TaxonomyStrategies
The Business of Organized Information

White paper

Why Semantics Matter

Version 1.1
May 20, 2013

Anyone who has spent time searching for information on the Web or at a Web site knows how frustrating the experience can be. More often than not the search returns zero hits, or thousands of hits that must be further sifted manually. Applying controlled vocabularies in vertical subject areas can transform content roulette into successful search experiences.

5253 Nebraska Avenue NW | Washington | DC | 20015 | USA

phone 415 377 7912

fax 202 450 2737

www.taxonomystrategies.com

info@taxonomystrategies.com

Why Semantics Matter

When you own a *Rembrandt* you can spell his name any way you want. But when you want to find a *Rembrandt* you better spell his name correctly. Vocabulary resources can help find the right artist even if their name is typed incorrectly. Users cannot type in the complex queries needed to find all the relevant items—but this could be done automatically. Complex queries are even more important when you search the entire web. So you find *Rembrandt* "the Dutch guy"—and not *Rembrandt* "the toothpaste".

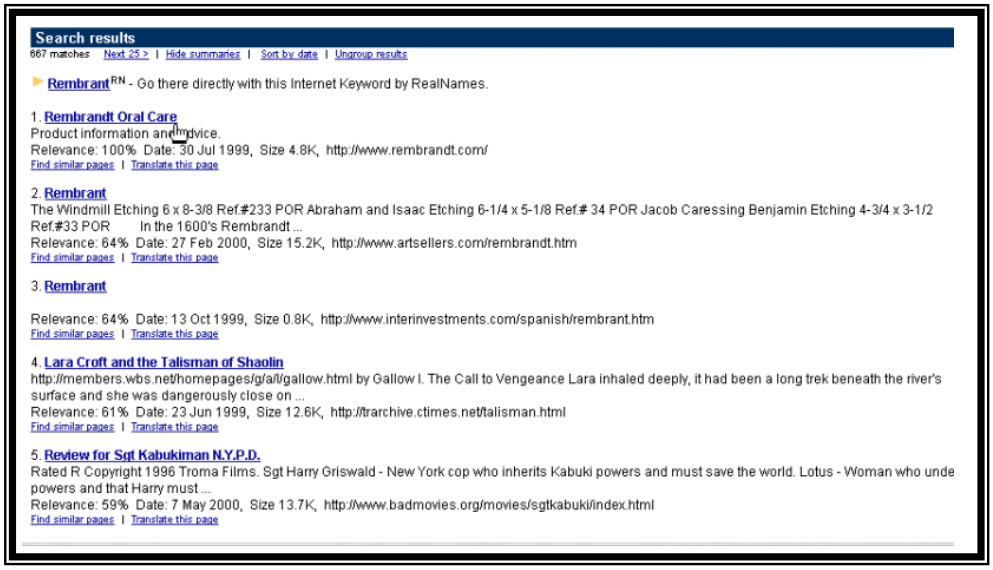

Figure 1 - Results from Search on "Rembrant"

Site Search

The reason that searching especially on Web sites is so poor is because users:

- Make typing mistakes,
- Have to guess by trial and error what terms were used when the content was created, and
- Are confused about what has been indexed for searching.

The simplest cause of search failures is character-based errors in the user's query. People inevitably make typing mistakes and spelling errors. They also have trouble with hyphenation and missing or incorrect punctuation.[1] A significant portion (19%) of search failures in retrieval systems has been attributed to such character-based errors.

The greatest cause of search failure (40%) is mismatches between query terms and the terms in the target content.[2,3] People have trouble choosing the right term when there are synonyms or terms with closely related meanings. They often enter acronyms or abbreviations.

---

[1] Young, C.W., Eastman, C.M., and Oakman, R.L., "An Analysis of Ill-formed Input in Natural Language Queries to Document Retrieval Systems." Information Processing and Management, Vol. 27, No. 6, 615-622, 1991.

[2] Norgard, B., Berger, M., Buckland, M., and Plaunt, C., "The Online Catalog: from Technical Services to Access Service." *Advances in Librarianship*, Vol. 17, 111-148, New York: Academic Press, 1993.

Searches also commonly fail when people do not understand how the content on a particular site has been indexed (20%). Not all search boxes provide full text search. Some sites index only metatags or titles. Some sites only index some of their pages. Some sites offer searches of product catalogs. In this case, people (even experts) have difficulty choosing among product, brand, and manufacturer names.

Most searchers on the Web have limited knowledge. Few users understand the syntax (which varies among search tools) that is required to enter a Boolean query. Few people understand how to effectively use advanced search features. And nobody reads help.

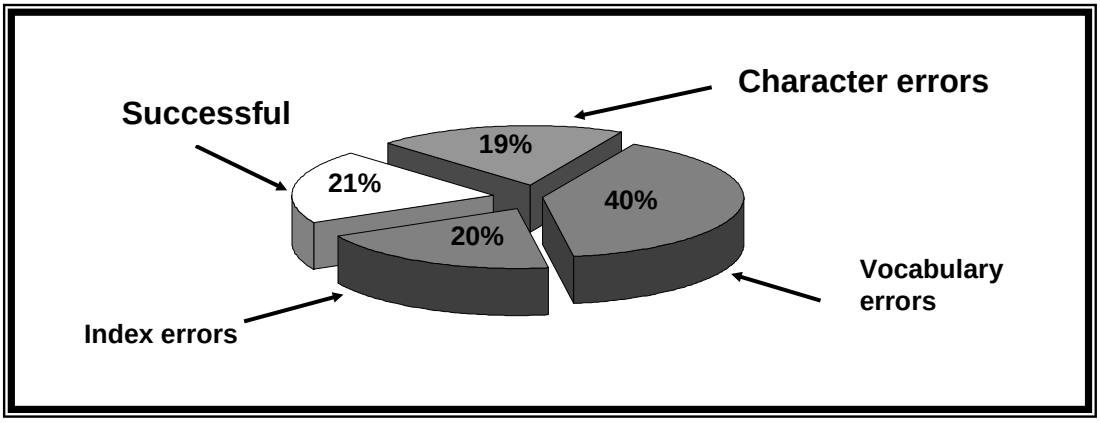

Figure 2 - Causes of Search Failure.

The problem is not the users, nor can it be solved by training the users in better searching techniques. The problem is not with search engine software either. The solution requires:

- Generating more consistent content to search on,
- Correcting user errors, and
- Mapping the language in which users type their queries to the language of the target content.

Search Alternatives

Various approaches are being used to address the unpredictable nature of site search on the Web. These include:

- Personalization,
- Analytics,
- Taxonomies, and
- Syndication.

Personalization applications collect information about users, typically by requiring them to submit information about themselves through a form fill-in registration. While self-registration generates user demographics, the Web site content still needs to be tagged with attributes that map to the user categories.

Analytics are an extension of the personalization concept. Analytics applications such as Google Analytics and KISS Metrics analyze Web site usage and automatically characterize users based on their

[3] Seaman, S., "Online Catalog Failure." *College and Research Libraries*, Vol. 53, 113-120, 1992.

behavior on the site. Analytics applications then suggest content or navigation paths based on previous usage trends, or based on demographics from other applications such as personalization. But users don't necessarily follow predictable and consistent pathways.

Taxonomies attempt to provide an optimized site map or information architecture that allows users to intuitively navigate to content, or directs users to the content the site owner wants them to see. Taxonomy generation tools automatically produce taxonomic maps based on linguistic analysis of content. But automatically generated taxonomies reflect the ambiguities of natural language, rather than an optimized knowledge map. Generating taxonomies is an art that can benefit from the best practices of library science.

Syndicating relevant new and updated content to subscribers, business affiliates, partners, and suppliers rather than have them search for it has become more common on the Web thanks to standards such as schema.org. But automatically delivering the right content to the right people at the right time requires subscriber profiles, well-categorized content, and mechanisms to define and manage rules for appropriately routing content.

Search Alternative	Application Shortcoming
Personalization	Content needs to be tagged with attributes that map to user categories
Analytics	Users don't follow predictable & consistent pathways
Taxonomies	Automatically generated taxonomies reflect ambiguities of natural language
Syndication	Requires subscriber profiles, well-categorized content, & managed rules

Table 1 - Shortcomings of Search Alternatives.

Search alternatives suffer from many of the same shortcomings as search itself. The source content and user profiles these applications have to work on may not have consistent structure or syntax so that machines can understand it. Even when the content conforms to a predictable standardized structure, the language is not controlled by consistent semantics. But the problem is not with the application software. The solution requires:

- Predictable standardized structures, and
- Consistent semantics to work on so machines can understand it.

Semantic Content Management

The vision for the Semantic Web is a rich web of linked information, with markup allowing machines to route relevant information to the audiences that value it most. To accomplish this vision, a lot of metadata will need to be added to content. This metadata will need to be complete and consistent, and the metadata will need to be kept up-to-date. This will need to be done without adding an army of indexers. Semantic content management systems will be needed to compile and maintain schemas and the controlled vocabularies for filling them, and to automatically process content at any time during its lifecycle—from creation through purging—to apply and update rich metadata. When content has been labeled with rich, accurate metadata we can call this content "intelligent" in the sense that it has been prepared to be used by search, personalization, analytics, taxonomy, and syndication applications.

Understanding the need for and value of rich metadata is not new. Consistent, in-depth indexing has been the business of information services for preparing research materials in high value or well-funded subject areas. What is new is the scope and scale of the application of rich metadata to a much wider variety of content objects—intranets, extranets, and even the Web. Clearly, this requires automated tools

that can readily mobilize controlled vocabularies and more complex knowledge representation schemes. The hard work is mining content to:

- Extract key information such as the mentions of people, organizations, places, and things;
- Infer what the content is about; and
- Format content objects with standard labels for effective exploitation.

Semantic content management requires changing the problem focus from applications to content preparation. How to make content more intelligent so that applications that use content work better.

Joseph A Busch is the Founder and Principal Consultant of Taxonomy Strategies. Taxonomy Strategies (taxonomystrategies.com) guides global companies, government agencies, international organizations and not-for-profits such as Nike, Oracle, the U.S. Environmental Protection Agency, the International Monetary Fund and Harvard Business Publishing in developing metadata frameworks and taxonomy strategies. Before founding Taxonomy Strategies, Joseph Busch held management positions at Interwoven, Metacode Technologies, the Getty Information Institute and PriceWaterhouse. He is a Past President of the American Society for Information Science and Technology (www.asis.org), and past member of the Board of Directors of the Dublin Core Metadata Initiative (dublincore.org).

