

The Canadian Content Management Industry

Technology Roadmap
(2003 – 2007)



ALL RIGHTS RESERVED.

No part of this document may be distributed, reproduced, recorded, photocopied, and entered into a spreadsheet or information storage and/or retrieval system of any kind by any means, electronic, mechanical or otherwise without the expressed written permission of AILIA Inc.

Copyright © 2004 AILIA Inc.

INTRODUCTION

In the summer of 2002, industry, university and government representatives from across Canada, with the support of Industry Canada and the National Research Council of Canada, formed the Steering Committee of the Language Industries Technology Roadmap. For the purpose of the exercise, language technologies consist of the technologies related to Content Management, Speech Processing, Translation and Training subsectors. This is a summary of the first phase report produced by the Content Management Technology Roadmap committee, whose goal is to assist Canadian Content Management firms to achieve a competitive advantage and maximize global market share.

THE TECHNOLOGY ROADMAP PROCESS

- Establish statements of the purpose and goals of the Technology Road Map (TRM) **(2003-2004)**;
- Define the scope of the Content Management Industry TRM. This includes key technologies and research **(2003-2004)**;
- Specify Content Management technology market drivers and their targets **(2003-2004)**;
- Define the needs of the Content Management industry and its customers **(2004-2005)**;
- Recommend Content Management technology research and development support strategies that should be pursued by ALLIA **(2004-2005)**;
- Define the skills and knowledge required by the industry's future work force in order to develop and implement the new technologies **(2004-2005)**;
- Implement ALLIA's Content Management technology research and development support strategies **(2004 –2007)**;
- Promote skill and knowledge development of Canada's future work force through business incentives and academic financial support **(2004 –2007)**.

WHAT IS CONTENT MANAGEMENT ?

Content Management research centers and firms develop technologies and applications that are used: to organize, categorize, and structure information resources so that they can be stored, published, and reused in multiple ways; to automatically create, interpret and analyze unstructured information (e.g., Word documents or emails) and semi-structured information (e.g., forms or Web pages with metadata); and to extract knowledge from information.

Content Management technologies and applications use statistical, rule-based and linguistic approaches. They make it possible for different categories of machines and people to process information more intelligently and more productively.

In this document the term information is used to refer primarily to textual information. The issue of multimedia content management is not addressed here. (For issues relating to speech processing, see the summary of findings produced by the Speech Processing Committee.) However, images are often accompanied by text (e.g., title or description), and thus Content Management technologies can be used to retrieve, classify and organize visual information.

THE CONTENT MANAGEMENT MARKET

According to a study conducted at Berkeley¹, five exabytes of new information were produced in the world in 2002, the equivalent of 2 500 000 new libraries the size of a typical Canadian university library. The rapid growth of information affects all industries. For example, financial institutions double or triple their storage capacities every year so they can stock the new information they generate, while life sciences companies expect that the quantity of data produced in their field will increase 100-fold over the next few years².

So much information is produced in the world today that processing it manually has become a completely unproductive approach. Ways to organize, interpret and analyze content more rapidly, more intelligently and more cost-effectively are desperately needed by private and public sector users looking to increase their productivity.

Among other things, Content Management tools help organizations:

- *Manage emails.* That means blocking spam, which accounts for 38% of the 31 billion emails sent each day in North America³, while letting legitimate mail get in. It also means making sure critical emails can be retrieved at the right time.
- *Improve customer service.* Internet customer service interactions are increasing by as much as 30% per year, according to the Giga Group. Using Content Management tools, routine messages are handled automatically, so customer service representatives have more time to spend on critical ones⁴.
- *Turn information into knowledge.* For example, in the Genome project, Content Management tools are used to look for co-occurring gene names within abstracts, identify the genes that frequently co-occur within documents and examine sentences describing different genes to discover the relationship that exists between them.
- *Increase their sales.* Semantic Web will make it easier for shoppers and their agents to find the products they are looking for (e.g., typing a query like "Late Georgian storage furniture" might return a photo depicting a "British chest of drawers, around 1800").
- *Protect assets.* Content Management tools can be used to detect defamatory or libelous content present on the Web, patent infringements, intellectual property leaks, etc.

Content Management technologies do not only make it possible for organizations to do the same things faster, better and cheaper. In many cases, they are also key to the development of *whole new markets*. For example,

- Paid Search, the fastest-growing segment in Internet advertising, depends, to be efficient, on advances in Content Management. (Paid Search grows at an annual rate of 50%, and in 2007 will become a US \$7 billion industry, according to US Bankcorp Piper Jaffray Inc.).
- Multimedia Messages (MMS), that is, virtual cards combining visual, textual and musical components, are predicted to replace Short Textual Messages (SMS), currently the source of more than 10% of operators' revenues, as the mobile industry's next killer application. This will happen with improvements in Content Management tools.

¹ Lyman, Peter and Hal R. Varian (2003), *How Much Information*, retrieved from www.sims.berkeley.edu/how-much-info on September 5, 2004.

² Gerr, Peter (2003), *Compliance: The Effect on Storage and Information Management*, retrieved from www.enterprisestrategygroup.com/documents/Report/Attachment2ID201.pdf on September 5, 2004.

³ Anonymous (2004), "Spam volume keeps rising" *CNET News.com*, September 1, retrieved from <http://makeashorterlink.com/?B2A022A39> on September 5, 2004. By itself, spam costs US \$ 10 billion a year in lost productivity.

⁴ Venetica (2004), *Achieving Customer Service Excellence, The Vital Role of Content Integration*, retrieved from www.dmreview.com/whitepaper/WID1008847.pdf on September 5, 2004.

The potential of Content Management tools explains that their sales are projected to grow at a rapid pace. IDC anticipates the market for new content management software licenses will grow to US \$3.8 billion by 2007, while the Meta Group estimates it will top US \$9 billion by the same year, "due to the heightened concern for proper compliance and legal risk policies/procedures⁵".

A sign of the growing importance of Content Management is the increasing presence of large firms in the sector. For example, Microsoft is working hard to make sure that "search will be nicer for customers⁶" in Longhorn, its future operating system; IBM has recently acquired Tarian, a producer of record-keeping technology; and OpenText has bought IXOS for its email archiving technology.

THE CONTENT MANAGEMENT INDUSTRY

The Content Management industry is defined by the following fields, technologies, resources, standards and applications. (The list is not complete.)

FIELDS

Content Analysis	Knowledge Mining
Content Generation	Knowledge Representation and Reasoning
Content Structuring	Natural Language Processing
Document Management	Text Analysis
Document Retrieval	Text Data Mining
Information Extraction	Text Understanding
Information Retrieval	Web Content Management
Knowledge Discovery	Written Language Input
Knowledge Management	

TECHNOLOGIES AND RESOURCES

Automatic Categorization	Machine Learning technologies
Automatic Clustering	Named Entity Recognition
Automatic Pattern Matching	Natural Language Processing
Automatic Summarization	Normative Structure Management
Content Clustering	Ontologies
Content Scanning	Parsing
Converters	Part-of-Speech Taggers
Corpora	Relevance Ranking
Formatters	Rule Extraction
Grammar Checking and Parsing	Search Algorithms
Grammars	Semantic Analysis
Hierarchical and Relational Databases	Spell Checking
Indexing	Text Categorization
Inference Engines	Text Classification
Information Extraction	Terminologies
Information Retrieval	Text mining
Intelligent Agents	Thesauri
Knowledge Representation	Topic Maps
Language Analysis	User Interfaces
Lexicons	Visualization
Metadata Extraction	XSL

⁵ Boulton, Clint (2004), "Latest ECM Match: Stellent Snaps up Optika", *Enterprise*, January 12, retrieved from www.interentnews.com/ent-news/article.php/3297991 on September 7.

⁶ Microsoft's technical evangelist Robert Scoble.

STANDARDS

HTML
RDF
SGML
XCES

XML
XSLT
OWL
Webdav

APPLICATIONS

Automatic Categorizers
Avatars
Content Creators
Content Scanners
Customer Relationship Management Systems
Document Management Systems
Email Management Systems
Expert Systems
Indexers
Knowledge Discovery Applications
Language Checkers

Normative Structure Management Applications
Personal Agents
Question Answering systems
Records Management Software
Search Engines
Summarizers
Text Analysis Systems
Thesauri and Ontologies Management Systems
Translation Aids
Web Content Management Applications

KEY CANADIAN FIRMS⁷

- Axonwave
- Blast Radius
- BorderWare Technologies
- CaseBank Technologies
- Chamblon Systems
- Cogilex R&D
- Cognos
- Convera
- Copernic
- Corel
- Delphes Technologies
- Documens
- Druide Informatique
- eManage
- Entrust (formerly Amikanow)
- Ever America
- GBS Design
- Hummingbird
- Idilia
- INM International
- inSystems
- iUpload
- Ixiasoft
- John Chandieux
- 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

⁷ A 2003 AILIA sponsored survey of the Content Management industry showed that the following Canadian firms are Content Management technology providers. However this list is not exhaustive. A complete survey will be conducted in 2004-2005. Please contact AILIA at communication@ailia.ca to add your company to that list.

KEY CANADIAN RESEARCH INSTITUTIONS⁸

- 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)
- Language Technologies Research Centre (Université du Québec en Outaouais)
- National Research Council Canada
- 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)

CANADA'S CONTENT MANAGEMENT INDUSTRY STRENGTHS AND WEAKNESSES

Canada is a significant player in Content Management. Canada's competitive advantage has different sources such as a large pool of experts in language processing, the presence of dynamic research centers, various government funding opportunities, a large demand for bilingual services, the strength of Canadian integrators and so on. However, the industry suffers from some weaknesses among which:

- Many Content Management trained employees have been downsized in the recent high-tech meltdown and many have left or are leaving Canada and/or the Content Management field altogether.
- When Canadian Content Management firms are financed by venture capitalists or others, the support they receive is generally (a lot) lower than that received by American competitors.
- Some critical language resources, e.g., corpora, knowledge bases, are not available, especially in languages other than English.
- Valid statistics about Content Management submarkets are often hard to come by or numbers conflict one another.
- Government procurement policies are an impediment to the development of small and medium firms in the sector.

⁸ See note above.

PRELIMINARY RECOMMENDATIONS OF THE CONTENT MANAGEMENT COMMITTEE

The Content Management Committee's preliminary recommendations include:

- Build an inventory of Canadian Content Management technology providers, researchers, and adaptors.
- Create new language resources needed by the Content Management firms and research centers.
- Build a Content Management information portal. Among other things, the portal should contain news on what's going on in the sector in Canada and around the world, provide a forum where industry, university and government players can discuss strategies to increase Canada's competitive advantage in the industry, and help locate language resources.
- Participate in committees where *de jure* and *de facto* standards that affect the future of Content Management, e.g., XML and OWL, are produced and ensure the dissemination of information about those standards.
- Provide a forum to interact with the other language technology roadmap committees (Speech Processing, Translation, and Training) and AILIA members.

BENEFITS OF JOINING THE LANGUAGE INDUSTRY ASSOCIATION (AILIA) FOR CONTENT MANAGEMENT ORGANIZATIONS AND PROFESSIONALS

Canada is considered a top supplier of language products and services but its market share is seriously challenged. Stakeholders in the field have therefore decided to act by doing all that is needed to give the sector a fresh start. The Association de l'industrie de la langue/Language Industry Association (AILIA) has been created to meet this challenge.

By joining AILIA you can receive a monthly market intelligence newsletter, obtain news on trade missions, inform other members of your products and services, participate in business development activities, acquire business contacts, as well as participate in defining and implementing AILIA's Content Management technology research and development support strategies.

Sign up today at <http://www.ailia.ca> !