The Value of Indexing

A White Paper Prepared for Factiva, a Dow Jones and Reuters Company
By Jan Sykes, Information Management Services
February 2001

Executive Summary

Finding particular documents after they have been reviewed and stored has been a challenge since the advent of the printed word. "Findability" is emphatically more important as we deal with information overload in general and with the specific need to quickly find relevant background information to support business decisions in a networked environment.

Because time is arguably the most valuable asset in today's economy, information users value tools that help them (1) quickly find the information they are seeking and (2) manage the quantity and quality of information they manipulate and work with on a regular basis. Although the term "indexing" may lack the cachet of some other terms we use to describe current information organization and management concepts, indexing is fundamental to precise information organization and retrieval, especially when dealing with large sets of documents.

Power users find great value in using a known, granular indexing language that can surface the most relevant items and filter out items of peripheral or no interest. Web architects and interface designers can likewise take advantage of indexing labels to present only the information meeting certain requirements for users who do not wish to learn the indexing structure or taxonomy. The user finds what is needed while the indexing language is used behind the scenes and is transparent to the user.

The importance of indexing in developing a **content navigation strategy for corporate intranets or portals** and the value of high-quality indexing when retrieving information from external resources are reviewed in this white paper. Some **general background information on indexing** and the use of controlled vocabularies (or taxonomies) are included for a historical perspective. **Factiva Intelligent Indexing**—which incorporates the best indexing expertise from both Dow Jones Interactive and Reuters Business Briefing—is described, along with some **novel customer applications** that take advantage of Factiva's indexing to create or improve information products delivered to users. Examples from the Excite and Google web search engines and from Dow Jones Interactive and Reuters Business Briefing are included in an Appendix section to illustrate how indexing influences the amount and quality of information retrieved in a specific search.

The Value of Indexing to Information Users

Why are behind-the-scenes indexing activities so important to users of information services? Now, more than ever, with the increasing quantities of content available to users, and with prospects of these quantities growing exponentially, information users cannot jeopardize their productivity by spending lots of time hunting for the information they require to do their work. Companies are grappling not only with the growing amount of information, but with providing desktop information access to more users for more applications. The business implications are tremendously positive if all users can quickly and without undue frustration find the right information to incorporate into their work processes and decision-making. As corporate web sites and extranets become more critical to e-business expansion, companies face the same challenge of making sure their customers and business partners can easily find information at those sites to close sales and build loyalty.

Information users will be able to find documents that have been accurately and consistently indexed from a controlled vocabulary. This is true whether working with print or electronic information resources and is particularly relevant in today's wired environment. Controlled language access always facilitates efficient information retrieval.

Users can also be confident of exhaustive coverage of a topic and thus, exhaustive retrieval, when the content set is properly indexed—reducing concerns about missing significant pieces of information. In a paper presented at the ASIS 1996 Annual Meeting, Dr. Bella Hass Weinberg (Division of Library and Information Science, St. John's University, Jamaica, NY) observed that approximately 10% of human-assigned index terms do not occur in full text. She also noted that subsequent studies have demonstrated that controlled vocabulary indexing enhances full text retrieval by 10%.

With commercial information retrieval services such as Dow Jones Interactive and Reuters Business Briefing, the user is further empowered to focus the search results on documents in a particular language, from specific publications, published in a defined time frame, or which contain additional concepts or company names. These additional indexing capabilities and system features provide the user with highly-targeted, relevant information to support business decisions.

The Value of Indexing in the Intranet or Portal Architecture

Users of organically-grown intranets frequently express frustration with how much time it takes to find items—both when searching for known items and when browsing to see if items on a particular topic exist in the system. To address this "findability" issue, web architects strive to refine their labeling and indexing systems to help users locate the information they are seeking. Buyers of portal software products are also keenly interested in the search and retrieval capabilities of portal products, with the goal of increasing productivity by giving employees access to information as they need it in the course of their work.

Browsing and search functions are much enhanced if the indexing and topic hierarchy, or taxonomy, make sense to the user and are customized to reflect the content of the source documents. The topic hierarchy must be substantive enough to describe the content of documents from all departments of an enterprise and in many cases, external content feeds as well.

TFPL, an international information management company, quotes Tom Koulopoulos, President and founder of the Delphi Group, as saying that "taxonomies are chic" in its summary of the recent EBIC 2000 Conference. The report also observes: "In order to rationalise information retrieval within unceasing quantities of content, taxonomies have moved to centre stage as one means of providing context and structure to search and exploit the information and data that will help drive e-business forward." (http://www.tfpl.com/ebic2000/EBIC2000fr1.htm).

Building the topic hierarchy is the largest single expenditure of initial corporate portal development, according to a Forrester research study (Building an Intranet Portal, Forrester Research, January 1999). Forrester estimates this cost to be in the neighborhood of half a million dollars for creating a topic hierarchy that spans half a million intranet web pages. Companies that do not yet have a corporate taxonomy, or those who are finding that their taxonomy must become more robust to handle an increasing volume of content, may reap both monetary and time savings in licensing all or parts of an established, tested, broad-based taxonomy.

Current Applications for Indexed Information

Some Factiva customers with sophisticated information and knowledge management systems are creating or updating internal proprietary taxonomies for labeling documents and reports in their electronic information repositories. Others are investigating the economics of creating an indexing structure/taxonomy or investing in automated tools for this purpose.

Factiva has developed Factiva Intelligent Indexing, its proprietary hierarchically organised taxonomy of over 1300 industry, geographic and news subject terms and 300,000 company codes, which is universally applied to the vast range of over 7,000 information sources on Factiva services. For Factiva's customers, the challenge of "findability" in ever-growing bodies of internal information, supplemented with external content, has prompted some of them to partner with Factiva to develop solutions for finding and retrieving information from a variety of information repositories, archives, catalogs, databases, and content feeds.

Because customers are dealing with these controlled vocabulary and indexing challenges in parallel with the development of Factiva Intelligent Indexing, Factiva is providing information management solutions in the following scenarios:

- Customer receives feed from Factiva of all items published in a defined set of journals. Factiva indexing remains on the documents. A form on the customer intranet allows users to define topics of interest within that subset so that updates are personalized for each user. The customer is selecting specific articles from a push product leveraging Factiva indexing.

- Customer has been using its taxonomy to categorize items from various news feeds for a business intelligence product. The customer wants to get out of this business so is working with Factiva to map their taxonomy to the Factiva taxonomy. Factiva will house content sets (including some content not available via Factiva), then filter and deliver this information to the customer intranet. Factiva is constructing additional complex queries based on its taxonomy, where necessary, to match all codes required by the customer.

- Customer wants to co-mingle internal and external content by having its search engine crawl external information as well as internal repositories. The product is being designed so that the customer's intranet search engine sends out two search queries—one to Factiva and one to crawl internal sites. Results can be presented separately or co-mingled. Factiva can offer specialized editorial help to build mappings between the customer's and Factiva's search queries.

- Factiva plans to soon supply its expansive company index to the customer and maintain it so the customer can use standardized company indexing, based on Factiva's success with supplying the Dow Jones Interactive company index to customers. Customer sets up sub-topics for company research. Sub-topics are mapped to news subject indexing. Updates are pushed to the customer's intranet.

Additional opportunities exist to license Factiva Intelligent Indexing for proprietary applications. The indexing structure is flexible and robust enough to be considered the standard taxonomy for business and news information products. There are few information companies who rival Factiva's experience and expertise in handling and indexing a broad range of multilingual, multimedia content from various information providers. Factiva's model of integrating a large variety and quantity of information resources into a single library, searchable with a single powerful indexing language, foreshadows what many customers are trying to accomplish in intranet and portal environments. In addition to supplying external content, Factiva can license tools to help users find the information they need from internal resources as well.

Background and History of Indexing

In order to have a deeper appreciation for indexing being a critical tool for finding information in today's intranets and corporate portals, it is important to understand a bit of the background and history of indexing. Indexes have been used for years to help information users quickly locate pieces of information of interest to them. We are all familiar with one type of index, the one typically placed at the back of a book. A book index refers the reader to subjects covered in the book and is usually an alphabetical list of subjects or names followed by the page numbers where reference is made to the subject or name.

With growing amounts of published information, indexes specific to subject disciplines such as chemistry, engineering, art and architecture, and business, were created to help users locate information from a variety of print resources in those disciplines. The subject indexing process consists of conceptual analysis of a document followed by translating that analysis into a particular vocabulary. Relevant terms from a designated, controlled vocabulary are attached to individual document records and aggregated into a master index to speed up the information retrieval process and help users locate relevant information. The usefulness of the subject index depends on:

- the indexer's ability to analyze the subject matter of the document,
- his/her knowledge of the discipline, and
- knowledge of the needs of users of that particular body of information.

Early print indexes were mostly created to accompany citation lists or abstracted publications. With a limited amount of searchable text, assignment of appropriate index terms was critical for connecting the user with documents on the topic of interest. It is ironic that assignment of appropriate index terms is now just as important to help users find precisely what they are searching for in an increasing volume of documents and an increasing amount of full text information.

As more and more information became available in a machine-readable format and as online bibliographic databases became popular, indexing practices flourished. The use of human intelligence in the analysis and organization of information was complemented with machine-aided indexing programs. Now, as web resources proliferate, computerized indexing using metatags, controlled vocabularies, and subject headings has become more sophisticated and more widely adapted. However, the most accurate retrieval of information usually occurs from those bodies of content in which there is some human involvement in the review of concepts and development of rules, as well as in monitoring accuracy and completeness of the controlled vocabulary.

Background of Controlled Vocabularies or Taxonomies

As noted earlier, the President and founder of the Delphi Group has been quoted as saying that "taxonomies are chic." What exactly is a taxonomy?

A controlled vocabulary is an indexing language, i.e., a standardized set of terms and phrases authorized for use in an indexing system to describe a subject area or information domain. The terms, *controlled vocabulary*, *thesaurus*, and *classification structure* are used interchangeably. The indexing structure can properly be called a taxonomy when the structure is hierarchical. A controlled vocabulary may be as simple as an alphabetic list of terms appropriate for describing the subject area. Thesauri are more frequently carefully constructed sets of terms connected by "broader-than," narrower-than," and "related" or cross-reference links. These links show the relationship between related terms and provide a hierarchical structure or taxonomy that permits searching at various levels of specificity from narrower to broader. Some examples of highly-regarded thesauri and taxonomic structures include:

- INSPEC Thesaurus (IEE Publishing & Information Services)
- Medical Subject Headings (MeSH) (National Library of Medicine)
- Proquest Controlled Vocabulary (http://www.proquest.com/hp/Support/)

Traditionally, indexers manually assigned terms from the controlled vocabulary to a document. With machine-aided-indexing programs, sets of rules were established to, at least partially, automate the indexing process. The number of indexing terms applied and the level of specificity of the index terms depends on editorial guidelines established by the publisher.

The person seeking information can refer to the same thesaurus as is used for assigning index terms and select those terms that are likely to produce relevant results from the database or information system being searched. Even if the searcher does not use the thesaurus in advance, there will be an improvement in precision of documents retrieved when the terms used in the search happen to be index terms or controlled vocabulary terms. The concepts will be retrieved whether or not the words appear in the text. The searcher can also examine the indexing applied to documents retrieved, particularly those that are of highest interest, to extract indexing terms for further, iterative searching.

An important purpose of an indexing language is to control for synonyms. In a Boolean-based retrieval system, searchers use text words or key words OR-ed together to cover the various ways authors describe a concept. However, selecting indexing from a well-developed controlled vocabulary for formulating a search query should accommodate synonyms. A single entry from the controlled vocabulary represents a particular concept no matter how it was referred to in the original article.

The amount of electronically accessible full text information is so immense, and is growing so fast, that users need all the help they can get in accessing it. Using a sophisticated controlled vocabulary to index content can provide tremendous benefits in helping the user with precise, targeted retrieval. A controlled vocabulary system, created and maintained by persons familiar with the subject area and the types of documents covered in the domain of information, is dynamic and will evolve as the domain of information evolves.

Fielded data and Indexed information—Explanation of terms.

In most information retrieval systems, individual documents are divided into explicit segments or fields to assist the user with precise retrieval. The types of document included in the repository of information determine the fields identified. For business and news articles, fields commonly identified include: title, author (or byline), publication name, publication date, company name, ticker symbol, etc. If the documents are company profiles, designated fields might include company name, ticker symbol, SIC (or NAICS) industry classification codes, city, state, zip code, name of president, sales, and number of employees. Information is extracted from these fields and put into specific indexes, such as ticker symbol or zip code indexes, to serve as additional useful access points for the information seeker.

Hierarchical Indexing Benchmark

Medical Subject Headings (MeSH), the hierarchical classification scheme of some 19,000 main headings and codes used for indexing databases produced by the National Library of Medicine, must be cited when looking for "best practices" in indexing. The Medline database is a premier biomedical database and is the electronic counterpart to Index Medicus, International Nursing Index, International Dental Literature. MeSH indexing available within Medline is a key feature of the database.

From 6-15 subject headings are assigned for each article, with up to 3 assigned for major emphasis of the article. Articles are indexed to the most specific term available to allow for very precise subject searching. Subheadings, terms which cover general, frequently discussed aspects of a subject such as *adverse effects* or *therapy*, are combined with MeSH terms to indicate the specific focus of an article.

A particularly powerful feature designed into Medline allows users to "explode" a category of terms in a hierarchy from general to specific to retrieve all of the articles on the general term and all of the specific terms listed underneath. "Explode" is distinct from the concept of truncation in that the terms do not have to begin with the same string of characters to be retrieved. "Exploding" a term allows the information requestor to search a term and all levels of its narrower terms.

The Medical Subject Headings are continually revised and updated by subject specialists responsible for areas of the health sciences in which they have knowledge and expertise. The staff collects new terms as they appear in the scientific literature or in emerging areas of research; define these terms within the context of existing vocabulary; and recommend their addition to MeSH. They also receive suggestions from indexers and other professionals.

This indexing structure has stood the test of time and is widely acclaimed for the accuracy and precision in retrieval that it allows. MeSH should be considered the gold standard and a benchmark for evaluating indexing structures in other disciplines.

Precision & Recall in Searching

The information science literature, particularly articles regarding information retrieval from online services, often refers to the concepts of precision and recall. Precision refers to the amount of relevant items retrieved, while recall refers to the total number of items retrieved. There is generally an inverse correlation between precision and recall, i.e., the higher the total number of items retrieved, the lower the precision or relevance. The caliber of indexing and editorial guidelines about the level of specificity to which indexers classify items also influence precision and recall. Most indexing is done at the greatest level of specificity, although some publishers choose to index at broad concept levels.

Different user audiences have different expectations regarding recall. For example, Public Relations departments may wish to see every single item that mentions their company. It is more common for users to opt for precision in retrieval—a smaller number of highly relevant items.

Optimally, information systems should be designed to allow the user to move easily from high recall requests to high precision requests. Hierarchical indexing schemes facilitate this flexibility in moving from broad to specific retrieval. The user wants to feel confident that the query terms are retrieving all of the articles on a given topic and that the retrieval system is powerful enough to focus on the subset of articles that contain all of the concepts specified.

Factiva Intelligent Indexing

Indexing practices are not new as evident from the long history of indexing print documents and later documents in electronic information systems. Yet, at Factiva there is renewed emphasis on refining indexing techniques as a means of countering information overload and improving precision in retrieval. Factiva Intelligent Indexing now blends the best of both Dow Jones Interactive and Reuters Business Briefing indexing traditions and experience. It also blends the best of state-of-the-art automation for some of the indexing process with human intelligence to analyze and classify articles according to what the article is about.

Indexing at Dow Jones Interactive evolved in several distinct phases—from manual indexing of proprietary content to sophisticated automated, rules-based company and topic indexing—as a result of customer demands. Reuters has a long tradition of indexing all content using a controlled vocabulary and companion alpha-numeric indexing or coding hierarchy. The indexing scheme used for Reuters Business Briefing content is based on the hierarchical indexing used in Finsbury Data Services' Textline database; Finsbury was acquired by Reuters in 1986.

In early 2000, Factiva expanded and modified the Reuters Business Briefing indexing hierarchy to build the new Factiva Intelligent Indexing hierarchy initially comprised of over 650 industry topics, 280 news subject, and 370 geographic labels. The industry classification structure is loosely aligned with the North American Industry Classification System (NAICS), and Factiva industry codes can be mapped to NAICS codes as well as other well-known standards. Five levels in the industry coding hierarchy allow users to search at broad or very granular levels. For example:

I3	Engineering and Metals Manufacturing
I35	Motor Vehicles and Parts
I351	Motor Vehicle Manufacturing
I35102	Commercial Vehicles
I3510201	Light Truck Manufacturing

Factiva editors are sensitive to the fact that some traditional industry hierarchical coding schemes do not cover emerging industries and technologies particularly well. Factiva's Intelligent Indexing structure for industries can easily be expanded to incorporate these new entities.

Codes for nations, and sub-nations are based on ISO classifications. News subject codes are arranged in a 4-level hierarchy and are classified in the following groups: corporate, economic, market, general and political, international political-economic organizations, and content types.

Approximately 300,000 companies are identified by Factiva company codes, forming a comprehensive, global index of public companies and privately-held companies. Unquoted subsidiary companies are assigned a unique code and are included in the company index if there is demand from customers. Searching with the company symbol for Computer Peripherals, Inc. vs. a free text search for the same company is an example which immediately and vividly illustrates the value of using the company index for precision in retrieval.

Dow Jones Interactive indexing codes are mapped to Factiva Intelligent Indexing codes as are Reuters Business Briefing indexing codes. This new, enhanced coding hierarchy will be used to index all Factiva products and it is anticipated that there will be on-screen displays of the hierarchies so users can drill down from top-level searches to more detailed queries. Experienced searchers will be able to integrate the Factiva codes and indexing terms into their search queries.

In summary, Factiva Intelligent Indexing is the common lookup language for:

- A comprehensive, global company directory
- Content published in over 20 languages
- Content from a variety of information providers
- Consistently-described Company, Industry, Geographic, and News Subject topics.

As noted in the earlier section on "Current Applications for Indexed Information," Factiva customers are already leveraging the sophisticated Factiva indexing structure to build and personalize or customize information solutions for users. In addition, they benefit from enhanced, precise retrieval from the Dow Jones Interactive and Reuters Business Briefing services.

An Appendix to this white paper includes sample searches using two popular Internet search engines and the same search performed in Dow Jones Interactive and Reuters Business Briefing. The searches in Dow Jones Interactive and Reuters Business Briefing use both free text terms and index terms to highlight the differences in retrieval and demonstrate the improvement in results when using index terms.

Recommended Resources

Aitchison, Jean, Gilchrist, Alan. Thesaurus Construction: A Practical Manual (ASLIB, 1972; distributed by Chicorel Library Publishing Group, NY)

Lancaster, F.W., Vocabulary Control for Information Retrieval. Second Edition. (Information Resources Press, Arlington, Virginia, 1986).

Rosenfeld, Louis and Morville, Peter, Information Architecture for the World Wide Web. (O'Reilly & Assoc., Sebastopol, CA, 1998).

Weinberg, Bella Hass, "Complexity In Indexing Systems – Abandonment And Failure: Implications For Organizing The Internet" (ASIS, 1996 Annual Conference Proceedings).

The Value of Indexing

Appendix to White Paper

Sample Searches Using Web Resources and Professional Information Resources

To illustrate the value of in-depth indexing when trying to quickly pinpoint articles on a specific topic in large collections of information, searches were performed on the Internet using Excite® and Google® search engines, and then on Reuters Business Briefing and Dow Jones Interactive. On Reuters Business Briefing and Dow Jones Interactive, free text terms were used in the first examples and then compared with search results using indexing terms appropriate to each service.

The Search Topic:

To help measure both the quantity and quality of items retrieved, searches were conducted on the following topic: "news about mergers and acquisitions in the food industry"

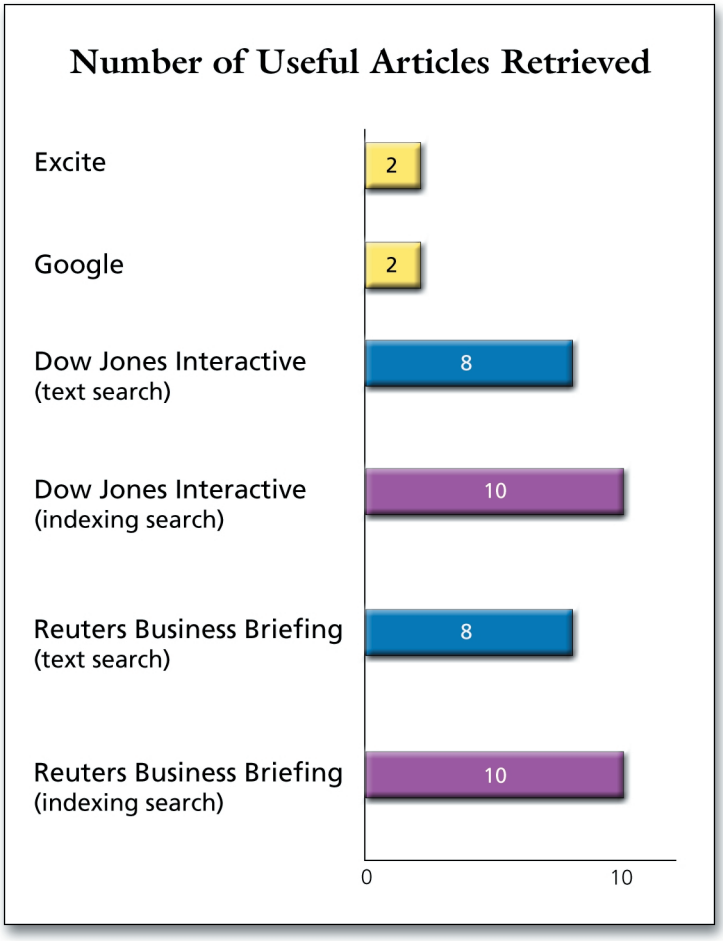
The Services Used:

The search was run in Excite and Google, two popular search engines, and in Dow Jones Interactive and Reuters Business Briefing ***Search***, two professional business and news information retrieval services.

Results from Excite:

Excite (www.excite.com), searches its crawler-built database for documents containing the exact words typed into the search box. It also finds related concepts, having learned the related concepts from analysis of the documents retrieved and analysis of links to related documents.

The top 10 articles retrieved in a search on the mergers or acquisitions in the food industry were examined. While all articles had some mention of mergers or acquisitions, there were only two in the top 10 dealing with news about mergers and acquisitions in the food industry.



Number of first ten search results actually relevant to food industry mergers and acquisitions.

The other provided news about mergers in the food industry, but the page and the entries were not dated. One item of peripheral interest was a directory of consultants specializing in the food industry; a couple of consultants specialized in M&A work for the food industry. Two items were broken links.

Results from Google:

A significant number of items (more than 1300) were found using the Google search engine (www.google.com) on the same topic: mergers and acquisitions in the food industry. Google employs their proprietary technology of ranking importance of web sites by frequency of links and returns pages to which others have linked most frequently. Of the first 10 items retrieved, three were links to tables of contents, five were articles that did not address both parts of our topic, and two were directory listings—both of which were of peripheral interest in that both were lists of consultants specializing in the food industry.

There is no doubt that searching the Web often lead the user to very interesting documents and interesting resources. However, it must also be acknowledged that searching the Web usually delivers mixed results at best. In general, the user must spend a significant amount of time culling through the items retrieved to find articles or documents that address the specific search topic, and then must determine the timeliness and authority of the material retrieved.

Results from Dow Jones Interactive:

A free text search was performed on Dow Jones Interactive for information on mergers and acquisitions in the food industry, published in the current month and last month. In free text searches, the system found exact character matches (accommodating truncation or variable word endings) in titles and the text of articles.

The titles retrieved appear to match our topic. When the top 10 items were examined, it was determined that two articles were not about mergers and acquisitions in the food industry, although both sets of terms were included. The two articles both referred to acquisitions by a company with customers in the food industry. With free text searching the user is dependent on authors using the same words as are used in the search query since there is no context or concept analysis.

The search was performed on Dow Jones Interactive, this time using Factiva's Intelligent indexing terms. It is noteworthy that retrieval increased by about 1,000 items. In this search we did not limit the search to exact word matches; rather we benefit from the indexing applied as the documents are analyzed and go beyond finding specific words we identified to finding concepts which include our terms, synonymous terms, and related terms.

The top 10 items all are about mergers and acquisitions in the food industry. The search retrieves more documents than the user may wish to examine. However, the number of documents can easily be reduced and the search further focused by adding date restrictions or by specifying companies and additional concepts.

Results from Reuters Business Briefing:

A similar pattern emerged when we performed our search using Reuters Business Briefing. A smaller number of articles are retrieved with the free text search than when we use indexing terms.

We also found random articles that have both sets of terms, but not in the context of mergers or acquisitions in the food industry. Using the indexing terms assures the most comprehensive, on-target retrieval. Relevance has a subjective aspect, depending on the framework from which the user approaches a topic. As noted earlier, additional delimiters can be added to reduce the quantity of items retrieved and further focus the search results to be more relevant to the user.

Asia/Pacific
+612.9373.1500

Japan
+81.3.3432.4287

Europe/Middle East/Africa
+44(0)20.7542.3344

The Americas
+1.800.369.7466
+1.609.452.1511

E-mail
solutions@factiva.com