Search@

# Taxonomy in Knowledge Management and Search

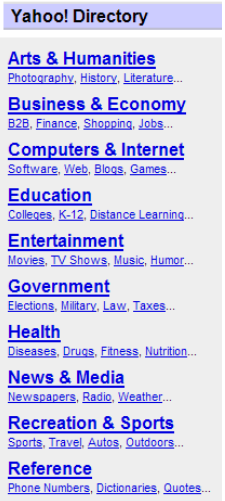
*Finding information is the core of our business …it should be the core of yours too.*

## Taxonomy is a Cornerstone of Knowledge Management (KM)

Knowledge management (KM) deals with the capture and exploitation of knowledge to enhance an enterprise's business. Knowledge can be classified as explicit and tacit knowledge. Explicit knowledge is knowledge that has been captured and represented in a format understandable by users and applications. Common examples are the management structure, products and services, and business processes and tasks in an enterprise. Explicit knowledge can be represented in hierarchical structures, known as taxonomy.

Tacit knowledge comprises experiences known only to individuals. Tacit knowledge is difficult to capture, since asking knowledge workers to codify their knowledge in a structured way is practically impossible. In the big data era, we can exploit the fact that knowledge workers communicate their knowledge with others through writings, e.g., memos, messages, emails and documents. Text mining enables a KM system to extract knowledge, such as names, locations, events and topics, from written texts and organize them in the taxonomy.  By keeping track of the mined knowledge over time, important trends can be observed.

## What is Taxonomy?



Yahoo! Directory

Arts & Humanities
Photography, History, Literature...

Business & Economy
B2B, Finance, Shopping, Jobs...

Computers & Internet
Software, Web, Blogs, Games...

Education
Colleges, K-12, Distance Learning...

Entertainment
Movies, TV Shows, Music, Humor...

Government
Elections, Military, Law, Taxes...

Health
Diseases, Drugs, Fitness, Nutrition...

News & Media
Newspapers, Radio, Weather...

Recreation & Sports
Sports, Travel, Autos, Outdoors...

Reference
Phone Numbers, Dictionaries, Quotes...

Taxonomy by itself is a valuable knowledge asset for management, governance, and information access

Taxonomy is a directory of concepts important to an enterprise. Yahoo! Directory and Open Data Project are well known examples. Taxonomy represents explicit knowledge such as management structure, products, services and tasks within the enterprise. By gathering and connecting all important business concepts, taxonomy serves as a global knowledge map upon which business concepts can be defined, described, visualized, shared and communicated consistently. Taxonomy provides a unified framework for guiding the transformation of implicit knowledge into explicit knowledge. Taxonomy is a valuable corporate asset, whether it is used alone (as a knowledge organization or browsing tool) or embedded in other applications (e.g., KM search engine).

## How to Develop Taxonomy?

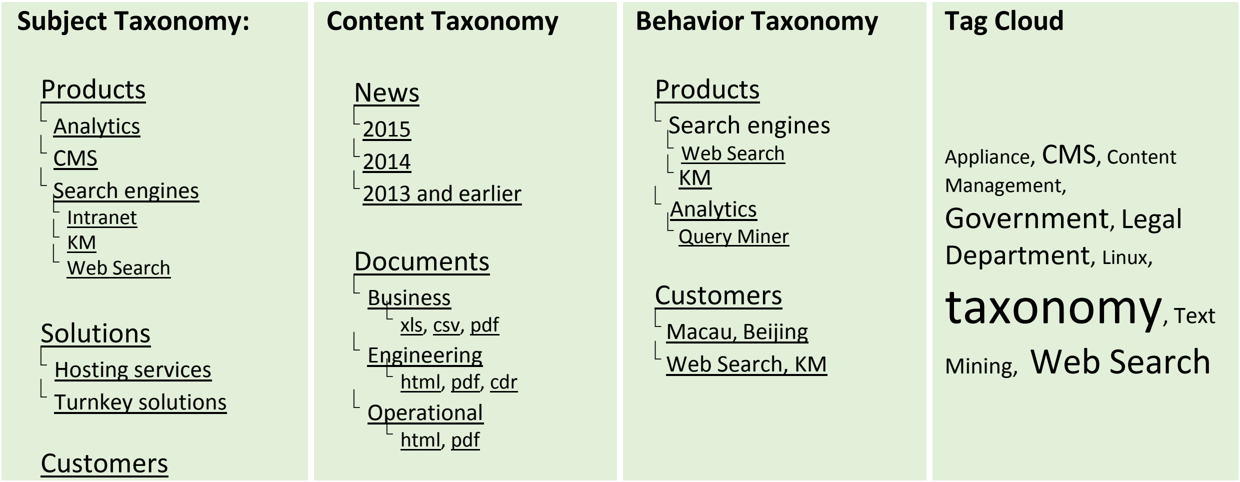
Taxonomy can be constructed with different methods, and these methods are not exclusive to each other; they can be used alone or combined to construct the taxonomy.

Subject-based taxonomy is developed through conceptual analysis. Knowledge engineers identify important business concepts and organize them into subject hierarchies. A business concept is included into the taxonomy as long as it is important to the enterprise. It does not have to associate with any document or data at the beginning. For example, an "Overseas Offices" category can be created without any sub-categories, data or documents, but as the enterprise expands overseas it can be expanded and populated with new data and documents.

Content-based taxonomy is created by extracting important features from business contents and iteratively organizing the contents bottom up. Knowledge engineers study the types of documents (e.g., management, legal, sales, engineering, support documents, etc.), important entities, time periods, etc., to create the taxonomy. An advantage of content-based taxonomy is that some concepts may not be known from conceptual analysis but rather discovered from documents. For example, the types of engineering documents depend on the software packages used and hence may not be known in conceptual analysis.

Behavior-based taxonomy is dynamically generated based on how the contents are used by people. For example, the taxonomy could contain only popular categories and recent topics minded from business contents. The rationale is simple: popular and recent data are more important to the business, so instead of listing dozen of product models in the taxonomy, only the most popular and recent ones are listed.

A tag-cloud is dynamically generated based on usage data. It represents concepts in a flat list, using only color and size to express the popularity of the concepts. It is like a one-level behavior-based taxonomy, and thus cannot represent the relations between concepts. Many websites displays a tag cloud for each page as a quick summary of the page, and users can click on a tag to get a list of pages about that tag.



## Maintaining the Taxonomy

Taxonomy is expensive to create, and is even more expensive to maintain. Fortunately, automated tools can significantly reduce the maintenance cost and enhance taxonomy quality. At one end of the spectrum, a completely automatic method employs automatic clustering methods (e.g., k-means) to group similar documents into clusters and then iteratively group similar clusters into more general clusters to form the taxonomy. While automatic clustering is inexpensive, the result is not perfect. It could generate too many or too few clusters, unnatural cluster names or clusters with themes that do not match business semantics.

At the other extreme, domain experts define the taxonomy through conceptual analysis, and content authors have to specify which category or categories a page belongs to using metatags  For example, documents with metatags category_0=”Documents”, category_1=”engineering” and category_2=”pdf” will be assigned to the category Documents -> Engineering -> PDF. The process is time consuming, and it is difficult to keep the taxonomy in sync with the dynamic businesses of an enterprise.

A semi-automated approach combines automatic clustering and manual fine-tuning iteratively. Automatic clustering is first applied to the dataset to get a rough cut of the clusters. In manual fine tuning, knowledge engineers can use an analytic engine to analyze the clusters, try out alternatives, and apply any of the following operations to improve the taxonomy:

- Rename a cluster to make it more meaningful to users
- Merge two clusters if they are deemed very similar
- Split a cluster if it is too broad
- Create a new cluster and populate it with documents from other categories
- Delete an existing cluster and reassign the affected documents to other categories

The analytic engine can analyze the clusters, obtain statistics of the clusters and simulate the result of an operation to help knowledge engineers decide how to improve the taxonomy:

| *Information from Cluster Analysis* | *Help to …* |
|---|---|
| High frequency terms, discriminative terms | Better naming, splitting and merging of clusters |
| Display exemplary documents in a cluster, which are documents at or near the center of the cluster | Better understanding of the cluster theme for better cluster naming |
| Cohesiveness of a cluster, which is the average distance between the documents in the cluster | Clusters with low cohesiveness may be split into tighter sub-clusters |
| Distinctness of a cluster, which is the distance of the cluster from its neighboring clusters | Clusters with low distinctness can be merged with neighbor clusters |

## What Makes a Good Taxonomy?

As a business operates, new data, new documents and new business concepts spring up. A rule of thumb is to at least evaluate the taxonomy quality yearly and decide if enhancement is needed. If the taxonomy is static for many years and still appears to be valid, the chance is that the taxonomy is too small and too general. It works but the enterprise is likely not enjoying the benefit of a more comprehensive taxonomy.

A good taxonomy should have meaningful category names. Each category should have high coherence and distinctness. Documents in a category should have a common theme that aligns with a business concept in the enterprise. The size and access statistics of a cluster are also important hint of the quality of a category.

## Taxonomy and Search

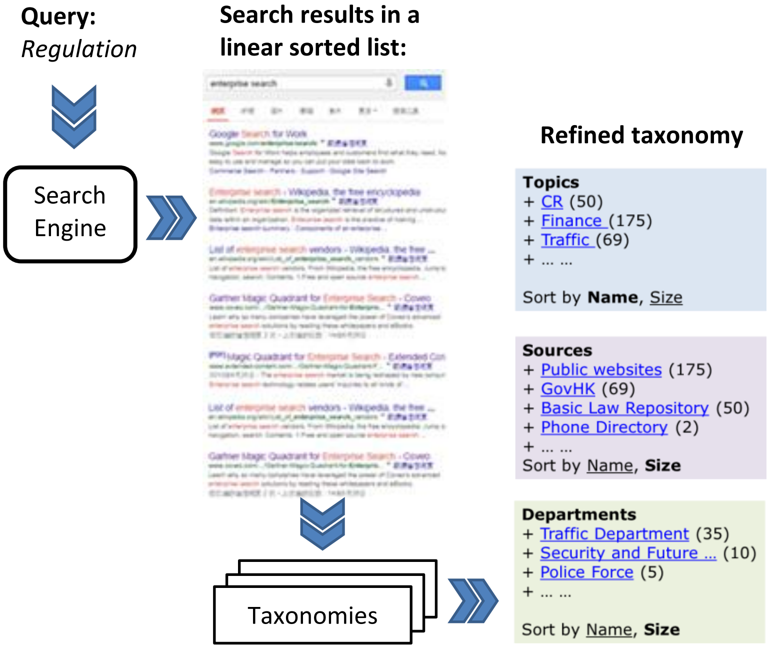
Most search engines (e.g., Google and Bing) have been very successful in returning relevant results to the users. However, they are not good at returning results with diversified and complete coverage. On Google, a search on “regulation” returns 8 results about dictionary definition of the word, and one result about regulation of bank and one about environmental protection. Bing returns similar results. This is not acceptable in an enterprise search engine,. A lawyer defending for a client or investigating a patent application would definitely want to get all of the relevant information covering every aspect of the case. Knowledge makes or breaks a case.

A traditional search engine returns only some top results in a single ordered list. This gives the user a very narrow view of the possible results. To ensure enough coverage, users have to issue many queries to retrieve different aspects of the queried topic. This not only takes time but at the end users only know what they have found, *but not what they have missed!* This is not an ideal situation for investigative and knowledge intensive tasks in an enterprise.

When you search for “regulations”, the search engine will return some popular and highly relevant regulations in your company, but it would not show you all types of regulations pertinent to your query. It will inevitably miss some less popular but useful results. Your company may have many types of regulations beyond your imagination: advertising, employment, environmental, privacy, safety and health regulations, to name just a few. You do not know what you missed.

The integration of taxonomy and search solves this problem. In addition to the regular linearly ordered result list, KM search will find all of the relevant results (not just the top results) and map them into the taxonomy. The taxonomy is refined to hide all nodes that do not contain any result. While the complete taxonomy depicts the knowledge space of the entire enterprise, a refined taxonomy shows the knowledge space around your query topic. You can easily visualize how the results are mapped to the enterprise knowledge and how they are distributed. While categories holding large number of results are important, categories holding a few results may contain the gemstone that you are looking for. Taxonomy transforms an otherwise limited ordered result list into multi-dimensional information space around your query.

Taxonomy is usually displayed in an expandable/shrinkable tree view. However, if it looks too complicated to your users, it can be hidden and utilize taxonomy to support faceted search. For example, a "regulations" category may consist of many sub-categories, such as environmental and financial regulations. Each of the sub-categories can be considered a facet of the "regulation" category. When a search is conducted on "regulations", in addition to showing top results within the regulations category, buttons labelled with sub-category names and number of results can be shown, say, on top of the result list. Users can see what composes a "regulations" category and drill down to each product sub-category to narrow down the result. In a sense, the taxonomy is shown level by level as the user drills down the taxonomy hierarchy.



## What Suntek can do for you?

### *Customization and implementation*

Suntek can analyze your business requirements and implement the KM solution that suits and integrates into your business applications. Implementation includes the development of:

- Taxonomy and ontology
- KM search and browsing tools to support knowledge intensive tasks
- Document Management System to support the complete document life cycle
- Communication tools to facilitate the collection and exchange of tacit knowledge

Suntek can develop case-based, task specific portals providing legal professionals with information adapted to individual or community preferences. The knowledge base can be collaborative expanded, maintained and shared, and along the way improving both the richness and quality of the knowledge. External knowledge sources can be continuously monitored and any important changes will be communicated to the legal professionals.

### *Solution options*

Suntek can develop and deliver KM solutions in several forms:

- Install KM solution on the customer's hardware and data centers
- Provide turnkey solutions including hardware and KM solutions sized to meet your stringent performance requirements
- Build, operate and transfer (BOT) KM solutions to your premises
- Hosting of KM solutions in Suntek's own data center with 7x24 operation, monitoring and support.

**Related reading:** *Suntek Enterprise Knowledge Portal*. White Paper, Suntek Computer Systems.