DE LA NÉCESSITÉ DE REPENSER LA GESTION ET L'ANALYSE DE L'INFORMATION TEXTUELLE DANS LES ORGANISATIONS

Suzanne BERTRAND-GASTALDY
École de bibliothéconomie et des sciences de l'information
Université de Montréal
Case Postale 6128, Station A
Montréal, Québec, CANADA H3C 3J7

Jean-Guy MEUNIER et Louis-Claude PAQUIN
Centre ATO•CI
Université du Québec à Montréal
Case Postale 8888, Station A
Montreal, Québec, CANADA H3C 3P8

RÉSUMÉ

Les différents types de logiciels disponibles sur le marché ont grandement amélioré la gestion des documents et de l'information textuels. Cependant, aussi bien les logiciels documentaires que les logiciels de gestion des documents saisis en mode image et les logiciels de repérage en plein texte n'apportent que des solutions partielles aux besoins des organisations. Ils ne tiennent pas compte de la diversité de structure des multiples données à consulter ni de l'ensemble des tâches à effectuer sur les textes. De plus, ils ne contribuent que très peu à faciliter les opérations de haut niveau de lecture et d'analyse qui, pourtant, consomment beaucoup du temps des professionnels et des décideurs. Cela tient d'une part à une incompréhension des processus d'interprétation, processus éminemment subjectifs, d'autre part à un manque de sensibilité à la culture organisationnelle et aux difficultés d'appropriation des technologies nouvelles. À la suite de divers projets effectués pour le compte d'organisations publiques et para-publiques, une équipe du Centre ATO•CI rattaché à l'Université du Québec à Montréal, a été amenée à proposer un mode d'intervention quelque peu différent. Sa méthodologie s'appuie sur une analyse statistico-linguistique des textes doublée d'une enquête cognitive et sur la mise au point de chaînes de traitement adaptées aux habitudes de lecture des utilisateurs. Les caractéristiques souhaitables d'un système de gestion intelligemment assistée de l'information laissent le contrôle ultime des opérations à l'être humain et le favorisent. La structure de réalisation qui s'est peu à peu dégagée des différentes expérimentions nécessite une forte implication des futurs utilisateurs et la formation d'un consortium d'organismes qui se partagent les risques d'une réalisation sortant des sentiers battus.

INTRODUCTION

L'avènement de l'ordinateur et surtout des micro-ordinateurs a augmenté de façon considérable la capacité des organisations et des individus à générer de l'information.

Mais si la plupart des documents sont depuis plusieurs années mis en forme avec l'aide de l'ordinateur, leur gestion - depuis leur production jusqu'à leur consultation en passant par leur analyse, leur diffusion et leur stockage - est loin de répondre aux besoins réels des utilisateurs dans le cadre des tâches pour lesquelles l'information est demandée et aux exigences d'efficience dans un contexte de compétitivité et de rareté des ressources matérielles. Bien souvent, en effet, ce sont encore des documents papier qui sont manipulés (on évalue à 92% la part de ceux-ci par rapport aux autres supports) et, lorsque des solutions informatiques sont adoptées, elles ne sont que partielles et se limitent à certaines sous-fonctions.

Nous passerons brièvement en revue les logiciels habituellement proposés pour nous attarder ensuite à quelques-uns des problèmes qui doivent être résolus pour améliorer la performance qualitative des systèmes informatiques consacrés à la gestion des documents textuels et de l'information dans le sens entendu plus haut. Nous montrerons comment, à l'occasion de plusieurs réalisations dans les organisations, une équipe du Centre ATO•CI[1] a pu explorer un espace de solution à ces problèmes. Ces diverses interventions ont abouti à la formulation de principes généraux et de recommandations quant à la méthodologie de gestion de l'information, à la mise en oeuvre exploratoire de cette méthodologie sous forme de chaînes de traitements, à la définition des caractéristiques souhaitables d'un environnement informatique unifié qui supporterait cette méthodologie et enfin à une proposition sur structure de réalisation des projets dans ce domaine.

1. LES PROGRÈS ACCOMPLIS DANS LA GESTION DE L'INFORMATION GRÂCE AUX LOGICIELS RÉCENTS

1.1 Les logiciels documentaires

La plupart des organisations sont dotées de logiciels conçus pour tenir compte d'un type particulier de documents et d'un type de service administratif: repérage de documents d'archives, interrogation de bases de données bibliographiques, emprunt de livres de bibliothèques, systèmes d'information de gestion, etc. La plupart ont été développés il y a plusieurs années pour remplir des fonctions bien spécifiques (acquisition, classement, prêt, conservation, préservation de la confidentialité, etc.) sans souci d'intégration des différents services d'information, à une époque où l'accès au plein texte relevait encore de l'utopie et où l'on n'envisageait pas de politique globale de gestion des ressources informationnelles (Karivalo, 1990).

Sur le plan logiciel, la structure interne des données de chacun des systèmes empêche le partage de l'information; les stratégies d'interfaces sont variées et, la plupart du temps, cryptiques de sorte que la manipulation des systèmes requiert souvent une période d'entraînement longue et intensive. Sur le plan conceptuel, chacun de ces systèmes a déterminé sa propre grille d'analyse et ses propres catégories d'accès (Bertrand-Gastaldy, 1990b: 74). Ceci constitue un obstacle à la collecte ponctuelle de renseignements. Or, la nature des tâches accomplies par les professionnels et les

enjeux qui y sont reliés exigent un accès rapide et précis à plusieurs types d'informations situées sur des systèmes différents et ce par la personne qui a un problème à résoudre, une décision à prendre, un dossier à évaluer, etc.

À ces systèmes sont venus s'ajouter, dans certains cas, des logiciels de repérage en plein texte et des systèmes de GED (gestion électronique des documents)[2] .

1.2 Les logiciels de gestion des documents saisis en mode image

Le succès des systèmes de gestion électronique des documents[3] s'explique par le fait qu'ils résolvent la plupart des problèmes liés à la manipulation du support papier. Ils réunissent sur un même support les documents composites autrefois dispersés dans plusieurs systèmes de stockage. Tout en préservant la présentation visuelle des documents originaux, ces systèmes réduisent considérablement les coûts de stockage. Les coûts et les délais de manipulation connaissent également une diminution importante. Les systèmes de GED centralisent la documentation et offrent, par conséquent, l'assurance que le document consulté ou dupliqué est toujours le plus récent, que la confidentialité est respectée et que seules les personnes autorisées peuvent apporter des modifications, ce qui garantit l'intégrité des données. De plus, l'ergonomie de la consultation ne change pas trop les habitudes par rapport au support papier. On peut agrandir ou rétrécir les documents sur l'écran, les faire pivoter, les envoyer à un télécopieur, etc. Enfin, il est souvent possible d'ajouter des annotations ou même des messages vocaux (Benmergui-Perez, 1989; Chevreau et Kelly, 1989).

Cependant ils automatisent surtout les tâches effectuées par du personnel de bureau sur des supports physiques. L'accès au contenu qui mobilise une grande partie du temps des employés de plus haut niveau pose les mêmes problèmes que lorsque l'information est sur support papier; il faut fournir des mots-clés pour décrire le ou les thèmes principaux traités dans les documents et renoncer à repérer directement l'information spécifique, selon de multiples points de vue. On voit cependant apparaître des logiciels de GED interfacés avec des systèmes de repérage en plein texte[4] qui travaillent sur les textes codés en ASCII après reconnaissance optique des caractères (ROC).

1.3 Les logiciels de repérage en plein texte

Aux États-Unis, le marché du repérage de l'information textuelle a presque atteint un stade de maturité, d'après Delphi Consulting Group (1992) qui a dénombré 107 000 sites où sont installés des logiciels. La croissance de ce marché est considérable si l'on en juge par l'analyse que ce groupe en a faite[5]. Conçus à l'origine d'après les logiciels de repérage des données bibliographiques, il évoluent vers un niveau plus élevé d'interactivité, des capacités de sélectivité plus étendues et vers une convivialité plus grande. Dans certains cas, le repérage s'appuie sur des analyses statistiques et permet de réinjecter une réponse pertinente à titre de nouvelle question. Différents opérateurs sont fournis pour travailler sur les chaînes de caractères elles-mêmes (masque, troncature, etc.) et sur leur position dans la phrase. Quelques logiciels offrent de plus des possibilités de navigation hypertextuelle: l'utilisateur peut alors, comme avec le support papier, s'appuyer sur l'organisation logique des documents en sections, chapitres, paragraphes, illustrations et tableaux. Mais, pour être exploitable électroniquement, cette organisation logique doit avoir été préalablement décrite et reliée au contenu des documents.

Cependant la plupart de ces logiciels de repérage en plein texte[6] n'offrent pas la possibilité de retrouver autre chose que des mots du texte, des chaînes de caractères. À la limite, la mise à disposition brute de très nombreux textes enregistrés sur support informatique accroît les problèmes d'accès à l'information plus qu'elle ne les résoud. En effet, l'ambiguïté inhérente au langage naturel empêche la formulation de requêtes précises et un repérage vraiment efficace avec pour conséquence que les utilisateurs sont inondés de textes non pertinents. De plus des phénomènes courants comme l'anaphore, l'ellipse, l'implicite et la paraphrase nuisent au repérage de tous les textes pertinents. La segmentation des textes en paragraphes et en phrases réduit l'abondance, mais ne constitue pas une solution suffisante aux problèmes de bruit et de silence.

Certes, il existe des logiciels qui tiennent davantage compte de la nature linguistique du matériau à traiter. Ils tentent de retrouver par delà les chaînes de caractères de véritables unités conceptuelles. Des traitements morphologique et syntaxique permettent d'extraire automatiquement des séquences de mots et d'isoler, parmi celles-ci, celles qui sont réputées être les plus chargées sémantiquement, les syntagmes nominaux. La précision des résultats des interrogations s'en trouve considérablement améliorée et le taux de rappel également, pour peu qu'existent des dispositifs linguistiques prenant en compte des phénomènes de paraphrasage ou que des représentations lexico-sémantiques s'attaquent à la synonymie, à l'antonymie et à l'hyperonymie, entre autres. En général, les meilleurs résultats s'arrêtent cependant à la reconnaissance de termes complexes susceptibles de dénoter des notions importantes dans le domaine de référence à condition que celles-ci soient «bien formées»[7]. La possibilité d'explorer les textes dans une perspective autre que terminologique et de les analyser en fonction d'objectifs divers est quasiment absente. En dernière analyse, la plupart des logiciels constituent une "boîte noire" qui a pour fonction unique de mettre les utilisateurs en relation avec les textes ou passages de textes contenant telle ou telle expression ou traitant de tel ou tel sujet.

2. QUELQUES RÉALITÉS NÉGLIGÉES PAR LES CONSTRUCTEURS DE SYSTÈMES:

Les solutions informatiques que nous venons de présenter à grands traits sont loin de répondre aux muliples besoins des professionnels oeuvrant dans les organisations en ce qui concerne la gestion de l'information. Son volume, son rythme d'accroissement, l'instabilité de son contenu, l'interdépendance des différents documents qui réfèrent les uns aux autres, la variété et l'hétérogénéité des supports, des sources, de la durée de vie de l'information et, finalement, la

multiplicité des motifs qui mènent à sa consultation requièrent des approches plus novatrices.

2.1 Diversité des données à consulter

La résolution des problèmes quotidiens nécessite la récupération d'informations disponibles sur plusieurs serveurs, diversement structurées comme celles des banques de données bibliographiques, des banques de données terminologiques, des dictionnaires ou encyclopédies électroniques, du courrier électronique, d'articles de revues, de lois, de règlements, de procédures, etc.

On donc a un besoin urgent de systèmes qui mettent les utilisateurs en relation directe avec de multiples sources de données, au moyen d'une interface commune et conviviale (Belkin *et al.*, 1991).

2.2 Diversité des tâches à effectuer sur les textes

À l'examen des logiciels récents que l'on classe soit dans le créneau de l'informatique documentaire soit dans celui de la bureautique intelligente, on peut observer une tendance à l'intégration progressive des fonctionnalités offertes autrefois par plusieurs logiciels différents dédiés chacun à l'un ou l'autre aspect de la chaîne documentaire: la production, la gestion et l'exploitation des documents (Bertrand-Gastaldy, 1990a). Toutefois, les organisations sont loin de disposer du soutien nécessaire à un accomplissement harmonieux, intégré et économique des multiples tâches qui mènent de la création d'un document à son élimination et ce, dans un environnement souvent multilingue où le travail est effectué sous un mode coopératif.

Ainsi les logiciels de repérage en plein texte qui s'appuient sur des analyseurs linguistiques pour l'indexation et/ou le repérage n'intègrent pas d'outils pour assister la rédaction des documents; ce sont des correcteurs linguistiques dédiés qui le font. On se retrouve dans la situation où une analyse complexe doit être mise en branle pour dépister certaines caractéristiques qui auraient pu facilement être encodées lors de la production des documents si les concepteurs des logiciels de production connaissaient les besoins des logiciels de repérage.

Les différentes phases de la production des textes par plusieurs auteurs, avec ses cycles de consultation, de traduction, de révision, d'approbation, etc. ne sont pas couvertes. De même, la validation des éléments autres que le contenu est laissée pour compte: aucun automatisme ne permet de vérifier l'uniformité de la terminologie employée, la lisibilité en fonction du public visé, la conformité à une politique éditoriale. L'insertion de modules n'est pas prévue pour assister la création de thésaurus ou de bases de connaisances de plus en plus nécessaires pourtant dans les systèmes dits "intelligents", malgré l'intérêt qui émerge pour ce genre d'applications dans des publications récentes (Schmitz-Esser, 1990; RIAO Conference Proceedings 1991). Les documentalistes, terminologues et ingénieurs cogniticiens sont contraints d'attendre "que les outils informatiques d'analyse de contenu des textes soient à la portée de tous" (Ranjard, 1991).

De plus, la vision qu'ont les concepteurs de logiciels des interactions entre les utilisateurs et leurs textes est extrêmement réductrice; ils ne s'intéressent qu'à une petite portion des motivations de consultation: le repérage de passages en fonction d'une question thématique. Pourtant de nombreuses questions factuelles et stylistiques seraient posées si l'infrastructure le permettait. On oublie aussi tout le travail d'annotation et de synthèse qui est effectué une fois les passages pertinents retrouvés et qui pourrait, au moins en partie, bénéficier de l'assistance de l'ordinateur.

La barrière entre les divers systèmes de gestion même électronique des documents est, on le voit, encore bien réelle.Cette étanchéité regrettable des logiciels les uns par rapport aux autres a été soulignée par Duchastel (1991: 601), à propos des tâches complexes de lecture et d'écriture qui "nécessitent la mise en oeuvre d'un grand nombre de nos facultés". Il ajoutait:.

> Cette multiplicité se reflète dans le profusion des solutions informatiques proposées (traitements de textes, correcteurs, dictionnaires, analyseurs). Cependant, ces progiciels sont rarement pensés dans un cadre d'intégration. Tant que l'usager ne cherche qu'une aide ponctuelle pour effectuer une tâche spécialisée, il trouve généralement des systèmes adaptés à cette demande. C'est dans la mesure où un même usager requiert une aide globale pour effectuer un ensemble de tâches complexes de lecture et d'écriture que devient urgente leur intégration dans un cadre méthodologique complet.

On notera que la solution préconisée réside dans l'intégration méthodologique et non pas dans l'intégration des logiciels pour lesquels la modularité et, par conséquent la compatibilité, semble suffisante.

2.3 Importance des tâches de haut niveau

On aura remarqué que certaines des tâches mentionnées, l'écriture et la lecture, constituent des activités cognitives de haut niveau, comme d'ailleurs les activités d'indexation et de condensation qui en constituent un aspect particulier. Elles mobilisent une grande partie des ressources humaines et financières dans une organisation. Ainsi, on a calculé que les professionnels peuvent passer plus de 60% de leur temps à lire des documents papier. D'après une étude américaine (Killen), les employés de bureaux consacrent plus de temps à la lecture, à l'analyse et à la génération d'information qu'à la simple manipulation des supports.

Donc il y aurait des gains de productivité importants à réaliser dans l'automatisation de ces tâches, automatisation malheureusement rendue impossible par la complexité et le caractère "privé" des interprétations.

2.4 Complexité de la lecture et de l'analyse des textes

Pour interpréter un texte, il faut d'abord aller au-delà de sa matérialité. C'est, en effet, le contenu signifiant, le sens, le discours que tente de rejoindre un lecteur. Celui-ci veut accéder à l'acte cognitif d'un locuteur qui a composé, organisé et communiqué des propositions, des idées, des

arguments, une histoire. Pour cela, il effectue parallèlement des traitements cognitifs complexes au niveau perceptif, linguistique, sémantique et contextuel, ce qui l'amène à projeter toute une série de connaissances préalables sur les signes graphiques et à les catégoriser de diverses manières.

Puisque chaque individu construit le sens au fur et à mesure de la lecture, comme le montrent bien les théories sémiotiques de l'interprétation (Eco, 1985), mieux vaut l'outiller pour faciliter ce processus plutôt que de le déposséder au profit d'un automate qui ne pourra de toute façon que construire des représentations rudimentaires. En effet, s'il est impossible de doter un ordinateur de toutes les connaissances et habiletés nécessaires pour "comprendre" un texte, il est cependant réaliste de concevoir des outils capables d'assister l'utilisateur dans la transformation des données en éléments sémantiquement structurés, en connaissances. Il sera alors plus aisé pour ce dernier de manipuler, de classifier, de relier et d'interpréter de tels éléments que de simples chaînes de caractères sur lesquelles les opérations sont très limitées, car elles ne sont que les porteurs matériels de l'information.

2.5 Caractère "privé" de l'interprétation des textes

Mais l'interprétation est une activité subjective par essence qui dépend à la fois de la structure cognitive du lecteur et de ses objectifs de lecture.

> Quiconque veut comprendre un texte a toujours un projet. Dès qu'il se dessine un premier sens dans le texte, l'interprète anticipe un sens pour le tout. À son tour, ce premier sens ne se dessine que parce qu'on lit déjà le texte, guidé par l'attente d'un sens déterminé. C'est dans l'élaboration d'un tel projet anticipant, constamment revisé, il est vrai, sur la base de ce qui ressort de la pénétration ultérieure dans le sens du texte, que consiste la compréhension de ce qui s'offre à lire [...]. Ce processus est donc le renouvellement incessant du projet qui entretient le mouvement de la compréhension et de l'interprétation. (Gadamer; 1976: 196)

Autrement dit, il faut toujours situer la gestion et l'analyse d'un texte dans un horizon d'action. Le texte ne se laissera jamais épuiser par une approche unique. Son interprétation peut varier d'une personne à une autre, d'un moment à un autre. Et même s'il se présente dans une langue spécifique, par exemple, le français, même si l'ensemble des expressions linguistiques qui le constituent sont relativement stables et susceptibles d'être décrites selon une certaine grammaire, son contenu est toujours plurivalent.

> Le texte contient une composante d'indétermination. Ce n'est pas un défaut, mais bien une condition fondamentale de la communication du texte; elle permet la participation du lecteur à l'intention du texte. (Iser 1985: 15)

Un système d'information devrait donc s'adapter aux différents points de vue des utilisateurs.

En raison de la nature sémiotique de l'objet textuel et du postulat herméneutique de la multiplicité des interprétations possibles d'un texte, il est donc impossible de construire un système de lecture et d'analyse automatiques de textes. Cela n'est d'ailleurs pas souhaitable, puisque la "lecture" change en fonction de l'évolution des besoins et de l'état des connaissances des lecteurs humains:

> [...] different persons, in different occupations may possess different world views and make different demands upon sources of knowledge as a consequence. For example, some occupations may require no more than 'recipe knowlege' for their effective performance; others, falling short of a need for 'expert ' knowledge, may demand more in the nature of 'reasoned opinion' and, hence, a greater need for access to sources of information. (Wilson, 1984: 200).

Or, les systèmes actuels de gestion de l'information ne permettent pas une catégorisation du vocabulaire ni une analyse "sur mesure" adaptée à des objectifs de recherche bien particuliers. Pourtant, au sein de l'organisation, chacun devrait trouver son compte dans la manipulation de l'information, du cadre supérieur au personnel de secrétariat en passant pas le rédacteur, le réviseur, le chargé de projet, l'agent d'information, etc. (Trowbridge, 1988).

Pour cela, il faudrait envisager un système qui soit un "adjuvant" à l'activité cognitive de l'être humain (Meunier, 1992) et qui laisse la maîtrise ultime du traitement de l'information entre les mains de l'expert. C'est en ce sens qu'on devrait plutôt envisager des systèmes de gestion et d'analyse intelligemment assistées de la documention textuelle qui réconcilieraient deux formes de lecture: il s'agirait de procurer au lecteur des instruments à l'aide desquels son expertise pourrait être mise à profit, en même temps qu'ils lui garantiraient une capacité de lecture augmentée en termes de volume, de rigueur, bref de systématicité (Paquin et Beauchemin, 1988). Afin de respecter le caractère "privé" de toute interprétation, l'outillage mis à la disposition du lecteur devrait être paramétrable, se plier à ses objectifs de lecture. Or, même si quelques systèmes actuellement disponibles intègrent une certaine catégorisation, ils ne le font pas à tous les niveaux: graphique, morphologique, lexical, syntaxique, sémantique, pragmatique. Avec de telles informations supplémentaires, il deviendrait possible d'envisager des traitements mixtes (à la fois linguistiques, cognitifs et statistiques) sur de grands corpus sans aller à l'encontre des parcours interprétatifs des individus.

2.6 Importance des modes de lecture développés en fonction de la culture organisationnelle

Dans les organisations, ces parcours interprétatifs sont autant le fait de groupes que d'individus en particulier. Les analyses des textes sont dirigées à la fois par les structures de textes et par les tâches à accomplir. Bien souvent les systèmes d'analyse livrés, même s'ils ont été précédés d'une analyse de besoins, sont développés en dehors de tout contact avec les utilisateurs et uniquement en fonction de

modèles théoriques qui ignorent les contextes et les situations particulières d'utilisation. Si l'effort requis des utilisateurs pour s'adapter à de nouvelles technologies est trop grand et qu'en retour cette dernière ne parvient pas à faciliter leur tâche, elle sera rejetée malgré les avantages objectifs qu'elle peut apporter. En somme, les développements technologiques qui ne prévoient pas de modes d'appropriation par les utilisateurs sont voués à l'échec.

2.7 Acculturation en ce qui concerne les technologies d'exploitation des informations textuelles

La manipulation consciente de l'information textuelle avec des outils aussi nouveaux pose un certain nombre de problèmes d'acculturation. Le surlignage de passages pertinents, les annotations marginales qui aboutissent à la lente élaboration d'une synthèse ne se prêtent pas aux mêmes manipulations que les catégories surimposées aux textes exploitables par toutes sortes d'algorithmes, y compris les calculs statistiques, dans un système ouvert. Il faut une certaine formation et un certain temps d'assimilation pour maîtriser toutes les opérations possibles. D'ailleurs même les individus familiers avec les outils informatiques traditionnels ont du mal à utiliser un système qui fait appel à la créativité et surtout à des connaissances spécialisées sur ce qu'est un texte et ce qu'est une analyse de texte. C'est pourquoi, on ne peut pas se contenter de livrer un système; il faut d'abord former les utilisateurs, les rendre conscients de leurs stratégies d'appropriation des textes, notamment par des exemples d'application qui se rapprochent le plus de leurs besoins. Ceci nécessite de part et d'autre, un lent apprivoisement des cultures respectives de l'organisation et de la firme auteur du système.

3. LES SOLUTIONS PRÉCONISÉES PAR UNE ÉQUIPE DU CENTRE ATO•CI

Depuis plusieurs années, des chercheurs du Centre ATO•CI[8] interviennent dans les organisations pour développer des systèmes d'analyse de textes par ordinateur. Au fil de leurs réalisations, ils en sont venus à proposer une méthodologie d'analyses mixtes (de type statistico-linguistique et cognitif) des textes, implantée sous forme de chaînes de traitement adaptées aux différents contextes d'utilisation. Ceci les a conduits à réfléchir aux caractéristiques souhaitables d'un système intégré de gestion intelligemment assistée de l'information et aux conditions de sa réalisation. Ce sont les trois volets que nous allons maintenant exposer.

3.1 La méthodologie

Analyse statistico-linguistique des textes doublée d'une enquête cognitive

Il nous faut d'abord procéder à l'analyse d'un échantillon des corpus de l'organisation. On ne prend pas connaissance d'une lettre comme on le fait d'un rapport de recherche, d'une directive, d'une loi, etc. À chaque genre de texte correspondent une structure d'information et un type de lecture qui projette la connaissance préalable de cette superstructure. Il est, à notre avis, naïf de penser qu'on peut indexer ou repérer de l'information de la même façon dans tous les textes, ce que nous confirment d'ailleurs l'examen des produits dérivés de ces textes et les observations des employés dans l'exercice de leurs tâches.

La technique d'analyse de textes n'est pas seulement employée pour les textes primaires, mais aussi, lorsqu'ils sont disponibles, pour les textes secondaires, produits d'une analyse humaine préalable et pour les outils documentaires utilisés à cette fin (thésaurus et plan de classification). C'est ainsi qu'à travers les traces laissées par les indexeurs dans leurs résumés, leur choix de mots-clés ou de rubriques de classification, nous pouvons déceler des tendances et des anomalies qui nous renseignent sur les processus plus ou moins conscients de condensation. En somme, en confrontant les textes primaires et les textes secondaires, nous cherchons à découvrir les propriétés des éléments d'information retenus par les indexeurs par opposition aux propriétés des éléments laissés de côté. Il va de soi qu'une telle observation nécessite un ou des logiciels capables de catégoriser de multiples façons les unités lexicales et textuelles. Cette technique est employée actuellement dans deux projets [9].

En parallèle, nous menons une enquête cognitive auprès des indexeurs; nous les interviewons, nous les observons dans l'exécution de leurs tâches, nous leur demandons d'expliciter leur démarche, et aussi de commenter les résultats de nos analyses. Puis nous revenons aux textes pour trouver la confirmation de leurs dires. Cette double approche permet de mesurer l'écart entre les affirmations et la pratique réelle, d'enrichir nos intuitions de départ, de provoquer aussi de la part des indexeurs une prise de conscience qui les amène à revoir et à normaliser leurs habitudes, de leur plein gré et en toute connaissance de cause. La richesse des résultats obtenus par les analyses statistico-linguistiques pique leur curiosité et suscite la demande de formation à l'utilisation des logiciels.

Du point de vue de la recherche, cette interaction nous conduit à observer davantage de propriétés que celles auxquelles se sont intéressées les quelques études de ce genre déjà effectuées (Grunberger, 1985). Alors que ces dernières ont surtout porté sur la fréquence des mots-clés et leur position dans le paragraphe, nous vérifions plusieurs indices statistiques, dont la valeur discriminante, nous tenons compte aussi de l'appartenance au sous-domaine du savoir dans lequel le texte peut être classé, de la position des termes à la fois dans la micro-structure et la macro-structure aussi bien dans le texte intégral que dans le résumé, de la pertinence de certaines informations (comme, dans notre projet de système expert d'aide à l'analyse des jugements, l'intitulé d'une loi, les lois ou articles de lois cités, les parties au litige pour déterminer la classification, les variations selon le domaine de droit et selon la provenance du jugement, etc.). Ces études contribuent à enrichir la connaissance encore très rudimentaire que nous avons des tâches cognitives extrêmement complexes de lecture et d'analyse, en vue de la classification et de l'indexation dans ce cas précis. Elles abordent sous un angle différent des réflexions entreprises par quelques rares auteurs

comme Beghtol (1986), David (1990), Farrow (1991) et Endres-Niggemeyer (1990).

Cette modélisation de l'analyse humaine en vue de la classification et de l'indexation peut bien sûr être étendue à d'autres types de lecture effectuée pour atteindre d'autres objectifs: repérage de textes écrits avec une perspective de prospective, classement en fonction de variables thématiques, stylistiques, détermination des auteurs les plus prolifiques, des centres de recherche les plus productifs dans un champ d'intérêt nouveau pour l'utilisateur, représentation des termes associés à telle ou telle problématique, etc. Il s'agit de comprendre ce que les lecteurs font avec leurs textes et de déceler les marqueurs les plus pertinents pour réduire la masse d'information, la présenter dans certains cas de façon synthétique, donc faciliter la prise de connaissance du contenu en fonction de l'horizon de lecture. Davantage de temps peut alors être consacré aux opérations cognitives très complexes.

Les chaînes de traitement

Une fois les habitudes de lecture mises en évidence et une fois validés les paramètres, des chaînes de traitement peuvent être mises au point. Les analyses appropriées sont découpées en opérations qui sont effectuées séquentiellement à l'aide de plusieurs logiciels. Une interruption des opérations automatiques est prévue chaque fois que l'utilisateur doit intervenir. Cette façon de faire rend plus aisée la reproduction des analyses sur des textes nouveaux. Nous envisageons de formaliser les modèles après validation dans un système expert, ce qui nous permettrait un contrôle plus fin et une plus grande sensiblité au contexte. C'est ainsi qu'on espère modéliser non seulement les traitements mais les parcours interprétatifs des utilisateurs (Paquin, 1992).

Les systèmes que nous développons répondent bien à ce que Lee (1985) recommandait pour les entreprises: des systèmes d'aide à la décision qui ne remplacent pas l'être humain, mais l'assistent et augmentent ses capacités, avec des modules qui peuvent être combinés différemment pour répondre à des situations différentes.

L'approche adoptée nécessite une architecture logicielle ouverte dont les fonctionnalités sont développées au fur et à mesure des besoins. SATO (Système d'Analyse de Textes par Ordinateur)[10] constitue la boîte à outils de base pour pré-traiter les textes, les catégoriser (de façon automatique, assistée ou manuelle selon la complexité des connaissances mobilisées), les fouiller sur les mots ou sur leurs catégories, les partitionner en domaines et effectuer des calculs statistiques de base. Nous avons aussi recours à SPSS pour implanter les analyseurs statistiques. Enfin, la modélisation en système à base de connaissance sera faite sur ACTE (Atelier Cognitif et TExtuel) qui est en cours de développement au centre ATO•CI.

3.2 Caractéristiques souhaitables d'un environnement informatique unifié

Les études réalisées en étroite collaboration avec les intervenants dans les organisations nous ont amenés à réfléchir aux caractéristiques que tout système de gestion intelligemment assistée de l'information devrait posséder:

- Grande capacité de traitement: pour traiter, sans dégradation de performance, des banques de documents textuels dépassant plusieurs fois le gigaoctet.
- *Catégorisation à différents niveaux:* pour répondre à la complexité variable des traitements à effectuer.
- *Modularité* : pour qu'il soit possible de combiner à volonté les analyseurs et les chaînes de traitement, en fonction des besoins, le contrôle de l'échange entre unités de traitement s'effectuant au moyen d'un système de communication de type "black board".
- *Paramétrisation* : qui va de pair avec la modularité pour adapter les traitements à la qualité souhaitée des résultats et aux investissements en termes de temps et de coût qu'on veut bien consentir.
- *Compatibilité* : des données soumises aux traitements et issues de ceux-ci pour être admissibles aux différents modules.
- ***I**ntégration* : c'est, d'après les résultats de l'étude d'Andersen Conseil en 1990 sur la gestion de l'information dans les années 1990 la question qui, avec celles de connectivité, de réseaux et de communication, figure au premier rang - et de loin (61%) - des principaux problèmes techniques dont se préoccupent le plus les responsables de la gestion de l'information.
- *Assistance intelligente* : pour encadrer une partie des tâches cognitives qui sont normalement associées à la gestion de l'information: la production et plus spécifiquement l'écriture, la normalisation, la catégorisation, la description, l'analyse, la classification, l'indexation, l'extraction de connaissances, le repérage, la diffusion, etc.
- *Interactivité* : l'utilisateur gardant le contrôle ultime des opérations, celui-ci doit pouvoir choisir l'une des stratégies offertes ou encore développer - et sauvegarder pour réutilisation - ses propres modèles pour naviguer dans la base de documents et réaliser les opérations cognitives désirées.
- *Convivialité* : la convivialité va de pair avec l'interactivité, car si le système est trop difficile à utiliser, il est délaissé.
- *Navigabilité* : opérer sur de multiples documents ou portions de documents, y appliquer des descriptions et des catégorisations, les annoter, les rechercher au moyen d'outils d'aide au repérage, rechercher de l'information pertinente, tout cela nécessite des moyens de naviguer au moyen de liens hypertexte et hypermédia.

3.3 Structure de réalisation des projets de mise sur pied de systèmes de gestion de l'information

La méthodologie exposée plus haut ainsi que l'implantation de systèmes dotés des caractéristiques énumérées exigent beaucoup en termes de ressources humaines et financières. La structure de réalisation qui s'est

peu à peu dégagée des différentes interventions de l'équipe du centre ATO•CI s'appuie d'une part sur une forte implication des futurs utilisateurs, d'autre part sur la formation d'un consortium d'organismes qui se partagent les risques d'une réalisation sortant des sentiers battus.

Implication des futurs utilisateurs à tous les stades du projet et transfert d'expertise

Les futurs utilisateurs sont impliqués à tous les stades du projet: enquête préliminaire, élaboration du cahier des charges, design, développement par prototypage, mise au point du système, évaluation.

Le système est conçu, comme nous l'avons dit plus haut, à partir de corpus de documents sélectionnés par les participants de l'organisation. Afin que ceux-ci puissent mieux définir leurs besoins, une formation leur est donnée tout au cours du projet pour qu'ils assimilent la technologie de l'analyse de textes assistée par ordinateur et prennent progressivement en charge la gestion de l'information dans leur milieu de travail jusqu'à ce qu'ils acquièrent une autonomie totale. Le transfert d'expertise prend d'abord la forme de cours, puis de séances d'animation et finalement de travaux supervisés dans l'organisation même. Cette façon de procéder a fait ses preuves dans tous les projets menés jusqu'ici par l'équipe et rejoint une nouvelle théorie de développement des systèmes d'information qui tient compte à la fois du caractère évolutif du processus et de l'importance de la communication entre les concepteurs, les développeurs et les utilisateurs. Cette approche permet d'inclure un processus continu d'évaluation et de réajustement du produit en cours de route, plutôt qu'en bout de piste seulement, et d'y ajouter des critères aussi importants pour tous les participants que la croissance professionnelle et la reconnaissance par les pairs, ainsi que la pérennité de l'utilité et de la qualité du produit (Sonnenwald, 1992).

Nous avons pu constater (Bertrand-Gastaldy *et al*., 1990) que la mise à disposition de technologies nouvelles pour exploiter le texte et la formation *in situ* stimulent la créativité des utilisateurs qui découvrent des applications nouvelles, des façons différentes de travailler avec les textes et réclament des systèmes dont les fonctionnalités sortent des applications habituellement commercialisées (ce que Delphi Consulting Group (1992: Intro-4) appelle: "new generation of information management that is more versatile and comprehensive than its predecessors"):

> As this technology is accepted into the corporate information management systems, users are discovering their capability beyond traditional applications. It is these advanced applications that will finally propel text retrieval into the main stream of an integrated electronic document management system." (Delphi Consulting Group, 1992: TR-9)

Les futurs utilisateurs découvrent qu'ils peuvent exploiter les gisements documentaires pour toutes sortes d'activités qu'il était tout simplement impossible d'effectuer auparavant: normalisation des textes sur le plan terminologique et stylistique, documentation systématique des décisions, mise à disposition de renseignements autrefois dispersés dans de nombreux dossiers, présentation de la documentation en fonction d'utilisations différentes. Cela va bien au-delà du simple repérage que permettent les systèmes documentaires traditionnels.

Développement par consortium

Un projet de développement d'un environnement informatique intégrant toutes les caractéristiques énumérées plus haut implique non seulement l'adaptation des logiciels existants à des normes industrielles, mais aussi des développements d'envergure autant pour tenir compte de la masse des documents que les organisations ont à traiter que pour raffiner la «profondeur» d'analyse requise. Un tel projet présente un risque à la fois technologique et organisationnel qui ne peut être assumé par les moyens traditionnels où une entreprise reprend les travaux des universitaires pour les traduire et les encapsuler dans des développements logiciels robustes. La seule solution possible qui pourrait être envisagée consiste en un partenariat entre des universités, des entreprises productrices de logiciels et des entreprises utilisatrices comme des banques et des services para-publics, selon une formule encouragée par le gouvernement du Québec et préconisée par une étude de l'Observatoire français des industries de la langue pour répondre aux besoins qui ne sont pas satisfaits par les logiciels existants:

> Les utilisateurs pourraient s'associer entre eux pour financer partiellement le développement de produits dans le cas où aucun de ceux disponibles ne répondrait à leurs besoins. Mais le développement de ce type de projet coûte très cher [...]. Les fournisseurs pourraient ainsi investir dans une recherche-développement à risque partagé. (Observatoire français des industries de la langue: 100)

CONCLUSION

Malgré les progrès réels enregistrés dans la gestion de l'information textuelle, nous avons pu constater un certain nombre de lacunes qui proviennent autant des logiciels que des méthodologies de développement et d'implantation dans les organisations. L'option que nous préconisons de fournir un accès différencié au contenu textuel à des utilisateurs qui s'impliquent dans le développement et la prise en charge de leur système s'oppose aux options faciles de recherche en plein texte ou d'indexation automatique de type "boîte noire". Cette option ne peut d'emblée emporter l'adhésion, car elle suppose une prise de conscience préalable des processus cognitifs complexes de lecture et d'interprétation. Le marché actuellement en émergence a un potentiel important mais qui est relativement mal compris. C'est pourquoi la formation de consortiums réunissant des intervenants convaincus des bénéfices d'une gestion intelligemment assistée est peut-être le seul moyen de contribuer à un changement de paradigme dans la façon d'envisager le rapport des utilisateurs aux textes.

BIBLIOGRAPHIE DES SOURCES CITÉES

Beghtol, C. (1986). Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation*; 42(2); June 1986: 84-113

Belkin, N.J.; Marchetti, P.G.; Albrecht, M.; Fusco, L.; Skogvold, S.; Stokke, H.; Troina, G. (1991). User interfaces for information systems. *Journal of Information Science*; 17; 1991: 327-344.

Benmergui-Perez, M. (1988). Charting the uncharted. *Office Equipments & Methods*; November 1988: 26-29.

Bertrand-Gastaldy, S. (1990a). L'évolution de la gestion de l'information documentaire sous l'impulsion des nouvelles technologies. *Terminogramme; Bulletin d'information terminologique et linguistique* ; 55; mars 1990: 25-31.

Bertrand-Gastaldy, S. (1990b). L'indexation assistée par ordinateur: un moyen de satisfaire les besoins collectifs et individuels des utilisateurs de bases de données textuelles dans les organisations. *ICO Québec; intelligence artificielle et sciences cognitives au Québec;* 2(3); septembre 1990: 71-91.

Bertrand-Gastaldy, S.; Paquin, L.-C.; Dupuy, L. (1990). The need for information and knowledge management. In: Hans Czap et Wolfgang Nedobity, eds. *TKE'90: Terminology and Knowledge Engineering*; Proceedings of the Second International Congress on Terminology and Knowledge Engineering, 2-4 October 1990, University of Trier, Federal Republic of Germany. Frankfurt: Indeks Verlag; 1990: 509-517.

Chevreau, J.; Kelly, T. (1989). Paperless report. *Office Equipments & Methods*; January-February 1989: 42-46.

David, C. (1990). *Élaboration d'une méthodologie d'analyse des processus cognitifs dans l'indexation documentaire*. Montréal: Université de Montréal, Département de communication; septembre 1990. (mémoire de maîtrise).

Delphi Consulting Group.(1992) *Information Management: The Next Generation; Conferences and Seminars on Electronic Management Systems* ; 1992.

Duchastel, J. (1991). Pour une méthodologie d'aide à la lecture et à l'écriture. *Actes du coloque "Les industries de la langue: perspectives des années 1990, Montréal, 21-24 novembre 1990*. [s.l.]: Office de la langue française / Société des traducteurs du Québec, 1991: 583- 601.

Eco, U. (1985). *Lector in fabula ou la Coopération interprétative dans les textes narratifs*. Paris: Grasset; 1985.

Endres-Niggemeyer, B.(1990). A procedural model of abstracting, and some ideas for its implementation. In: Hans Czap et Wolfgang Nedobity, eds. *TKE'90: Terminology and Knowledge Engineering; Proceedings of the Second International Congress on Terminology and Knowledge Engineering*, 2-4 October 1990, University of Trier (FRG), Frankfurt: Indeks Verlag; 1990: 230-243.

Farrow, J. F.(1991). A cognitive process model of document indexing. *Journal of Documentation*; 47(2); June 1991:149-166.

Gadamer, H. G. (1976) *Vente et méthode*. Paris: Seuil; 1976.

Grunberger, M. W. (1985). *Textual Analysis and the Assignment of Index Entries for Social Science and Humanities Monographs*. New Brunswick, NJ: Rutgers University; 1985. 136 p. (Ph.D thesis).

Iser, W. (1976). *The Art of Reading A Theory of Esthetic response*, Baltimore 1976: John Hopkins University.

Karivalo, M.(1989). Training for information management in a company. *Information Services & Use*; 9; 1989: 341-346.

Lee, R.M. (1985). On information system semantics: expert vs. decision support systems. *Social Science Information Studies*; 5; 1985: 3-10.

Meunier, J.-G. (1992). SATO: un philologue électronique. *Documentation et bibliothèques*; 38(2); avril-juin 1992: 65-69.

Observatoire français des industries de la langue (1991) *Utilisations et utilisateurs en produits et services des industries de la langue* . Québec: OFIL; 1991.

Paquin, L.-C. (1992). La lecture experte" *Technologie, idéologie et pratique*, numéro spécial consacré au colloque "Intelligence artificielle et sciences sociales"; 10 (2-4): 209-222.

Paquin, L.-C.; Beauchemin, J. (1988). Apport de l'ordinateur à l'analyse des données textuelles. In: RELAI: Recherche en linguistique aappliquée à l'informatique. Actes du colloque "La description des langues naturelles en vue d'applications informatiques". Université Laval, 7-9 décembre 1988. Québec: Centre international de recherche sur le bilinguisme; 1989: 197-210.

Perriault, I. (1989). SITE: la documentation technique sur supports optiques. *Archimag*; 23; 1989: 81.

Ranjard, S. (1991). L'indexation manuelle: une valeur ajoutée. *Archimag*. Hors série; novembre 1991.

RIAO 91 Conference Proceedings (1991). Intelligent Text and Image Handling, Universitat Autonòma de Barcelona, Barcelona, Spain, April 2-5, 1991. 2 vol.

Schmitz-Esser, W. (1990). Thesauri facing new challenges. *International Classification* ; 17 (3/4); 1990: 129-132.

Sonnenwald, D. H.(1992). Developing a theory to guide the process of designing information retrieval systems. *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and*

Development in Information Retrieval, Copenhagen, Denmark, June 21-24, 1992 : 310-317.

Trowbridge, R. (1988). Lost, stolen or strayed; you can escape the horror of the lost document. *Office Equipment & Methods*; 1988 November: 34-38.

Wilson, T.D (1984). The cognitive approach to information-seeking behaviour and information use. *Social Science Information Studies*; 4; 1984: 197-204.

[1] Le Centre d'Analyse de Textes par Ordinateur, Cognition et Information est rattaché à l'Université du Québec à Montréal

[2] Aussi appelés DIP (Document Image Processing) en anglais.

[3] Des logiciels comme Desktop Document Manager, Inspire VisionQuest, Optix, etc. sur le marché nord-américain et Taurus en France.

[4] Des logiciels comme BRS, BasisPlus, MicroQuestel, etc.

[5] "Text retrieval was a $118+ million market in 1990. Both the PC and mini/mainframe markets are growing at an impressive rate. The PC market revenue is growing at a 45% CAGR. The mini/mainframe market revenue is growing at a 35% CAGR.

The market is expected to reach the critical 300+ million mark in approximately 2-3 years".(Delphi Consulting Group, 1992 :TR-12)

[6] Des logiciels comme Book Manager, Basis +, Open TEXT, TOPICS, ConQuest, Elexir, Isys, Zyindex, etc., et plus près de nous : CEDROM, Édibase, Seconde, etc.

[7] Des logiciels comme ALETH de la firme GSI-ERLI et SPIRIT de la compagnie SYSTEX.

[8] Outre les auteurs, cette équipe comprend Gracia Pagola et François Daoust. Il nous faut aussi mentionner l'apport passé de Luc Dupuy.

[9] Le premier projet "Conception d'un système expert d'aide à l'analyse (tri, classification, indexation) des jugements" est subventionné par le CEFRIO (Centre Francophone de Recherche en Informatisation des Organisations), SOQUIJ (Société Québécoise d'Information Juridique) et le ministère des Communications du Québec; il bénéficie aussi de l'appui de l'Université de Montréal et de l'Université du Québec à Montréal. Le prototype fait l'objet d'une démonstration au colloque ICO 93.

Un autre projet, subventionné par le CRSHC(Centre de recherche en sciences humaines du Canada) et mené par Suzanne Bertrand Gastaldy et Luc Giroux à l'Université de Montréal s'appuie en partie sur la même approche. Il porte sur "Les processus cognitifs et la représentation des connaissances dans les bases de données documentaires." (#410-92-0713).

[10] Le logiciel est développé par François Daoust, du Centre ATO.CI.