

L'industrie canadienne de la gestion de contenu

Carte routière technologique
(2003 – 2007)



TOUS DROITS RÉSERVÉS.

Il est interdit de distribuer, de reproduire, d'enregistrer, de photocopier, de mettre dans un tableur ou dans un système de mise en mémoire et de récupération de l'information le présent document, par quelque moyen que ce soit, y compris par procédé électronique ou par tout autre moyen, sans l'autorisation écrite explicite d'AILIA Inc.

Copyright © 2004 AILIA Inc.

INTRODUCTION

À l'été 2002, des représentants d'entreprises, d'universités et du gouvernement de diverses régions du Canada ont formé, avec l'appui d'Industrie Canada et du Conseil national de recherche du Canada, le comité directeur de Carte routière technologique de l'industrie de la langue. Pour les besoins de l'exercice, on a déterminé que les technologies de la langue se composent des technologies liées aux sous-secteurs de la gestion de contenu, du traitement de la parole, de la traduction et de la formation. Le présent document est un sommaire du rapport de la première étape portant sur les technologies de traitement de la parole. La mission du comité de la Carte routière technologique de la gestion de contenu de l'Association de l'industrie de la langue / Language Industry Association (AILIA) est la suivante :

PROCESSUS DE CRÉATION DE LA CARTE ROUTIÈRE TECHNOLOGIQUE

- Élaboration d'un énoncé de vision décrivant la raison d'être et les objectifs de la Carte routière technologique (2003-2004);
- Définition de l'envergure de la carte routière technologique de la gestion de contenu, ce qui comprend les technologies clés et la recherche (2003-2004);
- Détermination des forces motrices du marché de la gestion de contenu et de leurs cibles (2003-2004);
- Recommandation de stratégies à adopter par AILIA pour le soutien de la recherche et du développement dans le domaine de la gestion de contenu (2004-2005);
- Définition des habiletés et des connaissances que devra posséder la main-d'œuvre de demain pour mettre au point et utiliser les nouvelles technologies de gestion de contenu (2004-2005);
- Mise en œuvre de stratégies à adopter par AILIA pour le soutien de la recherche et développement en matière de gestion de contenu (2004-2007);
- Encouragement du développement des habiletés et du savoir de la main-d'œuvre canadienne de demain par des mesures incitatives dans les entreprises et un soutien financier aux études (2004-2007).

QU'EST-CE QUE LA GESTION DE CONTENU ?

Les centres de recherche et les entreprises du secteur de la gestion de contenu mettent au point des technologies et des applications utilisées: pour organiser, catégoriser et structurer les ressources informationnelles, de manière à ce que l'on puisse les emmagasiner, publier et réutiliser de diverses manières; pour automatiquement créer, interpréter et analyser l'information non structurée (ex.: documents Word ou courriels) et l'information semi-structurée (ex.: formulaires et pages webs avec métadonnées); et pour extraire des connaissances de l'information.

Les technologies et les applications de gestion de contenu sont créées à l'aide d'approches à base de règles, d'approches statistiques et d'approches linguistiques. Elles permettent à différentes catégories de machines et de gens de traiter l'information plus intelligemment et avec une productivité accrue.

Dans ce document, le terme *information* renvoie principalement à de l'information textuelle. Le thème de la gestion de contenu multimédia n'est pas abordé en ces pages. (Pour en savoir plus sur les questions relatives au traitement de la parole, voir la synthèse produite par le comité de la Carte routière technologique du traitement de la parole.) Cependant, comme du texte accompagne souvent les images (ex.: titre ou légende), on peut utiliser les outils de gestion de contenu pour repérer, classer et organiser l'information visuelle.

LE MARCHÉ DE LA GESTION DE CONTENU

Selon une étude menée à Berkeley¹, cinq exaoctets de nouvelle information ont été produits dans le monde en 2002, l'équivalent de 2 500 000 nouvelles bibliothèques de la taille d'une bibliothèque universitaire canadienne typique. La croissance rapide de l'information touche toutes les industries. Par exemple, chaque année, les institutions financières doublent ou triplent leurs capacités de stockage pour pouvoir emmagasiner la nouvelle information qu'elles génèrent. De leur côté, les acteurs de l'industrie des sciences biologiques anticipent que la quantité d'information produite dans leur domaine centuplera d'ici cinq ans².

Tant d'information est produite dans le monde que traiter celle-ci manuellement est devenu complètement inefficace. Des façons d'organiser, d'interpréter et d'analyser les contenus plus rapidement, plus intelligemment et plus économiquement sont désespérément requises pour accroître la productivité des organisations des secteurs public et privé.

Entre autres, les outils de gestion de contenu aident les organisations à :

- *Gérer le courrier électronique.* Cela signifie bloquer les pourriels, lesquels représentent 38% des 31 milliards de messages envoyés chaque jour en Amérique du Nord³, tout en laissant les courriels légitimes parvenir aux organisations. Cela veut aussi dire assurer que les courriels importants puissent être recouverts au bon moment;
- *Améliorer le service-client.* Le nombre d'interactions électroniques entre les organisations et leurs clients augmentent de 30% par année au moins selon le Giga Group. Le recours aux outils de gestion de contenu permet de traiter les communications routinières automatiquement. Les représentants peuvent dès lors passer plus de temps sur les messages les plus importants⁴;
- *Transformer l'information en connaissances.* Par exemple, dans le projet sur le génome humain, on utilise les outils de gestion de contenu pour rechercher les cooccurrences de noms de gène présentes dans les résumés d'articles, déterminer quels noms de gène figurent souvent simultanément à l'intérieur des documents et examiner les phrases décrivant différents gènes pour découvrir les liens qui les unissent;
- *Accroître leurs ventes.* Le Web sémantique aidera les internautes et leurs agents à trouver les produits qu'ils recherchent (ex. : taper une requête comme « meuble de rangement *Late Georgian* » pourra permettre de repérer une photo représentant « une commode anglaise, autour de 1800 »);
- *Protéger les actifs organisationnels.* Les outils de gestion de contenu peuvent être utilisés pour détecter l'existence de contenu web diffamatoire, de contrefaçons de brevet, de fuites d'éléments de propriété intellectuelle, etc.

Les technologies de gestion de contenu ne permettent pas seulement aux organisations de faire les mêmes choses plus vite, pour moins cher et mieux. Dans de nombreux cas, elles sont aussi la clé du développement de marchés complètement neufs. Par exemple :

- Le marché de la vente de mots-clés (*paid search*), le segment qui croît le plus vite en matière de publicité internet, dépend, pour s'épanouir, des progrès réalisés en matière de gestion de contenu. (Les ventes de mots-clés augmentent de 50% par an et, en 2007, devraient atteindre les 7 milliards de \$ US, selon la société US Bankcorp Piper Jaffray.);

¹ Lyman, Peter and Hal R. Varian (2003), *How Much Information*, consulté le 18 septembre 2004 à l'adresse www.sims.berkeley.edu/how-much-info.

² Gerr, Peter (2003), *Compliance: The Effect on Storage and Information Management*, consulté le 18 septembre 2004 à l'adresse www.enterprisestrategygroup.com/documents/Report/Attachment2ID201.pdf.

³ Anonymus (2004), "Spam volume keeps rising" *CNET News.com*, 1er septembre, consulté le 18 septembre 2004 à l'adresse <http://makeashorterlink.com/?B2A022A39>. À eux seuls, les pourriels coûtent 10 milliards de \$ US en productivité perdue.

⁴ Venetica (2004), *Achieving Customer Service Excellence, The Vital Role of Content Integration*, consulté le 18 septembre 2004 à l'adresse www.dmreview.com/whitepaper/WID1008847.pdf.

- Selon les prédictions, les messages multimédias (MMS), c'est-à-dire des cartes virtuelles combinant des images, du texte et de la musique, remplaceront les textos (SMS) comme application phare de l'industrie du cellulaire (les textos génèrent plus de 10% des revenus des fournisseurs téléphoniques). Cela arrivera grâce aux améliorations apportées aux outils de gestion de contenu.

Le potentiel des outils de gestion de contenu explique que l'on prédise une augmentation rapide de leurs ventes. IDC prévoit ainsi que les ventes de licences d'utilisation de nouveaux logiciels de gestion de contenu bondiront à 3,8 milliards de \$ US d'ici 2007. Pour sa part, le Meta Group estime qu'elles pourraient dépasser 9 milliards de \$ US cette même année « à cause de la nécessité croissante pour les organisations de se conformer aux lois et d'adopter des procédures/politiques les protégeant des poursuites⁵ ».

Un indice de l'importance croissante de la gestion de contenu est la présence grandissante dans le secteur de grandes sociétés. Par exemple, Microsoft fait des efforts significatifs pour assurer que la recherche de contenus soit « plus agréable » avec Longhorn, son prochain système d'exploitation; IBM a récemment acquis Tarian, un producteur de technologies de tenue d'archives; et OpenText a acheté IXOS pour ses technologies d'archivage de courriels.

L'INDUSTRIE DE LA GESTION DE CONTENU

La gestion de contenu se définit par rapport aux champs, aux technologies et ressources, aux standards et aux applications dont une liste encore incomplète suit.

CHAMPS

Analyse de contenu	Forage de connaissances
Génération de contenu	Représentation des connaissances / raisonnement
Structuration de contenu	Traitement du langage naturel
Gestion de documents	Analyse textuelle
Recherche de documents	Forage de données textuelles
Extraction d'information	Compréhension de textes
Recherche d'information	Gestion de contenu web
Recherche de connaissances	Annotations écrites
Gestion des connaissances	

TECHNOLOGIES ET RESSOURCES

Catégorisation automatique	Technologies d'apprentissage machine
Agrégation automatique	Reconnaissance d'entités nommées
Reconnaissance de forme automatique	Traitement du langage naturel
Résumé automatique	Gestion de structures normées
Agrégation de contenu	Ontologies
Balayage de contenu	Parsage
Convertisseurs	Catégorisation grammaticale
Corpus	Classement par pertinence
Formatage	Extraction de règles
Correction grammaticale et parsage	Algorithmes de recherche
Grammaires	Analyse sémantique
Bases de données hiérarchiques et relationnelles	Correcteurs orthographiques
Indexation	Catégorisation de textes

⁵ Boulton, Clint (2004), "Latest ECM Match: Stellent Snaps up Optika", *Enterprise*, 12 janvier, consulté le 18 septembre 2004 à l'adresse www.interentnews.com/ent-news/article.php/3297991.

Moteurs d'inférences
Extraction d'information
Recherche d'information
Agents intelligents
Représentation des connaissances
Analyse de la langue
Lexiques
Extraction de métadonnées

Classification de textes
Terminologies
Forage textuel
Thésaurus
Cartes conceptuelles
Interfaces usagers
Visualisation
XSL

STANDARDS

HTML
RDF
SGML
XCES

XML
XSLT
OWL
Webdav

APPLICATIONS

Catégoriseurs automatiques
Avatars
Créateurs de contenu
Bayeurs de contenu
Gestion de systèmes de relation client
Systèmes de gestion de documents
Systèmes de gestion de courriels
Systèmes experts
Indexeurs
Applications de recherche de connaissances
Correcteurs

Outils de gestion de structures normées
Assistants virtuels personnalisés
Systèmes de question réponse
Systèmes de gestion de fichiers
Moteurs de recherche
Systèmes de résumés automatiques
Systèmes d'analyse textuelle
Systèmes de gestion de thesaurus et d'ontologies
Aides à la traduction
Outils de gestion de contenu web

SOCIÉTÉS CANADIENNES CLÉS⁶

- Axonwave
- Blast Radius
- BorderWare Technologies
- CaseBank Technologies
- Chamblon Systems
- Cogilex R&D
- Cognos
- Convera
- Copernic
- Corel
- Delphes Technologies
- Documents
- Druide Informatique
- eManage
- Entrust (formerly Amikanow)
- Ever America
- GBS Design
- Hummingbird
- Idilia
- INM International
- inSystems
- iUpload
- Ixiasoft
- John Chandiox
- Lingua Technologies
- Messaging Architects
- Minacs Worldwide
- Neural Machines
- MultiCorpora
- MXI
- Nomino Technologies
- North Sea NMT
- Novator
- nStein
- OpenText/Ixos
- Palomino
- Readplease
- Roaring Penguin Software
- SonicBoomerang
- Tarian/IBM
- Vircom
- Zi Corporation

⁶ Une revue supportée par AILIA en 2003 a prouvé que ces entreprises canadiennes étaient liées aux technologies de gestion de contenu. Cependant, cette liste n'est pas exhaustive. Une revue complète sera menée en 2004-2005. Contactez AILIA à communication@ailia.ca pour ajouter votre compagnie à cette liste.

INSTITUTIONS DE RECHERCHE CANADIENNES CLÉS ⁷

- Centre for Pattern Recognition and Machine Intelligence (Concordia University)
- Centre interdisciplinaire de recherche sur les activités langagières (Université Laval)
- Computational Linguistics at Concordia Lab (Concordia University)
- Department of Computer Science (University of Toronto)
- Dalhousie Natural Language Processing Group (Dalhousie University)
- Department of Computing Science (University of Alberta)
- Department of Computer Science (University of Manitoba)
- Knowledge Acquisition and Machine Learning Group (University of Ottawa)
- Laboratoire d'analyse cognitive de l'information (Université du Québec à Montréal)
- Laboratoire de Recherche Appliquée en Linguistique Informatique (Université de Montréal)
- Laboratoire en informatique cognitive et environnements de formation (Télé-Université)
- Centre de recherche en technologies langagières (Université du Québec en Outaouais)
- Centre national de recherche du Canada (CNRC)
- Natural Language Lab (Simon Fraser University)
- OLST Observatoire de linguistique Sens-Texte (Université de Montréal)
- School of Computing Science (Carleton University)
- Stat-NLP Group (University of Waterloo)
- VRQ Groupe sur le traitement des langues naturelles (Host: Université du Québec à Montréal)

FORCES ET FAIBLESSES DE L'INDUSTRIE CANADIENNE DE LA GESTION DE CONTENU

Le Canada figure parmi les joueurs majeurs de la gestion de contenu. L'avantage concurrentiel du Canada trouve ses origines dans un vaste bassin d'experts en traitement de la langue, la présence de centres de recherche dynamiques, diverses possibilités de financement gouvernemental, une demande importante de services bilingues, la force des intégrateurs canadiens, etc. Cependant l'industrie présente aussi quelques faiblesses parmi lesquelles on note:

- la diminution du bassin d'emplois qualifiés en gestion de contenu, en raison de la crise récente dans les technologies de pointe et de l'exode consécutif de professionnels hors du pays et / ou du secteur;
- l'investissement en capital de risque généralement moindre ici qu'aux États-Unis;
- le manque de diverses ressources langagières telles que les corpus ou les bases de connaissances, en particulier pour les langues autres que l'anglais;
- la difficile identification de statistiques valables concernant les sous-secteurs du marché de la gestion de contenu ou l'obtention de données chiffrées divergentes;
- le frein au développement des petites et moyennes entreprises constitué par la politique gouvernementale d'approvisionnement dans le secteur.

⁷ Voir la note précédente.

RECOMMANDATIONS PRÉLIMINAIRES DU COMITÉ SUR LA GESTION DE CONTENU

Les recommandations préliminaires du comité sur la gestion de contenu incluent :

- Dresser un inventaire du domaine de la gestion de contenu au Canada, fournisseurs de technologie, chercheurs et intégrateurs;
- Créer les nouvelles ressources langagières requises par les compagnies et les centres de recherche;
- Développer un portail d'information sur la gestion de contenu. Le portail devrait proposer, entre autres, des nouvelles sur ce qui se passe dans le secteur au Canada et à travers le monde, un forum où l'industrie, les universités et le gouvernement pourraient discuter des stratégies propres à augmenter l'avantage concurrentiel du Canada dans le domaine, ainsi que de l'aide pour localiser les ressources langagières;
- Participer aux forums où des standards de droit et de fait, comme XML ou OWL, susceptibles d'affecter l'évolution de la gestion de contenu, sont produits et propager l'information les concernant;
- Fournir un forum permettant l'interaction avec les autres sous-comités de la Carte routière technologique (traitement de la parole, traduction et formation) et, plus largement, avec les membres d'AILIA.

POURQUOI LES ORGANISMES ET LES PROFESSIONNELS DE LA GESTION DE CONTENU DEVRAIENT-ILS DEVENIR MEMBRES D'AILIA?

Le Canada est considéré comme un fournisseur de premier plan en biens et services langagiers, mais la concurrence se fait de plus en plus vive. Les intervenants du domaine ont donc décidé de passer à l'action et de prendre les moyens nécessaires pour assurer un nouveau départ au secteur. C'est pour relever ce défi que l'Association de l'industrie de la langue / Language Industry Association (AILIA) est née.

Devenir membre d'AILIA vous permet de recevoir un bulletin mensuel de veille propre à l'industrie, d'être informé à propos des missions commerciales, de tenir les autres membres au courant de vos produits et services, de participer aux activités de développement du secteur, d'établir des contacts d'affaires, ainsi que de participer à la définition et à la mise en place, par AILIA, de stratégies d'assistance à la recherche et au développement dans le secteur des technologies de gestion de contenu.

Inscrivez-vous dès maintenant à <http://www.ailia.ca> !