Texto ! juillet 2008, vol. XVIII, n°3

SEMANTIQUE DU WEB *VS* SEMANTIC WEB ?

— Le problème de la pertinence

François RASTIER
CNRS / ERTIM-INaLCO

[à paraître dans Valette, M. (2008), éd. Textes, documents numériques, corpus.
— Pour une science des textes instrumentée, *Syntaxe & Sémantique*, n°9]

Résumé : Le programme du *Web sémantique* entend remplacer le « Web des documents » par le « Web des données » et prolonge ainsi le programme classique de la représentation des connaissances. En revanche, pour une *sémantique du Web* inspirée par la linguistique de corpus, les connaissances résident dans les textes et les documents qui les véhiculent, et ne peuvent en être abstraites sans perdre leur valeur contextuelle et leur pertinence. Cela conduit à recontextualiser la notion même de donnée, ainsi qu'à problématiser le rapport entre données et métadonnées.

Mots-clés : sémantique, donnée, métadonnée, connaissance, pertinence.

Abstract: The Semantic Web programme aims to replace the "Web of Documents" by the "Web of Data", thus prolonging the classical programme of knowledge representation. In contrast, a corpus-linguistic inspired Web Semantics/ situates knowledge within texts and the documents that convey them. Data cannot therefore be abstracted without losing their contextual valeur and pertinence. This leads to a recontextualisation of the notion of "data" and a rethinking of the relationship between data and metadata.

Key words: semantics, data, metadata, knowledge, keyness.

1. Ambitions et crédibilité du Web sémantique

On sait que le Web fonctionne d'après trois standards : le protocole HTTP, l'adressage par les URL et le langage HTML. Tim Berners-Lee, qui dirige le W3C, instance qui préside aux destinées du Web mondial, a présenté depuis 1994 le Web sémantique comme une extension du Web qui le transformerait en un espace d'échange de documents permettant d'accéder à leurs *contenus* et à effectuer des *raisonnements*. Cela exigerait une représentation du contenu des documents par des ontologies pourvues d'une sémantique dénotationnelle (le Web sémantique n'en reconnaît pas d'autre) ; l'ensemble des contributeurs au Web sémantique, et bientôt l'ensemble de ceux qui mettent en ligne des contenus, doivent donc respecter une infrastructure commune figurée d'abord par le fameux « cake » de Tim Berners-Lee, présenté à la conférence XML 2001 :

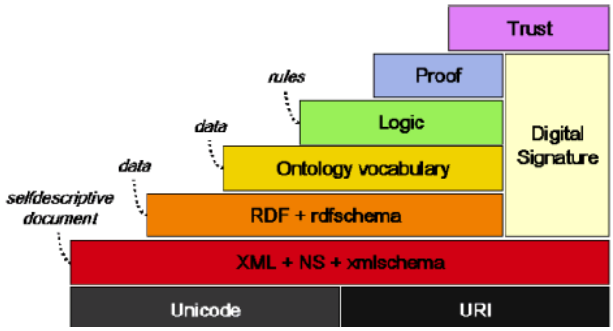

Figure 1 : Le « Cake » du Web sémantique

Cette infrastructure est aujourd'hui jugée effectivement normalisée jusqu'au niveau des ontologies : elles fournissent notamment le vocabulaire de ces métadonnées pour représenter le contenu des documents de la même manière qu'un thésaurus, composé de termes (ou concepts) et non de mots. Notons bien qu'au-dessus du deuxième niveau, on perd ainsi tout contact avec les textes et les langues, en passant à des langages (« formels ») de représentation. En promouvant le *Web sémantique*, le W3C entend remplacer le « Web des documents » par le « Web des données » (cf. Tim Berners-Lee, 2007). En utilisant des ontologies, il s'agit de s'affranchir de la complexité des documents et de leur diversité linguistique et sémiotique.

En accord avec l'objectivisme de la philosophie du langage issue du positivisme logique, la donnée est alors conçue comme une simple chaîne de caractères (ex. la donnée *pêche*, qui peut être reliée soit à *poisson*, soit à *fruit* ; Berners-Lee, *loc. cit.*). Il serait discourtois d'insister sur l'indigence banale de cette conception des données.

Rassurantes, car présentées comme purement pratiques, les recommandations du W3C ont vocation à devenir des standards. Or, l'adoption de standards « de bas niveau » comme HTML, ou Unicode voire XML n'entraîne aucunement que l'on doive ériger en standard des langages de représentation comme RDF ou OWL, sauf à céder benoîtement à la tentative de coup de force du W3C en faveur du « Web sémantique ».

Il serait plus discourtois encore d'insister sur les enjeux économiques du Web sémantique : on comprend parfaitement que le Département US du Commerce soutienne le Web sémantique, car la normalisation des contenus accessibles sur le Web se concrétise par des ontologies généralement faites de mots anglais écrits en majuscules, et réputées toutefois représenter des « métadonnées » permettant d'accéder aux documents.

Les standards de métadonnées sont l'un des trois éléments-clés de la Stratégie données en *réseau centré* (*Net-Centric Data Strategy*) du Département de la Défense des États-Unis, arrêtée en décembre 2001 et rendue publique en mai 2003 (cf. Stenbit éd., 2003). Rendant obligatoires certains types de métadonnées, cette stratégie consiste à contrôler un réseau définitoirement non centré : « to ensure that all data are visible, available, and usable », pour en finir décisoirement avec le Web caché qui échappe au contrôle. Par ailleurs, « all posted data will have associated metadata » (Stenbit éd., 2003, Avant-propos, p. 4), de manière que le volume des données privées soit drastiquement réduit (divisé par deux, alors que les données communes seraient multipliées par trois ; cf. p. 10). La centralisation du réseau ainsi réalisée (pour réaliser la *Net-Centricity*) permet alors « a completely different approach to *warfighting and business operations* » (Appendice A, p. 2, mes italiques).

On comprend ainsi que les soutiens militaires n'aient pas manqué depuis septembre 2001 : le Web sémantique se veut en effet collaboratif, chaque fournisseur de contenu devant mettre en ligne ses bases de données selon un format unique qui les rendra interopérables et permettra par là-même d'y accéder — par exemple, pour découvrir de nouveaux médicaments (selon Tim Berners-Lee, 2007). L'intelligence, au sens économique et militaire du terme, a tout à gagner à cette transparence coopérative.

Les enjeux politiques et économiques ne doivent pas faire oublier les conséquences épistémologiques de ce programme. On peut s'interroger sur la cohérence du « cake », qui

graphiquement évoque les délices étagés de la tranche napolitaine : il s'agit vraisemblablement d'une simple juxtaposition éclectique, mais cet éclectisme est orienté par les objectifs dont le statut scientifique reste douteux. Quiconque est un peu frotté de sémiotique visuelle aura en effet reconnu dans les gradins du « cake » les marches d'un *gradus ad Parnassum*, qui nous conduit d'*Unicode* à *Trust* (au sens de *confiance*, comme dans *In God we trust*, plutôt que de *monopole*). Bref, on édicte des standards, puis on les érige au rang de modèles théoriques, ce qui est caractéristique de la technoscience, non seulement instrumentaliste, mais instrumentalisée. Sir Tim Berners-Lee est ingénieur ; en proclamant opportunément en 2007 la formation d'une *Web Science*, Tim Berners-Lee évite cependant que les problèmes scientifiques soient posés et débattus hors de la communauté du Web sémantique, qui s'est auto-engendrée et se doit aussi de s'auto-évaluer.

Laissons les analystes futurs s'interroger sur l'unité de pensée entre le W3C et le *Department of Defense*. Internet est né d'une contradiction toute militaire entre sécurité du réseau et contrôle des informations. Le réseau devait être distribué pour pouvoir résister à toute tentative de destruction ; mais son succès même et son extension à l'économie et aux données privées l'a rendu difficile à contrôler. Or tout Appareil, économique ou militaire, et le *Department of Defense* n'est qu'un exemple éminent, se doit de maintenir une hiérarchie pour exercer son pouvoir et constituer sa légitimité : il projette donc nécessairement sa structure sur le monde qui l'environne, et nous avons déjà souligné par exemple la parenté théorique, au sens le plus métaphysique, entre les ontologies et les organigrammes (l'auteur, 2004b). Le « cake » du W3C restitue à sa manière une hiérarchie de hiérarchies (les ontologies, les DTD XML, etc.) et permet de subsumer par différents niveaux de métadonnées puis de « données » la diversité incontrôlable des documents, comme des langues et des systèmes de signes qu'ils mettent en jeu.

Aussi le titre de cette étude met-il en scène de manière trop simple deux conceptions différentes qui ne s'opposent pas directement et ne sont pas commensurables. Ne nous attendons pas à un combat de David contre Goliath : le Web sémantique est un programme politico-technique, alors que la sémantique du web est un projet méthodologique et un domaine d'applications fondées sur une sémiotique des corpus.

S'il bénéficie de soutiens influents, le Web sémantique rencontre aussi un facile assentiment, car il concrétise un ensemble de conceptions reçues qui appartiennent à la tradition de l'Intelligence artificielle classique et que nous allons questionner.

2. Habitudes de pensée et limites actuelles

2.1. Postulats de la représentation des connaissances

Le Web sémantique transpose dans un environnement nouveau et à une échelle inédite la problématique de la représentation des connaissances. Elle repose sur trois postulats qui la relient au cognitivisme orthodoxe.

(i) Les connaissances sont des représentations du monde empirique : l'image du monde comme « mobilier » ontologique s'est en général imposée dans les milieux pour lesquels le positivisme logique reste la référence implicite. Cependant, rien ne permet d'affirmer, avec le *réalisme naïf* revendiqué par des sémanticiens influents, comme John Lyons, que les objets du monde soient discrets, dénombrables, publics et partout identiques.

(ii) Les connaissances sont (relativement) indépendantes de leurs substrats sémiotiques, de telle manière que leur extraction ou leur représentation ne modifie pas leur contenu. Ce postulat fut celui de la théorie générale de la terminologie, élaborée par Wüster et reprise par le Cercle de Vienne : les termes sont considérés comme indépendants des langues[1].

[1] Cela suppose un dualisme difficile à accepter, car les concepts ne sont pas indépendants des textes où ils sont définis, configurés et remaniés. En effet, les connaissances ne résident pas dans des termes, mais dans des textes ; en outre, les discours scientifiques et techniques ne sont évidemment pas transparents ni indépendants des langues dans lesquelles ils sont élaborés, même si les normes internationales d'une discipline peuvent favoriser leur traduction. La représentation des

(iii) Les connaissances sont discrètes et formalisables au sens où elles sont représentables par un formalisme logique, en général la logique des prédicats. On retrouve ici la dualité de l'empirisme et du logicisme dans l'empirisme logique : la théorie de la dénotation assure l'ancrage empirique de la théorie, tandis que l'usage de l'organon logique est censé lui apporter une productivité conceptuelle[2].

Une approche non référentielle s'impose alors pour éviter ces postulats métaphysiques : l'alternative la plus cohérente à présent nous semble être la sémantique différentielle.

2.2. Les débats

Dans le domaine de la représentation des connaissances, on a élaboré des formalismes de représentation considérés comme adéquats, dans la mesure où ils conviennent à des applications peu ambitieuses. L'adoption de standards comme XML ou RDF permet une interopérativité de principe, mais ne résout pas le problème de la production, de l'identification et de l'évolution des connaissances.

Les débats portent en amont sur le problème de la réification des connaissances hors des contextes d'utilisation, ou complémentairement sur l'adéquation de leurs modes de représentation à leur utilisation effective.

Les tenants de la position réifiante s'appuient sur l'essor des ontologies, qui radicalisent la préconception objectiviste des connaissances. Les ontologies restent des *thésaurus* – celui de Roget a d'ailleurs explicitement servi de modèle à Miller et à ses collaborateurs pour WordNet. Elles en gardent les inconvénients notoires : une généralité qui ne leur permet pas de s'adapter aux points de vue sélectifs exigés par les tâches et un manque d'évolutivité qui exige une maintenance manuelle. Elles réduisent la langue à une nomenclature, qui ne rend compte ni des structures textuelles, ni des variations considérables de genres et de discours.

Même dans le domaine du Web sémantique, pourtant très lié aux ontologies, la perspective centrée sur les utilisateurs conduit à des constats résignés comme celui-ci : « Semantic Web researchers accept that paradoxes and unanswerable questions are a price that must be paid to achieve versatility » (Berners-Lee et coll., 2001). La variété des points de vue des utilisateurs et des régimes de pertinence propres à leurs tâches leur interdit de se satisfaire d'une norme unique au demeurant arbitraire : mais l'absence de contradiction reste un postulat absolu des ontologies, conformément aux lois d'identité, de non-contradiction et de tiers exclu qui fondent leur conception logiciste du monde.

Plus radicalement, les tenants de la cognition située et les ergonomes spécialisés en recherche d'information insistent sur la diversité imprévisible des applications et sur le fait que les formalismes ne sont que des supports à des parcours d'interprétations. Dès lors, les facettes définitoires d'un objet, quel qu'il soit, ne peuvent être fixées *a priori* : en d'autres termes, ce sont les pratiques qui définissent les propriétés pertinentes des objets.

Cette divergence peut aujourd'hui être tranchée empiriquement. En effet, l'étude de grands corpus, y compris techniques, a montré que les relations sémantiques qui organisent les ontologies différaient selon les discours et les domaines, au point que certaines relations sémantiques de base sont tout bonnement absentes de certains corpus pourtant étendus (cf. projet Safir conduit par un consortium Crim-Lip6-Edf).

Par ailleurs, l'expérience de Wordnet et EuroWordnet est instructive : ces ontologies se sont révélées inutilement complexes. Fondées sur les postulats psychologiques datés de Miller et Johnson-Laird (1976), elles ignorent des savoirs linguistiques élémentaires comme la notion cruciale de *morphème*, ce qui conduit à créer des sous-réseaux distincts pour les noms, les verbes et les adjectifs. Malgré des coûts sans précédent, les ontologies se révèlent peu utiles et sont ordinairement consultées comme des dictionnaires ou des aides à

connaissances passe donc par le recueil et l'étude de corpus scientifiques et techniques.

[2] Cette préconception logique du monde ne mobilise qu'une petite part de la logique bivalente et se tient à l'écart tant de la logique multivalente que de la logique modale. *A fortiori*, elle reste bizarrement à l'écart des mathématiques, car elle ne tient pas compte de leurs trois problèmes constituants que sont l'infini, le continu et les grands nombres.

la traduction. Enfin, à l'échelle du Web, la fusion des connaissances provenant de différentes ontologies reste problématique, du fait que, même au sein d'une même discipline, elles ne sont pas interopérables entre elles, malgré les consignes de standardisation.

2.3. Ontologies et Web sémantique

Dans les sciences de la communication et dans le domaine des traitements automatiques du langage, la séparation entre cognition et communication s'est classiquement traduite par le privilège donné à la représentation des connaissances, sans préoccupations particulières pour leur production, leur sélection et leur transmission. On extrait l'information, puis on la communique, la seule condition mise à la communication se limitant à *l'information packaging*, conçue comme simple emballage des connaissances.

Les ontologies sont l'aboutissement de cette conception héritée du positivisme logique : semblant faites par personne pour personne et donc indépendantes de tout point de vue, elles sont censées représenter un monde objectif, indépendant de toute langue et de tout système de signes, comme d'ailleurs de toute tâche. La nomenclature des objets du « monde » n'est évidemment pas problématisée, puisqu'elle repose sur l'évidence partagée ; dans le cas des ontologies « locales » ou spécialisées, l'inventaire des entités dépend simplement de l'état de l'art tel qu'il est admis.

Dans ce type de représentation, la différence entre les langues s'efface, de même que la diversité des discours et des genres qui leur sont liés : le format des connaissances dans les hiérarchies ontologiques reste celui des réseaux sémantiques : un thesaurus en *basic english*, dopé par des relations sémantico-logiques stéréotypées et d'ailleurs hétérogènes, comme l'hypéronymie, la méronymie, etc.

Tout entier dépendant de cette problématique, le « Web sémantique » reste tributaire d'un petit nombre de relations sémantiques universelles et pauvres. Comment peut-on supposer que la pertinence d'un mot soit liée à la position de son référent dans une hiérarchie ontologique ? Les inégalités qualitatives dans un texte sont sans rapport déterminable avec la position hiérarchique des entités : en général, comme les entités superordonnées sont triviales, plus un concept est superordonné moins il est discuté et donc moins il est pertinent. La pertinence, si on la définit comme un principe d'économie cognitive (selon Sperber & Wilson), ne privilégie alors que les concepts les plus triviaux, mais non ceux sur lesquels portent effectivement les débats.

En outre, la richesse sémiotique des documents numériques n'est pas ou peu prise en compte, car elle est inconciliable avec la problématique référentielle et ne contribue pas à la dénotation : or les indices de l'expression (typographie, codes de couleur, etc.) peuvent se révéler hautement discriminants.

Enfin, la variété des tâches d'application impose de pouvoir définir et faire varier des régimes de pertinence : aucune connaissance n'est indépendante d'une tâche. Comme toute pratique définit son régime de pertinence, c'est à une *praxéologie* (et non à une ontologie) de déterminer quelles sont les « informations-clés » dans les textes et les corpus.

2.4. Exigences pour la linguistique

Concernant le Web, l'enjeu majeur est évidemment l'amélioration des moteurs de recherche, l'adaptation des stratégies en fonction des tâches d'une part, de la nature des documents d'autre part.

Cela demande le recours à une linguistique *applicable* qui puisse traiter des textes, analyser leur sémantique et refléter leur diversité linguistique et sémiotique. La linguistique de corpus se voit ainsi dans la nécessité d'innover. Elle est issue d'une part de la linguistique computationnelle, qui pose des problèmes dérivés du cognitivisme chomskyen (génération de phrases, construction d'arbres syntaxiques, etc.) et de la lexicométrie (issue de la linguistique mathématique et des statistiques).

La linguistique computationnelle se heurte à des obstacles issus de la philosophie du positivisme logique (notamment par la séparation entre syntaxe, sémantique et

pragmatique). En revanche, la lexicométrie, en tant que « simple » méthodologie, ne défend pas de préconception du langage, ce qui la rend plus adaptable. Ces deux problématiques ont en commun de ne pas avoir de conception théorique du texte : pour la linguistique computationnelle, c'est une suite de phrases ; pour la lexicométrie, un ensemble de mots. Aussi ces disciplines restent-elles dépourvues quand elles se trouvent affrontées à la fois à des corpus, massifs, multilingues, polysémiotiques, dont l'abord dépend de multiples demandes sociales et culturelles.

C'est donc à une linguistique conçue comme science des textes et consciente de son appartenance aux sciences de la culture qu'il revient de faire des propositions d'unification et de remembrement. En tenant compte des demandes sociales et non simplement en appliquant des théories : la linguistique ne peut être véritablement appliquée que si elle est également *impliquée*. Elle se doit d'intervenir, même de façon auxiliaire, à diverses étapes : création des logiciels, constitution des corpus, balisages, expérimentations avec outils sur corpus (balisés), interprétation et discussion des résultats. À toutes ces étapes de la chaîne de traitement, des connaissances linguistiques, et plus largement sémiotiques, se révèlent indispensables.

2.5. Pour mettre fin à l'oubli des textes

La problématique ontologique de la représentation des connaissances reste sans doute tributaire d'un état de l'art obsolète, celui d'un temps où l'on *n'avait pas accès* au plein texte. Les thésaurus et autres classifications formalisées servaient alors à indexer les textes à partir d'une représentation statique de leur contenu présumé. Les inconvénients sont connus : coûts de construction et d'entretien considérables, pertinence insuffisante et non modulée en fonction de la tâche qui préside à la recherche d'information.

Le point de vue normatif repose sur des oublis méthodologiques voire épistémologiques qui affectent : (i) les contextes locaux et globaux des informations au sein des textes ; (ii) le contexte des corpus où les textes (et donc les informations) prennent sens ; (iii) les points de vue dont les informations dépendent et qui les ont configurées ; (iv) les collectivités auxquelles elles sont destinées. En bref, la soustraction des contextes est aussi une soustraction des usages dont dépend la notion même de pertinence.

Ces obstacles sont inévitables si l'on réduit les textes à des « ensembles de mots » sans tenir compte des structures, des genres, etc. En revanche, l'accès au plein texte permet désormais des réponses plus adaptées, dès lors qu'il est guidé par les propositions de la linguistique de corpus. En effet, les métadonnées que l'on accumule à présent n'ont tout de même pas pour fonction de permettre d'oublier les données !

3. Reconceptions

3.1. Élaboration dynamique

Nous formulons la proposition méthodologique de fonder toute représentation des connaissances sur l'analyse sémantique et sémiotique des corpus effectifs qui les manifestent : *les connaissances et les ontologies qui les "normalisent" doivent et peuvent être élaborées dynamiquement, en fonction des applications et de leurs corpus*. En effet, les « connaissances » sont des interprétations objectivées de textes et d'autres performances sémiotiques.

Chaque application définit dans son corpus un régime de pertinence propre. Aucun concept n'est pertinent en toute application. Par ailleurs, un des grands problèmes des ontologies est la définition de leur « nomenclature » : comment distinguer les concepts qui doivent y figurer, alors que tous les mots du lexique sont des candidats potentiels, sans parler des syntagmes phraséologiques. La pratique de George Miller montre qu'il n'a pas d'autre critère que le « bon sens », c'est-à-dire le préjugé du créateur d'ontologies[3].

[3] En 2002, il fait ainsi sortir de l'ontologie le franc, la lire et le mark, puisque ces monnaies

Si l'on admet que le lexique n'est pas organisé en une arborescence unique, car chaque discours et chaque genre a son lexique, on doit substituer à l'image totalisante du réseau unifié des zones locales organisées par des rapports de *profilage* plutôt que de subsomption : chaque concept est une *forme sémantique* qui se profile sur un fond. Certains termes lexicalisent des formes ou des parties de formes, d'autres des fonds. Par exemple, le mot *texte* en critique littéraire est un élément de fond, et non un concept : il sert de base compositionnelle à des expressions comme *texte balzacien*, mais il ne se trouve jamais dans le contexte de termes comme *notion* ou *concept*.

Par ailleurs, les formes sémantiques sont *valuées*, alors que les concepts d'une ontologie ne le sont pas : par exemple, dans un réseau comme WordNet, *carré pané* pourrait fort bien être le plus proche voisin de *caviar*. Or il est évident, et l'exploration des corpus le confirme, qu'on ne les rencontre aucunement dans les mêmes contextes (cf. Rastier et Valette, à paraître). Aussi la hiérarchie évaluative prime-t-elle la hiérarchie ontologique construite sans tenir compte des évaluations.

Les concepts peuvent être décrits comme des formes sémantiques propres aux textes théoriques : leurs lexicalisations diffuses ou synthétiques, leurs évolutions, de leur constitution à leur disparition (par extinction ou banalisation désémantisée), leurs corrélats sémantiques, leurs cooccurrents expressifs, tout cela dessine un champ de recherche qui commence à peine à être exploré.

L'alternative que nous proposons est celle de moteurs de recherche en plein texte qui tiennent compte des avancées de la sémantique textuelle, notamment : (i) la définition d'unités textuelles non strictement bornées et séquentielles (les passages) ; (ii) l'extension du principe différentiel de la sémantique au contraste de corpus, entre discours, genres, et sections de textes ; (iii) l'analyse des genres textuels en zones de pertinence différenciées.

L'enjeu est non pas la représentation mais la *production* de connaissances à partir de données massives non structurées issues du Web — ou, de préférence, de banques documentaires.

Enfin, la problématique de la représentation des connaissances doit être conçue dans le cadre d'une sémiotique. En effet, les textes ne sont pas de simples chaînes de caractères. Leur découpage, leur structure « logique », leur typographie, voire leurs balises, font partie de leur sémiotique. Par exemple, en philosophie classique, l'usage des majuscules désignait les concepts principaux. Au-delà, les textes scientifiques et techniques intègrent ce qu'on appelle improprement des *hors-textes* : figures, tableaux, diagrammes, photographies participent à la textualisation des connaissances et appellent pour leur traitement une sémiotique multimédia.

3.2. Les connaissances textuelles

Une connaissance est un *ensemble de passages* de textes (éventuellement multimédia) : dans leurs récurrences, le contenu de ces passages (les fragments) et leurs expressions (les extraits) sont en relation de transformation, ne serait-ce que par changement de position. Résultant de figements et de réductions de syntagmes, les mots sont une sorte très particulière de ces passages, et comme les autres passages, ils restent impossibles à interpréter sans recontextualisation.

En somme, la connaissance est issue d'une décontextualisation de certaines formes sémantiques saillantes et des expressions qui leur correspondent, qu'elles soient compactes (comme les lexicalisations) ou diffuses (comme les définitions). Les formes donnent l'illusion de l'indépendance, voire de leur idéalité, parce que les formes sont par définition éminemment transposables.

Toutefois, aucun mot ni aucun passage ne peut prétendre résumer un texte. Certes, définir une saillance, comme on le fait en faisant figurer en tête d'un article une liste de mots-clés, c'est donner une « instruction » interprétative : la clé n'est cependant pas une clé qui ouvre la serrure du sens, car il reste à construire dans l'interprétation. Aussi, les

n'avaient plus cours, et il y fait entrer l'*intifada* et le *bacillus anthracis* (cf. l'auteur, 2004b).

métadonnées doivent-elles garder trace du texte et du contexte et permettre d'y accéder – sans jamais pouvoir s'y substituer.

Puisque la problématique logico-grammaticale ne peut penser la textualité, les métadonnées utiles n'ont pas de statut logico-grammatical déterminable. En revanche, dans la problématique que nous adoptons, elles revêtent un statut philologique (pour documenter le texte) et herméneutique (pour permettre de l'interpréter). Les informations ne sont plus alors simplement assimilées à des connaissances : on n'appellera *connaissances* que les informations sélectionnées pour une pertinence interprétative. Il reste à les *comprendre*, c'est-à-dire à les relier entre elles en fonction de la structure du texte dont elles sont extraites et de l'objectif de la tâche en cours.

4. Propositions

4.1. Typologie des formes de pertinence

La communication scientifique n'est pas plus directe et pas plus claire que les autres types de communication. De toute façon, la prétention à la clarté n'exclut pas la nécessité de l'interprétation, même si l'herméneutique des textes scientifiques et techniques reste peu développée.

Ces textes se caractérisent par un usage notoire de l'indexation et une structure hiérarchique particulière. Concrétisant une inégalité qualitative, la pertinence résulte d'une valorisation : tel point du texte sera primé, et servira de point d'accès à d'autres, considérés alors comme secondaires. La pertinence affichée doit alors régir les parcours interprétatifs. Alors que les discours scientifiques se limitent en principe à des faits, la pertinence y introduit des valeurs qui concernent tant les faits eux-mêmes que le mode d'accès à ces faits. En effet, les connaissances sont des objets culturels et, à ce titre, elles ne peuvent être dissociées des valeurs.

4.1.1. La pertinence « objective »

Selon les parties du texte, on peut distinguer plusieurs types de pertinence qui introduisent des indices d'inégalité qualitative et donnent ainsi des indications de valeur.

A/ Le péritexte. — Tant par sa fonction que par sa structure sémiotique (capitales, corps, graisse) il établit des inégalités qualitatives : par exemple les titres ne se limitent pas à des résumés, mais sont des indications interprétatives.

Partie du péritexte, les mots-clés explicites placés en début de texte sont également des indications interprétatives qui pointent des formes sémantiques saillantes.

B/ L'intratexte (ou corps du texte). — Dans cette partie du texte, les unités sont moins normalisées. Les passages-clés peuvent être des mots, des syntagmes, des phrases, des paragraphes, etc. Leur caractérisation suppose des contrastes par des méthodes de linguistique de corpus, quantitatives notamment.

On relève traditionnellement la pertinence des mots singuliers : ils peuvent être isolés par un test probabiliste comme caractéristiques d'un passage ou du texte (cf. la fonction *thème* du logiciel Hyperbase).

La pertinence des passages reste plus importante mais moins étudiée : en raison des phénomènes de diffusion sémantique, les passages réunissent des faisceaux de corrélats (lexicalisations partielles d'une même forme sémantique) que l'on peut appeler des *paratopies*. Il faut alors définir des techniques de *zonage* minimal : l'unité textuelle retenue n'est plus le mot, mais le *passage*, si bien qu'un mot-clé n'est utile que s'il conduit à un passage-clé.

Dans tous les cas, la pertinence intrinsèque est construite par trois types de contrastes : entre passages du texte ; avec des passages d'autres textes du corpus ; avec le corpus choisi considéré dans son ensemble.

C/ L'infratexte. — Conventionnellement, le contenu de l'*infratexte* (les notes, ou la bibliographie, par exemple) est considéré comme faiblement pertinent. Mais c'est

« l'inconscient du texte » et la lecture experte peut y déceler des indices cruciaux, comme de simples références bibliographiques, qui situent l'ensemble du texte ou permettent d'en reconsidérer des passages.

4.1.2. La pertinence « subjective »

La paresse voudrait que l'on se satisfasse de la pertinence proposée : il est vrai que la complexité des rapports entre le péritexte et l'intratexte appelle des recherches propres. Mais l'on doit conserver une attitude critique, car la communication scientifique est aussi « indirecte ». Au-delà de la pertinence « affichée », il peut exister une pertinence cachée : le double langage existe aussi dans les domaines scientifique et technique.

Les textes sont inclus dans des pratiques sociales et leur production comme leur lecture dépend de tâches et de stratégies différenciées. Outre la pertinence objective, un autre régime de pertinence dépend de la lecture et de la tâche qu'elle concrétise : on peut la nommer pertinence « subjective ».

4.1.3. Pour une pertinence dynamique

La distinction entre pertinence objective et subjective n'est que temporaire. L'auteur propose, le lecteur dispose : parmi les indications proposées par l'auteur, il ne retient que les mots ou passages-clés qui correspondent à sa tâche, en soulignant des mots ou passages-clés qu'il désigne comme tels en fonction de sa tâche. Ni subjective, ni objective, la pertinence doit ainsi être construite dynamiquement en fonction : (i) de la structure du document, (ii) de ses spécificités telles qu'elles peuvent être déterminées par contraste avec son corpus de référence, (iii) enfin, de la pratique en cours.

4.1.4. Incidences sur la redéfinition textuelle des concepts

La caractérisation textuelle des concepts s'appuie sur les contextes locaux et globaux.
1/ Les indices contextuels locaux comprennent : (i) les lexèmes cooccurrents, les entités nommées adjacentes (noms d'auteurs notamment), les morphèmes, les ponctèmes ; (ii) les indices d'expression : typographie, balises.
2/ Les indices contextuels globaux comprennent : (i) la position des concepts dans le texte ; (ii) la spécificité des concepts et de leurs contextes immédiats, pour caractériser un texte ; (iii) la spécificité du texte dans son corpus de référence (de manière experte, on peut caractériser aussi un texte par les concepts absents).
3/ La position temporelle des concepts : l'évolutivité des concepts impose des études en diachronie (exemple : les travaux de Mathieu Valette sur un corpus de Gustave Guillaume étendu sur 40 ans).

4.2. Pour une sémantique du Web

Développée à partir de la sémantique des textes et de la philologie numérique, une sémantique du Web peut les mettre à profit pour adapter à la diversité des requêtes la diversité des réponses, qui seront pertinentes si elles reflètent la diversité des textes. Elle n'est à son tour qu'une étape de médiation pour constituer une *sémiotique comparée* des documents numériques.

On doit tenir compte tant au plan épistémologique que méthodologique des sources de diversité que la problématique actuelle du Web sémantique ne permet pas de traiter de manière satisfaisante.

1/ *La diversité des langues*. — Le Web est multilingue et le sera de plus en plus. L'hégémonie initiale de l'anglais a été renversée par la montée en puissance d'autres grandes langues. Les moteurs de recherche doivent donc gérer un multilinguisme croissant, ce qu'ils ne font pas encore de façon satisfaisante.

Par ailleurs, les représentations des connaissances devraient varier avec les langues : il

ne s'agit pas simplement de découpages différents des mêmes champs de réalité, mais encore de définitions différentes de ces champs comme l'attestent par exemple les contrastes « ontologiques » entre le chinois et l'anglais.

2/ *La diversité des discours et des genres*. — Les « concepts » qui peuplent les « ontologies » dépendent étroitement des discours et des genres. L'existence de communautés internationalement structurées a favorisé la constitution de discours disciplinaires plurilingues et la diffusion de genres comparables malgré les différences linguistiques : cela peut se traduire par des calques terminologiques, mais aussi par des modes de structuration textuelle, tant au plan du contenu que de l'expression. Toutefois, l'adoption de normes internationales limite la diversité linguistique, mais sans pouvoir l'annuler.

3/ *La diversité des styles.* — La formation et l'évolution des concepts sont l'objet d'importantes différences non seulement selon les disciplines, mais encore selon les auteurs. Le style philosophique de Deleuze définit par exemple un régime de transformations conceptuelles tout à fait différent de celui de Bourdieu. Les méthodes de la linguistique de corpus ont permis sur ce point des résultats qui confirment le bien-fondé d'un programme comparatif (cf. Loiseau, 2006).

4/ *Les inégalités qualitatives au sein des documents*. — Chaque genre, chaque texte singulier définit un régime de pertinence qui prime certaines formes sémiotiques relativement à d'autres. Cela impose de définir, ici encore avec les méthodes de la linguistique de corpus, des techniques de détection des inégalités qualitatives.

5/ *La diversité sémiotique intrinsèque des documents.* — La distinction entre textes (multimédia) et documents doit être réduite voire supprimée, car le texte n'a pas un contenu indépendant de son expression, et le document ne peut être véritablement décrit en faisant abstraction de son contenu. Corrélativement, au plan épistémologique, les divergences de fait entre linguistique et philologie doivent être reconsidérées au sein d'une sémiotique générale de la communication. La sémantique du Web relève en effet d'une *sémiotique comparée* des documents numériques.

6/ *La diversité des tâches.* — Malgré le problème lancinant mais faussé de la réutilisabilité, les représentations des connaissances qui ne sont pas établies en fonction d'une application déterminée sont en général peu utilisables et guère réutilisables. Construire de telles représentations avec une ambition de généralité reste une tâche indéfinie sinon infinie, car les tâches déterminent en effet un régime de pertinence.

En revanche, la rencontre entre *l'horizon de pertinence* déterminé par la tâche et *les formes sémiotiquement saillantes* détectées par l'analyse contrastive du corpus de travail permet de qualifier les passages essentiels et de restreindre drastiquement les réponses en recherche d'information.

7/ *La diversité des statuts de fiabilité*. — Au plan pratique comme au plan éthique, la question de la fiabilité des documents ne doit pas être négligée, car le Web fourmille d'écrits apocryphes, de faux, sans parler de textes diversement négationnistes. Un écrit non authentique ne peut évidemment jouir que d'une autorité usurpée.

Cette question doit être traitée dans le cadre d'une réflexion sur les types de communication, les dimensions de la destination et de l'adresse, comme enfin de l'autorité et de l'authenticité. Le nombre de liens et le *page-ranking* ne définissent qu'une métrique conformiste de l'autorité. On ne peut véritablement parvenir à une recherche d'information fiable si l'on ne tient pas compte du degré de confiance que l'on peut attribuer aux documents : c'est là un point faible du Web 2, quand il fait de l'anonymat un principe — comme on le voit avec Wikipedia.

4.3. La complexité de toute donnée

Nous proposerons ici un modèle moins sommaire de la donnée, qui tienne compte de la dualité sémiotique irréductible entre expression et contenu, ou plus généralement entre *phore* et *valeur*. Cela s'étend à toute chaîne de caractères, du signe de ponctuation au chapitre, sans égard pour le modèle apocryphe du signe prêté à Saussure par les rédacteurs

du *Cours de linguistique générale* et contredit par les écrits autographes.

La dualité phore/valeur, qui constitue le corps sémiotique de la donnée, se trouve sous la rection d'une dualité de rang supérieur entre le *point de vue* et le *garant*. Le point de vue n'est pas un simple point d'observation : il est déterminé par une pratique et un agent individuel ou collectif ; dans un traitement de données, il dépend donc de l'application. Le garant est l'instance de validation qui fonde l'évaluation de la donnée : cette instance est une norme sociale qui peut être juridique, scientifique, religieuse ou simplement endoxale. En linguistique de corpus, le garant est l'autorité qui a présidé à la constitution du corpus ; certaines métadonnées documentaires, comme l'auteur ou l'éditeur, relèvent de cette instance.

Le point de vue est « subjectif » dans la mesure où il est occasionnel ; le garant, « objectif » dans la mesure où il est constitutionnel ou du moins constituant. La dualité du point de vue et du garant définit deux régimes de pertinence, particulière pour le point de vue et générale pour le garant. Puisque les données sont bien ce qu'on se donne, elles sont ainsi les résultats initiaux d'un processus d'élaboration — et leur traitement produit des résultats ultérieurs, dans un cycle susceptible de récursivité.

Dans les termes de la sémiotique des zones anthropiques (cf. l'auteur, 1996, 2001a, 2002), le corps (phore+valeur) de la donnée, en tant qu'elle est objectivée, relève de la zone proximale de l'environnement ; le point de vue, de la zone identitaire ; enfin, le garant, de la zone distale où se situent les instances de normativité. L'axe sur lequel se répartissent ces zones est celui de la *médiation symbolique*, alors que l'axe subordonné qui relie le phore et la valeur relève de la *médiation sémiotique* (cf. l'auteur, 2001a). Soit en bref :

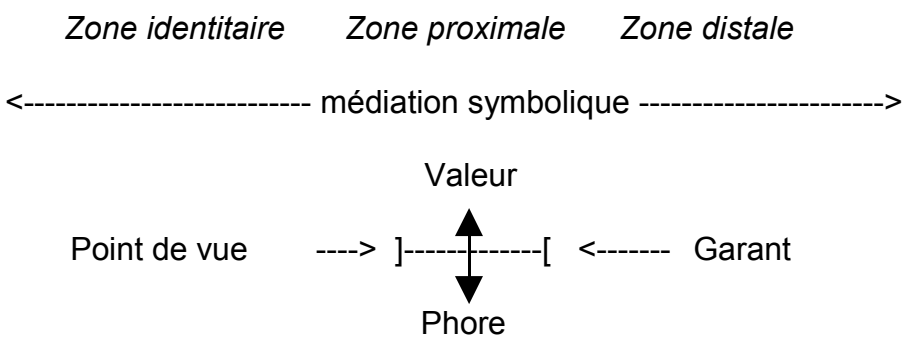

Figure 2 : Les quatre instances et les trois zones de la donnée

La structure de la donnée présentée ici dépasse les grandeurs proprement linguistiques et vaut pour d'autres grandeurs sémiotiques. Plus généralement, elle convient à tout objet culturel – et un corpus est évidemment un objet culturel. Constituante de la complexité propre aux objets culturels, la mise en relation interzone dont ils résultent leur permet de jouer un rôle de médiation entre les individus (ou les groupes) et leur environnement.

En ne percevant pas le caractère instituant de la valeur, du point de vue et du garant, en réduisant la donnée à la seule instance du phore, le positivisme ordinaire élude toute dimension critique et épistémologique. Recueil de données ainsi appauvries, un « corpus » sans point de vue ni garant n'est pas véritablement un objet scientifique mais un amas numérique inexploitable en tant que tel ; ainsi des pseudo-corpus recueillis par aspiration aléatoire de sites.

4.4. Problématiques métadonnées

Les indications philologiques élémentaires, quand elles sont retenues, sont aujourd'hui catégorisées comme des « métadonnées ».

La notion de métadonnée transpose, en l'atomisant pour ainsi dire, celle de métalangage,

issue de la logistique russellienne. Les ontologies sont d'ailleurs considérées comme des métalangages de description des documents, conditionnant l'accès à ceux-ci. Les métadonnées sont des données qui figurent dans l'en-tête du document (*header*), alors que les données proprement dites composent le corps du texte (*body*, ou plus précisément *intratexte*). Du « Web des documents » on passe ainsi au « Web des données », puis, pourrait-on dire, au « Web des métadonnées » : cette conception prévaut aujourd'hui avec le Web sémantique.

Une grande confusion règne toutefois, puisque l'on classe parmi les métadonnées toutes sortes de données incompatibles avec la théorie appauvrie du texte qui prévaut généralement : on juxtapose des indications simplement bibliographiques, comme l'auteur, l'éditeur, l'ISBN, le lieu d'édition ; des indications documentaires, comme le résumé ou les mots-clés ; des caractérisations textuelles globales, comme le genre.

Les théories linguistiques du péritexte, qui limitent le texte à l'intratexte, en séparant les titres, voire les notes, etc., n'ont fait qu'ajouter à la confusion : par exemple le titre sera considéré comme une métadonnée, alors qu'il est une partie inaliénable du texte. En règle générale, les données relèvent de la *linguistique interne*, les métadonnées de la *linguistique externe*, et, faute de réfléchir leur dualité, on ne peut théoriser le rapport entre données et métadonnées. Les problèmes négligés reviennent alors, réifiés, sous la forme de métadonnées. Par exemple, dans le domaine du multimédia, les textes eux-mêmes deviennent les métadonnées des images.

Sans trop croire à l'efficacité d'un moratoire sur les métadonnées, retenons que les métadonnées sont des critères globaux et les données des grandeurs locales qui en dépendent : au lieu de les séparer *a priori*, c'est à une théorie élaborée de la textualité qu'il revient d'établir systématiquement les corrélations entre métadonnées et données, pour restituer la complexité des textes.

La notion de métadonnée doit ainsi être critiquée et refondue : elle n'est pas un concept, mais une classe de problèmes hétérogènes ; par exemple, plus techniquement, on définit comme métadonnée une étiquette de colonnes d'une base de données relationnelle ou un attribut de classe d'un langage à objets, ou encore une variable d'un langage prédicatif.

Le succès de Google s'explique d'ailleurs par l'introduction d'un nouveau type de métadonnées (les liens qui pointent vers le document) et par une perspective praxéologique implicite qui représente le document en fonction d'un *point de vue* (de qui sélectionne le lien) et d'un *garant* (celui qui pose le lien et apporte ainsi une évaluation). Cela conforte la reconception de la notion de donnée que nous souhaitons poursuivre.

Un texte n'est pas un réservoir de connaissances qui pourraient être extraites par indexation et condensées en données résumant son contenu informationnel ; l'indexation donc n'a qu'une relative valeur de recherche et de classement. Prenons un exemple : à l'heure actuelle, dans les services de renseignement militaire d'un grand pays européen, des personnels extraient de documents Word des mots et expressions qu'ils transfèrent dans des feuilles Excel où ils sont classés en « ontologies ». Ces feuilles sont ensuite transmises à des analystes qui en font la synthèse sous la forme de Powerpoints présentés à l'état-major[4]. L'éloquence militaire prise certes le laconisme, mais toute modification systématique d'un texte en change le genre et donc l'interprétation. La sélection de ces passages minimaux que sont les mots et expressions reste de fait incontrôlable, puisque le recouvrement de deux indexations du même texte par la même personne s'établit en moyenne à 40%. La délinéarisation, la « compression » augmentent l'équivoque et créent l'ambiguïté.

Au Web sémantique, il faudra inévitablement substituer une sémantique du Web, car les besoins sociaux pour la recherche d'information, l'amélioration des moteurs de recherche, le *data-mining*, ne pourront être satisfaits que par une linguistique et une sémiotique de corpus permettant l'analyse des données textuelles et documentaires.

[4] Je n'invente rien : partir de Word, passer par Excel pour arriver à Powerpoint, tel est aujourd'hui le cycle d'extraction et de d'exploitation des « connaissances ».

N.B. : J'ai plaisir à remercier Évelyne Bourion, Carmela Chateau et Christian Mauceri.

Bibliographie

Amardeilh, F. (2007) *Web Sémantique et Informatique Linguistique : propositions méthodologiques et réalisation d'une plateforme logicielle*, Paris, Université Paris X (Thèse).
Berners-Lee, T. (1998) *Weaving the Web*, Harper, San Francisco.
Berners-Lee, T. (2007) Le web va changer de dimension. *La Recherche*, 413, pp. 34-38.
Ganascia, Jean-Gabriel (2006) *Information, communication et connaissance*, Paris, Presses du CNRS.
Loiseau, S. (2006) *Sémantique du discours philosophique : du corpus aux normes*, Thèse, Université Paris X.
Pédauque, Roger T. (2006) *Le document à la lumière du numérique*, Caen, C&F éditions.
Mauceri, C. (2007) *Indexation et isotopie : vers une analyse interprétative des données textuelles*. Thèse de doctorat, ENST-Bretagne et Université de Bretagne Sud. Rééd. : http:// www.texto-revue.net
Rastier, F. éd. (1995) *L'analyse thématique des données textuelles*, Paris, Didier.
Rastier, F. (1996) Représentation ou interprétation ? — Une perspective herméneutique sur la médiation sémiotique. In V. Rialle et D. Fisette (dir.), *Penser l'esprit : des sciences de la cognition à une philosophie de l'esprit*, Grenoble, Presses Universitaires de Grenoble, pp. 219-239.
Rastier, F. (2001a) L'action et le sens. — Pour une sémiotique des cultures, *Journal des Anthropologues*, 85-86, pp. 183-219.
Rastier, F. (2001b) *Arts et sciences du texte*, Paris, PUF.
Rastier, F. (2002) Anthropologie linguistique et sémiotique des cultures. In F. Rastier et S. Bouquet (dir.) *Une introduction aux sciences de la culture*, ch. 14, pp. 243-267.
Rastier, F. (2004a) Doxa et lexique en corpus - Pour une sémantique des « idéologies ». In *Actes des Journées scientifiques en linguistique 2002-2003 du CIRLLEP*, Reims : Presses Universitaires de Reims.
Rastier, F. (2004b) Ontologie(s), *Revue de l'Intelligence Artificielle*, Numéro spécial Informatique et terminologies, 18, pp. 16-39.
Rastier, F. (2005) Enjeux épistémologiques de la linguistique de corpus. In G. Williams (éd.), *La linguistique de corpus*, Presses universitaires de Rennes, pp. 31-46.
Rastier, F. (2007b) Passages. *Corpus*, 6, pp. 127-162. [aussi sur Texto ! : http://www.revue-texto.net].
Rastier, F. et Ballabriga, M., éds (2007) *Corpus en lettres et sciences sociales. — Des documents numériques à l'interprétation*, CALS, Albi. [aussi sur Texto ! : http://www.revue-texto.net].
Rastier, F. et Valette, M. (à paraître) De la polysémie à la néosémie, *Langue française*.
Stenbit, P. éd. (2003), *Department of Defense Net-Centric Data Strategy*, memorandum, Washington, Defense Pentagon, 30 p.
http://defenselink.mil/cio-nii/docs/Net-Centric-Data-Strategy-2003-05-092.pdf
Valette, M. (2006) La genèse textuelle des concepts scientifiques. Étude sémantique sur l'œuvre du linguiste Gustave Guillaume. *Cahiers de lexicologie*, 2006/2, n°89, p. 125-142. [aussi sur Texto ! : http://www.revue-texto.net].
Valette, M. et Slodzian, M. (2008) « Sémantique des textes et Recherche d'information », *Extraction d'information : l'apport de la linguistique*. In A. Condamines & Th. Poibeau (éds.), *Revue Française de Linguistique Appliquée* (volume XIII-1 / juin 2008), pp. 119-133.