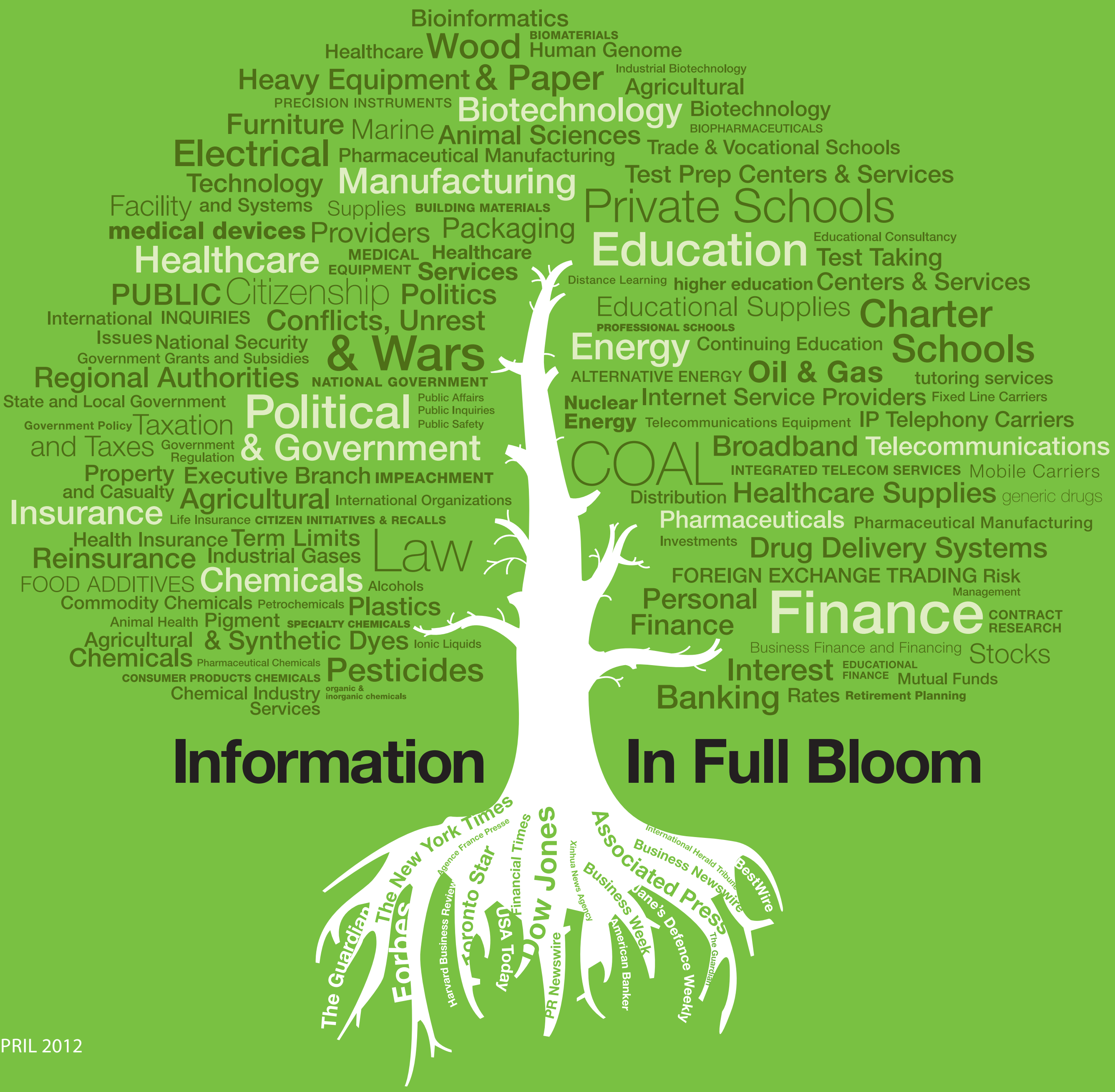
acquiremedia

The Business of Content Categorization

Mastering taxonomies, entity extraction and filtering


Bioinformatics
Healthcare Wood BIOMATERIALS Human Genome
Heavy Equipment & Paper Industrial Biotechnology Agricultural
PRECISION INSTRUMENTS Biotechnology Biotechnology BIOPHARMACEUTICALS
Furniture Marine Animal Sciences Trade & Vocational Schools
Electrical Pharmaceutical Manufacturing
Technology Manufacturing Test Prep Centers & Services
Facility and Systems Supplies BUILDING MATERIALS Private Schools
medical devices Providers Packaging Education Educational Consultancy
Healthcare MEDICAL EQUIPMENT Healthcare Services Test Taking
Distance Learning higher education Centers & Services
PUBLIC Citizenship Politics Educational Supplies Charter
International INQUIRIES Conflicts, Unrest PROFESSIONAL SCHOOLS
Issues National Security & Wars Energy Continuing Education Schools
Government Grants and Subsidies
Regional Authorities NATIONAL GOVERNMENT ALTERNATIVE ENERGY Oil & Gas tutoring services
State and Local Government Political Public Affairs Public Inquiries Public Safety
Nuclear Energy Internet Service Providers Fixed Line Carriers
Government Policy Taxation and Taxes Government Regulation & Government
Telecommunications Equipment IP Telephony Carriers
COAL Broadband Telecommunications
Property and Casualty Executive Branch IMPEACHMENT INTEGRATED TELECOM SERVICES Mobile Carriers
Agricultural International Organizations Distribution Healthcare Supplies generic drugs
Insurance Life Insurance CITIZEN INITIATIVES & RECALLS Pharmaceuticals Pharmaceutical Manufacturing
Health Insurance Term Limits Law Investments Drug Delivery Systems
Reinsurance Industrial Gases FOREIGN EXCHANGE TRADING Risk Management
FOOD ADDITIVES Chemicals Alcohols Personal Finance Finance CONTRACT RESEARCH
Commodity Chemicals Petrochemicals Plastics
Animal Health Pigment SPECIALTY CHEMICALS
Agricultural Chemicals & Synthetic Dyes Ionic Liquids Business Finance and Financing Stocks
Pharmaceutical Chemicals Pesticides Interest EDUCATIONAL FINANCE Mutual Funds
CONSUMER PRODUCTS CHEMICALS Chemical Industry Services organic & inorganic chemicals
Banking Rates Retirement Planning
The New York Times
Agence France Presse
Financial Times
Dow Jones
Xinhua News Agency
Associated Press
International Herald Tribune
Business Newswire
BestWire
The Guardian
Forbes
Harvard Business Review
Toronto Star
USA Today
PR Newswire
Business Week
American Banker
Jane's Defence Weekly
The Guardian

Information In Full Bloom

APRIL 2012

Acquire Media is a young company with a deep history processing electronic news feeds. The formats may have changed, new standards emerged, and technology has evolved but at its core, value is obtained when an individual or businesses can quickly identify information pertinent to their needs. The ability to achieve speed, precision and accuracy in digital content categorization and extraction is a skill that is developed and learned with time and experience.

Background

Beginning in 1985 the founders of Acquire Media worked together at a New Jersey based information delivery company, GARI Software. GARI was the backbone of real-time news integration at many major financial firms where they specialized in collecting feeds from a variety of originating news sources, standardizing into a single file format, and delivering these aggregated feeds to customers without slowing down performance. Recognizing the value in GARI Software's work, Dow Jones and Company purchased the company in 1990 where GARI continued its real-time news integration for Wall Street clients along with bringing Dow Jones to the forefront of online technology driving the backend of Barron's Online as well as the Wall Street Journal Interactive Edition.

Around this time a separate company, Individual Inc., was also working with electronic news feeds with a further emphasis on news categorization. Identifying news by industry and subject matter enabled customers to view and read news based on topics of interest. In 1996 Individual Inc. received its first patent (U.S. #5,537,586) addressing the selection of profiled text from a database of defined categories. When Individual, Inc. (founded 1989) and Desktop Data (founded 1988) merged in 1998 to form NewsEdge, it brought together the best of both organizations. NewsEdge now had the categorization and real-time delivery technologies to extend its reach over larger news volumes and enterprise-scale customers.

In 2001, Thomson Corporation acquired NewsEdge, integrating its core technology with its existing Dialog product line and categorization engine into Westlaw which is still in use today. Also in 2001, Acquire Media was formed with the core employee base from GARI Software. In 2007, Acquire Media purchased NewsEdge from Thomson, including its human and intellectual assets devoted to content classification for the previous eighteen years.

This history lesson is important to understand because the expertise and intellectual property from these entities now reside in Acquire Media today. Our clients directly benefit from our experience and leadership because our history is so deeply rooted in content classification and extraction.

Content Classification

The Acquire Media Metadata Enhancer (ACME) taxonomy is the product of our years of real world experience in the news business. It was created from the ground up in 2009 by Acquire Media, with assistance from Access Innovations, industry-veterans in the information management and database construction space.

The ACME taxonomy has 3 main components; Industry, Subject and Location. Today, we categorize each story we process with relevant, meaningful tags selected from over 1300 industries, 1250 subjects and 68,000 geographic locations. Within each is a 4 level hierarchy detailed enough for the most granular data refinements but simple enough to browse for discovering the information users need. The faceted structure of the taxonomy allows users to combine different categories like building blocks to create targeted filters for their specific information needs. Adding or removing categories can broaden or narrow research results giving users total control of the information flow.

ACME is a proprietary taxonomy unique to Acquire Media but the originating metadata provided by our information providers is just as vital. Whether you are accustomed to using a particular Dow Jones Newswire subject code, know the Reuters instrument codes (RICs) or use the Associated Press location codes in your filtering, Acquire Media retains all coding or metadata available in the news feeds that we process. But we don't stop there. Our Editorial Team will review all information provider metadata serving two key functions. The first is to map all proprietary provider codes to the ACME codes to ensure all users get access to the news they request as they request it. The other is to learn from those who know the content best – the content creators. Our stringent review involves a rigorous testing of provider metadata against active content to determine which codes can be further leveraged by the ACME taxonomy.

Our efforts notwithstanding, we know taxonomy is never complete as it is an ever-changing and growing process. As we engage with clients we might encounter a need to customize a branch of our taxonomy to filter content specific to a niche area of an industry or subject. Many of these needs are proprietary and Acquire Media approaches each instance uniquely and will confidentiality. However, the results are the same. Clients receive an efficient, flexible categorization process to filter their news to meet their needs completely and effectively.

Entity Extraction

Side-by-side with content categorization, Acquire Media has finely tuned entity extraction and indexing of semantic entities in news stories. During the past 10 years, we have built, expanded, refined and enhanced the news industry's most advanced semantic analysis engine, Metabot.

We built Metabot with special focus on accuracy and speed. Guided both by practical experience in the news industry and by current theory in statistics, machine learning, neural networks, pattern recognition, taxonomy development, and computational optimization, the Acquire Media semantic algorithms team has achieved a series of unmatched breakthroughs. Metabot produces industry-leading quality, at unsurpassed story-throughput rates, as measured in objective, head-to-head comparisons.

Metabot's capabilities include story classification by the ACME taxonomy and standard industry codes, entity extraction of company names, person names, locations, products, dates and times, and money, and story summarization. It is fully customizable according to each customer's business needs and preferences.

Metabot embodies Acquire Media's long experience in news. Every rule, exception, and nuance encountered over the years has been incorporated cumulatively into Metabot's ever-growing semantic network model. As a result, Metabot is the most accurate product industry-wide in processing real news -- and has been acclaimed in study after study by major customers.

Metabot is thoroughly integrated into Acquire Media's news syndication infrastructure, with indexing and retrieval, high-speed/low-latency distribution, and business analytics tools.

Commitment to Excellence

Acquire Media is the industry leader in news categorization and extraction and we are committed to delivering solutions best suited to the individually unique needs of our clients. Our knowledge and experience is vast and our employees are experts in their fields.

Our Editorial Team boasts a group of highly educated individuals with backgrounds in journalism, reporting, news database design, and taxonomic architecture. The majority of the team holds Masters of Library Sciences degrees while others hold advanced degrees in their field. With a minimum of 15 years each of hands on experience, this team hails from such organizations as Dow Jones Newswires, Ovid Technologies, Thomson-Reuters, Screaming Media/Pinnacor, Northern Light, Hearst Corporation and CBS Corporation.

Our internally managed, US-based overnight staff is tasked with operational review of news story selections, performs taxonomy maintenance and provides round-the-clock customer care. This staff includes editors, teachers and publishing professionals.

The semantic algorithms team at Acquire Media comprises experts in natural language processing, pattern recognition, pattern clustering, data mining, text parsing, adaptive neural networks, genetic algorithms, intelligent systems, numerical computation, combinatorial optimization, supercomputing, search engine internals, web spidering, indexing, data compression, large-scale databases, symbolic regression, and semantic taxonomy. Credentials include senior computer scientists, statisticians, and linguists, with doctorate and master's degrees from Princeton, Yale, Columbia, Harvard, Oxford, and other leading academic institutions. Prior experience of our team members includes work at Oxford University Press, Morgan Stanley, Lucent Technologies, NEC Telecom, ADP, Hearst Digital Media, Quintiles Transnational, and many smaller firms.

Conclusion

At Acquire Media we have taken the years of experience and expertise in our company to build robust engines and flexible processes to handle any content categorization and extraction requirement. We are proud of our services and the people that support them. Our goal is to surpass client expectations in capability, flexibility and customer service. This is part of our commitment to excellence and we always deliver.