



Heather Hedden

How SEMANTIC TAGGING Increases Findability

Findability is about making information easier to find. After all, if it cannot be found, it may as well not exist. Leading information specialists have been saying this for years, and now with the increasing volume of content and increasing pressures of time, money, and competition, more of us are finding this statement to be true. In addition to traditional controlled vocabulary-based indexing, information architecture has evolved to make browsing and navigation methods more effective, search engine capabilities have been improving to help us find the proverbial needle in the haystack, and bookmarking and social tagging have emerged to help us find our own content, and that we share with members of a social networking group.

The various methods of enhancing findability each have their limitations. Traditional document indexing/material cataloging and web information architecture do not go deep enough. Indexing is usually at the document level, and cataloging only works on the level of the material as a whole (books, sound recordings, video recordings, etc.). Information architecture aids in the navigation of a website, intranet, or portal, but in itself it is often not





Author	Theater	Production	Company	Resources
Birth date	District	Director	Name	Play
Death date	Location	Theater	Productions	Director
Birth Place	Capacity	Cast	Performers	Theater
Death Place	Style	# of Perfs.	Etc...	Production Co.
Nationality	Etc...	Lighting	(14 fields)	Character
Occupation	(18 fields)	Costumes		Scene
Awards		Etc...		Etc...
(38 fields)		(47 fields)		(45 fields)

Characters	Scenes	Texts
Plays	Where	Keyword
Age	When	Author
Author	Setting	Date Written
Performer	Subject	Date Published
Etc...	(41 fields)	Production
(30 fields)		(67 fields)

Give me scenes about AIDS written by South African authors in the past 5 years....

Alexander Street Press LLC has developed highly structured facets of tags for plays and scenes.

The Alexander Street Press' highly specific tag categories for Early Encounters in North America

sufficient for finding specific information. Search engines match user-entered keywords and phrases to those found within the texts or metatag fields of documents, but these are still just word matches and do not necessarily go after the meaning of a document. For example, many words are quite ambiguous, and search results would not be accurate on words such as “state,” “log,” or “screen”—even in combination with other words. Social tagging only involves files or webpages that the user and colleagues have already viewed or created. More significantly, though, social tagging tends to suffer from inconsistent application of tags, such as using both synonyms (movie, motion picture, film), singular/plural forms, and abbreviations (Corporation/ Corp., information/info).

New techniques and tools are being developed to address the shortcomings of these various approaches to finding information and to deliver better results in an increasingly competitive information industry. “Semantic tagging,” in the various ways that it is understood, is a term that describes many of these new (and some not-so-new) findability approaches. Semantic tagging is by no means an accepted concept with an agreed upon definition. Other than the obvious “tagging for meaning,” semantic tagging means

different things to people coming from different parts of the information management field. It may be used interchangeably with “semantic indexing” in contexts where “indexing” is used for “tagging.” Nevertheless, in the quest for better methods of findability, the term semantic tagging is starting to appear in descriptions of information services and products, blogs, online articles, and presentations.

SEMANTIC TAGGING IN PUBLISHED INDEXES

“Semantic information ... enables publishers to distinguish their content from their competitors,” explains Bill Kasdorf of Apex CoVantage, organizer/moderator of a preconference seminar on semantic tagging at the Society for Scholarly Publishing’s (SSP) annual conference this May in Boston. “In addition, great progress has been made recently in moving semantics beyond the theoretical: Actual publishers are actually doing it, and they’re actually getting real benefits from it.”

Some people would argue that semantic tagging is nothing new. It can be defined as the assigning of selected controlled vocabulary (aka taxonomy) terms, especially by trained indexers, to content items, such as articles, images, or other

documents, to reflect the meaning of the content. Human subject indexing is inherently semantic, because human indexers can discern the meaning of content. This has been done by periodical and other database index publishers for decades. Once the domain of large database publishing companies (H.W. Wilson, ProQuest, Gale, EBSCO, etc.), more affordable client/server and desktop software for taxonomy management, indexing, and web database publishing have enabled publishers of all sizes to engage in this form of semantic indexing. Meanwhile, the growing popularity of social tagging has made users more aware of the value of subject terms that reflect the meaning of a piece of content in comparison-free text word/phrase search.

Nevertheless, there are publishers that consider semantic tagging to be something more than mere controlled vocabulary-based human indexing; they are pursuing new techniques. This was evident in the participation in the SSP Boston conference’s semantic tagging seminar, Say What You Mean: How Semantic Tagging Makes Content More Discoverable, More Useful, and More Valuable.

One way that semantic indexing is distinguished from traditional subject indexing of documents is that it focuses on concepts rather than the documents

Search Results: **necrotizing fasciitis** [View Image Search Results](#)

[necrotizing fasciitis](#)
 - after cesarean delivery
 - classification
 - complications
 - course
 - diagnosis
 - differential diagnosis
 - disposition
 - effects on vulva
 - epidemiology
 - etiology
 - group a streptococcal
 - imaging studies
 - in diabetic patients
 - in gynecologic surgery
 - lab tests
 - **mortality >**
 - of hand
 - pathogenesis

1-3 of 3 Results

Necrotizing Fasciitis
 Schwartz's Principles of Surgery > Chapter 11. Patient Safety, Errors, and Complications in Surgery > Wounds, Drains, and Infection

Emergency Department Treatment and Disposition
 Emergency Medicine Atlas > Chapter 12. Extremity Conditions > Necrotizing Fasciitis

The overall mortality rate is usually reported in the 25 to 50 percent range. Bacteremia is...
 Tintinalli's Emergency Medicine > Chapter 152. Soft Tissue Infections > Necrotizing Soft Tissue Infections > Necrotizing Fasciitis (Polymicrobial Infection) > Pathophysiology

Silverchair search results, indexed to the chapter subsection level and utilizing a structured taxonomy

According to Zarnegar, “Tagging should be done at the smallest ‘atomic’ level that can stand on its own if taken out.” Whether the original source is a book, article, or pamphlet, subject indexing is often done to the paragraph level.

SEMANTIC TAGGING IN SEARCH

Turning to the area of automated search and retrieval, enterprise search engines, content management systems, and related discovery and data mining products that do not utilize human indexing, semantic tagging obviously plays a smaller role. Nevertheless, some of these vendors claim to offer semantic capabilities. In the competitive enterprise search space, new technologies are often based on either autocategorization (automatic indexing/tagging) or various text analytics techniques, such as pattern recognition or entity extraction. Most of text analytics is not semantic because it does not discern the meaning of words, but rather may classify words by part of speech (grammar). Various forms of autocategorization, on the other hand, may or may not have a degree of semantic technology involved.

as a whole. Panel presenter Stephen Rhind-Tutt, president of Alexander Street Press, LLC, explained that semantic indexing can answer complex questions of who, what, and when, such as “What battles during the Civil War resulted in more than 1,000 deaths?” Regular indexing merely answers the question “What documents discuss this battle?”

Specialized and multilevel facets (or metadata, depending on your perspective) of controlled vocabularies can be implemented to support semantically complex user queries, as done by humanities publisher Alexander Street Press. Its database of theatrical plays is indexed by the top-level facets, including playwright data, theater data, specific production data, theater company information, character characteristics, scene data, and play text data. Its Early Encounters in North America history database has nine controlled vocabularies, including author, source, year, place environment, flora, fauna, encounter, people, personal event, and cultural event. Setting up the controlled vocabulary and facets requires one to “go into the data and ask ‘what are the latent semantic issues that will be asked’ ... This needs to be discipline specific,” according to Rhind-Tutt. Finally, the content searched with faceted taxonomies and supporting interfaces needs to be sufficiently structured with metadata, tagging, or indexing

that precisely captures each subject in its appropriate facet.

Another way that semantic indexing is distinguished from traditional subject indexing of documents is that it focuses on pieces of content at a finer, granular level rather than the documents as a whole. This is an approach taken by medical research database developer Silverchair, as explained by its CTO Jake Zarnegar: “We apply semantic tags at any change of topic or concept in the data at any level—including articles, sections, paragraphs, tables, figures, equations, sidebars, videos, etc. Many taxonomic tagging systems deal with the entire data entity as one unit.” Using its internally developed TOTEM taxonomy management platform, Silverchair inserts taxonomy tags into the XML content.

Collexis Holdings' Research Profiles database with weighted subjects indicated in bar graphs

MAYO CLINIC

Research

[Home](#)
[Dept/Centers](#)
[Labs](#)
[Faculty](#)
[Publications](#)
[Postdocs](#)
[Grants](#)
[Search](#)

Research Profiles

Institutions

Loegering DA
 Rochester: A

Research Profile

Publications

Diseases & Pathologic Processes
 Ataxia Telangiectasia
 Leukemia, Myeloid, Chronic
 Leukemia, Myelocytic, Acute
 Leukemia
 Leukemia, T-Cell, Acute
 Leukemia, Lymphocytic, Acute
 Leukemia, Myeloid, Chronic-Phase
 Neoplasms
 Cancer
 Carcinoma
 Toxicity

Chemicals & Drugs
 Poisons
 Camptothecin
 Tyrosinase
 Etoposide
 Reactive Oxygen Species

Organisms
 Chickens
 Mice

Physiology
 Apoptosis
 Cell Survival
 Signal Pathways
 Cell Death
 Down-Regulation

Behaviors
 Role

Concepts & Ideas
 Sensitivity
 Dose-Response Relationship, Drug
 In Vitro

Companies Featured in This Article

Alexander Street Press, LLC www.alexanderstreet.com	Silverchair www.silverchair.com
Apex CoVantage www.apexcovantage.com	Teragram Corp. www.teragram.com
Collexis Holdings, Inc. www.collexis.com	TextWise www.textwise.com
Interwoven, Inc. www.interwoven.com	Thomson Reuters' Calais service www.opencalais.com
Northern Light Group, LLC www.northernlight.com	Zigtag, Inc. www.zigtag.com
Relevad www.relevad.com	

In cases where autocategorization search solutions or content management software come prepackaged with taxonomies or have a feature to build or automatically generate taxonomies (which only some vendors offer), there is a potential for what may be called semantic tagging. A simple taxonomy as used in information architecture with a hierarchy of category terms is not sufficient for effective autocategorization. What is needed is really more of a “thesaurus” style of taxonomy, whereby there is a cluster of synonyms or other equivalent terms (abbreviations, acronyms, spelling variations, grammatical variations, etc.) for each concept in the taxonomy. Thus, the taxonomy is comprised not merely of words, but of concepts which derive meaning (“semantics”) from their cluster of synonyms. Autocategorization products that provide integrated taxonomies include Interwoven, Inc.’s MetaTagger; Teragram Corp.’s Categorizer and Taxonomy Manager; and Northern Light Group, LLC’s Enterprise Search Engine, MI Analyst, and Analyst Direct. Northern Light supports what it calls “meaning extraction.”

Knowledge discovery vendor Collexis Holdings, Inc. makes use of taxonomies in what it calls semantic tagging by using weighted taxonomy terms. In the Collexis Knowledge Dashboard product, based on

statistical approaches including frequency, uniqueness, and data field location (such as title or text body), terms’ relative weights are displayed with bar graphs. According to Collexis COO Steve Leicht, who also presented on the SSP panel, semantic tagging “can include taxonomic tagging, ontology-based tags, topic maps, other controlled vocabularies, mixed statistical approaches, etc.”

While much of text analytics does not involve semantic analysis, the specialty of natural language processing (NLP) is often involved in such attempts. NLP has many other applications beyond semantic analysis and tagging, but it is being applied in that area as well. At the fourth annual Semantic Technology conference in San Jose, Calif., in May, the topic of semantic tagging was presented by TextWise, a developer of text extraction, search, categorization, and classification technologies using both NLP and statistics. In the presentation “Applying Trainable Semantic Vectors to Tagging, Search/Discovery, Bookmarking and Matching,” a panel of TextWise speakers explained how its Semantic Signatures function as tags for bookmarking or in generating tags to map/link an existing tag set.

Semantic tagging’s integration with search technologies is also being applied

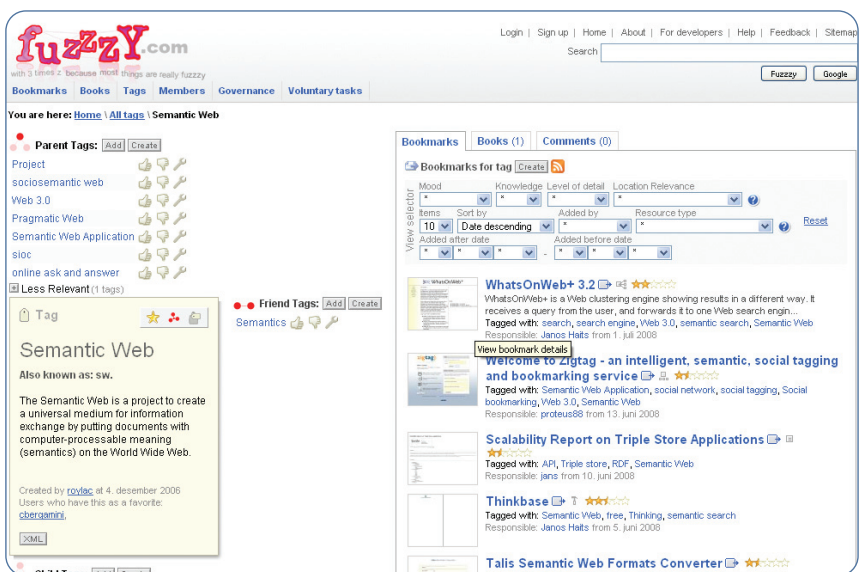
in niche service areas. For example, Relevad, whose tagline is “semantic keyword analytics,” provides hosted web service for online advertisement placing. Relevad claims a growing database of more than 8 million keywords and more than 500 million neighbor keyword meanings. Trovix, meanwhile, provides a web service of matching jobs to resumes utilizing complex scoring algorithms in combination with a “hierarchical knowledgebase” of U.S. cities, skills, positions, industries, and companies.

SEMANTIC SOCIAL TAGGING

The term “tagging” is most strongly associated these days with social tagging or social bookmarking, whereby people assign tags (terms or keywords) of their own choice to documents, blog posts, or webpages that they have created or have viewed to assist in locating the documents later, whether by themselves or by others. Better known tagging websites and services include Delicious, Flickr, and Technorati. There is generally no taxonomy or controlled vocabulary involved, as any words can be used as tags, although this is changing in some applications.

Fundamentally, this type of tagging is “semantic” as well, because humans

Fuzzy tagging/tag creation UI supporting parent tags (broader terms), friend tags (related terms), and child tags (narrower terms)



manually tag content for what it means. The problem is that this tagging is done based on what the document means to the tagger at the time of tagging, not necessarily what it means to other users or even to the initial tagger at a later time. Furthermore, any lists of the occurrences of a tag can be long, undifferentiated, and ambiguous. The term “semantic tagging” within the sphere of social tagging, therefore, is being used to refer to a method of imposing consistent and more refined meaning. In other words, utilizing some kind of a taxonomy. Such semantic social tags are also being called “rich tags.” Not only are the tags’ meanings clarified by synonyms, but there also may be links to related-term tags and the presence of glossary definitions for tags. In other words, semantic tags or rich tags are essentially terms in what is known to librarians as a thesaurus.

Social tagging sites/services that offer what they call semantic tagging include Zigtag, a Canadian startup, and individual-led projects Faviki and Fuzzzy (yes, with three z’s). Zigtag (in private beta as of this writing) is a sidebar plug-in, which differentiates itself from other tagging services by providing a “semantic dictionary” of more than 2 million tags. Tags are defined and synonyms are linked together. Faviki is a social bookmarking tool that provides terms from Wikipedia, extracted by the open DBpedia tool. This not only provides consistency, but also extensive definitions for each of more than 2.18 million Wikipedia resources. Fuzzzy, on the other hand, did not start with a prebuilt taxonomy, but user-created terms are entered into a shared tag set (thesaurus) and various relationships (broader, narrower, related)

are supported. Thus, Fuzzzy “enables global distributed tagging.” The organic tag set of Fuzzzy is built upon the Topic Map ISO standard and an underlying infrastructure with Web Services.

It isn’t just new kids with extra consonants pursuing social tagging however. Big, established content players are also getting involved. Thomson Reuters offers its open Calais Web Service, which ingests unstructured text and, using NLP, and returns RDF-formatted results identifying entities, facts, and events within the text. In May, Calais was made available as plug-in software for the Drupal publishing platform, Yahoo!’s new Searchmonkey service, and the WordPress blogging platform. The Calais plug-in for WordPress, called Tagaroo, returns tag suggestions based on text typed into a blog but gives users the option of choosing which they want to apply. Calais also offers licensed code to make one’s site part of the “Semantic Web.”


TAGGING AND THE SEMANTIC WEB

Finally, semantic tagging can be defined as tagging for the semantic web. This involves tags that make use of RDF (Resource Description Framework) specifications or OWL (Web Ontology Language) of the World Wide Web Consortium (W3C). This also implies being used for public webpages that can be accessed with semantic web browsers, rather than merely internal enterprise or library products or services. As such, a tag is more than a term; it is an object with its own attributes. According to Rhind-Tutt, “The difference between semantic indexing and standard indexing is that the former does more than simply apply subjects to terms. It includes the addition of meta-data

about tags that allows semantically indexed terms to interoperate with other similarly indexed terms.”

(This is discussed at more length in the blog post “Tagging and the Semantic Web”; see www.designmills.com/2008/05/20/tagging-in-the-semantic-web.)

While social tagging can be made more semantic, we have to remember that social tagging is not always about pure findability. The social aspect is about identifying what other people have labeled as interesting or noteworthy, especially if there is a rating aspect involved. For the semantic web, on the other hand, information findability is a major objective, as stated in W3C’s Semantic Web Activity Statement: “to create a universal medium for the exchange of data. It is envisaged to smoothly interconnect personal information management, enterprise application integration, and the global sharing of commercial, scientific and cultural data.”

Silverchair’s Zarnegar put it well: “Semantic tagging is best applied in areas when there is a qualitative ‘best answer’ to a user query (as opposed to a ‘most popular’ answer) ... If you look at industries where semantic tagging (and structured data) have found a foothold (aviation, medicine, genetics, chemistry, and others) you’ll see they are not areas where you want to go too far with iffy information!” 

HEATHER HEDDEN (HEATHER@HEDDEN.NET) IS AN INSTRUCTOR OF CONTINUING EDUCATION WORKSHOPS AT SIMMONS COLLEGE GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE, AND FOUNDER AND MANAGER OF THE TAXONOMIES & CONTROLLED VOCABULARIES SIG OF THE AMERICAN SOCIETY FOR INDEXING.

COMMENTS? EMAIL LETTERS TO THE EDITOR TO ECLETTERS@INFOTODAY.COM.