Lotus

White Paper

Building Enterprise Taxonomies with the Lotus Discovery Server

April 2001

A Lotus Development Corporation White Paper

Excerpted with permission from the forthcoming book on the Lotus Knowledge Discovery System by Wendi Pohs et al of Iris Associates © 2001 International Business Machines,
ISBN: 1-931182-03-5, to be published by IBM Press, 5650 El Camino Real, Suite 225, Carlsbad, CA 92008, 800-477-5665, 760-931-8615, www.ibmpress.net, custsvc@ibmpress.net.

Contents

An Overview of Taxonomies and Their Benefits

A taxonomy is defined as a set of ordered groups or categories. Taxonomies were initially employed in biology and other sciences to classify plants, animals, and other organisms into consistent groupings for study. Now enterprise taxonomies — comprehensive categorizations of information and expertise — are proving extremely beneficial to a wide range of organizations. This white paper explores the concept of enterprise taxonomies and offers advice for establishing a taxonomy within your organization.

Taxonomies at Work on the Web

The Yahoo Web taxonomy is probably the best current example of a working Web-based taxonomy. The Yahoo Web site considers itself an "online navigational guide" to resources available on the Web. Yahoo editors use this guide to analyze existing Web sites, and then add them to the Yahoo directory to help users find them easily — and also to circumvent using keyword searches. These editors create categories and give them general, descriptive names. Users browse the categories to find information without having to have prior knowledge of where or how the information was produced.

Taxonomies are a Powerful Tool for Businesses

Taxonomies can be used the same way within the business setting. They define common business language and can serve as a navigational aid to finding business information. They can unify legacy and new business systems by providing a single way to access both.

Before you create or organize a taxonomy, think about the problems you want your taxonomy to solve. Do you want to combine the data in legacy applications with new applications? Eliminate duplication? Facilitate reuse? Perhaps you want your engineers to start thinking about their tasks in new ways. In any case, a taxonomy enables you to establish an enterprise-wide directory that your employees can use to access all the information created by different functional groups in your organization.

Typically, taxonomies are used for browsing. Categories are created so that end users can find the information they need without typing in a search term. Category terms range from general at the higher levels to more specific at the lower levels. Users drill down through each level until they find a category that describes information they need. Because similar categories are grouped together, users can then look for related documents, or they may even discover additional information they were not aware of before.

For example, suppose you recently attended a presentation about employee benefits and want to see the presentation again. You know that your HR department has worked with your IT department to create a taxonomy of corporate HR information and you know you can browse this taxonomy at your corporate Web site. When you find the information you need, you might also find additional items of interest in the same category as the presentation, such as:

- Documents about individual benefits referred to in the presentation
- A profile of the person who gave the presentation, including background and contact information
- An internal Web site dedicated to employee benefits

You could then use the taxonomy to navigate to the presentation as well as to any of these related items. When used as a navigational menu, taxonomies provide effective reuse of corporate information.

The Benefits of Taxonomies

But what's the real business value proposition of taxonomies for management? A quick survey of the knowledge management literature uncovers the following measurable benefits:

- Effective information reuse
- Better organization of information
- Improved quality of information
- Economies of scale
- Full-text search enhancement
- Information discovery

Effective Information Reuse

Taxonomies help organizations avoid "reinventing the wheel" by providing all employees with access to useful, shared information such as proposals, marketing plans, and budgets. In this way, organizations can leverage the full power of information they already have — and avoid needless, costly duplication of effort. Taxonomies help organizations ensure that the work of business units or divisions is shared efficiently across the entire organization, enabling everyone to benefit from key knowledge, experience, and hard work.

Better Organization of Information

Taxonomy creation also lends itself naturally to efficient information organization. Typically, a cross-section of corporate groups will be involved in creating an initial taxonomy. Each of these groups submits data repositories to the administrator charged with creating the taxonomy. These repositories should represent the kind of data the group works with. For example, market research groups might contribute both textual and survey data, while finance groups might contribute spreadsheets.

Each functional group tends to take a second look at their data as part of this process, and often turns up redundant or outdated sources. These sources then become good candidates for archiving, freeing up valuable computer disk space and maintenance resources.

Improved Quality

Information quality also improves when the information is more easily accessed through a taxonomy. Functional groups are often reluctant to share information outside their functions. But these same functional groups feel more responsible when they know their data will be more generally accessible throughout the organization. They want other corporate functions to find their information and, by extension, their work, useful. Often, these groups review their existing authoring procedures before they create a taxonomy. For example, they might establish new editorial guidelines to improve overall quality, or they might enforce consistent standards for metadata.

Economies of Scale

Enterprise taxonomies provide an opportunity for all the different functional groups within an organization to think about describing data in the same way. Once a taxonomy is in place, groups do not have to spend time creating their own classification schemes and then trying to fit new information into an existing scheme.

If an organization uses automatic categorization, many documents can be added to the taxonomy quickly. A human indexer can typically index 30 to 50 documents per day, given an existing term list. The same indexer can index thousands of documents per day if the documents have been classified automatically. The indexer confirms what the software has done automatically, without having to read and assign index terms to each document.

Full-text Search Enhancement

Taxonomies complement full-text search systems. Research in online help systems has shown that users use full-text search if they know what they are looking for. They browse categories if they do not. They might also want to browse related categories around the subject they retrieved using a full-text search. For example, if a user is searching for a job on an Internet Web site like monster.com or dice.com, the user might start with a full-text search using words that describe the user's area of expertise, such as programming. If the user navigates a taxonomy, not only would he retrieve programming jobs, but he might also see development, engineering, or information technology jobs.

Information Discovery

One of the basic knowledge management tenets is that knowledge is in the eye of the beholder. Meaning cannot be derived only from the actual words in text; it requires human cognition. Taxonomies provide opportunities for workers to look at corporate data in new ways. Both managers and employees can see relationships and make connections that may not have been obvious to either before.

On the whole, taxonomies have strong value for management. By providing operational efficiencies in information organization and improvements in information quality, taxonomies enhance information flow. The result? Organizations that tap the powerful capabilities of taxonomies can become more productive, more innovative, and better able to leverage their knowledge and information.

Building a Taxonomy with the Lotus Discovery Server

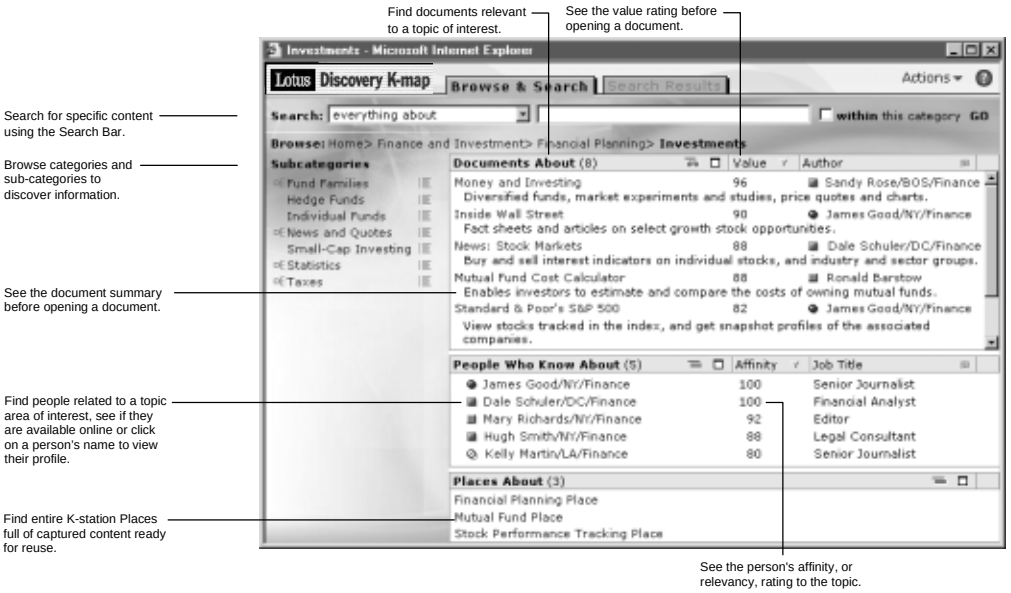
The Lotus® Discovery Server™ is a key component of the Lotus Knowledge Discovery System — which is comprised of two product offerings: the Discovery Server and Lotus K-station™, a knowledge portal. Discovery Server plays an important role in the overall Lotus knowledge management strategy, which provides collaborative e-business solutions that bridge people and knowledge to optimize an organization's business transactions.

This section takes a closer look at the capabilities of the Discovery Server and the technology that enables it to create powerful taxonomies for organizations.

Introducing the Lotus Discovery Server

The Discovery Server locates and organizes the relevant content and expertise required to address specific business tasks and projects. It analyzes the relationships among content, people, topics, and activity, and creates a Knowledge Map — called the K-map — of information that can be accessed, shared, and exchanged throughout the organization.

The Discovery Server uses an innovative affinities mining tool to determine relationships between categories in the taxonomy and user activities. Affinities are relationships between people and categories in the taxonomy. After the server creates the taxonomy, it determines the relationships by examining user actions on the documents in the categories. For example, if a person authors many documents in the same category, the system might propose an affinity between that person and that category. Reading documents, authoring documents, responding to documents, editing documents, and creating links to documents are the actions that the system continuously analyzes to create these relationships.



Users search or browse for the right information and expertise from the K-map user interface.

Fields and Tagging

Manually created taxonomies have traditionally relied on fields and descriptive tags (or metadata) to describe documents. An indexer or an author would read a document to find out what it was about, and would then describe a set of fields to reflect the way users were likely to look for the information. For example, a market research article might be described this way:

<Category>Marketing

<Document Type>Competitive Analysis

<Subject>Knowledge Management

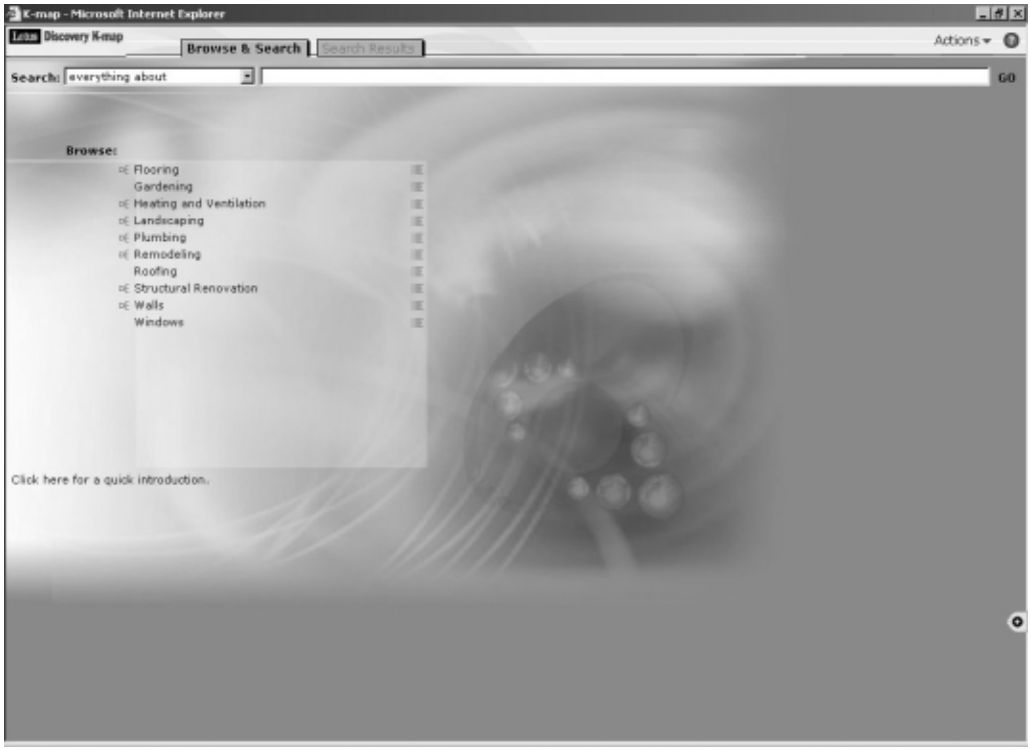
But the article might include information that also would be valuable to software developers. There is a chance they would never see the document if it was considered a "Marketing" document. Retrieval is limited to the information contained in the fields, even if the fields contain multiple values.

Automatic categorization systems use more information than human indexers when clustering documents into categories. Because of this approach, the categories the automatic systems create are less representative of an a priori mindset and more representative of the words in the texts. The most successful systems use a combination of both automatic and human categorization.

The K-map Builder Builds Taxonomies

The K-map Builder is the tool that the Discovery Server uses to automatically create and maintain its taxonomy, which is the K-map. The K-map Builder operates in two modes — clustering (used only to create the K-map) and categorization (used to add new documents and subcategories to existing categories).

The K-map Builder uses vector analysis of the words in documents to create groups of similar documents, called clusters. It uses a combination of EM (Expectation-Maximization) and K-means clustering techniques to build the initial clusters. These statistical algorithms are good at teasing out general themes in collections of documents, but the taxonomies they create almost always require manual reorganization.



This K-map displays the top level categories of an enterprise taxonomy.

How the K-map Builder Works

Based on the SABIO software developed at IBM Research Center at Almaden, the K-map Builder treats words within documents as points in a large, multi-dimensional space. Each dimension corresponds to a single word and the number of times it appears. When two documents share many of the same words and phrases, they will be relatively close together in this space, and will appear in the same document cluster. If desired, you can specify that the K-map Builder give existing keywords and words in title fields extra weight when it creates the clusters.

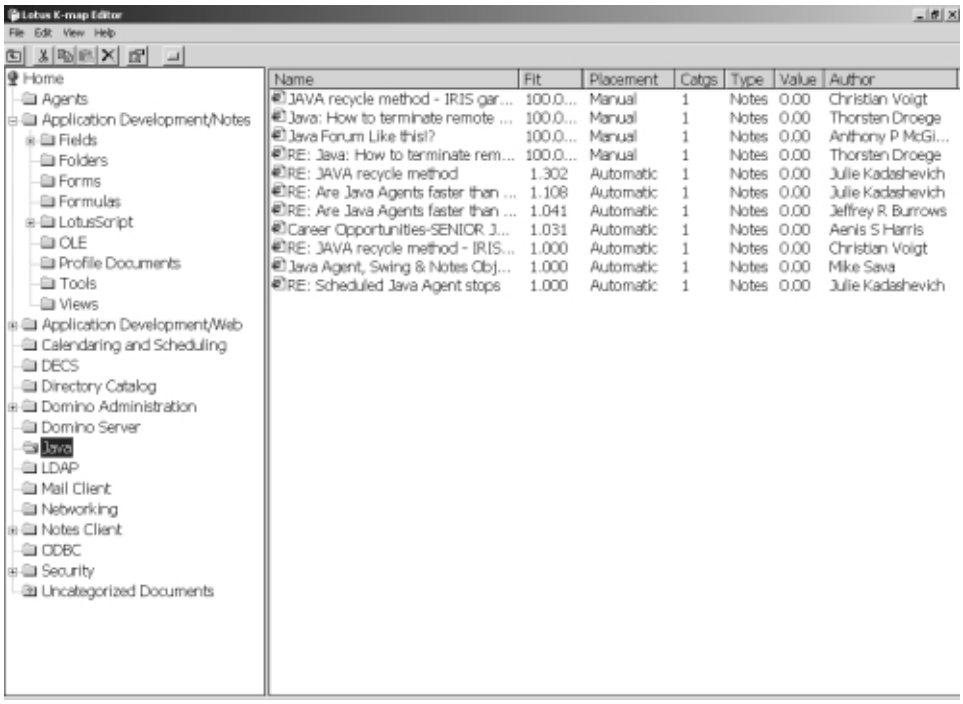
The K-map Builder makes two passes through the data — one to create the clusters, and another to create labels for the clusters. The K-map Builder chooses the top three most-frequently used nouns within the data to label the clusters. These nouns are chosen from a list of possible candidates, so documents in the clusters may not contain any of the words in the cluster labels. Clustering happens only once, when the K-map Builder creates a first draft taxonomy.

A Smarter Approach to Taxonomy Creation

Rather than create categories in advance, or import an existing taxonomy, you select a collection of representative documents for the K-map Builder to use as a training set. The K-map Builder will use the words in the documents in this training set to create the initial set of clusters and labels.

Labels are based only on the words in documents. When the K-map Builder creates an initial taxonomy, no human judgment is applied. The out-of-the-box K-map is the K-map Builder's view of your data after its clustering pass. After an initial taxonomy has been built, editors relabel, reorganize, and refine the categories. When the editors are satisfied with the organization and labels, automatic categorization begins. The K-map Builder learns from the editors' actions from this point on.

Since the K-map Builder creates the K-map automatically, it will not reflect an existing taxonomy. With a carefully chosen training set, it does create categories that can be merged and re-labeled to reflect an existing taxonomy. But it also uncovers new categories the human editor may not have considered when the existing taxonomy was created.



A view of the enterprise taxonomy using the K-map Editor.

Adding the Human Touch

No automatic process can predict precisely how an organization wants to structure its content. An automatic system can only build a taxonomy based on the content. The Discovery Server includes the K-map Editor software, which human editors use to modify and train a taxonomy, tuning its structure to meet the needs of their specific organization. Editors also use the K-map user interface to test their work.

Creating a good K-map is an iterative process. Editors move documents from one category to another, and often re-label categories to reflect use of business terminology used within a specific organization. The K-map is dynamic and learns by example. Editors do not have to create and refine rules, but they do have to analyze the content of the categories to ensure that they are meaningful for their organizations. Often, they'll

perform an information audit to identify the appropriate content and content experts. Editors may identify categories the K-map Builder did not and then populate these categories with representative documents. The K-map Builder does the work after that.

Turning Content into Categories

Once the initial set of clusters has been created and edited, the K-map Builder compares the words in new documents to the words in the existing categories, using the SVM (Support Vector Machine) classifier. If the new documents are similar to the documents already in existing categories, the new documents will appear in the same categories. If the new documents are not similar, the K-map Builder puts them in an "Uncategorized Documents" category for evaluation by a human editor later. The K-map Builder also creates new subcategories when the maximum number of documents per category level is exceeded.

Leading the Way to Effective Knowledge Management

Taxonomies are very effective in providing access to corporate information. They help eliminate redundancy across corporate functions and they improve overall information quality. A well-organized taxonomy reflects the needs of users searching for information and provides the much-needed context that helps them find relevant answers.

The Discovery Server has the capacity to evaluate high volumes of source text, newsfeeds, attachments, and Web source data — automatically and accurately classifying information, updates, and new content over time. The Discovery Server's categorization components — the K-map Builder, the K-map Editor, and the K-map user interface — are effective tools for building and maintaining enterprise taxonomies. Knowledge managers and editors use these tools to build large taxonomies from disparate repositories — establishing a new way of categorizing organizational knowledge for use by all.

For More Information

For more information on the Lotus Discovery Server, contact your Lotus Sales Representative or visit the following URLs:

Knowledge Management	www.lotus.com/km
Lotus Discovery Server	www.lotus.com/discoveryserver
Lotus K-station	www.lotus.com/kstation
The Lotus Notes Network and Iris Today webzine	www.notes.net
Lotus Development Corp. general information	www.lotus.com

The following resources are available for download at www.lotus.com/discoveryserver:

- The Lotus Discovery Server: Taking Advantage of the Collective Experience in Your Organization (white paper)
- Locating Organizational Expertise with the Lotus Discovery Server (white paper)
- Lotus Discovery Server product spec sheet
- Knowledge Discovery in the Information Age (business brochure)

The following resources are available for download at www.lotus.com/kstation:

- Lotus K-station Portal Overview (white paper)
- Lotus K-station Reviewers Guide
- Lotus K-station product spec sheet
- Lotus K-station Free 90 Day Trial Download

Lotus

An IBM Company

SUPERHUMANSOFTWARE™

www.lotus.com

Part No. CC7VWNA