**Documentation et bibliothèques**

# ICO : intelligence artificielle et sciences cognitives au Québec

## *Revue de liaison de la recherche en informatique cognitive des organisations*, vol. 2, no 3 (numéro thématique : « Gestion de l'information textuelle» ), septembre 1990

Claude Allen

Volume 38, numéro 2, avril–juin 1992

Analyse et gestion de l'information textuelle

URI : https://id.erudit.org/iderudit/1028619ar
DOI : https://doi.org/10.7202/1028619ar

Aller au sommaire du numéro

Éditeur(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (imprimé)
2291-8949 (numérique)

Découvrir la revue

Citer ce compte rendu

Allen, C. (1992). Compte rendu de [ICO : intelligence artificielle et sciences cognitives au Québec / *Revue de liaison de la recherche en informatique cognitive des organisations*, vol. 2, no 3 (numéro thématique : « Gestion de l'information textuelle» ), septembre 1990]. *Documentation et bibliothèques*, *38*(2), 127–130. https://doi.org/10.7202/1028619ar

Chronique

# ICO : intelligence artificielle et sciences cognitives au Québec

## Compte rendu de *ICO Québec*

**[Revue de liaison de la recherche en informatique cognitive des organisations], vol. 2, no 3 (numéro thématique : « Gestion de l'information textuelle »), septembre 1990**

**Claude Allen***
Montréal

Ils proviennent d'horizons aussi divers que ceux des communications, bibliothéconomie et sciences de l'information, lettres, histoire, philosophie, philologie, linguistique, psychologie, biologie, informatique et mathématiques. Ils oeuvrent actuellement - à titre de chercheurs, chargés de projet, analystes, conseillers, agents de recherche ou simples utilisateurs - dans le milieu universitaire, un centre de recherche, le milieu gouvernemental ou l'entreprise privée. Ils préconisent des approches théoriques ou méthodologiques différentes. Mais ils sont, malgré cette apparente diversité, préoccupés par un objet commun : les données textuelles. Nous avons désigné les collaborateurs du numéro thématique d'*ICO Québec* (septembre 1990), entièrement consacré à la gestion de l'information textuelle.

La revue *ICO Québec* est publiée par le Groupe interuniversitaire de recherche en informatique cognitive des organisations (GIRICO), avec la collaboration de la Direction générale des technologies de l'information, ministère des Communications du Québec. Pour cette livraison, Richard Parent, attaché à la Direction susmentionnée, a assumé le rôle d'éditeur intellectuel.

Précisons, afin de dissiper toute ambiguïté, que la gestion de l'information textuelle, telle qu'elle est définie et abordée dans ce contexte, s'applique la plupart du temps à des corpus de textes intégraux disponibles sur support ordinolingue.

On a fait allusion d'entrée de jeu au caractère interdisciplinaire des études consacrées à la gestion de l'information textuelle. On remarquera dans la suite de l'exposé l'effervescence des échanges entre spécialistes, le croisement des idées, l'interpellation mutuelle de la théorie et de la pratique. Des savoirs multidisciplinaires se constituent. Les frontières s'estompent.

Outre le texte d'introduction du philosophe et sémioticien Jean-Guy Meunier qui trace l'évolution de l'objet d'analyse de textes par ordinateur (ATO) et en définit la nature, ce numéro comporte trois sections que nous décrirons sommairement. Le parcours que nous proposons ici est libre, sélectif et nécessairement subjectif.

Comme le rappelle Richard Parent dans un texte de présentation générale du numéro, nul autre que Jean-Guy Meunier n'était mieux placé pour « aborder le sujet à la fois dans sa généralité et dans sa profondeur, vu l'influence qu'il a exercée depuis vingt ans pour stimuler la réflexion et la recherche en analyse de texte par ordinateur ».

Après avoir situé l'évolution de l'ATO dans le contexte des humanités et présenté les difficultés posées par une telle entreprise, Jean-Guy Meunier montre comment, « malgré les apparences de matérialité qui le portent », le texte électronique soumis à l'analyse ne constitue qu'une « forme limitée d'expression du discours ». Pour J.-G. Meunier, « c'est là, dans le discours, que se trouve le véritable texte analysé ».

Jean-Guy Meunier dresse ensuite un tableau sommaire des diverses représentations mises en oeuvre dans les systèmes d'ATO, rappelant avec Desclés (1989) que ces derniers fonctionnent « selon un mode compilatoire, [c'est-à-dire comme] une suite complexe de transformations de représentations en représentations » . Il distingue ainsi le texte manuscrit du texte électronique, le texte discursif, la représentation que l'ordinateur se fait à lui-même et enfin, le texte sémiotique. Ce dernier correspond à la traduction de la représentation de l'ordinateur sous une forme d'expression sémiotiquement accessible à l'usager.

Enfin, l'auteur conclut en évoquant les compétences requises et en soulignant le caractère multidisciplinaire de l'entreprise d'ATO.

### Gestion de la ressource informationnelle : le cas des textes

La première section, consacrée à la gestion de la ressource informationnelle, démontre l'importance - qualitative aussi bien que quantitative - des données textuelles au sein des organisations. Malgré cette prise de conscience, René Lortie fait voir le besoin de systématiser l'accès à l'information conceptuelle, afin de mieux exploiter la mémoire institutionnelle des organisations.

L'article du conseiller en informatique Hiep Dao souligne « l'absence de volonté politique, de budget, de matériel, de normes et de méthodologie pour le développement du traitement textuel » au sein des organismes publics. Il rappelle avec réalisme qu'une formulation claire de besoins s'impose pour l'obtention de budgets.

Hélène Lapointe (ministère des Communications) brosse un tableau général des étapes nécessaires à l'élaboration et à l'implantation d'un système de gestion de données textuelles (GDT). Jean Asselin (Conseil du Trésor) propose pour sa part une grille

* Actuellement étudiante au doctorat conjoint en communication (UdM, UQAM, Concordia), l'auteure possède une maîtrise en bibliothéconomie et en sciences de l'information. Elle est également chargée de cours à l'École de bibliothéconomie et des sciences de l'information (EBSI) de l'Université de Montréal.

détaillée d'analyse des besoins qu'il a élaborée en vue du développement de systèmes informatisés. Cette grille comprend trois dimensions qui méritent d'être examinées. Il s'agit de la nature des textes (données), des activités associées à ce que Jean Asselin désigne comme le « cycle de réalisation des textes » et enfin des utilisateurs (traitements cognitifs).

Pour la première composante, Jean Asselin distingue les caractéristiques externes des textes (taille, nombre d'unités, médium de stockage, provenance) et leurs caractéristiques internes (sujet traité, structuration du contenu, vocabulaire du domaine, niveau d'information).

Le « cycle de réalisation des textes » se découpe selon l'auteur en quatre phases: consultation, production, édition et diffusion. Les bibliothécaires pourront reconnaître ici les maillons de la chaîne documentaire. En effet, la consultation ou, selon les termes de Jean Asselin, la localisation, l'acquisition et la cueillette pourraient correspondre à l'acquisition ou au repérage suivant que l'on se situe à une extrémité ou à une autre de la chaîne. La production, qui désigne pour l'auteur les activités d'annotation, d'analyse, de synthèse et de rédaction, évoque le traitement analytique. Tandis que l'édition (saisie, correction, mise en forme) équivaudrait à l'enregistrement et au stockage, enfin, la diffusion (classement, calendrier de conservation, distribution) équivaudrait à la circulation. Ce parallèle que nous avons librement établi ne rend toutefois pas justice à la vision de l'auteur. Car c'est l'utilisateur même qui est ici visé; c'est lui qui exerce un rôle actif tout au long du processus.

Or cette interaction de l'utilisateur avec le texte crée de nouveaux besoins. Elle implique une série de traitements cognitifs qui doivent encore être précisés, afin d'être modélisés et rendus mécanisables. L'auteur est ainsi amené à suggérer la mise en place de projets pilotes mettant à la disposition des utilisateurs des outils génériques pour une expérimentation encadrée et devant contribuer à une modélisation des tâches en contexte réel.

Dans un article qui complète celui de Jean Asselin, Marielle Gévry, de la firme Informission Ltée, tente de « démystifier la réalisation des systèmes à base de connaissances ». Elle relate certaines techniques utilisées dans le cadre précis d'un projet appliqué au domaine des textes de loi et/ou règlements et identifie des éléments d'aide à la modélisation des connaissances.

Enfin, l'article de Richard Parent nous permet de faire le lien avec ceux des sections subséquentes. Cet auteur souligne le caractère distinctif des données textuelles par comparaison avec les données homogènes des bases de données conventionnelles. Si ces dernières peuvent facilement être reliées et organisées en matrices de références et en tableaux de données calculables, il en va autrement des données textuelles: leur hétérogénéité, leur variété requièrent des identificateurs informels, non objectifs. C'est ce qui fait dire à l'auteur que les systèmes informatiques ne peuvent plus, à l'ère de la micro-informatique, s'appuyer sur une simple enquête sur les besoins de l'entreprise menée une fois pour toutes en phase préliminaire de conception.

Une part accrue dans l'activité cruciale de modélisation conceptuelle est accordée à l'usager par le prototypage. Ici, Richard Parent fait valoir l'importance du « vocabulaire de domaine » et de l'« analyse de contenu » qui confèrent, selon lui, une position centrale aux experts de domaine. Il insiste enfin sur la nécessité de pouvoir disposer d'opérateurs spéciaux dans le traitement de l'information textuelle évoquant, à titre d'exemple, la construction de thésaurus ouverts. La gestion de l'information textuelle constitue une nouvelle source de productivité. Les experts de chaque domaine étant les mieux placés pour effectuer une modélisation conceptuelle, il importe de redéfinir les rôles respectifs des informaticiens et des usagers.

**Méthodes et applications**

C'est dans la deuxième section du recueil que les auteurs se penchent plus spécifiquement, selon une approche méthodologique et appliquée, sur la « gestion électronique des documents ».

Pour Maurice Gingras qui examine les défis posés par la nouvelle technologie, « la gestion électronique des documents est en voie de devenir pour les corporations l'élément clé d'une stratégie globale de gestion de l'information ». Complément à la gestion documentaire classique, celle-ci offre non seulement des outils de recherche et de repérage optimisés, mais aussi des outils d'annotation de textes lus. On verra que ce dernier concept - qui consiste essentiellement en la création et l'ajout de propriétés numériques ou symboliques sur les lexèmes ou les mots en contexte - se situe au coeur de l'approche SATO (Système d'analyse de textes par ordinateur).

Dans un article où il adopte délibérément un point de vue philosophique, François Daoust, concepteur du logiciel SATO et coordonnateur de l'équipe ITC (Ingénierie textuelle et cognitive) du Centre d'ATO (UQAM), examine de plus près l'*objet informatisable* qu'est le texte. Après avoir suggéré comment SATO, par sa démarche interactive, se démarque par rapport aux concordanciers de première génération, Daoust pose la distinction essentielle entre le texte abordé d'un point de vue informatique et le texte abordé par le lecteur. L'objet se présente dans le premier cas comme une *suite de caractères,* alors qu'il apparaît plutôt dans le second comme une suite de mots ou d'*unités langagières.* Cette constatation en apparence toute simple est absolument fondamentale. Le modèle que veut construire François Daoust vise « à représenter le texte comme une suite de mots correspondant à autant d'occurrences d'unités langagières » (formes lexicales ou lexique du texte). Discutant des avantages informatiques et métho-

dologiques de la représentation lexicale, l'auteur montre comment SATO permet de décupler les capacités de lecture et justifie du coup l'approche interactive qui y est développée. Le projet ACTe (Atelier cognitif et textuel), qui prévoit notamment l'intégration d'un moteur d'inférence (du type de celui de D_expert) dans SATO devrait, selon François Daoust, contribuer à l'approfondissement « du nouveau rapport au texte rendu possible par l'outil informatique ».

Selon un ordre de préoccupation différent, les auteurs Fernand Harvey et Claude Lamarre, de la Direction de la gestion documentaire, ainsi que Serge Garceau, de la Direction générale de la vérification, au ministère du Revenu du Québec, mènent une analyse comparative de la « gestion documentaire » et de la « gestion textuelle » au sein de cet organisme. L'étude conclut que bien qu'elles desservent des clientèles distinctes - l'une diversifiée, l'autre spécifique - les deux démarches sont complémentaires et doivent être harmonisées.

La chercheure Suzanne Bertrand-Gastaldy, de l'École de bibliothéconomie et des sciences de l'information (EBSI, Université de Montréal), compare également l'approche plus traditionnelle assimilée à la gestion documentaire, soit l'indexation intellectuelle déléguée à un spécialiste, avec l'approche propre à la gestion de l'information textuelle. Cette dernière approche peut être assimilée aux méthodes reposant sur des fondements statistiques et combinatoires ou encore, comme le démontre l'auteure, être assimilée avec plus de bonheur aux méthodes d'« indexation et de repérage assistés par ordinateur ».

Selon Suzanne Bertrand-Gastaldy, l'indexation déléguée est inadéquate pour le traitement des documents administratifs. Cette inadéquation est liée à la nature des documents administratifs, au processus même et aux résultats de l'analyse intellectuelle et, enfin, aux besoins des utilisateurs. Quant à l'accès direct au texte intégral, il présente, dans la majorité des logiciels commerciaux, les inconvénients majeurs de ne pas reconnaître les unités linguistiques et de ne se fonder que sur des propriétés statistiques et de positionnement.

Dans la suite de son exposé, l'auteure discute de la nécessité des caractérisations linguistiques, sémantiques et textuelles. Elle fait l'inventaire des nombreuses difficultés que pose un tel projet : ambiguïté des caractères, extraction des termes composés, catégorisation grammaticale, regroupement des variantes orthographiques (équivalences) et flexionnelles (lemmatisation), homographie, polysémie, relations lexico-sémantiques, pragmatique, etc. Enfin, l'auteure suggère, à partir d'une expérience menée avec SATO, que l'indexation assistée par ordinateur, parce qu'elle offre des possibilités de catégorisations personnalisées, semble constituer « pour le moment la meilleure façon d'exploiter au maximum la richesse des gisements textuels ».

**Approches linguistiques et cognitives**

La troisième section du numéro thématique d'*ICO Québec* regroupe quelques articles davantage traversés par la dimension de la recherche.

Ainsi, après la présentation du projet SAGÉE (Système d'aide à la gestion des évaluations environnementales), le chercheur Louis-Claude Paquin, en collaboration avec Luc Dupuy et Yves Rochon, expose quelques concepts fondamentaux du processus d'acquisition de connaissances à partir des textes, en mettant l'accent sur le traitement et l'organisation des groupes nominaux qu'ils effectuent à l'aide du logiciel SATO. Ces auteurs tracent enfin les perspectives de développement des archives textuelles de SAGÉE, soit : la hiérarchisation des objets textuels, l'analyse des processus de raisonnement et la mise au point de mécanismes d'indexation textuelle.

Léo Laroche rend compte d'un projet de recherche réalisé pour le ministère de l'Éducation du Québec, visant à la mise au point, à partir de SATO, d'un outil d'analyse informatique de la lisibilité de textes destinés aux élèves du primaire et du secondaire.

Le linguiste André Dugas, qui travaille pour sa part dans la perspective de la construction d'un dictionnaire électronique du français, passe en revue les facteurs dynamiques de la variation orthographique, de l'affixation et de la composition qui entraînent la création de néologismes verbaux.

Enfin, l'article de Sophie David et de Pierre Plante et, dans une moindre mesure, celui de Claude Ricciardi Rigault, développent et défendent une approche résolument linguistique de l'ATO. Ils sont tous trois chercheurs à la section RDLC (Recherche et développement en linguistique computationnelle) du Centre d'ATO.

Préoccupée par le problème de la représentation d'un corpus de textes scientifiques, Claude Ricciardi Rigault expose comment, après avoir expérimenté des outils informatiques exploitant les qualités associatives de la langue (LEXIMAPPE, LEXIQUEST), elle en est venue à vouloir développer un traitement sémantique à base morpho-syntaxique. Les deux principaux problèmes soulevés quant à la première méthode résident « dans l'hétérogénéité [des unités] (représentation d'objets et d'actions mêlées) et dans la non-identification des liens existant entre les noeuds (...) ». La chercheure, qui s'intéresse plus spécifiquement à la représentation des actions, donne une description succincte du module LIAISSE comprenant notamment une composante syntaxique (ALSF) et une dimension sémantique (SOFILEX).

L'article de Sophie David et de Pierre Plante établit d'abord en quoi l'approche lexico-textuelle est inadéquate pour la détection des unités complexes. Selon les auteurs, cette approche ne règle en effet ni les problèmes d'ambiguïté catégorielle, ni ceux de la lemmatisation. Elle

n'autorise pas non plus l'application de procédures de fouille, puisqu'elle ne parvient pas à distinguer les configurations structurales différentes. Une deuxième partie de l'article justifie la nécessité d'une approche morpho-syntaxique pour l'étude ou la détection automatique des synapsies (unités complexes). Ces dernières se définissent comme des unités polylexicales dont la structure relève de la syntaxe et qui occupent la « position noyau » d'un groupe nominal. Enfin, la présentation du logiciel de dépouillement terminologique TERMINO, qui permet le repérage automatique des synapsies, complète l'exposé.

**Un nouveau défi pour les bibliothécaires**

L'accès à l'information de nature textuelle est de plus en plus valorisé dans les organisations. Quotidiennement, une quantité considérable de connaissances et d'expertises sont stockées sous forme de textes ordinolingues intégraux, constituant ainsi une nouvelle forme de mémoire institutionnelle.

Le concept de gestion de l'information textuelle est aujourd'hui indissociable de la technologie informatique. Les chercheurs prennent en compte les effets de la nouvelle technologie et sont toujours plus sensibles aux besoins qui émergent chez les utilisateurs de ces ressources.

Comme le suggère François Daoust, « la clé du succès pour le développement de l'informatique textuelle repose [sur] une liaison étroite entre l'utilisateur qui emploie l'instrument, le chercheur en sciences humaines et l'informaticien de métier ». L'implantation de systèmes automatisés de GDT suppose la mise au point d'une méthodologie qui soit adaptée à la nature de l'organisation, qui respecte la nature et la variété des documents considérés et enfin qui rencontre les contraintes de temps humain.

Suzanne Bertrand-Gastaldy a relevé l'inadéquation de la solution traditionnelle d'analyse déléguée pour la représentation des textes de nature administrative. Dans la même perspective, François Daoust ainsi que Louis-Claude Paquin et ses collègues préconisent une approche interactive qui valorise les processus cognitifs mis en oeuvre lorsque l'utilisateur lit, sélectionne, comprend et représente l'information contenue dans les textes.

Ainsi, l'intervention du bibliothécaire ne doit plus être marquée par l'« autorité » - entendue ici dans le sens que lui donne N. Ford[1] - mais doit plutôt favoriser la « liberté » des utilisateurs. Cette approche prend en compte les diverses structures cognitives des utilisateurs et favorise le développement de leurs habiletés métacognitives.

On entrevoit toutefois, et c'est ici notamment que le bibliothécaire doit intervenir, la nécessité d'instaurer une approche méthodique. On entrevoit déjà un besoin d'organisation des connaissances, d'élaboration de plans de classement et de contrôle du vocabulaire. Richard Parent fait allusion à la construction de « thésaurus ouverts », Jean Asselin mentionne l'emploi d'« outils génériques ».

Les articles de ce numéro d'*ICO Québec* démontrent bien la complexité de l'entreprise. La gestion automatisée des données textuelles pose des problèmes bien particuliers. Si les données textuelles, du fait de leur hétérogénéité, sont semi-structurables (René Lortie), requièrent des identificateurs informels ou non objectifs (Richard Parent), elles n'en exigent pas moins paradoxalement, et ceci dans le but d'optimiser les processus automatiques effectués sur des corpus de plus en plus volumineux, des procédures formelles (S. David et P. Plante), entièrement automatisables, fiables et autonomes.

À ce propos, Suzanne Bertrand-Gastaldy a évoqué les différents niveaux d'ambiguïté des textes écrits. Sophie David et Pierre Plante ont montré les difficultés qui persistent pour lever l'ambiguïté lors de la catégorisation grammaticale et de la lemmatisation sans prise en compte de la syntaxe. Comment résoudre par ailleurs le problème de la « mise à plat » de l'information (C.R. Rigault) ? Et, comme le fait valoir Jean-Guy Meunier, il ne faut pas perdre de vue que « le texte ne constitue qu'une forme d'expression limitée du discours ».

Les niveaux d'intervention sont divers : l'approche SATO, sans évacuer totalement la composante syntagmatique (blocage de locutions, concordances, contexte) oeuvre davantage au niveau de la représentation lexicale ou paradigmatique. Le logiciel TERMINO, quant à lui, prend en charge le niveau morpho-syntaxique. La solution mixte d'indexation assistée par ordinateur est sans doute celle qui répond le mieux pour l'instant aux considérations d'ordre pragmatique.

Le défi, on le constate, est de taille. Il est essentiel que les intervenants de la bibliothéconomie et des sciences de l'information participent à ces recherches en gestion de l'information textuelle en mettant à profit leur expertise acquise en matière de gestion documentaire. En revanche, ils pourront prélever dans les savoirs multidisciplinaires issus de ces recherches les outils susceptibles de répondre aux nouveaux besoins d'information de leurs clientèles.

À la lumière de ces considérations, il paraît donc souhaitable que les bibliothécaires examinent lors de rencontres professionnelles futures les gestes concrets qu'ils sont appelés à poser en matière de gestion d'information textuelle.

1. N. Ford « Knowledge Structures in Human and Machine Information Processing - Their Representation and Interaction », *Social Sciences Information Studies*, 3 (1983), 209-222.