

# Searching for the Needle in the Haystack: Taxonomies, Tags and Targets

Michael Pelikan

The Pennsylvania State University,  
Technology Initiatives Librarian,  
University Libraries, University Park,  
PA 16802  
+1 814 865-5660  
mpelikan@psu.edu

James Leous

The Pennsylvania State University,  
Manager, Research Programmer,  
Academic Services and Emerging  
Technologies, Information  
Technology Services, University  
Park, PA 16802  
+1 814 863-7206  
leous@psu.edu

Richard Pearce

The Pennsylvania State University,  
Director, Business and Finance for  
Auxiliary and Business Services,  
University Park, PA 16802  
+1 814 865-3061  
rhp1@psu.edu

Margaret E. Smith

The Pennsylvania State University,  
Manager in Information Technology  
Services, University Park, PA 16802  
+1 814 863-8125  
mes8@psu.edu

Russell Vaught

The Pennsylvania State University,  
Associate Vice Provost, Information  
Technology Services, Affiliate  
Professor of Information Sciences  
and Technology, University Park,  
PA 16802  
+1 814 863 3746  
rsv@psu.edu

## ABSTRACT

The Penn State Taxonomic Tags group, with representatives from Information Technology, Business Administration, and the Penn State Libraries, was formed to examine whether a taxonomic set of tags, systematically applied across the university's Web pages, could (a) make finding specific pages easier from among the University's greater than 500,000 Web pages, (b) simplify Web content management tasks and (c) prove useful over time as search engines continue to evolve and despite whether open source or commercial (and often, proprietary) search algorithms are employed. The University has had broad experience with several search engines, and currently holds a University-wide license for the Google Appliance. The Tags Group has developed recommendations that it believes will address issues found in the current environment and yet remain useful during and after what it expects will be the increasing adoption of Content Management Systems across Penn State University.

## Categories and Subject Descriptors

H3 [Information Storage and Retrieval

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGUCCS'04, October 10–13, 2004, Baltimore, Maryland, USA.  
Copyright 2004 ACM 1-58113-869-5/04/0010...\$5.00.

## General Terms

Management, Performance, Design, Standardization.

## Keywords

Taxonomies, metadata, web search engines, controlled vocabularies, content management systems

## 1. FROM PENNSTAC TO THE WEB

In the years between 1956, when PENNSTAC (Penn State University's first computer system) was built, and 1994, when <http://www.psu.edu> first appeared, ongoing and fundamental change in how academic and administrative information were stored and retrieved in the university environment became a way of life.

The Internet, developed in the late 1960's and widespread by the mid-1980's, provided a network enabling worldwide computer connections. Microcomputers had emerged during the 1970's, and by the late 1980's, an increasing number of them (many of them "personal computers") had network connectivity.

In 1989, Tim Berners-Lee,[1]then at the European Organization for Nuclear Research (CERN, in French the *Organisation Européenne pour la Recherche Nucléaire*, is located in Geneva, Switzerland) conceived the World Wide Web as a means for the high-energy physics community to share information between dissimilar computers. His idea was to create an application called a *Web browser* that would run on a desktop computer that was connected to the Internet.[2] The browser could display information retrieved over the Internet from a *Web server* that ran on a central computer.

The first implementations of Berners-Lee's idea were developed in 1990. Early browsers and servers were released in 1991. They were effectively limited in their use to the high-energy physics community. It would take a few years more for the Web, as we know it now, to become reality.

A major turning point occurred in 1993, with the development and release of the first truly *user-friendly* Web browser, *Mosaic*, by Marc Andreessen at the University of Illinois's National Center for Supercomputing Applications (NCSA) [3]. One of Mosaic's great strengths was that it was available in versions for all the major operating systems of the time (e.g., Windows, Mac OS, UNIX). The ease and simplicity with which Mosaic allowed non-information professionals to locate and access networked information provided a crucial catalyst to the emerging importance of the Web.

## 1.1 Penn State's First Web Page

Penn State first appearance on the Web occurred much as early pages appeared elsewhere – quietly. In a low key, informal, indeed, *unofficial* move, a junior IT staff member created the University's first home page. With the server software from the University of Illinois in hand, this staff member decided that Penn State ought to have a presence on the World Wide Web, just as a few other universities did. He installed the software on a Penn State system and <http://www.psu.edu/> became available around the world.

Penn State was on the Web, but it was several months before anyone but diehard “techies” knew it. At Penn State as elsewhere, to those who had installed it, the Web was just “cool technology.” They simply wanted to get to know it, and share it with their counterparts elsewhere. Few understood the important role it was to play.

So complete has been the transformation of the information landscape in the decade since 1994 that it has become almost trite to refer to the change as “revolutionary.” And yet, this technology has become ubiquitous, and its impact has touched nearly everything we do. This happened, at Penn State and elsewhere, because Web software was relatively easy to install, was independent of computer type, and produced an information environment inherently decentralized in design. Just about anybody could do it and just about everybody did.

And now, as of this writing in August of 2004, although the exact figure remains unknown, the Penn State presence on the World Wide Web, including “official” and “unofficial” Web pages, numbers somewhere between 600,000 and 1,000,000 pages.

## 1.2 From Really Cool to Strategic Tool

The Web marched quickly from cool technology to strategic tool. By the late 1990's, Penn State's executives understood the strategic importance of this technology, and IT professionals increasingly sharpened their efforts to examine how the Web could serve Penn State's mission more effectively.

Prior to the Web, the retrieval of specific information stored electronically required the mediation of a professional. Rigid underlying record structures of the information systems in use made it difficult for a non-computer professional to get the

precise figure or specific record needed. Often it took successive attempts get precisely what was wanted, or what one could get was out of date. The result of these successive attempts was often piles of printout paper that failed to answer the particular question.

## 1.3 No Panacea

The Web changed much of this, or perhaps more accurately, it *appeared* to change much of this. The Web made it easy for anyone to retrieve *lots* of information, to browse it, to swim around in it until the desired piece of information was either found, or the original objective forgotten. From the perspective of the burgeoning class of information providers, it became almost trivially easy to put up Web pages. The number of pages skyrocketed.

Very quickly the hapless information seeker was, once again, buried in too much information. The new twist in the situation was that this information was not only out of date (as ever), but very possibly not coherently structured. As a result, one still had trouble finding the information needed to answer a particular question.

In a decade, we have gone from a dearth of information to a glut of information. It would appear that we have traded an old problem for a new one, but in fact, the consequences for the common information seeker remain the same: it's still too hard for the average user to get the information he or she is looking for.

## 1.4 The PSU Tags Group and its Context

In the fall of 2003, Russel Vaught, Penn State Associate Provost for Information Technology, formed a small group, dubbed the Taxonomic Tags group, following a series of discussions he held with representatives of information technology-intensive departments at Penn State. He asked the group, representing Information Technology, Business Administration, the Penn State Web and the Penn State Libraries, to examine the question of whether a taxonomic set of tags, systematically applied to Penn State University web pages, could (a) make finding specific Penn State pages easier, and (b) offer any benefits to the university's ever-expanding, massively decentralized efforts to manage web content across the Penn State web presence.

Since then, the Tags group has been examining approaches to making information on the Penn State Web both more readily accessible and more coherent, to enhance its usefulness to individuals within and outside Penn State. The University has had broad experience with several search engines, and currently holds a University-wide license for the Google Search Appliance<sup>1</sup>. A notable feature on the university's information landscape is the Penn State Portal, which is individually customizable by Penn State users. There is a growing body of content in this Portal and its adoption as a personal home page is becoming more prevalent.

These two features – a recognized search engine on the one hand, and a nascent, Content Management System with an individually customizable user interface on the other – formed the bookends for the Tags group's mission. Given these tools at hand, the group began to examine what could be done, from the standpoint of purposefully designed information storage, to provide an optimal,

---

<sup>1</sup> <http://www.google.com/appliance/>

underlying structure supporting information retrieval on the Penn State Web.

## 2. TAGS, TAXONOMIC & NOT

Taxonomies are structures for information. Familiar examples of taxonomies include those used in the natural sciences in the classification of plants, animals and rock. The Library of Congress Subject Headings that structure the information in the library catalog are another example of taxonomy. The Tags group began to look at whether a *taxonomic representation of Penn State, expressed as a controlled vocabulary contained in html tags*, could provide effective underlying structure to Penn State Web content.

Search engines systematically *crawl* through Web pages to harvest specific information and organize it according to the rules established by the search engines' designers. To varying degrees, search engines employ the textual content, both visible and hidden in tags, to index the web pages they search.

When a user makes a query, the search engine employs its index to retrieve the stored URLs for pages that have content matching the query. In contrast to queries against a database in which data is structured in discrete fields, there is often no underlying structure to the textual content in the Web pages a search engine crawls. HTML Tags are one way to provide a structure for textual content in a Web Page. Consistently applied, such tags can identify page content in a way that is potentially useful to search engines – if the search engine is designed to “notice” such tags and factor them into its indexing.

### 2.1 The Influence of Commercial Pressures upon Search Engine Design

When the Tags group first started looking at these issues, the primary search engine in place at Penn State made heavy use of tags in creating its index of pages. The Tags group felt that providing a taxonomic vocabulary in a consistent tag set could make a significant improvement in the probability that a search would retrieve a page with the desired content on it. The tag set would be used to describe organizational elements at the university, to identify persons and their roles, as well as the functional purposes for the Web pages themselves.

Within an organization, there are clear purposes to be served by ensuring that particular search terms retrieve specific pages. A search on “Admissions” should reliably get a user to top-level page access to the university Admissions department. In the larger, external realm, in which commercial and non-commercial sites are thrown together, the challenge facing search engine designers becomes much more complex.

With the presence of commercial content on the Web comes commercial pressure – it is in the interests of a company to have its web pages high in the ranking in any result set produced for a potential customer by a search engine. In a kind of “arms race” not unlike the never-ending contest between thicker armor and more powerful armor-piercing munitions, it became commercially imperative for competing ventures to elevate their Web pages in the rankings produced by search engines. Commercial ventures often did this by employing keywords contained in tags intended as targets for the search engines to index.

Concurrently, Penn State was preparing to replace its primary internal Web search engine. After a wide examination of various search engines that included trials of a number of offerings and consultation by a wide cross-section of people within the University, the Google Search Appliance was chosen. This is a site-specific version of the popular and well-regarded Google search site.

### 2.2 The Google Fly in the Ointment

The Google Search Appliance became the University-wide search engine in late 2003. One of the features of Google is its relative imperviousness to attempts to raise the relevance ranking of a given web page artificially. It is apparent that in order to decrease the ability of ventures to fool the Google search engine into giving prominence to their pages, Google has considerably decreased the importance of tags in shaping its search results

At first glance, this would seem to make the systematic creation and application of tags a less important element in improving the probability that searches will return desired content. Closer examination shows, however, that a well-designed taxonomy still has value in making Penn State's web space both more useful and coherent.

## 3. SEARCH ENGINES AND DATABASES

A comparison of *search engines* and *database management systems* is useful in offering clues to designing information storage to enhance the precision of information retrieval. Databases structures are granular at the field level. In other words, one searches for entries in the database by posing queries against one or more fields in the database. These queries can be a single word or more complex, with phrases, Boolean operators, wild cards for characters, or simultaneous searches against multiple fields. A search for *Dodge* in a field designated to store “Vehicle Manufacturer” will only retrieve vehicles made by Dodge.

While not widely realized by the general Web-searching public, *search engines* perform their queries quite differently. Typically, Web search engines, depending upon their characteristics and those of the pages they index, are granular at the document level, that is, at the Web page level. Search terms are matched against the entire contents of the Web page rather than targeted at a specific field. This can make the context of a term uncertain. A search for *Dodge* may find a car, but it may also find *Dodge*, Kansas, an artful *dodge* (e.g., trick), or a move to keep getting tackled in football. There is no taxonomic structure to provide context to the search term as there is in a search against a database field.

The problem then becomes how to make Web page text more database-like in order to reduce ambiguity. There may be several ways to do this. For example, document types found in academia often have a highly predictable structure. Journal articles have predictable places for title, author, citations, etc. A search engine can take advantage such predictability of form to parse the context for the search term. To search for the citations in an article, for example, a system designed to look for text that follows a heading such as “Works Cited” will prove to be, in practical terms, just about as precise as searching for entries in a specific field in a database.

*Citeseer*,<sup>2</sup> a highly regarded specialized search engine that gets about two million hits per day, employs such tactics. Its capabilities are based upon the predictability of the structure of academic articles.

In administrative areas, the Tags group's goal is to help Penn State organize *enterprise information* in such a way that both browsing and searches are more likely to provide the information needed for a particular task. Searches are hampered by the lack of coherent, predictable structure across the Web pages of administrative and academic units. Often, even *unit names* are of little help. There is a good business case for providing accurate, predictable information retrieval to a university's internal and external audience. This has been recognized and acknowledged in action by the National Associate of College and University Business Officers (NACUBO), which has developed a standardized set of tags to describe a university administration. [4] This tag set can be extended to meet local needs. The Penn State Tags group's goal in the administrative area is to represent the information available in the Penn State Administrative Web space in a well organized and intuitive classification structure.

### 3.1 But Where are the Constraints?

Although tags can provide needed structure, adopting a tag set addresses only part of the problem. Unless you constrain the terms, values, or entries that are stored in the tags, you have simply provided a predictable place for unpredictable terminology to appear. For example, if you could assign a Web page the subject of "automobiles" (by having that word appear between <DC.Subject> tags) then you could perform a search on the term *dodge* in pages whose subject is "automobiles" and find pages related to the car of that name. This would be true even if the word "automobile" was not visible in the Web page's text (referring instead, for example, to "cars and trucks"). The result would be a more precise search that was more likely to find pages that were relevant to the query. It is for this reason that the application guides related to metadata element standards such as those developed by Dublin Core Metadata Initiative urge the use of controlled vocabularies.

Dublin Core (DC) comprises a comparatively simple set of descriptive metadata elements for use in Web pages<sup>3</sup>. All DC elements are optional, and all can be repeated if necessary to accommodate multiple terms. Institutions often extend Dublin Core, adding customized elements to accommodate their organizational needs.

DC's relative simplicity, its extensibility, and the optional, repeatable nature of its element set all conspire to make the effective application of DC far trickier in practice than one might first guess. The complications are compounded dramatically once you begin to search across multiple sets of records in DC, containing either dissimilar terminology for like things, or similar terminology for unlike things. The matter of context becomes central to understanding what a given term stored in a metadata element means. The term "mushroom" has one meaning in proximity to the term "fungus", and entirely another in proximity to the term "cloud."

Contextual subtleties pervade the issues that the Tags group was formed to address. Still, a simplified, controlled vocabulary might well be developed to describe the academic space just as the NACUBO keywords can be used as a basis for the administrative space. It is within the technical realm of possibility to extend the equivalent of a library catalog's authority control<sup>4</sup> into a well-coordinated metadata encoding and searching system.

### 3.2 Content Management Systems

The Tags group quickly realized that the emerging importance of content management systems, added to this already heady mix of tag sets, vocabularies, search engine algorithms and information representation, presented a new set of challenges, but also, perhaps, some opportunities. Content management systems often create pages dynamically, on demand. This can quickly confound a Web search engine designed to index unstructured text floating in an html page. For example, Penn State IT administrators discovered that multiple passes made by a search engine's crawler can result in multiple, separate index entries for a single web page – one for each pass of the crawler. In the case of the Google Search Appliance, which is licensed for a particular number of entries, this could quickly result in both a lower probability of finding the right content and an escalation in cost because of the much larger resulting number of entries. At Penn State, this situation resulted in a decision by the search engine administrators to stop indexing dynamically generated pages of this type.

While the adoption of content management systems is going to make Web content more current, it may well also make the challenges of planning the storage of information for retrieval more complex. The Tags group believes that the development of an underlying taxonomy to provide structure to pages becomes more important with the adoption of content management systems. The challenges of retrofitting coherent tagging into half a million Web pages is daunting. Content management systems provide both require and, we believe, provide a practical opportunity and means to accomplish the implementation of a tag-based taxonomy.

At the same time, there is a clear motivation for vendors and the open source community to continue efforts to develop search engines that employ new and improved strategies for this emerging environment. Searches may well have to be targeted against the database used by the content management system itself rather than against the resulting web pages. The Tags group has also considered using a search against LDAP entries extended to include externally defined keywords. Multithreaded searches of a type similar to Apple's Sherlock also merit investigation, as do other forms of federated or broadcast search methods that can target multiple metadata repositories and aggregate the results into a single result set.

## 4. GOOGLE – WHAT WE KNOW

Each different search engine uses a particular strategy. This may be configurable or not, but understanding how your institution's search engine works can lead to strategies to improve its capability to find the desired Web pages predictably.

---

<sup>2</sup> Citseer, <http://citeseer.ist.psu.edu>

<sup>3</sup> Dublin Core, <http://dublincore.org>

---

<sup>4</sup> Authority control is the capability that enables one to find Tom Sawyer whether one searches for the author "Mark Twain," or "Samuel Langhorne Clemens".

Google, like most search engines, uses a *spider* program that *crawls* the Web space starting with a particular page. In the case of Penn State, that page is our university welcome page, <http://www.psu.edu/>. We have set the Google Search Appliance to crawl our Web twice per week. With each pass, it indexes between 500,000 and 600,000 Penn State pages. The Google PageRank™ system then uses proprietary algorithms to create a weighted index to facilitate search queries. The workings of the Google PageRank™ system are highly proprietary, indeed, a secret closely guarded by Google, not to be shared with Google’s competitors, nor its users, nor paying customers for that matter. The Tags group can only make inferences, therefore, about how the Google PageRank™ system works.

Nevertheless, the Tags group believes that Google has improved the quality of search results most Penn State users achieve, compared to our previous search engine. Users are, in general, more satisfied with it, and the number of complaints about searching Penn State Web pages has decreased since adoption of the Google Search Appliance. Still, the fact that Google uses a proprietary ranking algorithm does pose problems.

On the one hand, it makes it harder for us to force pages to the top of a result set. To the extent that we can do so, we must do so by *intentionally fooling the system*. These difficulties in manipulating the results weighting algorithm decrease the probability that the search will produce a result set that quickly finds and displays the desired URL. Google claims that such manual manipulation (also known as “*page tweaking*”) is rarely successful. There are, however, examples that crop up from time to time in which individuals have been able to manipulate the Google system<sup>5</sup>.

## 4.1 Google – What We Think We’ve Figured Out

The Tags group believes that Google evaluates page importance, in part, by examining how the page is linked to from other pages. It appears that a page’s importance is determined by the importance Google places on the pages that link to it. While this is significant when considered on the scale of the public web search site Google.com (indexing ~ 4.3 billion pages in June 2004), the Tags group suspects that Google’s algorithm may not optimally scale down to the size of a typical university or college level search environment.

Google excels at helping a potential purchaser find a specific product, especially if the source for the product is less important to the searcher than finding the product itself. A search on the term *red backpack*, for example, will immediately produce many pages of hits linking to red backpacks.

In many respects, Google’s job is made simpler by the sheer scale of the Web. When you index billions of pages, you are bound to get hundreds of highly relevant results on a search term like *red backpack*. In a university environment, even one as large as Penn State, the algorithm designed to find you a red backpack from billions of pages might not work effectively to retrieve the Department of Admissions from half a million pages and put it at

the top of the results. Nevertheless, there are some things that can be done from the perspective of page design to improve search results with Google.

Our observations of Google’s behavior suggest that in Google:

- Links from other pages are weighed heavily
- Page Title content (i.e. <TITLE>) is very important
- High Keyword density within a document elevates its ranking in a result set
- Keyword-laden links appear to elevate a page’s ranking, such as “[Office of Undergraduate Admissions](#)” as a link, rather than “To go to the Office of Undergraduate Admissions, [click here](#)”
- Keywords in Metadata tags are not that important

If we assert that the interconnections between individual colleges and units are not as plentiful as they are in commercial Web pages (as is demonstrated by our own examination of Web pages at Penn State), the middle three findings listed above (i.e. Title content, keyword density, and keyword-laden links) take on increased importance. Applying these findings, we believe, can lead to a higher probability that desired pages will be highly ranked in a search result set.

It’s our observation that Web page Title tags are often left blank (or the page has a non-meaningful title). A page title that is descriptive of the content of the page appears to be very important in determination of rank for the page. The text used within the page can also be used in this way. The text used for links appears to be especially important in the rating of the page that is linked to. Links should be descriptive of the content on the page being linked to – that is, rather than having a link say “Click here”, it is better to label it with terms descriptive of the link’s target.

## 5. SUMMARY

Despite the current Google algorithm, the Tags group believes that determining a taxonomy and vocabulary for university Web pages will prove effective, over time and technological generations, in allowing users to make more efficient use of university Web pages.

Most recently, the Tags group has been discussing a phased approach that would address Penn State’s current information retrieval environment while attempting to pave the way for its foreseeable successor, which the group believes will be an environment in which Penn State Web-accessible content will be spread across some yet-unknown number of Content Management Systems. A comprehensively centralized approach to content management is probably out of the question at Penn State for a number of reasons ranging from the technical to the organizational. Therefore, we can foresee a federated content management environment. Many of the University’s web pages, while generated from database-housed content, will be static targets for a search engine that operates upon the textual content of the pages. At the same time, searchable data in the fields of the databases from which the pages are generated could include non-displaying metadata to serve as targets for queries.

<sup>5</sup> A search on particular pejorative terms, for example, have been known to result in the retrieval of the home page of particular candidates for high political office, at least until Google put a stop to it.

Such an approach suggests that we should explore the adoption of a meta-search front-end for the Penn State Web search system. A search term could be submitted both to a web search engine that would have crawled and indexed the static or rarely-generated University web pages, and at the same time would also submit the search term as a query against a targeted set of database fields containing keywords, acronyms, synonyms, etc. A results page could be generated listing the two sets of results, identifying one as the results of a web search, and the other as a database search.

Finally, the Tags Group is united in a shared sense of resistance to targeting university efforts at any particular, proprietary commercial product.

## 6. REFERENCES

- [1] Berners-Lee, Tim. New Web Will Enable Scientists to Share Data Across Disciplines. *Chronicle of Higher Education*, 49, 22 (February 7, 2003), A25 or online at <<http://chronicle.com/weekly/v49/i22/22a02502.htm>>.
- [2] CERN: <http://welcome.cern.ch/>
- [3] NCSA: <http://www.ncsa.uiuc.edu/>
- [4] NACUBO: <http://www.nacubo.org/>