique des contenus via les ontologies, c'est-à-dire le respect d'une sémantique formelle.

Tout ceci n'est possible que si l'ensemble des contributeurs au Web sémantique respectent une infrastructure commune[20]. Cette infrastructure est souvent présentée à travers le « Cake » de Tim Berners-lee (*cf.* figure 2). Dans ce schéma qui veut décrire les besoins du Web sémantique, on trouve une organisation en couches de différents langages :

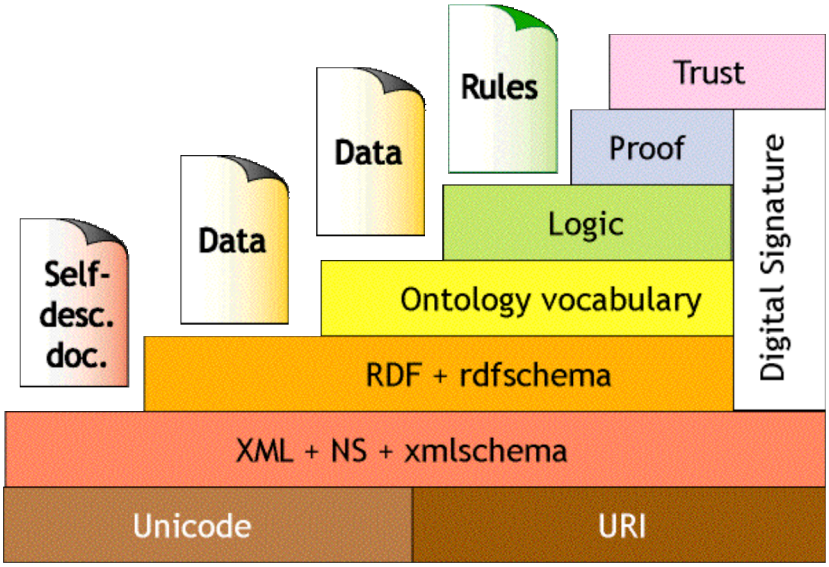

Fig. 2 – Les couches des langages du Web sémantique selon le W3C[21]

- Dans le premier niveau, la norme UNICODE de codage des caractères qui s'impose pour toutes les langues du monde et le mode d'adressage des pages (*Unified Resource Identifier*) du Web déjà en œuvre pour le Web normal forment le socle du Web semantique.
- La couche suivante, « XML + NS + xmlschema », fournit les langages nécessaires à l'écriture des documents sur le Web, à savoir, le langage XML, les espaces de noms (NS pour *name space*) pour permettre de former des groupes de balises indépendants et unique (en préfixant les noms des balises) et les Schéma XML qui permettent de décrire la structure d'un document XML et d'en typer son contenu. À ce niveau, on a les bases nécessaires (en termes de langages normés) à la mise en place et l'échange de documents textuels sur le Web.

[20] Dans le cas du Web, cette infrastructure repose sur un réseau d'ordinateurs et le respect du protocole de transport HTTP, de l'adressage via les *Unified Ressources Locator* (URL) et du langage HTML.

[21] http://www.w3.org/2000/Talks/1206-xml2k-tbl/

- La couche « RDF + rdfschema » correspond à un changement de paradigme annoncé : elle permet de passer d'un *document* à des *données* (de *selfdescriptive document* à *data*). Cette couche permet de décrire des concepts et des relations entre eux (avec des triplets, {concept, relation, concept}) et donc de décrire des informations complexes sur le monde. Ces informations peuvent être associées à un document grâce à ce même langage et être des métadonnées – appelées aussi annotations formelles – de ce même document. En d'autres termes, elle assume ce que l'utilisation des Schémas XML de la couche précédente ne fait pas totalement, à savoir le remplacement du texte par sa représentation conceptuelle (*Cf. infra*).
- La couche « Ontology vocabulary » est, comme son nom ne le précise pas complètement, la fourniture des langages qui permettent de représenter des ontologies et de raisonner dessus (en particulier, faire des inférences sur les héritages de propriétés de concepts grâce à la relation de subsomption et sur les types de concepts). Ce raisonnement est fait avec un classifieur – c.-à-d. un programme informatique, ici de la famille des *logiques de description* – qui parcourt l'arbre ontologique. Selon le nombre de constructeurs disponibles pour décrire les classes et les propriétés, OWL est découplé en 3 langages distincts, OWL-Light, OWL DL et OWL-full. La justification de ces trois langages est que la décidabilité – la capacité qu'aura l'algorithme qui traite ces langages de conclure sur les inférences décrites – baisse avec la richesse d'expression. On choisit ainsi son langage en fonction des besoins d'expressivité et de décidabilité. Les ontologies exprimées en OWL fournissent des métadonnées pour la couche précédente.
- Les 3 couches suivantes, « logic », « proof », « trust » ne sont pas encore normalisées et permettraient respectivement, *a*) d'appliquer explicitement des règles d'inférences que l'on se sera donné sur les concepts de l'ontologie, *b*) de se donner les mécanismes nécessaires à faire la preuve des algorithmes et *c*) de se donner les mécanismes nécessaires à calculer et exprimer la croyance en une assertion au sein d'une application. Ces couches reprennent les intentions de certains travaux de l'Intelligence artificielle.
- Enfin, le rectangle de droite représente les méthodes de signatures électroniques et (ce n'est pas sur cette version du schéma) la disponibilité de méthodes de cryptage. Ces deux addenda complètent le « cake » pour fournir les mécanismes d'échange de documents et données de tous types.

4.3 Au-delà du documentaire

Il est remarquable que, contrairement aux discussions passionnées sur les ontologies, la représentation proposée ci-dessus ne fasse pas l'objet de discussions interdisciplinaires plus intenses. Elle est pourtant soumise aux mêmes tensions que celles-là, entre une nécessaire formalisation technique et une représentation de la structure de nos connaissances.

Le Web sémantique, en effet, est plus qu'une infrastructure, c'est aussi un certain nombre d'attendus, d'hypothèses, sur l'usage de ces infrastructures. Ceux-ci peuvent être regroupés en trois fonctionnalités différentes :

- *Les services Web*. La notion de service Web désigne essentiellement une application (un programme) mise à disposition sur Internet par un fournisseur de services et accessible par les clients à travers des protocoles Internet standards. L'idée est de transformer le Web en un dispositif distribué de calcul où les programmes (services) peuvent interagir de manière « intelligente » en étant capables de se découvrir automatiquement, de négocier entre eux et de se composer en des services plus

complexes. La description d'un service Web inclut tous les détails nécessaires à cette interaction. Si tout n'est pas résolu, loin s'en faut, en ce domaine, des réalisations se développent dans des domaines comme les prévisions météorologiques, la réservation de voyage en ligne, les services bancaires. Dans ces domaines, le repérage des ressources, le partage des référentiels entre les machines et l'utilisateur passent par une acceptation négociée et normalisée de l'espace des concepts partagés, de l'ontologie. De ce point de vue et parce que les acteurs partagent un même point de vue, souvent depuis longtemps, ils partagent la même ontologie. Qui plus est, ici, cette ontologie ne sert pas à annoter des textes mais des ressources structurées. Nous n'y reviendrons pas.

- *Le Web sémantique comme infrastructure pour les systèmes à bases de connaissances.* Ici, le Web sémantique fournit son infrastructure à des applications qui utilisent des ontologies pour faire des inférences. Les ontologies y sont des modélisations du domaine, comme toujours, mais ne sont pas les métadonnées qui ont servi à indexer un document.
- *Le Web sémantique pour annoter des documents.* Ici, les ontologies sont utilisées pour annoter formellement un document et permettre ainsi de l'indexer puis de faire de la recherche d'information (RI) dessus. Cette RI peut amener à faire, si nécessaire, des inférences sur ces annotations.

Le troisième point, sur l'annotation formelle des documents, est celui qui pose le plus de problèmes. Dans le Web, les textes ne sont soumis à aucune contrainte informatique, sauf celle du langage HTML. La conséquence de cette liberté est que le seul moyen d'accès aux textes est celui des moteurs de recherche fondé sur la recherche « plein texte », avec bien des avantages, mais aussi les difficultés que l'on connaît : incapacité à prendre en compte les synonymies, à retrouver des textes du même sujet, c'est-à-dire ayant des contextes d'énonciation similaires ou partageant les mêmes concepts.

La réponse du Web sémantique est d'indexer les textes avec les concepts d'une ontologie partagée par une large communauté. La recherche se fait alors sur les concepts avec lesquels les textes sont indexés. Cet index est une représentation formelle du document. Cette représentation se substitue au *texte* pour la phase de recherche et même plus : le *texte* est enrégimenté à la représentation formelle, ou, pour reprendre le raisonnement de la section 1.4, il s'agit d'une tentative de « grammatisation », pratiquement inédite si on excepte les initiatives, aux ambitions beaucoup plus modestes, de classification universelle des bibliothécaires ou d'indexation des documentalistes.

Comme noté précédemment, cet enrégimentement[22] est un véritable changement de paradigme : le *document* devient une représentation formelle de connaissances dont on contrôle la sémantique (formelle) pour l'inférence et le calcul par des agents implémentant des services.

On le retrouve dans le « cake » de la figure 2. Au-dessus du niveau documentaire, les « data » ne sont plus sujettes – ou soumises – à l'interprétation. Celle-ci a déjà été effectuée et est consignée dans l'organisation ontologique élaborée pour annoter.

---

[22] Quine parle d'« enrégimentation » dans des écrits où il s'intéresse justement à la question de l'existence d'un sens logique aux choses qui serait préservé ou pas durant la traduction (*Cf.* 3.3) et qui, dans le cas d'une réponse positive, justifierai que la langue cible soit enrégimentée dans la logique du premier ordre. Il remet en cause cette position et rejoint directement nos questionnements.
Pour une présentation des thèses de Quine : Bonnay Denis, Laugier Sandra, Quine et la logique sauvage, document de travail, Institut d'histoire des philosophies des sciences et techniques, PICS « Logique et rationalité ». Papiers 2002 <http://www-ihpst.univ-paris1.fr/pics/fusion1910.doc>

Mais le *texte* laisse ouvert toute la richesse de l'interprétation tandis que l'ontologie, en privilégiant un (ou deux au plus) contexte, bloque toute autre capacité d'interprétation de l'homme comme de la machine. C'est le présupposé du Web sémantique et d'autres applications des ontologies dans une recherche d'automatisation des traitements.

Pour ne pas avoir de mauvaises surprises, il faut, soit assumer ces contraintes, soit préférer d'autres possibilités d'automatisation existant du côté de l'ingénierie documentaire ou des études sur les corpus. Mais, dans tous les cas la naïveté doit être exclue, la technique ici comme ailleurs n'est pas neutre.

4.4 Questionner le Web sémantique

Pour les ontologies mais contre l'universalisme

Dans le contexte des systèmes à base de connaissances (SBC), des ontologies sont en développement pour fournir les « connaissances du domaine » de systèmes faisant des inférences et assumant le caractère réducteur de la formalisation.

Ainsi, en médecine, des ontologies sont construites, justement pour ce qu'elles « savent faire », par exemple proposer une représentation précise décontextualisée – c.-à-d. ne représentant qu'un point de vue – d'un patient. Cette représentation est évidemment réductrice de ce que contient le dossier médical du patient mais c'est parce qu'elle est réductrice que cette représentation permet au médecin un rappel rapide, une représentation résumée de ce qu'il sait sur le patient. À partir de là, il est possible ensuite de proposer un codage médico-économique[23] du patient prenant en compte d'autres critères que les seuls médicaux. Ce type d'application n'est valide du point de vue des connaissances mise en jeu que parce que la représentation faite ne prétend pas représenter le patient pour des échanges médicaux et qu'elle assume son caractère réducteur (et irréversible !) aux seules informations jugées utiles pour du codage. Les ontologies développées dans ce contexte ont la portée de la médecine occidentale pour laquelle elles sont développées et la légitimité de l'application qui doit s'en servir – p. ex. l'aide au codage.

En résumé, un consensus se fait autour des ontologies dans des communautés d'intérêt et de pratiques, que ce soit la médecine, les voyagistes ou des entreprises de production industrielle impliquées dans les mêmes réseaux de partenaires... Les ontologies, par leur dynamique, aident à expliciter un consensus là où les communautés ont envie de travailler ensemble.

Du Basic English au Global English

Un autre problème posé par certaines interprétations du Web sémantique est sa relation au langage. L'ambition de certains est explicitement de doter le Web sémantique d'une base de connaissances inter-langues de portée mondiale qui couvrirait toutes les connaissances possibles. Qu'il s'agisse de WORDNET[24], de CYC, ou d'autres projets d'ontologie générale, l'hypothèse implicite est alors que les concepts préexistent aux mots et que les relations qui structurent le lexique ne sont pas différentes de celles que l'on conçoit entre les référents, les différences culturelles étant jugées « accidentelles ».

Ceci revient à poser que toutes les langues représentent les mêmes choses, les mêmes « essences », et à considérer comme synonymes des mots de langues différentes. Ce modèle

[23] Le « codage » des patients dans les milieux hospitaliers, est une activité obligatoire qui vise à munir les tutelles d'un représentation médico-économique et bientôt médicale des patients dans le buts d'études économiques et épidémiologiques.

[24] http://wordnet.princeton.edu/

de la langue comme représentation, et celui de l'ontologie qui en découle, remontent à une longue tradition logico-grammaticale au fondement des sciences du langage.

Mais les producteurs d'ontologies générales et ceux des membres de la communauté du Web sémantique qui les suivent ne semblent guère s'interroger sur les postulats qui sous-tendent leurs travaux. Ce manque de réflexion conduit certains à voir dans la disparition des langues un phénomène positif et dans la prédominance de l'anglais, le résultat de la sélection naturelle. C'est ainsi que le concept de *Global English*[25] gagne du terrain, porté par l'enthousiasme des nouveaux communicateurs d'une diaspora mondiale communiquant par le Web. Le *Global English* est conçu comme une nouvelle langue issue de la dialectisation des anglais britannique et américain, dotée d'une interlangue fonctionnelle, contrôlée, pivot conceptuel de tous les dialectes. Les effets pernicieux de la babélisation seraient ainsi maîtrisés : le *Global English* mettrait fin à plusieurs siècles de recherches infructueuses d'une « langue parfaite » puisqu'il offrirait en même temps la liberté d'une langue naturelle (tous les pidgins possibles de l'anglais issus d'une créolisation généralisée) et le garde-fou d'une langue artificielle (un nombre fini de mots concepts univoques) qui servirait d'interlangue conceptuelle.

Le *Global English* s'inscrit dans la tradition des grandes utopies linguistiques et reprend notamment l'héritage du *Basic English* d'Ogden et Richards[26], très en vogue dans les années trente. Censé supplanter l'esperanto, adopté par la BBC et promu par l'UNESCO après la guerre, le *Basic English*, établit la liste des huit cent cinquante mots-concepts nécessaires et suffisants pour toute communication. Présenté comme un « langage sténographique pour la pensée », il revendique le principe de compositionnalité cher à la logique positiviste et se présente comme un outil simple, efficace, non ambigu visant à éliminer les incompréhensions induites par le langage courant, entaché d'ambiguïté et d'émotivité.

L'objectif avoué d'Ogden et Richards était de planter les « semences morales et mentales » dans l'esprit des populations analphabètes du monde. Armés d'un dictionnaire de poche conceptuel écrit dans la meilleure langue possible, fruit de la sélection naturelle, les nouveaux missionnaires entendaient faciliter le commerce, les sciences et l'industrie, répandre les valeurs chrétiennes de l'Occident. Ce programme dit « orthologique » répondait au principe de moindre effort. Plus de temps perdu à apprendre des langues : le *Basic English* fournissait l'interlangue universelle qui permettait de penser juste et efficace.

WORDNET ne propose rien de bien différent avec ses *synsets* et son *Inter Lingual Index*. Néanmoins, ce qui n'enlève rien à la critique, ce vocabulaire contrôlé a l'avantage de pouvoir intervenir directement dans des systèmes techniques. Il est ainsi utilisé, par exemple, pour gérer automatiquement la répartition des contenus des espaces publicitaires selon l'indexation des pages Web.

Dernier avatar d'une langue auxiliaire internationale, le *Global English* procède de la même philosophie que ses prédécesseurs (synthèse du logicisme, du pragmatisme et du cognitivisme), en revanche il bénéficie d'un contexte technologique et économique autrement favorable.

Questions en suspens

Au vu des défauts d'argumentation dont souffre le Web sémantique, et aussi sans doute compte tenu du peu de réalisation de recherches d'information en grandeur réelle, en

[25] Crystal David, *English as a Global Language*, Cambridge: CUP, 1997.

[26] Ogden, Charles Kay, *Basic English: A General Introduction with Rules and Grammar*, Paul Treber & Co., Ltd. London, 1930.

particulier sur la possibilité, pour les ontologies, d'être un moyen pertinent d'entrer dans le *contenu* d'un texte, d'autres pistes sont suggérées et suivies. Citons-en trois :

- Des travaux essaient d'améliorer les modes de recherche « plein texte ». Ces travaux font une analyse des corpus sur lesquels porte la requête et peuvent proposer de l'élargir à des termes proches, repérés par une analyse statistique ou syntaxique.
- Le contexte social d'élaboration de structures conceptuelles a été peu questionné par les travaux du Web sémantique et c'est une direction qu'il faut certainement suivre pour produire des ressources que les utilisateurs puissent investir dans leur activité et sur lesquelles des raisonnements heuristiques puissent s'exercer. Dans une perspective de « documents pour l'action », certains proposent d'intégrer une dimension communicationnelle à la construction des ontologies.
- Enfin, la question des corpus textuels mériterait plus d'attention qu'elle n'en a fait l'objet jusqu'ici. Ils sont pourtant au centre des analyses indispensables pour automatiser ou supporter par un artefact la prise en compte des documents numériques. Par exemple, des méthodes de construction d'ontologies se fondent sur une analyse de corpus textuels élaborés durant l'activité pour laquelle on veut développer une aide. Pour mieux argumenter les choix des corpus comme leur élaboration, la question de leur genre textuel semble une piste à explorer. Cela participe d'une sémiotique de corpus qui permettrait d'imaginer d'autres façons de prendre en compte les documents numériques.

5 *Texte*, *document*, et médiation

Le développement de notre réflexion sur les relations entre *texte* et *document* dans un contexte numérique a fait resurgir deux dialectiques, très anciennes dans l'histoire mais renouvelées par les fonctionnalités des outils informatiques et souvent masquées par la brutalité du changement rappelée en section 1 :

- D'une part, les relations entre objets linguistiques (les mots, les phrases, les discours écrits) et objets sémiotiques (les images, fixes et animées, les sons) ; les premiers sont inclus dans les seconds, les seconds peuvent être désignés par les premiers, mais pourtant les uns ne se réduisent pas aux autres ; les possibilités nouvelles de manipulation des signes et des signaux apparues avec l'informatique, réunies sous le terme de « multimédias », obligent à approfondir cette dialectique ; Ceci nous oblige à réviser la notion de *texte* dans sa relation au *document*.
- D'autre part, la représentation des relations sociales et leur calculabilité ; la structuration des documents et la construction des métalangages ne sont pas apparus avec l'informatique, mais les opportunités de calculs ouverts sur ceux-là sont inédites par leur échelle et leur sophistication. Nous devons alors revoir la question de la médiation, qui sera développée plus précisément dans un prochain travail collectif.

Ces dialectiques traversent nos propos et ont largement sous-tendu les discussions au moment de l'élaboration de ce document. Chacun des trois modèles implicites repérés en première section les règlent à sa manière, initiant en réalité une conception différente du *document*.

5.1 Réviser les relations *texte* / *document*

Dans le premier cas, celui d'un modèle « DTD », le problème se réduit à une question de représentation et la calculabilité n'est mobilisée que pour permettre de modifier celle-ci. Le *document* est considéré comme le *contenu*, texte, image ou son, dont l'informatique peut faire varier la *forme*, pourvu que sa structure soit préservée. Ce modèle est issu d'une tradition

plutôt littéraire. Mais alors, le questionnement se déplace sur la notion de *texte*, linguistique ou sémiotique, qui ne peut échapper à une mise en forme. Et nous avons vu que les réponses butaient sur bien des difficultés et que toutes les pistes n'ont pas encore été suivies comme celle, parmi bien d'autres, de la dimension temporelle.

Dans le second modèle, celui des « Schémas », le problème documentaire est en quelque sorte inversé. Ici le *document* est plus présent dans la structure dont on doit repérer les invariances pour les préserver, que dans les données, dans ce cas le *contenu*, qui peuvent lui faire prendre sens selon une déclinaison particulière. Le terme traditionnel de « formulaire » illustre bien la logique à l'œuvre, même si elle a pris une toute autre dimension avec l'informatique. Ici, la calculabilité et le droit font bon ménage pour la définition du *document* dont la principale caractéristique doit être l'authenticité, matérialisée par sa structure et la maîtrise de la variabilité de son contenu. Il semble que des progrès pourraient être accomplis à condition de prendre en compte les structurations particulières traditionnelles de l'objet document sans les renvoyer à un objet générique abstrait. Néanmoins, poser ainsi la question d'invariants dans un document revient à privilégier ses caractéristiques sémiotiques au détriment des caractéristiques langagières comme nous l'avons montré dans l'exemple de la traduction.

Le troisième modèle, celui du Web sémantique, le plus radical dans ses ambitions, est aussi celui qui a suscité le plus de discussions au cours de la rédaction de ce texte. On peut se demander dans ce modèle si la notion de *document* a encore un sens. Le problème se déplace, en effet, sur les communautés et les interprétations, partagées ou non, de concepts réifiés par des mots. Si, dans un contexte donné, il y a accord sur une « ontologie », alors c'est ce système cognitif global qui génère une sorte de documentarisation, dans laquelle un document n'est plus qu'une représentation éphémère d'un besoin ponctuel de connaissance. Mais, pour nombre de chercheurs, ce système relève de l'utopie si la communauté dans laquelle il est construit n'est pas rigoureusement normée ou si le problème cognitif posé n'est pas strictement encadré. Dans ce cas, il y aurait une homologie entre structure sociale, utilité et structure documentaire mesurée par un système d'ontologies. De telles communautés ou de tels usages sont utiles pour des systèmes précis de connaissances ou pour gérer des activités normées, mais il ne s'agit pas de l'organisation cognitive générale et, sans doute heureusement pour l'humanité, car elle est contradictoire avec le fonctionnement même du langage. Aussi paraît-il indispensable d'avancer plus avant dans les relations entre le social et le sémantique pour approfondir ce modèle.

5.2 Reconsidérer la médiation

Confrontés à des traditions plus anciennes, les modélisations documentaires actuelles peuvent s'analyser de deux façons :

- Il s'agit d'une part à la fois d'une continuité et d'un saut qualitatif pour les outils documentaires. Les professionnels du document (traduction, édition, documentation, archivistique, etc.) y retrouvent des notions qui leur sont familières, même si elles ont parfois changé de nom et les plus à la pointe d'entre eux sont très actifs au côté des informaticiens dans le développement et l'application des normes, dans les indexations, dans les mises en ligne et dans les débats et les mises en pratique. De ce point de vue il s'agit d'une formidable montée en puissance de leur fonction traditionnelle.

  En même temps, ils peuvent être déroutés par la puissance des modélisations en cours qui découpent, reconfigurent les objets qu'ils manipulaient, jusqu'à effectuer à leur place des opérations qui leur paraissaient ne pouvoir échapper à une intervention humaine. Mais ce saut, s'il transforme considérablement leur métier au point d'en

éteindre certains et d'en faire émerger d'autres, ne remet pas en cause la logique de leur fonction.

- Par ailleurs, il s'agit d'un outil de lecture-écriture en réseau, reliant les bureaux de tous les internautes. Ainsi, les liens hypertextes et hyper-médias, les agents intelligents permettraient à chacun, en liaison avec les ressources du réseau d'écrire son texte qui viendrait naturellement lui-même alimenter le patrimoine commun. Mais ce système coopératif universel suppose que la communication sociale puisse s'abstraire de médiations. Sans doute, cela est-il possible dans certaines communautés très normées. Sans doute aussi cela donne l'occasion d'ouvertures nouvelles, inédites, éphémères ou germes de réorganisations. Pourtant, il porte aussi le risque du chaos, de la confusion ou de la manipulation.

Pour le dire autrement, on peut considérer l'ambition des systèmes documentaires numériques comme celui d'une vaste bibliothèque intelligente de configuration de documents ou alors comme un texte s'écrivant à l'échelle de la planète à partir d'une bibliothèque qui ressemblerait à « la bibliothèque de Babel » de Borgès[27]. Dans le premier cas, le projet est difficile et risqué mais cohérent avec la constitution humaine des savoirs, dans le second, il nous paraît une prétention démesurée conduisant à la confusion ou l'imposition des représentations car faisant l'impasse sur le rôle de la médiation. En effet, l'accès direct, « transparent », aux ressources suppose une décontextualisation radicale des représentations et par conséquent l'impossibilité de leur donner un sens réellement partagé, sauf dans des communautés suffisamment structurées et normées où le contexte est déjà inscrit dans l'implicite collectif.

L'examen des transformations des rapports entre *texte* et *document* pointe en effet un certain nombre de questions, trop vives pour nous en soustraire, qui conditionnent des réalités politiques, culturelles, sociales de grande envergure. En voici quelques-unes :

– Quels liens voulons-nous conserver avec la culture documentaire dont notre société est issue, souhaitons-nous rompre avec elle, la transformer, en inventer une autre ? Quels principes guident aujourd'hui les grands programmes qui se mettent en place par le concours des industriels et des acteurs publics ? Quelle est la valeur des modèles revendiqués par les uns et les autres ?
– Où mène l'idéal d'une culture structurée par des protocoles de plus en plus uniformisants ?
– Quels sont les enjeux liés à l'utilisation de tel ou tel modèle, de tel ou tel protocole ? Peut-on analyser les situations et les paradigmes de leur utilisation ?

---

[27] « La bibliothèque de Babel » est une nouvelle se trouvant dans le recueil *Fictions* de l'écrivain Jorge Luis Borges.
Il s'agit d'une bibliothèque de taille infinie dont toutes les salles hexagonales sont disposées d'une même manière. Elle est composée d'une infinité de livres ayant tous le même format, placés dans des étagères comprenant toutes le même nombre d'étages et recevant toutes le même nombre de livres. Chaque livre a le même nombre de pages et de signes écrits au hasard ; l'alphabet utilisé comprend toujours vingt-cinq caractères.
On peut donc dire que la Bibliothèque contient tous les ouvrages qui ont déjà été écrits ainsi que tous les autres, parmi une infinité de livres sans aucun contenu lisible (puisque chaque livre peut n'être constitué que d'une succession de lettres ne formant rien de précis dans aucune langue). Celle-ci est habitée par une race d'hommes qui ne connaît que ce monde, à la recherche qui du livre ultime, qui d'une révélation, qui de la Vérité.
Cette nouvelle est une métaphore de la littérature et montre une grande influence de la kabbale <http://fr.wikipedia.org/wiki/La_bibliothèque_de_Babel/>.

– Qui peut et doit décider de ces enjeux ? Peuvent-ils être débattus ou seront-ils tranchés, de fait, par ceux qui ont le pouvoir de configurer les dispositifs ou par un jeu d'acteurs tellement éclaté que personne n'en maîtrise le sens ?

Le risque serait que les questions ici posées disparaissent, non parce qu'on leur aurait apporté une réponse, mais simplement parce que les conditions pour les poser auraient disparu. Ce texte vise donc à sortir quelques-unes de ces questions de l'impensé où elles baignent : c'est une urgence de le faire et c'est la responsabilité politique des scientifiques et des institutions scientifiques d'y contribuer sans tarder.

Annexe : un document médical suivant plusieurs paradigmes

Dans cette annexe, nous illustrons les trois « philosophies » de représentation exposées en section 1.4 montrant les différents paradigmes de représentations auxquelles elles aboutissent et quels types d'information elles permettent d'enregistrer. L'exemple utilisé un est extrait d'un compte rendu d'hospitalisation (CRH) qui apparaît tel qu'il est décrit dans le paradigme XML documentaire.

1 Paradigme XML documentaire

Dans ce paradigme, le texte est conservé et la contrainte qui s'y applique est l'organisation logique du texte, grammatisé. La figure 1 propose donc l'extrait de CRH qui nous servira de fil rouge.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE CRH SYSTEM " ExemplePedauque.dtd">
<CRH>
   <ExamenClinique>
      Surcharge pondérale (78 kg, taille à 1,69), saturation à 98 % en air ambiant.
      Apyrétique. Tension artérielle à 14-8 . Aucune symptomatologie pulmonaire en dehors
      d'une dyspnée d'effort (à  la montée de 2 étages).
   </ExamenClinique>
   <ExamenComplementaire>
      EFR normales. ECG normal. Echographie cardiaque : très minime
      décollement péricardique. Epreuve d'effort normale avec performances juxta maximales.
      Examens à  la recherche d'une péricardite en attente.
   </ExamenComplementaire>
   <Conclusion>
      Douleurs thoraciques semblant être provoquées par des péricardites. Le patient sera
      revu la semaine prochaine en Hôpital de Jour afin de lui faire part des résultats du bilan
      et du diagnostic. Un avis gastro est prévu également la semaine prochaine.
    </Conclusion>
</CRH>
```

Fig. 1. – *Un extrait d'un compte rendu d'hospitalisation respectant une grammaire XML.*

Dans ce contexte, la structure du CRH est décrite dans une « définition de type de document » ou DTD. Cette DTD décrit principalement les balises, leur situation l'une par rapport à l'autre, le type des éléments entre les balises – p. ex. #PCDATA précise que c'est une chaîne de caractères. Ainsi, la figure 2 affiche la DTD qui spécifie l'extrait du CRH et qui permet de vérifier qu'il est *formellement valide* (*Cf.* 3.2).

```
<!ELEMENT CRH (#PCDATA | ExamenClinique | ExamenComplementaire | Conclusion)*>
<!ELEMENT ExamenClinique (#PCDATA)*>
<!ELEMENT ExamenComplementaire (#PCDATA)*>
<!ELEMENT Conclusion (#PCDATA)>
```

Fig. 2 – *Définition de type de document respectée par le document XML de la figure 1.*

L'apparence du texte XML ainsi spécifié est calculée par une feuille de style XSL et pourrait apparaître comme proposé dans la figure 3.

[…]

ExamenClinique

Surcharge pondérale (78 kg, taille à 1,69), saturation à 98 % en air ambiant. Apyrétique. Tension artérielle à 14-8 . Aucune symptomatologie pulmonaire en dehors d'une dyspnée d'effort (à la montée de 2 étages).

ExamenComplementaire

EFR normales. ECG normal. Echographie cardiaque : très minime décollement péricardique. Epreuve d'effort normale avec performances juxta maximales. Examens à la recherche d'une péricardite en attente.

Conclusion

Douleurs thoraciques semblant être provoquées par des péricardites. Le patient sera revu la semaine prochaine en Hôpital de Jour afin de lui faire part des résultats du bilan et du diagnostic. Un avis gastro est prévu également la semaine prochaine.

Fig. 3. – *Affichage du compte rendu XML.*

2 Paradigme Schéma XML

Dans ce paradigme, le texte est conservé pour les seules parties dont l'interprétation est jugée primordiales ; les grandeurs physiques sont typées par des Schémas XML. La partie contextuelle, le texte, est réduite au maximum, ici à la partie interprétation des données médicales. Par rapport à l'extrait précédent, cela montre le typage d'un certain nombre de grandeurs (taille, poids, etc.) et la conservation des expressions correspondant aux interprétations. En particulier, les valeurs dites « normales » sont bien des interprétations d'examens dont les résultats ne sont pas visibles ici. La figure 4 montre le CRH tel qu'il pourrait être dans ce second paradigme.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE CRH SYSTEM " ExemplePedauque.dtd">
<CRH>
   <ExamenClinique>
      Surcharge pondérale (<poids unite="kg">78</poids>, taille
        <taille unite="cm">169</taille>), saturation à  <saturation>98</saturation>% en air
      ambiant. Apyrétique. Tension artérielle à  <tension>14-8</tension>.
   </ExamenClinique>
   <Conclusion>
      EFR normales. ECG normal. Echographie cardiaque : très minime
      décollement péricardique. Epreuve d'effort normale.
       Douleurs thoraciques semblant être provoquées par des péricardites. Le patient sera
      revu la semaine prochaine en Hôpital de Jour afin de lui faire part des résultats du bilan
      et du diagnostic. Un avis gastro est prévu également la semaine prochaine.
   </Conclusion>
</CRH>
```

Fig. 4. – *Un extrait d'un compte rendu d'hospitalisation respectant une grammaire XML. Par rapport à la figure 1, cet extrait correspond à ce qui serait conservé dans le paradigme des Schémas XML.*

La figure 5 donne un aperçu du Schéma XML qui permet de décrire l'organisation logique du document et le type les grandeurs numériques décrites dans le CRH (poids, taille, saturation, tension).

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified">
   <xs:element name="CRH">
      <xs:complexType>
         <xs:sequence>
            <xs:element ref="ExamenClinique"/>
            <xs:element ref="Conclusion"/>
         </xs:sequence>
      </xs:complexType>
   </xs:element>
   <xs:element name="ExamenClinique">
      <xs:complexType mixed="true">
         <xs:choice minOccurs="0" maxOccurs="unbounded">
            <xs:element ref="poids"/>
            <xs:element ref="saturation"/>
            <xs:element ref="taille"/>
            <xs:element ref="tension"/>
         </xs:choice>
      </xs:complexType>
   </xs:element>
   <xs:element name="poids">
      <xs:complexType>
         <xs:simpleContent>
            <xs:extension base="xs:integer">
```

```
            <xs:attribute name="unite" use="required">
              <xs:simpleType>
                <xs:restriction base="xs:string">
                  <xs:enumeration value="kg"/>
                  <xs:enumeration value="g"/>
                </xs:restriction>
              </xs:simpleType>
            </xs:attribute>
          </xs:extension>
        </xs:simpleContent>
      </xs:complexType>
    </xs:element>
    <xs:element name="saturation">
      <xs:simpleType>
        <xs:restriction base="xs:integer">
          <xs:maxInclusive value="100"/>
          <xs:minInclusive value="0"/>
        </xs:restriction>
      </xs:simpleType>
    </xs:element>
    <xs:element name="taille">
      <xs:complexType>
        <xs:simpleContent>
          <xs:extension base="xs:integer">
            <xs:attribute name="unite" use="required">
              <xs:simpleType>
                <xs:restriction base="xs:string">
                  <xs:enumeration value="cm"/>
                </xs:restriction>
              </xs:simpleType>
            </xs:attribute>
          </xs:extension>
        </xs:simpleContent>
      </xs:complexType>
    </xs:element>
    <xs:element name="tension" type="xs:NMTOKEN"/>
<xs:element name="Conclusion" type="xs:string"/>
</xs:schema>
```

Fig. 5. – *Schéma pilotant la structure du CRH affiché figure 3 et typant un certain nombre de grandeurs physiques correspondant aux examens clinques.*

L'apparence du texte peut-être, comme précédemment, prise en charge par une feuille de style XSL et on pourrait avoir l'affichage proposé figure 6 qui ne diffère pas, en dehors de la réduction du contenu, de celui de la figure 3.

[…]

ExamenClinique

Surcharge pondérale (78 kg, taille à 1,69), saturation à 98 % en air ambiant. Apyrétique. Tension artérielle à 14-8 .

Conclusion

EFR normales. ECG normal. Echographie cardiaque : très minime décollement péricardique. Epreuve d'effort normale

Douleurs thoraciques semblant être provoquées par des péricardites. Le patient sera revu la semaine prochaine en Hôpital de Jour afin de lui faire part des résultats du bilan et du diagnostic. Un avis gastro est prévu également la semaine prochaine.

Fig. 6. – *Affichage du compte rendu fondé sur un Schéma XML.*

3 Paradigme ontologique

Dans ce paradigme, l'ontologie est réputée fournir les éléments – c.-à-d. fournir un index – pour représenter le texte qui peut être simplement abandonné. Dans ce cas, l'index est décrit en RDF et les balises (préfixées par « med ») sont choisies dans les concepts de l'ontologie (*Poids, Taille*, *Tension*, *Saturation*, *DouleursThoractiques*, *Pericardite*, ainsi que la relation *ProvoquePar*).

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE rdf:RDF SYSTEM ".dtd">
<rdf:RDF>
   <rdf:Description about="Examen-clinique">
      <med:Poids>78</med:Poids>
      <med:taille>169</med:taille>
      <med:tension>14-8</med:tension>
      <med:saturation>98</med:saturation>
   </rdf:Description>
   <rdf:Description about="Conclusion">
      <rdf:Description about="DouleursThoraciques">
         <med:ProvoquePar>Pericardite </med:ProvoquePar>
      </rdf:Description>
   <rdf:Description
</rdf:RDF>
```

Fig. 7. – *La représentation du compte rendu avec un point de vue ontologique.*

Il n'y a pas de représentation spécifique de l'index du compte rendu, à savoir les concepts avec lesquels il est étiqueté et finalement remplacé. Une représentation graphique est proposée figure 8.

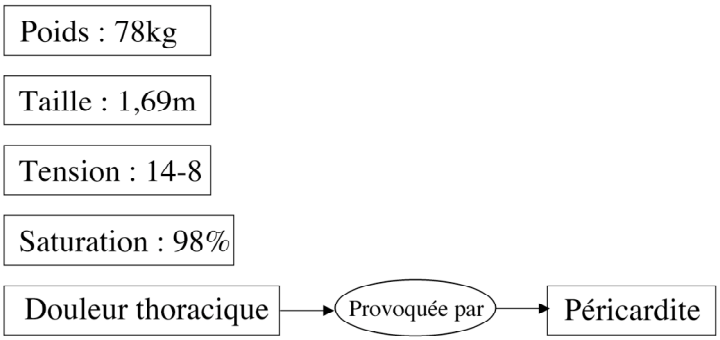

Fig. 8. – *L'affichage graphique de la représentation ontologique.*

4 La non neutralité de la technique

Pour chacun des paradigmes, nous avons choisi de ne représenter qu'une partie précise de l'information de référence (le texte de l'approche XML documentaire) liée aux fonctionnalités du paradigme en cours, dans le but de montrer les tendances forte qui vont de pair avec la technique :

– dans le cas du document XML « documentaire », l'ensemble des informations textuelles sont conservées ;
– dans le cas du Schéma XML, la « philosophie » voudrait que le compte rendu soit généré à partir des données stockées dans une base. Toutes les données pertinentes et disponibles sont affichées et la partie textuelle incontournable – celle qui décrit l'interprétation du médecin – est stockée et re-proposée telle quelle à l'affichage. Évidemment, rien n'empêche de conserver un document XML complet et de typer les grandeurs que l'on sait et veut typer. C'est dans ce sens que vont un certain nombre de travaux de normalisation en médecine même si peu ou pas de systèmes sont en réalité développés suivant ces recommandations ;
– dans le cas de l'ontologie, on interprète le compte rendu et l'on retient les concepts importants en même temps qu'on fait disparaître les modalités de croyance. Le « on » est chargé d'un travail complexe qui extrairait de textes, par définition peu ou pas structurés, les concepts médicalement importants. Pour la partie des résultats d'examens cliniques, ils ont été renseignés de façon structurée. On peut imaginer une ontologie qui conceptualiserait le fait que le médecin fait des hypothèses, qu'il privilégie l'une ou l'autre jusqu'à plus ample informé et l'on rejoint alors le Web sémantique qui propose que la question des croyances soit abordé dans la couche « trust » des langages du Web sémantique (*Cf.* 4.2). Mais cela amène une complexité telle – pas seulement algorithmique – qu'il est difficile de développer les outils qui permettraient ce passage du textuel au conceptuel[28].

A ce jour, rien n'est vraiment résolu, en informatique médicale, pour savoir quel type de technique sera privilégiée : la médecine est un domaine où la tension entre les textes et leur formalisation est forte, où de nombreux acteurs – les politiques y compris – négocient des normes dont on ne sait pas lesquelles seront finalement et réellement mises en œuvre. C'est encore un domaine parmi d'autres, où l'adage discuté ici, « la technique n'est pas neutre », prend tout son relief.

[28] C'est ce qui a été fait dans le cadre du projet MENELAS mais les résultats et difficultés ont amené à réfléchir à des systèmes plus simples <http://www.biomath.jussieu.fr/Menelas/Ontologie/>.