

Sam H. Kome. Hierarchical Subject Relationships in Folksonomies. A Master's Paper for the M.S. in I.S degree. November, 2005. 32 pages. Advisor: Jane Greenberg

The growth in digital resource repositories flickr and del.icio.us, mirrors the growth of Folksonomies to support resource classification and access. Despite this phenomenon, little is known about the effectiveness of folksonomy for retrieval and organization.

Little is also known about their structure and the types of semantic relationships among folksonomy terms. This study analyzes folksonomy metadata for hierarchal semantic relationships via a content analysis of approximately 2000 folksonomy tags in over 600 individual entries. The terms were classified into groups and analyzed for hierarchical relationships. The results indicate that hierarchical relationships are part of Folksonomies. The conclusion briefly explores the potential value of thesauri for Folksonomy development, and the value of Folksonomies to thesauri developers.

Headings:

Categorization/Psychology

Cognition

Hierarchy/Linguistics

Psycholinguistics

Taxonomy/Folk

HIERARCHICAL SUBJECT RELATIONSHIPS IN FOLKSONOMIES

by
Sam H. Kome

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

November 2005

Approved by

Jane Greenberg

Introduction

Each day technology enables access to more information; without appropriate categorization, however, this information cannot be fully utilized. The apparent potential for technology to automate the categorization process has tantalized yet challenged librarians and researchers for decades. Although automatic categorization can be useful now in some contexts (Liddy 2001), to date the best automatic categorization systems cannot match the accuracy of their human counterparts (Greenberg 2004). The systems lack nuanced understanding of context and relevance.

Librarians see that the time requirements of cataloging new volumes are too high to contend with the volume of digital resources. Aside from libraries, every organization with information resources faces similar issues, and alternatives are being sought.

Authors and users are alternatives to experts; to date, each has proved to help solve the scale problem, but evidence of their ability to provide information of consistently adequate quality is limited. Examples of author and user categorization schemes are popping up in networked applications where human subject enthusiasts provide classified content, and computer programs connect and display the commonalities found therein, as in the websites flickr and del.icio.us. These ‘folksonomies’ are large repositories of metadata, but there is little agreement on their utility for traditional search and retrieval.

This study analyzes folksonomy metadata for semantic relationships and quality of description, and presents the possibility of integrating these data into traditional cataloging.

The current disagreement over the efficacy of folksonomy metadata is based in part on an assumption that folksonomies lack hierarchical structure (Rosenfeld 2005). While search and retrieval are known to be facilitated by structured subject headings, the conventional wisdom has it that descriptive ‘tags’ which form the basis of a folksonomy’s organizational structure are flat, unlike the subject headings assembled by professional catalogers. Proponents celebrate this perceived feature (Shirky 2005), but it may not be an accurate perception. Decades of research into human cognition and categorization activities have found that categorization is a fundamental human cognitive activity, examples of category systems exist across cultural and lingual differences, and they share numerous traits including hierarchical organization. If folksonomy tags are indeed elements of hierarchical structures, then folksonomies could contain the benefits of thesauri for resource discovery: search and browsing, with increased speed of development and reduced cost. To better understand the information framework and implications of folksonomies, this study performed a content analysis of approximately 2000 folksonomy tags in over 600 individual entries. The data were classified into groups by library science catalogers, following which the presence of hierarchical relationships between the user terms were enumerated.

Background

The study was informed by an examination of the literature on folk taxonomy, folksonomy, the mechanics of delicious, and hierarchy in the area of library and information science. These four topical areas are reviewed below and provide the basis for this research.

Folk Taxonomy

Brent Berlin's (1966) study of plant naming in the Tzeltal dialect found a taxonomy that was, "casually collected, non-systematic, incomplete, or anecdotal" (p. 273). In cognitive linguistic psychology experiments Eleanor Rosch (1978) noted patterns related to language acquisition and cognitive categorization activities that meshed with Berlin's findings. Both Berlin and Eleanor Rosch noticed that the entries in folk taxonomies followed patterns with respect to language and description, regardless of subject area. Rosch's study of categorization (Rosch 1975) noted as a first principle that the world which we are trying to describe has a highly correlational structure. For instance, feathers correlate with birds and wings, not with cheese. The next important principle is that we classify information with the goal of maximizing information for the least cognitive effort. This goal is attained when we construct information categories that are optimally mapped to the underlying structures. If the underlying structures are highly correlational, then it follows that hierarchical information structures should map well to them, thereby maximizing our gain with the least cognitive effort. Put another way, we inherently classify information into highly correlational structures, including hierarchies, without extensive training in library science. Anderson (1991) successfully modeled these 'predictable structures'¹ with numerical methods. Therefore one can expect the presence of correlational structures including hierarchies in folk classification schemes.

One could make an additional prediction of folksonomy properties from Rosch's work on prototype theory (Rosch 1981). This is a theory which stipulates that humans assign objects to categories by comparison to a prototype object. Prototypes are "typical"

¹ In the sense of categories' use to predict set member features.

examples, such as a sparrow is considered a typical bird; more so than an ostrich (sparrows and ostriches are narrower terms to 'bird').

Prototypes often form the bases of information structures (Lakoff 1987), from which superordinate and subordinate classes are determined. Today's folksonomies could reflect this by:

- presenting hierarchical structures amidst the tags
- the most frequently used tags being not the highest nor lowest in a hierarchy. They will instead be terms which typify a concept to which users are assigning resources.

Folksonomy

In September of 2003 Thomas Vander Wal coined the term 'folksonomy' in a listserv discussion hosted by the Asilomar Institute for Information Architecture (now the Information Architecture Institute). The discussion concerned the websites del.icio.us and flickr, both of which (still) exist to enable subject enthusiasts to categorize information. Vander Wal's (2005b) definition first:

"Folksonomy is the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval. The tagging is done in a social environment (shared and open to others). The act of tagging is done by the person consuming the information."

Despite using a blend of "folk" and "taxonomy", Vander Wal is quite clear that the neologism 'folksonomy' was not conceived in an effort to convey the idea that the shared categorization results would contain hierarchical structure in the way of taxonomies (2005b): "It is not collaborative, it is not putting things in to categories, it is not related to taxonomy (more like the antithesis of a taxonomy), etc." On-line

discussions and articles now predominantly maintain that the categorized data do not resemble taxonomies, that they are 'flat' with respect to structure. (Rosenfeld 2005)

What can be seen of this relatively new version of folk classification are typically collections of information resources organized into subject categories which are created in an ad hoc fashion by any users who have access to the world wide web. In most cases, users can see but are not constrained to use each other's category choices. The subject categories or keywords are known as "tags". A web site at URL <http://del.icio.us> is often proffered as an example of a "broad" folksonomy, i.e., a categorization effort where terms may be contributed by anyone, as opposed to a narrow folksonomy wherein only the original content submitter defines the descriptive terms. Del.icio.us serves as a website bookmark manager and sharing tool. A fuller description of del.icio.us can be found in Mathes (2005). For the purposes of this study, del.icio.us will be used as a representative folksonomy.

The Mechanics of del.icio.us

Del.icio.us users may classify the URLs they submit with any number of tags (including none). They are not constrained to use known words or known tags. No vocabulary controls exist with the exception that the system requires tags to be single words. They may work around this limitation by simple concatenation. Anecdotal evidence suggests that there is no common convention; users substitute various punctuation characters for spaces, or simply elide words together.

Each user accumulates their own set of terms, a.k.a. "tags" and information resources. In a broad folksonomy like del.icio.us, for each resource they can also see the tags used by anyone else who also happened to contribute it. It should be emphasized that the user is not necessarily influenced by the term choices of other contributors. This

is an implementation level detail. At the time of this writing, del.icio.us did not display the term choices of other contributors for consideration. The choice not to do so is a deliberate design decision which reflects a tenet of modern folksonomy; tag diversity is good. Berlin observed that over time folk taxonomies became increasingly over-differentiated, i.e., there were more and more words for the same concept. Vander Wal describes this phenomenon as an observable trait of the modern folksonomies (2005b). There are also well defined group similarities in tags, due more to coincidence and cultural effects than to collaboration, which is not facilitated by the user interface.

Within the set of tags provided for a URL by an individual user, the conventional wisdom is that equivalence relationships are the norm. It is this property which causes some to argue that the neologism, “folksonomy” is inapt; taxonomy requires hierarchy and the possibility of mutual exclusivity between terms. Rosenfeld argues that it is this lack of structure in folksonomies that limits their usefulness for retrieval or administration, but that as a component of a broader "metadata ecology", they could have the useful purpose of enabling or encouraging users of information to create descriptive metadata. In this way users can contribute to the development of controlled vocabularies, which he believes are still needed.

Hierarchy

The semantic relationships between the words chosen for the subject heading lists in a document classification scheme typically comprise a hierarchy. Hierarchical semantic relationships between concepts can be used to both broaden and focus browsing and search activities. Hierarchies are recognized as having two types, hyponymic and meronymic. (Bodenreider, 2004). Genus – species are typical hyponyms. Across the genus felis , cats share numerous traits, but rufus is distinct from domesticus and all other

species. Words used to describe terms in this relationship include “is a”, or “is a kind of”. Class – instance hierarchies are a form of hyponym. An example of this is “University of North Carolina” as an instance of the class, “University”. Meronyms are “part of” relations, where concept A is part of concept B. “Grammar” is a part of “language”. For further information on this subject, see the ANSI/NISO Z39.19 standard (NISO 2005). For examples of variations, see the ALA Subject Analysis Committee’s report, Appendix C. (ALCTS 1997).

With respect to hierarchy in taxonomies, the distinction made between sets and features (Kay 1971) is kept here, namely that the basic structural components of taxonomies/hierarchies are sets, not properties or features. Descriptive language in tags can be perceived as hierarchical, but no standard applies. For example this standard causes “small” and “smallest” to be ignored.

Subject heading lists in library science are normally precoordinated and hierarchical terms taken from controlled vocabularies (Mann 2000; Lancaster 1986). In the case of folksonomy, the tags serve as the subject headings, they are not terms taken from a controlled vocabulary, nor is their ordering considered important. Creators pick tags which they think will remind them of aspects of the information resources. When they use the tags for retrieval, they can use the tags they themselves entered, or choose from the collective list.

The presence of hierarchical relationships in the either tag list provides structured mapping to the underlying information resources.

Summary

Modern folksonomies share mode of origination, internal structure, and many content properties with folk taxonomies. While Brent Berlin's original work focused on biological taxonomies, Rosch, Kay, Lakoff (1987) and others contributed to generalize his observations to describe human cognition. The resemblances suggest that the folk taxonomy literature is relevant to folksonomy. The cognitive economy and prototype theories suggest that subject enthusiasts will generate at least shallow hierarchies when describing features in their world -- in this case information resources. If they do, and if the tags can be mapped to traditional vocabulary, these user generated metadata could be exploited to help break Liddy's "metadata bottleneck" (2001).

Research Question

To date there has been little quantitative analysis of folksonomy tags. The analyses that have been conducted have at least confirmed that for each information resource (URL), a very few tags are used with high frequency, and there are a large number of tags with a very low shared usage -- a "long tail" on the curve. This phenomenon is consistent with the use of uncontrolled vocabularies for classification. This study considered the absence of controlled vocabularies and specifically sought to answer:

- Are there hierarchical subject relationships among the tags?
- Are tags adequate to distinguish between broad subjects?
- If yes to both, what are the implications for human or auto categorization efforts?

Methodology

A content analysis was conducted to investigate these questions. This section describes the research procedures, provides details regarding the sampling and classification, and explains the enumeration of hierarchical relationships.

Procedures

To evaluate semantic relationships between user generated descriptors, this study examined in detail a corpus of URLs with their user-generated subject tags.

A sample was selected from the web site del.icio.us, a popular folksonomy site as measured by numbers of registered users. The number of users is a secret, but certainly in the thousands. Del.icio.us provides an application programming interface (API), which enables convenient custom access to specific aspects or facets of the database. In many cases these are not easily or directly accessible through the web-based user interface. The API was employed in order to access, organize, and store higher volumes of information than would be possible to do by hand, clicking on links.

The API was accessed using python source code called "delicious.py", version 0.2.5 (Timmermann 2005). The script provides the base functionality required to query user, tag, and url information from the web site. Furthermore this study created a custom wrapper for delicious.py, called ds.py, or "DS". DS contains routines for the automated extraction of data from del.icio.us by repeated calls to the API.

The information structure at del.icio.us is represented by file system directories. File system directories are used to organize users, tags, URLs, "general" URLs, and more. Users register usernames with del.icio.us in order to post bookmarks. Each registered username has a directory under the root directory and each user's tag is contained in a subdirectory therein. As an example: "http://del.icio.us/skome/library" will

display the URLs for which user "skome" has used the tag, "library". Tags also have their own global directory, e.g; "http://del.icio.us/tag/library". URLs similarly have directories, given as "/url/(MD5 encoding of URL)". An important additional component of the site is a directory /rss/ under which all of the above content may be found, encoded in RDF format. A discussion of RDF is beyond the scope of this paper, suffice to say that it is a convenient format to parse. Wherever possible, this is the resource DS uses.

For the purposes of this study, each record at del.icio.us is treated as a "post" created by a registered user of del.icio.us. Each post consists of several components, any of which can be used as criteria for retrieval:

- URL: a unicode (text) representation of a uniform resource locator, e.g.;
'http://www.census.gov'
- Description: Free text entered by the del.icio.us user who created the record at del.icio.us. This is the text displayed as the title of the item.
- Tags: The list of single word descriptors entered by the user who created the record at del.icio.us
- Extended: Free text entered by the user who created the record at del.icio.us. This text is displayed as an item description.
- Date/Time: The date and time the del.icio.us record was created.
- User Name: The del.icio.us username responsible for creating the record.

The routines in DS perform the following tasks:

1. Retrieve URLs by tag

Given a tag as an input, DS can output a text file containing the maximum allowed number of URLs (31).

2. Retrieve Tags by URL, with tag frequencies

Given a URL as an input, DS can output a text file containing the URL and a list of each tag found to be associated with that URL, accompanied by the frequency of use for each tag.

3. Retrieve Whole Post by Tag

Given a tag or list of tags, DS can output text files containing whole posts. For example, given the tag, "flounder", DS will generate one text file containing the (at the time of this writing) one (1) post tagged by that term. Given the tag, "python", for which there are more than one hundred posts, DS will retrieve only the 31 which are available as RSS feeds at any given moment. For the purposes of this study that selection is presumed to be of the more recently posted records. Appendix A shows a sample of the data retrieved.

Sampling

Using DS to gather posts, two collections of posts were created, the general and the subject-specific. General posts were taken from the most popular URLs, i.e.; those with the highest user counts. These are URLs which can be found in the /popular/ directory of del.icio.us. Subject specific posts consisted of those where the tag, 'mathematics' was used; these can be found in the '/tag/mathematics/' directory.

The two collections initially totalled 60 URLs. A few of these were held in common (i.e.; general posts tagged, 'mathematics'). In these cases the URL was removed from the mathematics collection. The remaining URLs in each collection were shuffled and 15 were chosen from each list. The two sets of 15 were combined into a list of 30 URLs and shuffled together. The shuffling was performed by a software library built into the python programming language (see <http://docs.python.org/lib/module-random.html>).

The resulting file consisted of 30 URLs in random order with regards to subject specificity.

Classification

In order to establish a measure of tag accuracy, 10 library science students were enlisted to classify the URLs. Eight students were ultimately recruited on the basis of having completed INLS 151 at the UNC School of Information and Library Science. In this course students learn,

“formal systems for description, access, and subject cataloging including AACR2, MARC, Dewey Decimal classification, Library of Congress Classification, and subject headings.” (UNC)

Each student was offered ten dollars to compensate for an estimated one hour of work. Each student was assigned a login name and password for a web-based form through which they were to perform the classification task (Appendix B, Figure 1). None of the students were made aware by the investigator of each other's identities. They were able to complete the task at their own pace and in the place of their choosing, provided that place had a computer with an internet connection and a web browser.

The list of 30 URLs given to each student was presented on the web-based form as a list of integers from 1 to 30. (Appendix B, Figure 2). The integers were themselves hyperlinks to each of the URLs. Alongside each link was a checkbox and a text field. Instructions on the form indicated that the cataloger should click on each number to display an information resource. In some cases the link led to a single document in Adobe Acrobat Portable Document Format. In other cases to a web page or site.

Catalogers were instructed to check the box for each link if they determined that it should be classified under a Mathematics subject heading. The notes field was provided

for any comments or questions they might have. When finished, the student clicked on a button to submit the form. The information for each student was recorded, associated with their login name. The student's actual name was not recorded with their results.

Enumeration

To determine the presence of hierarchical relationships between tags, a file was generated that contained entries for each URL (from those examined) and the tags retrieved for each of the posts found for that URL. The number of posts and tags varied independently; some URLs had the maximum number of posts (30); others only one. Some posts had no tags, others had more than a dozen. The tags were considered for the purposes of this study to not be pre-coordinated; the user's order entry was ignored and the tags were retrieved in alphabetic order, ascending. A spreadsheet was generated and the number of posts and tags per each URL was recorded. (Appendix B Figure 3)

Working through the final file, the investigator recorded the number of hierarchical relationships per post in a second spreadsheet (Appendix B Figure 4). For the purposes of this study, hierarchical relationships were determined according to the definitions of semantic relationships set forth by ANSI/NISO Z39.19-2005. The number of hierarchical relationships was determined by a heuristic:

- 1) Starting with the first tag in the list, compare the tag with the next tag.
- 2) If no other tags exist, stop.
- 3) If the next tag is a superordinate or subordinate term, add one to the count for the current post.
- 4) The super-sub relationship judgement was guided by the list of hierarchical relationships in the ALA's ALCTS Subcommittee on Subject Relationship and Reference

Structures Checklist of Candidate Subject Relationships for Information Retrieval (ALA 1997).

5) Repeat the evaluation for each subsequent tag in the post.

6) Take the next (n-th) tag in the list. Repeat from step 2, above.

The result is an enumeration of relationships among the tags at the top level.

Example:

URL

“<http://ww.nikonsmallworld.com/gallery.php?grouping=year&year=2005&imageid=1>”

The first post (out of 30) contained the following 6 tags: “art, competition, images, micro, photo, science”.

“Art” is neither a broader (BT) nor narrower (NT) for “competition”. In fact, art is a discipline, and competition is a process. These are differing *kinds* of concepts, and so no hierarchical relationship is possible (NISO, 47)

Moving to “art” and “images”, the relationship is process to product, which is a type of associative relationship.

The first (and in this case only) hierarchical relationship found in this post is “image(s)” as a broader term for “photo”. Note that stemming for plurals was used, but not for other word endings. Word endings such as “ization” can indicate a process, which could flag an associative relationship.

The result of the count was a spreadsheet having a count of relationships for each URL by tag by post. Only posts with more than one tag were considered in this enumeration. A result of 0 indicates that multiple tags were present – the post was eligible – but no hierarchical relationships were found. A missing result indicates that 0 to 1 tags were present, therefore no hierarchical relationships were *possible*.

Analysis

Data for 30 URLs was retrieved from del.icio.us comprising 647 posts which in turn contained a total of 2,114 individual tags, for an average of 3.27 tags per post. 459, or 71% of the posts contained more than one tag, which is the pre-requisite for finding inter-tag relationships. These posts held 1968, or 93% of the total tags. 46% (210) of these posts contained tag relationships; a total of 415.

Classification of URLs: General and Mathematics

The results of the students' cataloging efforts (Appendix A, Figure 5) were compared and tested for agreement with the initial classification at del.icio.us, and for inter-rater reliability. The students agreed with each other 85% percent of the time at a 95% confidence interval. They agreed with the initial classification 85% of the time for each URL. Comments left by the students indicated their wishes for sub-headings of mathematics such as "Math Education", and "Statistics". There was one and only one URL that had been tagged mathematics which none of the catalogers indicated. None of the comments indicated any difficulty performing the task.

The tag, 'mathematics' was confirmed to be sufficient to differentiate URLs into a group separate from the general URLs.

Grouping URLs by Tag Counts:

Mathematics URLs consistently came with fewer tags than their general counterparts, but they were also posted less frequently. The sum of tags for general URLs came out higher, as seen in Table 1. The average number of tags for general posts was more than twice that of math posts. An independent T-test confirms the likelihood that the difference in means is due to the existence of two distinct groups.

Table 1: Tag Counts

M:Min	1
M:Max	126
M:Mean	37.466667
P:Min	80
P:Max	152
P:Mean	103.466667
T	0.0000

Grouping URLs by Post Counts:

Mathematics URLs returned fewer posts, with a mean post rate only 45% of the rate for general URLs, as seen in Table 2. An independent T-test confirms the likelihood that the difference in means is due to the existence of two distinct groups, and the indications are that the general URLs have more posts than mathematics URLs.

Table 2: Post Counts

M:Min	1
M:Max	30
M:Mean	13.3333
P:Min	28
P:Max	30
P:Mean	29.8
T	0.0001

Grouping URLs by Tag Relationships:

This test sought to discover a higher number of tag relationships for mathematical resources. Analysis of the tag relationships was conducted by summing the number of tag relationships across posts per URL, then calculating the probability a random sample would find smaller differences

Table 3: Relation Counts

M:Min	0
M:Max	20
M:Mean	5.7333
P:Min	9
P:Max	79
P:Mean	22.2667
T	0.0061

between "general" and "mathematics" resources using an independent t-test. The results show that random sampling would lead to smaller differences 59% of the time; not a

statistically significant result. In the discussion below, mitigating factors will be presented which may have heavily influenced this result.

Grouping URLs by Tag Relationship per Tag Counts.

This test sought to explore the effect of differing numbers of tags per post. The tag relationship counts were divided per post by the number of

tags, obtaining the number of relationships per tag. The resulting numbers were summed by URL, and averaged by the number of posts per URL. The goal was to assess the number of tags used to generate a given number of

Table 4: Normed Relation Counts

M:Min	0.0000
M:Max	56.0000
M:Mean	9.3095
P:Min	1.7500
P:Max	25.7300
P:Mean	11.2798
T	0.6214

relationships; for some posts a greater number of tags yielded few if any relationships, and so forth. The independent t-test found less than a 38% likelihood that random samples would find smaller differences than those present in the data set. The general URLs appear to be tagged in a hierarchical fashion slightly more often than those in the narrower knowledge domain of mathematics.

Discussion

The data provide several new insights into the nature of modern folk taxonomies. Hierarchical relationships in tag sets were common, found in 45 percent of eligible posts across all URLs without regard to subject area groupings. That percentage climbs when only posts with higher than the average number of tags per URL are considered. For some URLs, the majority – up to 90% -- of tags (per post) are members of hierarchies. While the subject domain of these high performing URLs could not be authoritatively classified within the scope of this study, they appeared to be elements of narrow knowledge domains.

Agreement of the catalogers with the tag, “mathematics” is an example of how bottom-up and top-down categorization schemes can produce equivalent results. To the extent that there was disagreement, the comments generally indicated the desire for a subordinate heading; not that the tag was a misclassification.

The groupings revealed no significant difference in proportion of relationship counts between general URLs and URLs of a specific subject, but the significance tests were affected by a relative lack of posts in the mathematics resources.

The logic behind choosing the popular URLs was that the most often posted URLs would relate to the most general subject matter. The term 'mathematics' was chosen because it is a superordinate term under which are several species level terms (calculus, geometry, algebra, etc). The tagging behavior was seen to be measurably different between the groups, with more tags and more tag relationships (raw count) given for general URLs. This finding depends on the fact that there were fewer mathematics posts; another experiment with equivalent post numbers could yield different results.

Limitations

Accessing any of this information via the API was subject to some important limitations. First of all, the API will return a limited number of records; far fewer than are actually stored at the site. The number of records is not documented, indeed the API is not comprehensively documented and this is a limitation of this study. This study cannot provide a reference to the exactly how the API behaves. For the duration of this study, the maximum observed number of records retrieved by any query to the API was 31.

A related limitation, one which is simply a consequence of the dynamic nature of the data, is that repeated queries will return differing results. Repeatability can only be achieved by storing results locally and working with the local cache. Additional information regarding the API can be found at del.icio.us/doc/api.

Observations during the tag and relationship enumeration processes point to potential biases in the data which were very likely significant to the outcome. The first such observation is the predictable bias towards computer science related URLs. Web users are by definition computer users, and a very large majority of computer science professionals are web users. Of the general URLs, it appeared that computer science-related links were over-represented. The second observation was the relatively higher number of tags and tag relationships for URLs which were apparently relevant to computer science resources. "Apparently relevant" is used because computer science was not chosen as a subject for evaluation by the student classifiers; mathematics was. That said, URLs with such tags as, "programming", "database", "javascript", "DOM", etc, appeared to have comparatively higher tag and relationship counts than others in the general cohort. If it is the case that *any* specific knowledge domain within a folksonomy would present higher rates of tag (subject) hierarchies, and there is a strong presence of another specific knowledge domain (computer science) in the general results, this would tend to prevent differentiation via rate comparison of hierarchical relationships among tags.

Conclusion and Future Research:

The results found that the consensus opinion that folksonomies generate flat categorization schemes is not accurate. This finding depends on the user interface and

API implementations, which allow for the retrieval of tag lists that are cross linked to information resources such that superordinate and subordinate search terms can be seen and used simultaneously, as in a structured subject heading list. While this is not perhaps the traditional view of the subject heading list, the presence and visibility of hierarchical terms, provide very similar value to a formal list, minus the strong visual patterning humans generally find valuable (Mann 2000).

Today's subject enthusiasts' classification efforts indicate that users in a collaborative computing environment can create valuable metadata. Greenberg (2001) found that retrieval augmented by automatic query expansion could increase recall and precision. The query expansion system required synonyms, narrower and broader terms; grist which could be mined from sites like del.icio.us. It may be beneficial for thesauri developers to examine the semantic relationship found in Folksonomies for terms and term relationships that might be incorporated into controlled vocabulary tools, specifically thesauri. The converse is true as well: as more subject thesauri become available in electronic format and people consider their use for digital repositories, it may make sense to consider their value for environments where Folksonomies are being developed.

Further Research

The depth of any of the relationships (one/two/three/n steps) could show that the folk classification schemes are closer to structured subject heading lists than to unstructured lists. The depth and types of relationships are also important data for comparison to taxonomies.

The relationship inventory contains several outliers which indicate that there are types of URLs (information resources) which are more commonly tagged hierarchically.

Which types?

The inventory suggests the existence of differentiable user types, for example these two posts from two different users for the same URL:

1. “crazy, writing”
2. “christianity, essay, god, reference, religion, secular”

User #1 used two tags, very casual language with low semantic value and no hierarchical relationships. User #2 used six tags, more precise language with hierarchical terms. How may these user types be automatically identified?

Could a system of authority grants (Russell 2005) be deployed such that taggers who generate higher rates of hierarchical (or otherwise useful) tags are given more weight by a search engine?

The percentage of hierarchical relationships per URL increases with the average number of tags per post. Is there a relationship between the number of tags and the quality or accuracy of topical representation? The catalogers agreed with the tag, “mathematics” despite relatively low tag rates. Are there patterns of tag relationships according to subject matter? For example, what are the hierarchical, associative, or equivalence tag relationships for URLs which bear on a particular medical protocol? Are these different from the relationships for generic URLs?

Folktaxonomies are growing quite rapidly and illustrate a means of metadata that is being adopted with enthusiasm by resource authors and users/viewers. This development provides a new information landscape and one that may provide a means of

addressing some of the most pressing challenges of how to better organize and classify digital resources in the future.

Appendix A

<http://www.nikonsmallworld.com/gallery.php?grouping=year&year=2005&imageid=1>

Class: general

Post 0/30:6

art competition images micro photo science

Post 1/30:2

cool photos

Post 2/30:2

photos science

Post 3/30:3

art photography science

Post 4/30:11

art cool design gallery images photo photography photos pic science small

Post 5/30:2

macro photography

Post 6/30:2

photos science

Post 7/30:1

photography

Post 8/30:9

gallery images insects macro photography photos pictures science world

Post 9/30:8

art cool gallery images interesting photo photography science

Post 10/30:3

kids photography science

Post 11/30:1

photography

Post 12/30:1

photography

Post 13/30:9

art competition cool design photo photography photos pictures science

Post 14/30:3

art photography science

Post 15/30:3

collection microscope pic

Post 16/30:2

fotos proposta

Post 17/30:1

photos

Post 18/30:3

art cool science

Post 19/30:1

photography

Post 20/30:4

Gallery Images Photo Science

Post 21/30:4

art design photography science

Post 22/30:0

Post 23/30:2
art photography

Post 24/30:2
photography science

Post 25/30:6
competition gallery images micro photos small

Post 26/30:4
art photography photos science


Post 27/30:2
images science

Post 28/30:3
art photography science

Post 29/30:3
Cool Fotos Images

Average Number of Tags/PopPost: 3

Appendix B



School of
Information and Library Science
The University of North Carolina at Chapel Hill

INLS 392: URL Soup

ID:

password

Instructions

Enter the ID and password provided to you.

On the next page, click on the numbers in the left hand column. A new browser window will open with a new web page. Use your knowledge of cataloging to decide whether or not the subject of each link is mathematical in nature.

More precisely, would the information be of value to someone engaged in a mathematical field? Would a math student, professor, researcher, theorist find it useful, interesting, inspiring?

If the material is judged to belong to a mathematics subject heading, check the box next to the number.

The notes field may be used to provide an alternative subject heading, ask a question, make a comment, etc.

Your time and effort are very much appreciated!
Sites chosen at random, no responsibility will be accepted for aesthetic or other offenses.

[Home](#)

Figure 1: Classification Login with Instructions



School of Information and Library Science
The University of North Carolina at Chapel Hill

INLS 392: Tag Analysis: URLs

Click on the numbers in the left hand column to open the link in a new window.

Use your knowledge of cataloging to decide whether or not the subject of each link is mathematical in nature.

If the material is judged to belong to a mathematics subject heading, check the box next to the number.

The notes field may be used to provide an alternative subject heading, ask a question, make a comment, etc. Your time and effort are very much appreciated! Links chosen at random, no responsibility will be accepted for omissions or other offenses.

Would the information be of value to someone engaged in a mathematical field? (you'd a math student, professor, researcher, science teacher, etc.) (yes/no, increasing, tagging?)

URL id	Mathematics?	Notes
1	<input type="checkbox"/>	
2	<input type="checkbox"/>	
3	<input type="checkbox"/>	
4	<input type="checkbox"/>	
5	<input type="checkbox"/>	
6	<input type="checkbox"/>	
7	<input type="checkbox"/>	
8	<input type="checkbox"/>	
9	<input type="checkbox"/>	
10	<input type="checkbox"/>	
11	<input type="checkbox"/>	
12	<input type="checkbox"/>	
13	<input type="checkbox"/>	
14	<input type="checkbox"/>	
15	<input type="checkbox"/>	
16	<input type="checkbox"/>	
17	<input type="checkbox"/>	
18	<input type="checkbox"/>	
19	<input type="checkbox"/>	
20	<input type="checkbox"/>	
21	<input type="checkbox"/>	
22	<input type="checkbox"/>	
23	<input type="checkbox"/>	
24	<input type="checkbox"/>	
25	<input type="checkbox"/>	
26	<input type="checkbox"/>	
27	<input type="checkbox"/>	
28	<input type="checkbox"/>	
29	<input type="checkbox"/>	
30	<input type="checkbox"/>	

[Home](#)

Figure 2: Data Entry Screen (numbers are hyperlinks) .

	p1	p2	p3	p4	p5	p6	m1	p7	p8	p9	p10	m2	m3	m4	m5	m6	m7	p11	p12	m8	m9	m10	p13	m11	p14	m12	m13	p15	m14	m15
post1	6	4	1	4	4	3	2	7	0	2	0	4	8	3	7	4	2	0	3	1	4	2	0	4	6	3	6	2	4	5
post2	2	2	3	2	8	0	1	19	3	2	3	4	2	7	5	3	2	2	2	1	2	3	4	3	1	6	4	1	1	
post3	2	6	5	1	4	3	2	2	2	7	3	2	3	3	5	3	2	5	3	2	8	0	1	2	2	1	1	1	1	
post4	3	6	5	4	1	0	1	5	1	4	1	5	2	3	3	3	4	3	3	4	4	11	1	3	2	4	2	1	1	
post5	11	2	2	7	6	4	1	7	4	1	3	3	2	2	1	1	0	1	1	0	1	1	0	4	3	14	2	5	2	4
post6	2	1	1	0	0	4	3	2	5	2	8	2	1	6	7	2	6	7	2	6	5	2	2	6	5	2	2	0	2	1
post7	2	10	0	3	5	9	6	3	4	1	3	6	1	1	0	17	1	0	17	7	1	2	7	1	2	1	3	1	1	
post8	1	3	0	2	1	1	8	6	4	1	2	2	6	1	3	6	1	3	6	1	1	4	1	1	4	4	0	1	1	
post9	9	3	1	2	3	13	3	2	2	0	2	1	5	3	4	1	3	4	1	3	4	1	11	1	2	2	12	1	1	
post10	8	0	3	6	2	5	1	3	2	6	1	5	2	2	3	5	2	3	5	11	1	2	11	1	2	1	8	1	1	
post11	3	4	2	4	3	4	1	6	4	4	1	2	2	3	0	2	3	0	2	3	3	1	3	3	1	5	7	4	1	
post12	1	4	3	4	2	7	1	2	1	8	1	6	3	4	2	2	3	4	2	2	4	3	1	6	2	1	6	2	2	
post13	1	6	3	2	3	7	4	3	0	5	1	3	1	1	1	4	1	1	4	5	1	1	4	5	1	1	2	3	3	
post14	9	2	8	5	1	4	1	1	0	15	4	11	1	1	3	2	1	3	2	0	2	1	1	1	1	1	9	0	0	
post15	3	2	3	8	0	9	5	3	0	6	4	2	1	1	1	2	2	7	4	0	6	6	3	1	1	1	1	1	1	
post16	3	3	3	7	4	3	1	0	3	2	0	10	2	1	1	2	1	1	2	1	1	2	0	6	0	2	1	1	1	
post17	2	2	1	1	3	6	1	5	3	2	4	2	8	3	4	5	3	4	5	1	7	3	7	1	7	1	2	1	2	
post18	1	1	1	4	1	4	2	5	7	2	2	4	4	3	3	4	4	3	4	4	5	19	2	14	1	1	1	1	1	
post19	3	2	3	5	14	4	1	3	4	1	1	1	1	3	0	2	3	0	2	3	3	1	5	7	4	1	1	1	1	
post20	1	6	2	2	1	5	1	1	3	0	2	3	3	1	1	7	1	1	7	2	2	1	4	3	1	1	1	1	1	
post21	4	1	1	0	1	1	1	4	3	1	1	8	1	1	1	8	2	6	4	1	7	1	1	7	1	1	1	1	1	
post22	4	3	2	2	4	0	2	1	3	7	2	2	2	2	3	2	2	3	2	4	3	2	11	3	1	1	1	1	1	
post23	0	3	2	10	6	3	4	2	8	7	1	10	1	1	1	1	3	2	4	3	2	11	3	1	1	1	1	1	1	
post24	2	2	3	10	0	7	3	1	2	1	3	4	3	1	5	2	1	5	2	4	7	3	2	3	1	1	1	1	1	
post25	2	4	1	0	1	2	1	3	6	3	4	3	4	1	3	1	1	3	1	0	7	5	1	30	2	1	1	1	1	
post26	6	3	2	2	1	2	6	2	3	1	2	5	5	2	2	4	2	2	4	3	2	6	5	1	1	1	1	1	1	
post27	4	3	8	3	4	1	2	1	8	2	5	5	5	1	0	4	1	0	4	1	0	4	3	7	1	1	1	1	1	
post28	2	3	4	0	1	2	4	3	1	5	3	3	3	0	1	1	1	1	1	0	3	10	1	4	0	1	1	1	1	
post29	3	14	5	4	4	4	4	4	1	2	3	3	3	6	1	1	7	0	2	6	4	3	5	6	2	2	2	2	2	
post30	3	2	6	1	3	3	2	4	4	3	2	2	2	1	7	0	1	7	0	2	4	10	3	2	2	2	2	2	2	
TotalTags	103	104	80	112	92	115	24	100	80	87	108	53	126	16	7	4	61	86	99	1	4	2	100	106	134	11	87	152	15	45
TotalPosts	30	29	30	30	30	30	8	28	30	30	30	23	30	4	1	1	30	30	30	1	1	1	30	30	4	30	30	6	30	
AvgNumTags	3.43	3.59	2.67	3.73	3.07	3.83	3.00	3.57	2.67	2.90	3.60	2.30	4.20	4.00	7.00	4.00	2.03	2.87	3.30	1.00	4.00	2.00	3.33	3.53	4.47	2.75	2.90	5.07	2.50	1.50
TotalMultiTag																														
MaxTags																														

Figure 3: Raw Tag Counts

	p1	p2	p3	p4	p5	p6	m1	p7	p8	p9	p10	m2	m3	m4	m5	m6	m7	p11	p12	m8	m9	m10	p13	m11	p14	m12	m13	p15	m14	m15
p1	1	0	0	0	0	0	0	1	2	0	0	1	1	1	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p5	2	0	0	0	0	2	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p6	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	0	4	1	0	0	0	0	0	0	0	0	0	0	0
p7	0	1	0	0	1	2	1	1	1	0	0	3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
p8	2	0	0	0	1	2	0	1	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p9	2	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	1	0	0	15	0	0	0	0	0	0	0
p10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	1	1	1	1	1	1	1
p11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p12	1	0	0	0	1	0	0	0	0	0	3	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
p13	3	2	0	0	1	0	0	1	2	0	1	2	1	0	0	0	0	0	0	0	0	6	0	0	1	0	0	0	0	0
p14	1	0	4	0	0	0	0	0	0	0	6	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p15	0	0	1	0	2	0	3	1	2	1	0	2	1	0	0	1	1	0	1	1	0	0	0	1	1	1	1	1	1	1
p16	0	0	1	1	2	0	0	0	2	0	2	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p17	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0
p18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
p19	0	0	2	0	4	1	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
p20	1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p21	2	0	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	3	0	0	0	0	0	0	0	0
p22	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
p23	0	0	0	3	2	0	0	0	0	4	3	2	2	0	0	0	0	0	0	0	0	1	0	2	0	0	0	0	0	0
p24	0	0	1	5	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	6	1	0	0	0	0	0	0	0
p25	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
p26	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	1	0	0	0	0	0	0
p27	0	0	3	1	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	2	1	1	0	0	0	0	0</	

References

- ALA Subcommittee on Subject Relationships/Reference Structures. (1997). Appendix B, final report to the ALCTS/CCS subject analysis committee. Retrieved November 17, 2005 from <http://www.ala.org/ala/alctscontent/catalogingsection/catcommittees/subjectanalysis/subjectrelations/finalreport.htm>
- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3), 409-429.
- Armstrong, S. L., Gleitman, L. R., & Gleitman, H. (1983). What some concepts might not be. *Cognition*, 13(3), 263-308.
- Berlin, B. (1992). *Ethnobiological classification : Principles of categorization of plants and animals in traditional societies*. Princeton, N.J.: Princeton University Press.
- Berlin, B., Breedlove, D. E., & Raven, P. H. (1973). General principles of classification and nomenclature in folk biology. *American Anthropologist*, 75(1), 214-242.
- Berlin, B., Breedlove, D. E., & Raven, P. H. (1966). Folk taxonomies and biological classification. *Science*, 154(3746), 273-275.
- Berlin, B., Breedlove, D. E., & Raven, P. H. (1974). *Principles of tzeltal plant classification; an introduction to the botanical ethnography of a mayan-speaking people of highland chiapas*. New York: Academic Press.
- Chan, L. M. (1989) "Inter-indexer consistency in subject cataloging". *Information Technology & Libraries*, 8(4), 349-358

- Connell, T. H. (1995). Subject searching in online catalogs: Metaknowledge used by experienced searchers. *Journal of the American Society for Information Science*, *American Society for Information Science*, 46(7), 506-518.
- Dunlop, C. E. M., & Fetzer, J. H. (1993). *Glossary of cognitive science* (1st ed.). New York, N.Y.: Paragon House.
- Greenberg, J. (2001). Automatic query expansion via lexical-semantic relationships. *Journal of the American Society for Information Science and Technology*, 52(5), 402-415.
- Kay, P. (1971). Taxonomy and semantic contrast. *Language*, 47(4), 866-887.
- Lakoff, G. (1987). *Women, fire, and dangerous things : What categories reveal about the mind*. Chicago: University of Chicago Press.
- Lancaster, F. W. (1986). *Vocabulary control for information retrieval* (2nd ed.). Arlington, Va.: Information Resources Press.
- Liddy, E. D., Sutton, S. A., Paik, W. Allen, E., Harwell, S. Monsour, M., Turner, A, & Liddy J. (2001). Breaking the metadata generation bottleneck: preliminary findings. *JCDL 2001*: 464.
- Mann, Thomas "Is Precoordination Unnecessary in LCSH? Are Web Sites More Important to Catalog than Books? : A Reference Librarian's Thoughts on the Future of Bibliographic Control" *Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium : Confronting the Challenges of Networked Resources and the Web*, Washington, D.C., Nov. 15-17, 2000, sponsored by the Library of Congress Cataloging Directorate, edited by Ann M. Sandberg-Fox. Washington, D.C.: Library of Congress, Cataloging Distribution Service, 2001 http://www.loc.gov/catdir/bibcontrol/mann_paper.pdf

- Mathes, A. 2005 Academic Works: Computer Mediated Communication. Folksonomies - Cooperative Classification and Communication Through Shared Metadata: <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html> Accessed 11/01/2005
- National Information Standards Organization Z39.19 Committee. (2005). ANSI/NISO Z39.19 -2005 guidelines for the construction, format, and management of monolingual controlled vocabularies. Retrieved November 17, 2005 from http://www.niso.org/standards/standard_detail.cfm?std_id=814
- Rosch, Eleanor. 1981. Prototype classification and logical classification: Two systems. *New trends in cognitive representation: Challenges to Piaget's theory*, ed. by E. Scholnick, 73-86. Hillsdale, NJ: Lawrence Erlbaum.
- Rosch, E., Lloyd, B. B., & Social Science Research Council. (1978). *Cognition and categorization*. Hillsdale, N.J.; New York: L. Erlbaum Associates; distributed by Halsted Press.
- Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7, 573-605.
- Rosenfeld L, 2005 Metadata Ecologies: http://www.louisrosenfeld.com/home/bloug_archive/000330.html Accessed 11/15/2005
- Ross, B. H., & Murphy, G. L. (1999). . *Cognitive Psychology*, 38(4), 495-553.
- Russell, T. (2005). Contextual authority tagging: Cognitive authority through folksonomy. Unpublished manuscript. Retrieved 11/16/2005, from <http://www.terrellrussell.com/projects/contextualauthoritytagging/conauthtag200505.pdf>

- Shirky, C. 2005. Clay Shirky's Writings About the Internet. Ontology is Overrated: Categories, Links, and Tags: http://shirky.com/writings/ontology_overrated.html
Accessed 11/01/2005
- Timmerman, F. (2005). Delicious.Py. World Wide Web: . Retrieved July 10 2005, from <http://delicious-py.berlios.de/>
- UNC School of Information and Library Science. (2005). SILS course descriptions. Retrieved November 16, 2005 from <http://sils.unc.edu/programs/courses/descriptions.html#151>
- Vakkari, P. (2002). Subject knowledge, source of terms, and term selection in query expansion: An analytical study. Proceedings of the 24th BCS-IRSG european colloquium on IR research, 110-123.
- Vander Wal, T. (2005a). Explaining and showing broad and narrow folksonomies. Retrieved November 16, 2005 from <http://www.vanderwal.net/random/entrysel.php?blog=1635>
- Vander Wal, T. (2005b). Folksonomy definition and wikipedia. Retrieved November 16, 2005 from <http://www.vanderwal.net/random/category.php?cat=153>