



**HAL**  
open science

## Interactions avec une base de documents : environnement numérique et perception sémiotique des changements indicibles

Jacques Labiche, Youssouf Saidali, Eric Trupin, Maryvonne Holzem, Pierre  
Beust

### ► To cite this version:

Jacques Labiche, Youssouf Saidali, Eric Trupin, Maryvonne Holzem, Pierre Beust. Interactions avec une base de documents : environnement numérique et perception sémiotique des changements indicibles. *Revue des Interactions Humaines Médiatisées (RIHM) = Journal of Human Mediated Interactions*, 2009, vol 10 (n° 2/2009), pp.67-84. hal-01066828

**HAL Id: hal-01066828**

**<https://hal.science/hal-01066828>**

Submitted on 22 Sep 2014

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Interactions avec une base de documents : environnement numérique et perception sémiotique des changements indicibles.

*Documents base interactions: digital environment and semiotic  
perception of inexpressible changes*

Jacques LABICHE (1), Youssef SAIDALI (1), Eric TRUPIN (1),  
Maryvonne HOLZEM (2), Pierre BEUST (3)

(1) LITIS, Université de Rouen  
[Jacques.Labiche@univ-rouen.fr](mailto:Jacques.Labiche@univ-rouen.fr) ; [Youssef.Saidali@univ-rouen.fr](mailto:Youssef.Saidali@univ-rouen.fr) ;  
[Eric.Trupin@univ-rouen.fr](mailto:Eric.Trupin@univ-rouen.fr)

(2) LiDiFra, Université de Rouen  
[Maryvonne.Holzem@univ-rouen.fr](mailto:Maryvonne.Holzem@univ-rouen.fr)

(3) Greyc, Université de Caen  
[Pierre.Beust@info.unicaen.fr](mailto:Pierre.Beust@info.unicaen.fr)

**Résumé.** Cet article présente un projet de recherche en cours visant à améliorer les interactions d'utilisateurs de différentes catégories professionnelles avec un système d'information dédié au droit du transport et de la logistique qui repose sur un corpus de textes réglementaires et de compte rendus de jurisprudence. L'objectif vise à concevoir et à mettre au point un environnement numérique de travail (ENT) aidant un public professionnel (entreprises de la filière logistique, juristes, risk managers, assureurs, avocats, ...) à la fois dans son interprétation et sa perception des changements indicibles mais néanmoins permanents au sein des sphères d'activités humaines. Nous placerons notre projet dans un cadre théorique actuellement novateur au sein des sciences cognitives, celui de l'énaction. Ceci nous amènera à proposer une approche résolument centrée utilisateur dans la conception de l'ENT. Nous poserons la question de l'appropriation des contenus dans le cadre des documents numériques et nous décrirons les spécificités de notre corpus de travail. Enfin nous terminerons par une description des premières spécifications du futur ENT.

**Mots-clés.** Environnement numérique, usages, corpus, jurisprudence.

**Abstract.** This article describes a in progress research's project aiming at improving the interactions between users of various professional categories and an information system dedicated to transport laws and logistics which rests on a corpus of statutory texts and jurisprudence's reports. The objective is to conceive and develop a

numerical environment of work (ENT) for helping a professional public (logistics companies, lawyers, risk managers, insurers, lawyers ...). in its interpretation and its perception of the inexpressible changes but nevertheless permanent within the spheres of human activities. We will place our project within a currently innovative theoretical framework within cognitive sciences, that of the enaction. This will lead us to offer resolutely a user centred approach in the design of the ENT. After asking the question of the contents appropriation within the framework of the numerical documents, we will describe specificities of our corpus. At last we will finish by a description of the first specifications of the future ENT.

**Keywords.** Digital environment, uses, corpus, jurisprudence.

## 1 Le passage au numérique : prouesses techniques versus appropriation des contenus

Avec l'apparition des techniques numériques et de l'internet, la distance entre la population et l'information économique-juridique tend apparemment à se réduire. De nombreux sites proposent aujourd'hui une vaste gamme d'informations générales ou spécialisées. Sur le plan juridique, par exemple, il est aujourd'hui possible d'accéder à un large pan de la réglementation et de la jurisprudence, qu'elles soient françaises ou étrangères. Mais ce rapprochement technique n'est pas pour autant signe d'une meilleure maîtrise et appropriation de l'information, bien au contraire. Les spécialistes en sciences de l'information qui s'intéressent à la gestion des flux ainsi qu'aux stratégies de mise en forme et de mise en circulation de l'information font le constat des écarts (des fossés) entre prouesses technologiques et appropriation des contenus par des utilisateurs de plus en plus hétérogènes au sein d'une économie mondialisée. La question de l'appropriation ne se résout pas en effet par le simple ajout de métadonnées aux documents numériques comme dans le projet du Web Sémantique où l'objectif annoncé par Tim Berners-Lee (Berners-Lee98) initiateur du projet et directeur du W3C, est d'enrichir (notamment au moyen des technologies développées autour du langage XML) les documents (à l'aide d'ontologies normalisées, soit automatiquement, soit en assistant leur auteurs) avec « des informations sur leur sémantique propre qui soient directement interprétables par des agents logiciels sans la supervision d'une interprétation humaine ». Ce positionnement fait l'hypothèse implicite que la valeur sémantique d'un passage de document n'est le fait que de son auteur, alors que c'est tout autant celui de son lecteur confronté à ses pratiques professionnelles. Il s'agit d'une voie aristotélicienne déjà fort ancienne, reprise par F. Bacon puis G. W. Leibniz au 17<sup>ème</sup> siècle, persuadés de la nécessité d'un système universel d'organisation des connaissances ; système ontologique de nature transcendante qui est donc indépendant des points de vue particuliers. Ces universaux cognitifs seraient alors à même de pourvoir à la circulation de l'information, celle-ci occultant la question liée à son interprétation. Comme l'a souligné R. T Pédaque<sup>1</sup>, « la réponse du Web sémantique est d'indexer les textes avec les concepts d'une ontologie partagée par une large communauté » sans que celle-ci ne soit d'ailleurs clairement définie, ni surtout que soient prises en considération la diversité et l'évolution des pratiques langagières au sein de sphères d'activités hétérogènes

---

<sup>1</sup> Acronyme développé en analogie avec le Réseau Thématique Pluridisciplinaire sur le Document (RTP DOC) sous lequel des chercheurs issus des départements STIC et SHS (dont les auteurs) ont tenté d'approfondir collectivement la réflexion sur le document numérique.

Constatant que les systèmes actuels (interface de dialogue, bases de données, ...) conduisent à une interaction Système/Utilisateur forcément appauvrie, parce qu'ancrée dans un environnement prédéfini, (Peschard 04) celui des réponses, sous forme de thésaurus qui réorientent la question de l'utilisateur : à des requêtes en langue naturelle le système répond en terme de requêtes acceptables par le système d'information. Nous jetterons ici les bases de la conception d'un environnement numérique de travail E.N.T. capable de s'enrichir d'apports successifs dus à des interactions de plus en plus denses et complexes au sein de sphères d'activités devenues numériques. Cela nous conduit à sortir de la problématique du mot-clé, ou du figement lexical (référentiels statiques), pour celle de la thématique des textes et de l'interprétation située. Celle-ci s'ancrera dans l'alternance d'innovation et de sédimentation (le substrat culturel), laissant libre cours à l'*imagination réglée* (Ricoeur 86) de l'utilisateur. Notre démarche pose donc la question des signes précurseurs et indicibles de changements tels qu'ils peuvent être perçus par un individu ou un groupe d'individus au sein d'une sphère d'activité. Notre manière de percevoir les transformations est altérée par notre mode de pensée occidentale, au sein duquel, selon le sinologue François Jullien, « la transition fait littéralement trou » (Jullien 2009). Dans le cadre ontologique peuplé d'entités discrètes et de relations fonctionnelles, nulle place pour apporter des réponses aux interrogations de Parménide (Platon [trad.] 1923) sur la nature des passages entre états.

Il faudra attendre le 17<sup>ème</sup> siècle et l'ouvrage de Giambattista Vico (Vico 1744) pour une articulation entre rationalité, intuition et imagination, mettant en avant la spécificité de l'esprit humain qui est de relier au sein de la communauté des hommes. C'est dans cet esprit que nous concevons un environnement numérique de travail, offrant au lecteur la possibilité de structurer ses thèmes et références (constitution de ses propres molécules sémiques) en fonction de son environnement en même temps qu'il constitue un corpus de textes par sa navigation intertextuelle. De ce point de vue expérientiel la place dévolue à l'Umwelt par Von Uexküll (1934) prend tout son sens : les sujets se constituent en même temps qu'ils constituent leur environnement. En précisant la notion d'environnement, Dionisi (2006) en retire un point fondamental pour notre propos : « *Cet univers commun, ou environnement, ou domaine de connaissance est soumis aux règles de l'enaction : il n'existe pas a priori, mais en tant que phénomène participant à l'expérience humaine sur le monde. Il apporte en particulier à l'expérience ses règles et sa complexité. Mais c'est la capacité des systèmes (concernés par l'expérience) à internaliser tous ces éléments sous une forme qui leur est propre, et à eux-mêmes s'intégrer au milieu par couplage, qui permet à la réalité d'émerger* ». C'est dans un champ donné, cette gestion dans l'immédiateté, hors toute rationalité (par exemple la marche en terrain accidenté), que F. Varela appelle un micromonde.

Il faut bien affirmer ici que cette théorie contrecarre la systémique puisque l'internalisation permet un couplage avec l'environnement et ainsi évacue la notion de système en tant que modèle, représentation d'un monde pré-donné. Par contre l'apport de ce couplage enactif est considérable. C'est le couplage de l'humain avec son environnement qui permet de parler de cognition incarnée indissociable du vivant et de l'histoire du sujet pensant.

C'est dans ce contexte de la révolution du passage au numérique qui modifie les usages professionnels et privés et qui interroge le statut même des media que nous abordons cette question. Les réflexions pluridisciplinaires menées autour du document numérique (Réseau Thématique Pluridisciplinaire sur le Document<sup>2</sup> du

---

<sup>2</sup> <http://rtp-doc.enssib.fr/>

CNRS de 2005 à 2007, Semaine du Document Numérique en 2004<sup>3</sup> et 2006<sup>4</sup>, ...) nous invitent en effet à nous interroger sur les usages du document induits par le passage au numérique, tout particulièrement sur l'interprétation textuelle à l'œuvre en navigation intertextuelle (herméneutique numérique) et à définir de nouvelles approches pour les échanges de contenus ainsi que pour les interfaces cognitives et interactives à mettre en œuvre pour l'accès à ces contenus, surtout lorsque la collection de documents est importante et augmente en masse. Ces nouveaux corpus devenus numériques sont ouverts à une lecture discontinue et à une navigation intertextuelle. Ils nous invitent à ne plus relier le sens aux textes, dépôts de connaissances à partir desquels doivent opérer des outils d'extraction, mais aux situations de production et d'interprétation. C'est dans ce contexte que nous nous inscrivons, car le passage au numérique invite à une lecture extensive (*versus* lecture scolastique) à partir de textes fractionnés puis recomposés à dessein. C'est dans cette discontinuité devenant à la fois réticulaire (Adam 06) et réflexive (les textes se réfléchissant les uns dans les autres) que nous aborderons la question des documents concernant le droit en transport et logistique pour la gestion des risques.

Nous nous intéressons ainsi à la conception d'un environnement numérique de travail E.N.T., sorte d'extranet, dédié aux usages de la filière transport et logistique. L'intégration d'un ensemble de ressources et de services interopérables en son tout, dédiés non pas à une collection de cas d'usages particuliers, mais justement à une sphère d'activité (transport et logistique) large et en évolution rapide, constitue une réelle nouveauté, voire une singularité. En effet nous faisons ici la différence entre un ENT qui convoque des facultés interprétatives de l'utilisateur et un système de GED qui s'apparenterait plus à une boîte à outils fermée. La mise en œuvre de ce dispositif est susceptible de contribuer à des évolutions notables de l'usage de documents réglementaires et, plus généralement, des activités des acteurs de la filière transport et logistique.

## 2 Un corpus stratégique mais encore difficile d'accès

Le corpus réglementaire est encore difficile d'accès malgré une forte demande sociale et économique tout particulièrement en transport et logistique. A ce jour, le corpus et la base documentaire de l'Institut du Droit International du Transport (IDIT) sont accessibles en ligne. Cette base s'adresse à des adhérents spécialistes du droit mais elle est difficilement utilisable par un novice dans le domaine juridique comme un transporteur, qui chercherait des informations pour la mise en place de conditions de transport de marchandises conformes à la législation en vigueur, par exemple. Cette base documentaire est associée à un thésaurus hiérarchisé « maison » pour améliorer son interrogation. Elle est renseignée manuellement à partir de décisions rendues par diverses juridictions françaises et étrangères depuis 1971, de revues papier ou en ligne ou d'après l'interrogation d'autres sources de données en ligne auxquelles l'IDIT a accès. Elle impose aussi la saisie manuelle de comptes-rendus (CR) de jurisprudence et de réglementation sous la forme de fiches. Cette captation de l'information et la veille présentent des difficultés majeures pour renseigner et mettre à jour le système d'information. Nous envisageons, suite à la numérisation (en cours) des collections papiers (des milliers d'articles et de décisions de justice relatifs à des risques et litiges en matière de transports), une aide à l'interprétation de contenus textuels. Pour cela, une analyse préalable du corpus est nécessaire. Nous détaillons ci-dessous les points de vue théoriques linguistiques que

---

<sup>3</sup> <http://www.univ-lr.fr/sdn2004/>

<sup>4</sup> <https://diuf.unifr.ch/event/sdn06/accueil.html>

nous comptons éprouver dans cette analyse à commencer par l'étude des spécificités propres aux contraintes normatives qui pèsent sur un genre textuel à forte valeur perlocutoire. C'est donc par une typologie de la structure argumentative en lien avec la présentation matérielle des données et les attentes d'adhérents hétérogènes (juristes, transporteurs, ...) que nous abordons cette étude de corpus.

Le SI de l'IDIT est conçu pour diffuser des informations aux adhérents (services payants) afin qu'ils puissent gérer dans les meilleures conditions leurs entreprises et sécuriser leurs activités. Ces acteurs du transport et de la logistique se doivent de rechercher et d'analyser des informations de plus en plus nombreuses. Ainsi le suivi et l'anticipation des cadres juridiques communautaires et nationaux sont des éléments incontournables. Or, la fragmentation de l'information relative au droit des transports et de la logistique qui couvre des domaines aussi variés que le droit commercial, le droit des sociétés, le droit de l'environnement, le droit administratif, le droit pénal, le droit social, rend difficile l'accès à l'information (information éparse, réglementation pléthorique, accès difficile et coûteux...), d'où la nécessité d'une mise en relief (signalément pour interprétation) de celle-ci.

### **3 Une approche nouvelle de la veille : herméneutique juridique et énonction**

Si la veille a été définie classiquement par l'un de ses pionniers français comme *le moyen pour l'entreprise de faire émerger les éléments stratégiques de la masse d'information disponible aujourd'hui* (Dou 97), la plupart des outils logiciels de Traitement Automatique des Langues (TAL) recherchent dans la discrimination de mots pertinents les signes annonciateurs des changements à venir. Ils s'en remettent à la décontextualisation (l'extraction), la lemmatisation (élimination de la variation suffixale), la cartographie basée sur des proximités syntaxiques (morphologie dérivationnelle, KPPV (K plus proches voisins) etc.) ou sémantiques (comme les relations hyper/hyponymie par marqueur ontologique « est un ») pour pouvoir séparer dans la masse des données *le bon grain de l'ivraie*. Les spécialistes de la veille, côté fouille de données ou de textes, parlent alors de KDD (Knowledge Discovery in Databases) ou de KDT (Knowledge Discovery in Texts) (Kodratoff 1999) comme d'une science qui découvre des connaissances contenues dans des textes. Qu'il s'agisse de fouille de données ou bien de textes, l'essentiel est de trouver des « modèles », des patrons, permettant de séparer l'information intéressante (signal pour l'action) du bruit. Nous nous proposons ici d'appréhender cette question sous un tout autre angle, celui de la situation d'un sujet interprétant dans le contexte d'une navigation numérique intertextuelle.

Consécutivement à une étude économique des services que peut rendre l'IDIT, portée par le pôle de compétitivité Logistique Seine Normandie (devenu Nov@log), deux types d'outils sont alors apparus comme nécessaires : un outil de veille globale pour renseigner le SI et un outil de diffusion. Le premier doit permettre, par exemple, de faire un état des lieux hebdomadaire selon un domaine précis (avec une requête) par interrogation automatique (à l'aide d'un moteur de recherche). Le second doit permettre une veille personnalisée sur le SI pour un adhérent, spécialiste ou non, du droit. Il s'agit de créer des alertes pour les adhérents hétérogènes de la base dans le domaine du droit des transports et de la logistique et plus particulièrement de la sûreté dans le transport.

Il convient de signaler le statut exemplaire de l'herméneutique juridique du point de vue de l'interprétation en contexte. La tâche d'interprétation d'une loi consiste à instancier cette loi dans chaque cas particulier. Cette instanciation (nous dirions

aujourd'hui l'interprétation contextualisée ou située) est d'ailleurs le thème central de la jurisprudence. Une loi ne demande pas en effet à être comprise historiquement mais doit se concrétiser dans sa valeur juridique (tant que sa fonction subsiste) à travers ses cas particuliers d'interprétation. Le juriste cherche en effet, comme le rappelle (Gadamer 76), à être fidèle à l'intention juridique de la loi en la mettant en rapport avec le présent. Il s'agit bien ici d'une démarche d'actualisation, de représentation au sens de rendre présent qui confère à l'herméneutique juridique une valeur alliant philologie et histoire (le sens d'un extrait de texte, d'un terme, relevant *de facto* de l'histoire de ses interprétations actualisées en contexte). Ce contexte ne saurait être anticipé tant la situation de chaque utilisateur de l'ENT lui est propre. Etant donné l'unicité de chaque cas de litige ou de mise en conformité avec la réglementation, nous ne pouvons disposer de connaissance *a priori* pour renseigner l'utilisateur de cet ENT. Son besoin d'information dans son intention globale doit être mis en rapport avec sa navigation locale en rapport avec son problème posé. Dans cet esprit, il nous semble intéressant de nous positionner scientifiquement avec une posture issue de la théorie de l'énaction établie à partir de l'observation du vivant par deux biologistes, Maturana et Varela, qui repose sur la propriété d'autopoïèse (auto constitution) propre au vivant, et sur l'affirmation que la connaissance est incarnée donc indissociable du vivant et de l'histoire du sujet pensant. La cognition n'est alors pas affaire de représentations mais d'actions incarnées. Il s'agit, selon nous, d'une nouvelle méthodologie de conception qui place l'interaction (le couplage) avec l'utilisateur au centre de la démarche et qui est basée sur un environnement numérique permettant de faire émerger des usages non prescrits *a priori* au cours des expérimentations à partir des différentes fonctionnalités proposées. Ceci est en rupture avec les méthodes de conception pour lesquelles les cas d'usages sont premiers. L'utilisateur disposera d'un ensemble de fonctionnalités ou d'outils qui lui permettra de construire son propre parcours interprétatif à partir d'informations rendues disponibles par cet ENT. Les outils dont nous disposons pour partie et que nous envisageons de développer (cartographie documentaire, traitement automatique de la langue, visualisation et navigation dans un grand ensemble de documents) seront intégrés et mis à disposition dans l'ENT afin que l'utilisateur puisse formuler ses requêtes et être aidé dans l'interprétation des résultats (textes réglementaires) qui lui sont proposés pour affiner et/ou reformuler sa recherche d'information. Finalement, la réalisation de cet ENT repose bien sur une nouvelle approche dans la conception d'interfaces interactives et cognitives où l'engagement cognitif de l'utilisateur sera concentré dans son activité et non plus déplacé dans la machine, par exemple par apprentissage par celle-ci de cas d'utilisation selon des finalités connues *a priori*.

### 3.1 Emergence de nouveaux usages des outils

L'approche herméneutique et énative dans la conception et l'intégration d'outils marque une différence de point de vue avec les méthodes évoquées ci-dessus en favorisant une démarche scientifique expérimentale. Nous mettons en avant l'expérimentation comme une boucle de conception où la modélisation n'est pas une étape initiale, pas plus que les évaluations (non nécessairement comparatives) sont des étapes finales. L'objectif ici n'est pas de chercher à faire mieux certaines tâches déjà réalisées avec des méthodes éprouvées mais plutôt d'inventer de nouveaux usages du TAL ainsi que de nouvelles façons d'utiliser des outils informatisés dans des recherches sur le langage. Il ne s'agit donc pas, comme dans la plupart des systèmes de Traitement Automatique des Langues, de proposer une fonctionnalité complexe (extraction de termes, de relation, classification automatique, annotation automatique) orientant le parcours interprétatif du lecteur

sur le texte, mais d'utiliser des fonctionnalités élémentaires qui permettent à l'utilisateur de faire émerger des fonctionnalités de plus haut niveau par combinaison. Des fonctions atomiques proposées émergeront alors de nouvelles fonctionnalités actualisées par l'interaction entre les utilisateurs et le corpus. L'idée est, par exemple, de proposer à l'utilisateur des rapprochements de contextes syntagmatiques (alignement) et de le laisser en inférer des classes sémantico-lexicales (apparaissant selon des patrons, des contextes particuliers).

Le but serait ici de combiner plusieurs outils de TAL existants au sein d'un ENT, en particulier :

- des outils centrés utilisateurs (Themeditor, Lucia, Proxidocs développés au laboratoire GREYC à Caen),
- des outils de statistique textuelle tels Lexico 3 (Salem), Hyperbase (Brunet), Nooj (Silberztein), permettant de construire une topographie des segments de textes (Lexicométrica 2007) (Viprey 05).

Cette expérimentation permettra ainsi de tester différents modes de présentation des données textuelles qui supportent une navigation interactive dans ces textes.

### 3.2 L'interprétation comme « énoncé de »

D'un point de vue expérimental, il s'agira de savoir comment un environnement numérique de travail, et le couplage qu'il induit, permettent l'émergence par énoncé d'une perception sémantique du corpus et ainsi un meilleur accès aux documents juridiques. Si on considère que le sens provient de la démarche outillée de l'interprétant face à un texte et à son intertexte, alors, une expérimentation mettant en oeuvre un environnement numérique de travail qui permet d'effectuer des traitements sur des documents électroniques participe à la co-production de sens pour l'expérimentateur. Dans le couplage personne-système les interprétations des utilisateurs et les traitements des machines ne sont pas en concurrence. Au contraire, nous les pensons comme complémentaires dans la mesure où l'activité d'une machine a pour objectif de produire dans l'interaction des traces qui vont participer aux interprétations du, ou des utilisateurs. Nous poursuivons ici l'idée de (Dionisi et Labiche 06) qui consiste à caractériser des *processus logiciels* impliqués dans des *processus expérimentiels*, eux-mêmes impliquant des *processus cognitifs*.

Nous nous situons dans l'opérationnalisation de protocoles expérimentaux. Il s'agit de faire un état de l'art des outils de navigation textuelle et intertextuelle existants, de les implanter, de permettre leur appropriation (ergonomie linguistique) et de faire une analyse précise et détaillée de leurs apports pour jeter les bases d'un environnement numérique de travail ENT. Cet ENT doit permettre de mettre en interaction des outils linguistiques et des outils de navigation, utilisables selon la démarche adoptée pour la construction dynamique d'une ressource termino-ontologique personnalisée.

Cette recherche s'insère dans le champ très large de la représentation d'information par navigation et du traitement automatique de la langue naturelle pour lequel il existe de multiples outils opérationnels. Notre positionnement, novateur et ambitieux, ne cherche pas à développer de nouveaux outils mais bien à combiner, séquencer, relier, rendre plus interactifs ceux qui existent déjà, en renouvelant les hypothèses et voies de recherche grâce aux apports combinés de l'herméneutique matérielle et de l'énoncé. La charge cognitive de production de sens résulte alors de l'histoire et du couplage des diverses actions qu'accomplit un être dans le monde (*comprendre* peut alors s'appréhender comme un *agir avec*). Il s'agit là, comme le remarque François Rastier (Rastier 05) d'un courant de pensée qui,



comme l'herméneutique matérielle, ne se présente pas comme une théorie globale mais comme une voie de recherche conduisant à un questionnement permanent, d'une part, des textes (domaine de l'interprétation comme *énaction de*) et, d'autre part, de la place qu'il convient de réserver aux outils informatiques dans le traitement des données.

#### 4 Une approche centrée utilisateur

Notre approche de l'accès aux documents se situe à l'opposé de celles défendues dans le cadre du Web Sémantique. Là où le Web Sémantique cherche à rendre le plus possible partagées de vastes ontologies qui synthétisent une connaissance pensée comme objective et devant convenir à tous les utilisateurs, nous préférons manipuler des ressources termino-ontologiques (bases de données terminologiques, représentations du contenu lexical etc.) propres à un utilisateur ou un petit groupe d'utilisateurs et liées à leur tâche, leurs besoins et de leurs centres d'intérêt. Il en découle une certaine *légereté* sémantique de ces ressources, au sens de (Perlerin, 04), dans la mesure où elles ne représentent que ce qui est important du point de vue de l'utilisateur et restent ainsi de taille raisonnable (par exemple une centaine de termes) ce qui les rend moins complexes à construire, à maintenir et à enrichir.

Cette approche centrée utilisateur conduit à opérer un certain renversement scientifique relativement aux ressources qu'utilisent les modèles de TAL. Premièrement, d'un point de vue très pratique, force est de constater que des ressources très généralistes, valables pour tout type de traitement envisagé ainsi qu'à destination de tout utilisateur potentiel, ne sont pas facilement disponibles (sous forme électronique pour des traitements automatiques) et encore moins gratuites. Deuxièmement, nous soutenons que l'idée même d'une ressource généraliste est illusoire car elle dépend inévitablement du contexte qui lui préexiste (le but recherché par le ou les auteurs ainsi que leurs spécificités socioculturelles). Le rapport de l'Action Spécifique 32 du CNRS/STIC en 2003 (Charlet & al 03) va également dans ce sens en précisant un obstacle au projet du Web Sémantique : la détermination et l'ajout, même de simples méta-données, n'est pas une activité naturelle pour la plupart des personnes.

La tradition logico-grammaticale et plus précisément la sémantique formelle et computationnelle cherchent à représenter et à produire, automatiquement ou pas, des formes le plus possible objectivées des significations et du sens. Dans la démarche centrée utilisateur, nous partons d'une position duale où nous considérerons que les traitements sémantiques appliqués à l'accès aux contenus des documents ont tout à gagner à être le plus possible subjectivés, tant du point de vue des ressources que du point de vue des résultats opératoires. Cette démarche nous paraît être une réponse au constat que dressent Didier Bourigault et Nathalie Aussenac-Gilles à propos de la variabilité des terminologies qu'il y a *autant de ressources termino-ontologiques que d'applications dans lesquelles ces ressources sont utilisées* (Bourigault & Aussenac. 2003).

Les ressources qui sont les plus importantes pour un utilisateur dans une instrumentation informatique pour l'accès aux documents sont celles qui doivent être produites de manière endogène dans une boucle d'interaction entre un outil logiciel, un utilisateur et des corpus. Dans cette boucle, chaque pôle est déterminant. Il en découle une importance significative des corpus utilisés qui du coup ne peuvent plus être considérés uniquement comme un réservoir de formes attestées sur lequel on tenterait de mettre en œuvre un calcul à base de ressources exogènes.

Le corpus utilisé est à l'origine des ressources lexicales construites et constitue en même temps le matériau d'expérimentation. L'accès personnalisé au contenu s'inscrit dans un processus interprétatif en aller-retour entre des outils (des logiciels d'étude), des corpus (des corpus d'étude) et des ressources personnelles, les uns étant conditionnés par les autres.

#### **4.1 L'indispensable intervention de l'utilisateur : l'exemple d'interprétation d'images de documents**

Dans notre approche herméneutique et énaïve du langage et dans le but d'une instrumentalisation en TAL, nous mettons l'accent sur l'interprétation plus que sur les connaissances. La priorité est donnée aux spécificités sociolinguistiques des utilisateurs (par exemple leurs centres d'intérêt, leurs habitudes terminologiques, leurs parcours interprétatifs). Ce qui a du sens pour les utilisateurs ne se réduit pas à une représentation et encore moins à une formalisation. Ce n'est pas le résultat d'un calcul, c'est une activité au centre d'une interaction homme-machine (activité qui de plus n'est pas forcément finalisée dans le temps). Ainsi, on remet en cause l'idée qu'un mot, une phrase, un texte ou un corpus ait du sens (ou non) pour défendre plutôt l'idée qu'ils font sens (ou pas) dans un couplage personne-système

Un petit historique du traitement d'image de documents (scan ou photo) illustrera notre propos. Alors que pour un lecteur humain, tout document, qu'il soit textuel, graphique ou mixte est un support direct d'information en ce sens qu'il peut en extraire immédiatement des informations, l'image, elle, n'est qu'un tableau de valeurs (pixels) que l'ordinateur ne peut décrypter immédiatement. Une vraie difficulté provient de ce que l'apprentissage de la lecture est prégnant pour le sujet humain, et qu'il lui est donc impossible d'explicitier simplement comment il procède (Labiche et al. 1993).

Dans les années 90, l'Intelligence Artificielle (IA) quoique en retrait par rapport aux espoirs fous (traduction automatique, diagnostic médical, jeux, ...) de ses débuts, reste un domaine en expansion dont on attend toujours beaucoup. L'informaticien, « maître » des algorithmes, doit pouvoir résoudre des problèmes aussi simples que la reconnaissance optique de caractères et de symboles cartographiques connus, ou bien la reconstitution de graphes de traits (type plans EDF ou France Télécom) à partir d'algorithmes de suivi de contours, de croissance de zones, de recherche de formes connexes etc., suivis d'une reconstruction à base de graphes. On dispose de tous les atouts pour développer des systèmes industriels de rétro conversion de plans techniques ou de documents textuels de bonne qualité en utilisant des systèmes à base de règles. Les grandes entreprises pressant le futur marché des données numériques, mettent en chantier d'ambitieux plans de rétro conversion de documents (BNF, France Télécom, EDF, IGN, DGI,...) dont le but est d'extraire automatiquement les données présentes sur les documents et de les stocker dans de grandes bases de données qui peuvent être topologiques (plans de masse, cadastre, ...). L'objectif est bien d'extraire ces données et leurs relations spatiales de manière à pouvoir construire les systèmes d'information géographique (SIG) correspondants. Les algorithmes de traitement d'images sont alors chaînés de manière à rechercher toutes les informations présentes. Avec ce choix du « tout automatique » l'utilisateur n'aurait qu'à configurer le système pour mettre en œuvre une application qui reconnaît, puis stocke toutes les données contenues dans le document.

Cette démarche du « tout automatique » perdure et est nécessaire dans beaucoup d'applications industrielles ; elle correspond à des types de problèmes suffisamment contraints et stables pour que les connaissances métier nécessaires puissent être codées « en dur » dans le dispositif. Par contre cette démarche s'est

avérée être un échec complet et n'a pas abouti à des systèmes opérationnels lorsque le problème à traiter nécessite des connaissances contextuelles importantes. Il apparaît alors une nouvelle catégorie de projets industriels pour lesquels l'intervention de l'utilisateur devient indispensable pour corriger les résultats ou faire un choix lorsque le système en est incapable. L'exemple caractéristique de cette approche concerne le traitement des formulaires du recensement de 1999 (Gilloux, 2001) pour lequel l'opérateur valide, ou non, les propositions du système concernant la reconnaissance de champs d'écriture manuscrite. Pour d'autres dispositifs dotés d'algorithmes d'évaluation des résultats, seuls ceux qui posent problème sont présentés à l'opérateur. Ces nouveaux dispositifs nécessitent une intervention humaine dans la boucle.

Après ce premier pas, il est apparu aux nombreuses sociétés (SWT, A2ia, ITESOFT, Jouve, etc.) qui proposent des services en ingénierie du document (formulaires, archives, manuscrits etc.) qu'il y avait un marché pour traiter des masses de documents d'un même modèle. De nouveaux dispositifs s'avèrent alors nécessaires : ce sont des systèmes de traitement comportant des outils logiciels reconfigurables pouvant être chaînés par le chef de projet, industriel compétent aussi bien en informatique qu'en ingénierie documentaire, et expert dans le domaine professionnel auquel se rattache la masse de documents. Dans ce cadre, l'intervention humaine est celle du concepteur de la chaîne de traitement. Il dispose également d'outils de test et d'interfaces permettant de visualiser et corriger si besoin les résultats intermédiaires. Dans l'entreprise, l'organisation du travail n'est pas fondamentalement remise en cause, les procédures métier restent identiques en leur principe, les utilisateurs finaux doivent accepter l'aide informatique et le maniement d'outils informatisés de plus en plus sophistiqués.

Cette problématique met l'accent sur les différents rôles des « utilisateurs » d'outils informatisés selon leur expertise et amène à s'interroger sur les capacités d'interaction avec les systèmes. La réponse à la nécessité de proposer des solutions informatiques susceptibles de prendre en compte des connaissances, une contextualisation, une mise en situation pour automatiser au mieux des services dédiés a été une évolution des logiciels informatiques vers des systèmes que l'utilisateur devenu également concepteur doit s'approprier pour pouvoir articuler les traitements informatique proposés. Il apparaît ainsi un glissement entre les rôles d'utilisateur et de concepteur et on peut légitimement se demander si les systèmes utiles, susceptibles de rendre des services « intelligents » car contextualisés ne doivent pas être conçus comme ces systèmes industriels qui acquièrent ainsi les capacités créatrices, imaginatives lors des interactions avec un utilisateur-concepteur compétent qui accepte de consacrer du temps à la re-conception d'un système informatique efficace.

#### 4.2 L'exemple de la cartographie documentaire

L'expérience de la cartographie documentaire dans le projet ProxiDocs (Roy 2007) est un exemple de mise en place d'un système interactif centré utilisateur avec ressources endogènes. Cette expérience sera reprise et mise à contribution pour la conception de l'ENT.

Le but du logiciel d'étude ProxiDocs<sup>5</sup> est de plonger son utilisateur (ou un petit groupe d'utilisateurs) dans des interactions qui offrent la possibilité de notamment mieux discerner l'homogénéité thématique d'un corpus, de mettre en évidence sa densité, d'en extraire les principales tendances thématiques de chaque document et de permettre un accès rapide à tel ou tel document ou passage de

---

<sup>5</sup> <http://www.info.unicaen.fr/~troy/proxidocs>

document. Au sein de ces interactions l'outil permet de produire et de naviguer dans des représentations graphiques personnalisées que l'on appelle des cartes.

Au préalable, l'utilisateur doit décrire un ensemble de termes qui sont ceux qui l'intéressent et les fournit en entrée au logiciel ainsi que son (ou ses) corpus. Ces termes peuvent être représentés selon deux modèles, soit sous forme d'une liste de graphies, soit sous la forme d'un dispositif de représentation sémique différentielle des significations des termes. A ces deux formes de ressources lexicales correspondent des outils interactifs qui permettent à des utilisateurs de les constituer de manière incrémentale. Avec ces ressources, ProxiDocs construit des cartes dynamiques et interactives (en 2 ou 3 dimensions statiques ou bien animées) ainsi que des visualisations des textes agrémentées d'une visualisation par coloriage des isotopies.

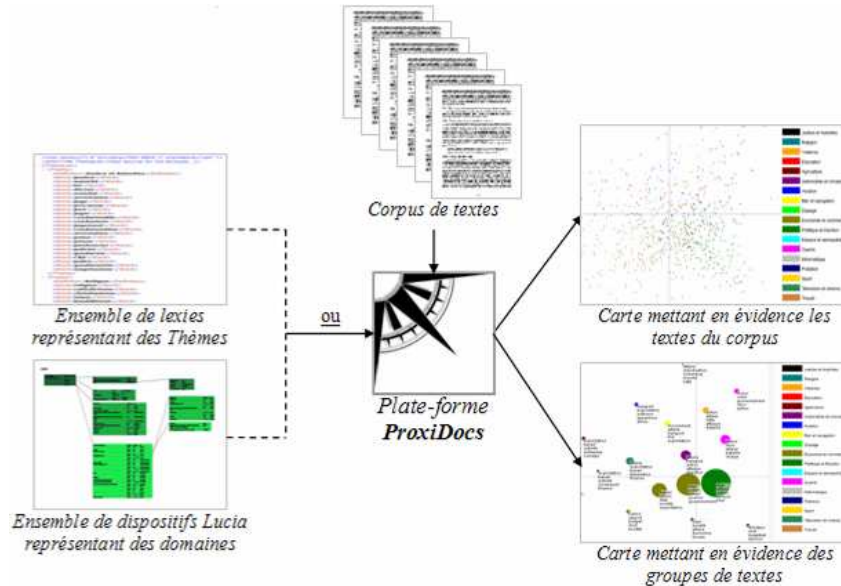


Figure 1 : Utilisation de ProxiDocs

Ces outils logiciels seront re-développés pour être intégrés à l'ENT et devront être complétés par de nombreux outils de TAL, de manipulation de corpus et de navigation que nous détaillons par la suite. Dans le couplage personne-système les interprétations des utilisateurs et les calculs des machines ne sont pas en concurrence car les uns n'ont en aucun cas le but de supplanter les autres. Au contraire, nous les pensons comme complémentaires dans le sens où l'activité d'une machine a pour objectif de produire dans l'interaction des traces qui vont participer aux interprétations du ou des utilisateurs.

## 5 Projet en cours : vers un ENT

D'un point de vue expérimental, il s'agit de savoir comment un environnement numérique de travail, et le couplage qu'il induit, permettent l'émergence par énonciation d'une perception sémantique du corpus et ainsi un meilleur accès aux documents juridiques. Si l'on considère que le sens provient de la démarche outillée de l'interprétant face à un texte et à son intertexte, alors, une expérimentation mettant en oeuvre un environnement numérique de travail qui permet d'effectuer

des traitements sur des documents électroniques participe à la co-production de sens pour l'expérimentateur.

### 5.1 Méthodes et outils

Dans notre stratégie d'amélioration de la navigation intertextuelle, nous proposons à l'utilisateur plusieurs approches pour naviguer dans l'ensemble des documents, visualiser, manipuler et organiser le résultat de ses recherches. Il pourra notamment s'appuyer sur l'historique de sa navigation, ses propres traces, mais aussi celles qui sont liées à sa sphère d'activité (collectif de travail). Il s'agira donc d'observer l'utilisateur dans son activité, et de lui permettre d'exploiter dynamiquement cette observation. Avec ses traces (volontaires ou involontaires) formalisées de manière à rendre compte non pas des événements seuls (clics souris), mais surtout des liens entre événements, processus « machine » et processus cognitifs (perception-action), nous ne cherchons pas à modéliser un comportement pour faire de la prédiction, mais à disposer d'outils de description et d'analyse de la navigation intertextuelle en situation réelle.

La visualisation et l'analyse des résultats de la recherche sont des étapes nécessaires qui s'inscrivent dans le processus global de recherche d'information. La perception de l'information est liée à la prise de décision dans le contexte d'utilisation de l'ENT proposé. Elle est donc initialisée par une tâche métier ou une motivation personnelle, elle reflète une culture et des contraintes organisationnelles, ou encore des normes sociales. Les tâches d'interprétation évoluent avec l'avancement du processus de reconnaissance/satisfaction en besoin d'information. Dans ce contexte l'utilisateur identifie les sources documentaires, formule des requêtes et examine les résultats ; il se retrouve ainsi au centre d'une boucle itérative « *formulation-analyse-visualisation-reformulation* » dans une complémentarité personne système comme celle de la figure 2 (inspirée de Kules et al 2008).

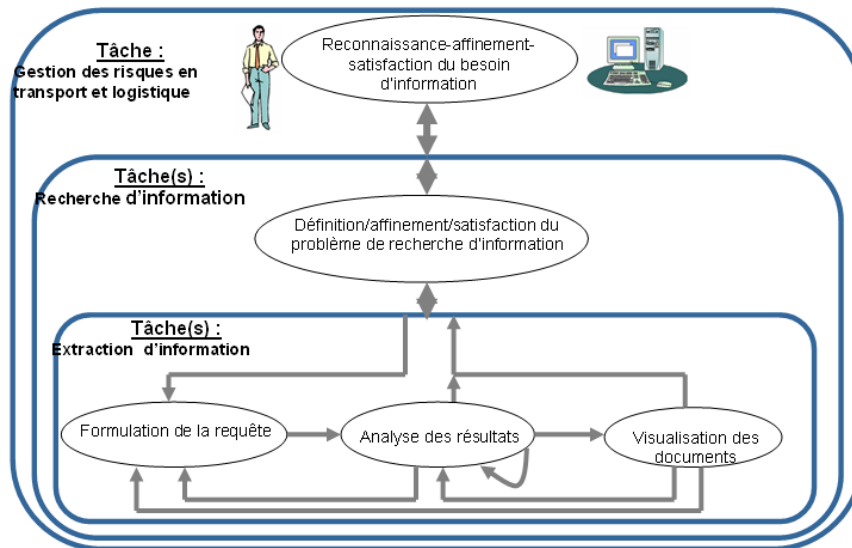


Figure2 : Complémentarité entre un agent humain éactif et un système

Le processus est donc initialisé lorsqu'un utilisateur identifie un besoin informationnel et tente de le satisfaire en entreprenant une ou plusieurs tâches de recherche. Il prend des décisions sur la ou les stratégies à adopter, les outils à

exploiter et le corpus ou partie du corpus à consulter. Chaque unité d'information découverte peut déclencher de nouvelles idées, suggérer de nouvelles directions et changer la nature même du besoin d'information (Bates 1989). On émet alors l'hypothèse que la gestion sous forme d'historiques de traces (incluant points de blocage et retours arrière) laissées par les différents utilisateurs peut aider à la découverte de nouvelles stratégies et de nouvelles informations.

Pour ce qui est de l'extraction d'information, chaque action implique un engagement cognitif et physique, et peut induire une évolution dynamique de l'interface ou des connaissances en émergence. Nous cherchons à faciliter le couplage et l'engagement de l'utilisateur en lui proposant des outils simples pour la manipulation/sélection/déplacement des documents résultats, ainsi que pour l'expression dynamique des requêtes. L'environnement numérique de recherche et de visualisation d'information que nous proposons de développer de façon incrémentale présentera alors des modalités de base :

- Définition interactive (graphique ou non) des requêtes
- Classification suivant des attributs et items définis dynamiquement par l'utilisateur ou le groupe d'utilisateurs de la sphère d'activité :
  - A partir d'une liste hiérarchique que l'utilisateur a la possibilité de reconfigurer (avec l'ajout ou la suppression d'éléments).
  - A partir de différents points de vue (Thèmes, catégories, dates d'édition, taille, numéro de fiche ..., etc.) exprimés et ré-organisables par l'utilisateur. Ces points de vue permettent à l'utilisateur d'ajouter des contraintes appropriées à son contexte de recherche (problèmes à résoudre et compétences). Pour simplifier la recherche, un passage rapide de la souris sur un point de vue peut donner le nombre de documents pouvant être sélectionnés à partir de ce point de vue.
  - A partir de tout ou partie de documents visualisés, l'utilisateur demande des documents *similaires*.
- Visualisation de l'ensemble des résultats de la recherche sous forme de :
  - Liste : Il s'agit d'une approche classique pour visualiser un ensemble de documents. L'ordre dans lequel les résultats sont présentés peut être déterminé ou non par des mesures liées à la pertinence du document au vu du contexte utilisateur.
  - Représentations 2D : On applique ici les techniques de visualisation de grands ensembles documentaires (Jacko & Sears06). La classification proposée par l'utilisateur sert de base pour une présentation animée sous forme de clusters, treemap, arbres hyperboliques, Zoom 2D, fish-eye, matrice... etc. L'utilisateur peut naviguer dans chaque catégorie, à chaque niveau de la hiérarchie et visualiser les résultats associés à la classe, en les organisant suivant ses propres paramètres.
  - Représentations 3D : En complément de la visualisation 2D, on vise avec une visualisation 3D une émergence de connaissances.
- Outils d'extraction et analyse de tout ou partie d'un document du corpus
- Affinement dynamique du contexte de la recherche dans une zone de filtrage

Notre démarche consiste à encapsuler sous forme de services l'ensemble de ces méthodes et outils, puis utiliser des représentations spatiales dynamiques sur une plate-forme générique en personnalisation de la visualisation et en intégration de modalités variées et hétérogènes.

## 5.2 Architecture

Notre plate forme en cours de développement est basée sur une architecture orientée services utilisant des grammaires XML et les Web Services. L'ensemble très riche de services proposés à l'utilisateur concerne l'acquisition des ressources (fiches), les prétraitements, le stockage, la classification, l'enrichissement sémantique, de même que l'exploitation dynamique et la navigation. Chacun de ces services est implémenté en WSDL (Web Service Description Language). Cette description dans un format XML permet de préciser pour chaque service, les méthodes pouvant être invoquées ainsi que la signature et les points d'accès. Ainsi formalisés, les services sont stockés dans un UDDI (Universal Description, Discovery and Integration); un annuaire évolutif de web services distribués qui facilite la publication et l'exploration. L'interopérabilité des services est rendue possible grâce à SOAP (Simple Object Access Protocol). Ce protocole d'échange inter-applications est également basé sur le langage XML, et s'appuie sur HTTP pour le transport des données. Enfin l'ensemble de la plate forme est orchestrée par l'ESB (Entreprise Service Bus) PETALS, suivant la norme JBI (Java Business Integration).

La plate forme ainsi déployée permet l'intégration et l'interopérabilité de composants et services hétérogènes. L'accès haut niveau aux différents services par l'utilisateur se fait par le portail présenté figure 3. Ce portail web utilise LIFERAY comme conteneur de portlets encapsulant les différents services. Tous les portlets sont agaçables et personnalisables dynamiquement par l'utilisateur.

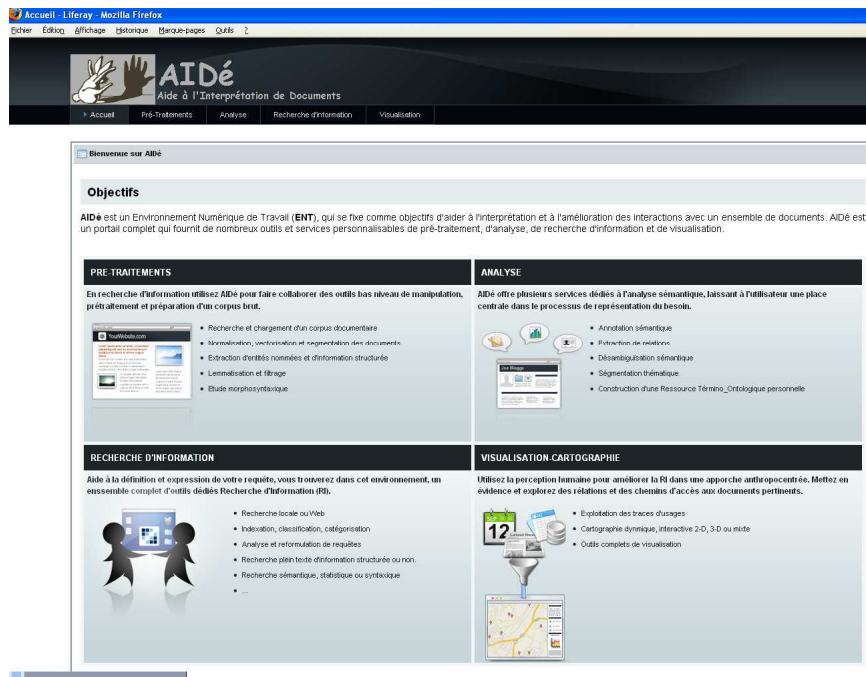


Figure 3 : Environnement personnalisable d'accès aux services

## 6 Conclusion

L'expérimentation de cet ENT auprès de plusieurs types d'utilisateurs nous semble nécessaire, car elle permettra de valider un certain nombre de nos postulats en s'appuyant sur une étude linguistique fine des interviews en première et deuxième personne menées avec ces utilisateurs professionnels au cours de navigation dans une base de documents. Nous mènerons cette expérimentation auprès des juristes attachés au droit du transport (clients de l'IDIT) et spécialistes en gestion des risques liés au transport auprès des assureurs.

Partant des postulats philosophiques, linguistiques et informatiques qui nous ont guidé pour élaborer ce prototype, nous donnerons la possibilité à un usager de construire son propre parcours interprétatif en fonction de ses pratiques sociales. Par l'actualisation de son interprétation en contexte issue de sa navigation intertextuelle, nous espérons œuvrer à une meilleure compréhension de ce que Thomas d'Aquin appelait au XIII<sup>ème</sup> siècle, l'intelligence interprétative, car *quel que soit ce qui est reçu, il l'est selon les modalités propres du récepteur* [Quidquid recipitur ad modum recipientis recipitur Thomas D'Aquin 1272 Liber de causis prop 10: Somme I,75,5c] Il nous a semblé impératif de mener cette réflexion conduisant à ce début d'expérimentation dans le contexte des corpus, devenus numériques, qui se sont ouverts à une lecture discontinue. Notre ENT ne cherche pas en effet à intégrer les intentions *a priori* de l'utilisateur dans la machine, mais l'invite en suscitant de nouvelles, à relier *in situ* les données textuelles et à paramétrer librement les outils mis à sa disposition<sup>6</sup>.

## 7 Bibliographie

Adam J-M. (2006) Autour du concept de texte. *Pour un dialogue des disciplines de l'analyse de données textuelles*, disponible à [http://www.cavi.univ-paris3.fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006\_JMA.pdf]

Bates M J. (1989). *The design of browsing and berrypicking techniques for the online search information*. Online review, 13, 407,-431

Berners-Lee,T. (1998) *What the semantic web can represent ?* W3C[http://www.w3.org/designissues/rdfnot.html]

Bourigault D. & Aussenac-Gilles N. (2003) Construction d'ontologies à partir de textes, *Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2003)*, Batz-sur-Mer, T2, pp. 27-50

Charlet,J, Laublet,P, Reynaud,G., (2003) *Web sémantique- Rapport de l'Action Spécifique 32 CNRS/STIC*

Dionisi D. (2006), *Proposition d'une méthodologie d'opérationnalisation informatique de l'approche enactive de la cognition*, thèse d'informatique soutenue le 11 décembre 2006 à l'INSA de Rouen.

Dionisi D., Labiche J. (2006). *Enaction et informatique : les enjeux de l'opérationnalisation technologique d'une théorie de la cognition*, in Actes du colloque ARCo 2006, 6 au 8 Décembre 2006 – Bordeaux

---

<sup>6</sup> Les auteurs de cet article sont membres d'un groupe de recherche autoqualifié d'interdisciplinaire (le groupev) auquel appartiennent également d'autres chercheurs des trois laboratoires normands mentionnés. Cet article est le fruit de notre collaboration fructueuse et désintéressée.



- Engisch K. (1953) *Die Idee der konkretisierung* : Abh der Heidelberger Akademie.
- Gadamer H-G. (1976) *Vérité et méthode : les grandes lignes d'une herméneutique philosophiques* Ed du Seuil
- Gilloux M. (2001) Une application industrielle de numérisation et de lecture automatique de documents : la saisie des questionnaires du recensement de la population de 1999., *CIDE 2001, 4e Colloque International sur le Document Électronique, Méthodes, démarches et techniques cognitives*, 24-26 octobre 2001 – Toulouse
- Jacko J.A., Sears, A. (2006). *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, 2<sup>nd</sup> Edition, Lawrence Erlbaum Associates.
- Jullien F. (2009) *Les transformations silencieuses*, Edition Grasset, Paris
- Kules B, Shneiderman B. (2008). *Users can change their web search tactics : Design guidelines for categorized overviews*. Information Processing and Management.
- Labiche, J., Ogier, JM., Balan B., Caston J. (1993). Stroop effect : an example of conflict analysis. *IEEE-SMC'93* Vol3; October 93, Le Touquet France, 458-462
- Lexicometrica* : Topographie et topologie textuelles, 2007 <http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special9/>
- Platon (trad1923) *Œuvres complètes, VIII, 1, Parménide*, texte établi et traduit par Auguste Diès, Paris, Les Belles Lettres.
- Perlerin, V. (2004) *Sémantique légère pour le document* Thèse d'informatique Université de Caen
- Peschard I. (2004) *La réalité sans représentation, la théorie de l'enaction et sa légitimité épistémologique*. Thèse de Philosophie. Ecole Polytechnique..
- Rastier, F. (2005). Sémiotique du cognitivisme et sémantique cognitive : questions d'histoire et d'épistémologie. *Texte*, mars 2005 Disponible sur : <[http://www.revue-texto.net/Inedits/Rastier/Rastier\\_Semantique-cognitive.html](http://www.revue-texto.net/Inedits/Rastier/Rastier_Semantique-cognitive.html)>.
- Ricoeur P. (1986) *Du texte à l'action : essais d'herméneutique*. Point Seuil
- Roy, T., (2007). *Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires*. Thèse d'informatique Université de Caen
- Varela, F J. (1989) *Invitation aux sciences cognitives*, Seuil
- Vico G (1744) *Principes d'une science nouvelle* (trad JL. Lemoigne) ed Nagel 1986
- Viprey, J-M. (2005) Philologie numérique et herméneutique intégrative, in *Sciences du texte et analyse de discours*, Adam J-M et Heidmann U (eds), Genève, Slatkine, pp 51-68.