PEER-REVIEWED PAPER

Elise Conradi

to_be_classified

A Facet Analysis of a Folksonomy

Abstract

This paper examines the use of the postulational approach to facet analysis to manually induce a faceted classification ontology from a folksonomy. An in-depth study of faceted classification theory is used to form a methodology based on the postulational approach, which is then used to facet analyze a dataset consisting of over 107,000 instances of 1,275 unique tags representing 76 popular non-fiction history books collected from the LibraryThing folksonomy. Preliminary results of the facet analysis indicate the manual inducement of two faceted classification ontologies in the dataset: a completed ontology representing the domain of books and an incomplete ontology representing the domain of subjects within the domain of books. The grouping of tags into theoretically based facets and conceptual categories give new insight into how users describe information resources. Furthermore, the relationships discerned in the ontologies are user-generated relationships between tagged information items, representing a new form of knowledge. Practical implications of the results are discussed in terms of potential areas in which user-generated metadata can enhance faceted structures in information architecture.

Folksonomies: Unstructured "Wisdom of the Crowd"

Since their inception on the web in 2003 with the tagging system Del.icio.us[1], folksonomies have become a popular way to categorize large amounts of information resources. Folksonomies emerge from the aggregation of textual labels called tags that are affixed to digital objects of various formats by either the creator or the users of the objects within sites that allow for tagging. Vander Wal (2005) distinguishes between broad folksonomies and narrow folksonomies, explaining that a "broad folksonomy has many people tagging the same object" whereas in a narrow folksonomy, an object is tagged "by one or a few people."

Quintarelli (2005) discusses broad folksonomies in terms of the Power Law distribution, stating that the "power law reveals that many people agree on using a few popular tags but also that smaller groups often prefer less known terms to describe their items of interest." Halpin et al. (2007) argue that the short head of the long tail in the Power Law distribution represents a consensus of what users find to be most important about each information resource. The study of folksonomies can therefore provide invaluable insight

[1] http://del.icio.us

to information organization professionals by unearthing what Weinberger (2006) has called the “wisdom of the crowd”.

Folksonomies have been criticized by those advocating top-down approaches to organizing information resources. It is argued that the uncontrolled vocabulary of tags causes too many recall and precision problems (primarily due to ambiguity, polysemy and synonymy) to make them useful as information retrieval tools, and that the flat structure of folksonomies prevent users from seeing valuable relationships between information items (Rosenfield, 2005; Petersen, 2006). In response to the latter critique, this paper illustrates how a facet analysis of a broad folksonomy based on the postulational approach can reveal underlying conceptual categories and facets to which the folksonomy’s aggregated tags belong. In this way, facet analysis techniques are used to manually expose a faceted classification ontology in the flat tag space, thus revealing user-generated relationships between information items.

Related Work

There are several studies and projects that have examined the use of faceted classification techniques for the organization of folksonomies. Weaver (2007) studied the tagging practices of a library community in order to glean facets to aid in information retrieval. Quintarelli, Resmini and Rosati (2007) introduced “Facetag”, a tagging system that allows users to choose tags within predefined facets in order to improve retrieval. Lichtblau, Trice and Wartik (2006) proposed a prototype social classification system, in which users describe services within a specific domain of the Department of Defense from seven different facets based on “the 7 W’s”[2]. Siderean launched the wonderfully named but short-lived site Fac.etio.us[3] in 2005, in which tags were automatically grouped into predefined facets. Other commercial enterprises combining the use of tags with facets are Buzzillions[4], Peter Van Dijck’s brainchild MeFeedia[5], and Raw Sugar[6], a “guided, tag-based search engine”.

Spiteri (2010) analyzes several of these attempts and concludes that, although “a number of studies exist in which facets have been applied quite successfully to social tagging applications, (…) none explain clearly the theoretical frameworks or methodologies used to derive the facets, nor do they address any strategies by which to enable end users to evaluate the usefulness and applicability of these facets” (p. 105). Indeed, with the exception of Weaver (2007), all of the above studies are largely focused on the improvement of faceted navigation and information retrieval through the

[2] 1. Who uses the service? 2. What does the service do? 3. On what does the service act? 4. To whom is the service generally directed? 5. Where is the service used? 6. When is the service used? 7. Why is the service used?

[3] Fac.etio.us, by way of the Internet Archive Wayback Machine: http://web.archive.org/web/20060526050202/demo.siderean.com/facetious/facetious.jsp

[4] http://www.buzzillions.com

[5] http://www.mefeedia.com

[6] http://www.rawsugar.com

placement of tags into predefined facets. The research illustrated in this paper takes a fundamentally different approach. Here, there are no predefined facets; focus is rather placed on discovering precisely which types of facets, conceptual categories and ontological relationships will emerge in a given folksonomy after being subjected to a facet analysis based on the postulational approach. Although the nature of this research is primarily theoretical, a number of practical implications of the results will be discussed in the final section.

The Faceted Classification Ontology

Central to the methodology used in the research presented here is faceted classification. The introduction and development of faceted classification in library and information sciences arguably represents a Kuhnian paradim shift within knowledge organization (Dahlberg, 1992: Xiao, 1994). Previously, although pragmatic by purview, library classificationists had been highly influenced by traditional philosophical classifications of knowledge, adapting the ontological view that knowledge can be divided into neat, hierarchical categories (Abrera, 1974, p. 21).

The first library classification systems all reflect a top-down one-place-for-everything ontological view of the universe of knowledge, most commonly depicted as an upside-down hierarchical tree-like structure. Like traditional library classifications, faceted classification is pragmatically based, but it is grounded in theory and it represents an entirely new ontological perspective within knowledge organization in which information resources can simultaneously be represented by a number of different perspectives. Its ontological structure has been shown to be both scalable and highly suitable to digital environments (Ingwersen & Wormell, 1992; Ellis & Vasconcelos, 1999) and the use of facets to organize information has become commonplace on the web today.

The underlying ontological model of a faceted classification system is primarily based on two distinct concepts existent within any particular universe[7]: facets and conceptual categories. Facet is succinctly defined by Classification Research Group member Mills (1960) as “the total subclasses (in a class) resulting from the application of a single principle of division” (p.8). Conceptual categories, on the other hand, represent broader characteristics that may be applied to all the classes in a given universe. It appears, however, that the concept of facet is sometimes confused with the concept of conceptual category. Furthermore, there is dispute concerning what facets may describe.

The former seems particularly to be the case in recent literature intended towards information organization professionals. In the third edition of Morville and Rosenfield’s oft-cited reference book, *Information Architecture for the World Wide Web* (2007), for example, the term facet completely replaces the term conceptual category in the discussion of faceted classification:

[7] Ranganathan defines universe as a “collection of entities, without any special arrangement among them, (and that is) under consideration in a given context” (Ranganathan, 1967, p. 80). It is here used as a synonymous for domain.

"(Ranganathan) suggested five universal *facets* to be used for organizing everything" (p. 221, my italics). Likewise, an article in the peer-written information architecture magazine *Boxes and Arrows* describes "the fundamental *facets* that Ranganathan developed" (Steckel, 2002, my italics). The misapplication is also found in (Uddin & Janecek, 2007), (Rabourn, 2003) and (Redmond-Neal, n.d.).

Although it is quite conceivable that the misapplication of the terms is intentional for the sake of simplicity, the distinction between the two has important implications. Facets are used to differentiate between aspects of each individual class in a universe, while categories differentiate between aspects of all of the classes equally in a universe. If facets are assumed to be equal to categories, one loses the distinction between the level of universe and the level of classes in the universe, thus requiring facets to differentiate from a more general level. Figure 1 illustrates the simplified relationships between foci, facets and conceptual categories that make up a faceted classification ontology:

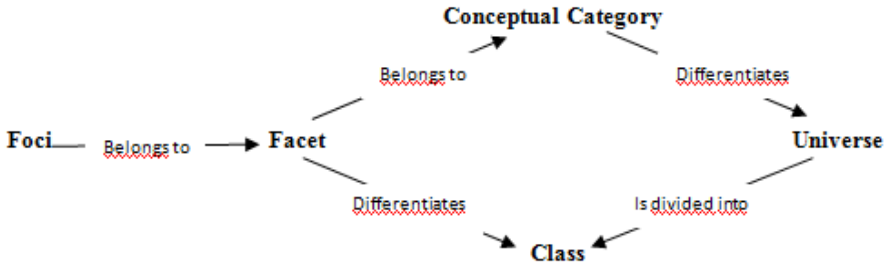

Figure 1: Ontological relationships between foci, facet, conceptual category, universe and class

The concept of facet has also caused confusion regarding what they may or may not describe. Can facets be facets of *anything*? Broughton (2006) writes about the purist view of faceted classification, which maintains that facets should be regarded as facets of subjects, as they are in library classifications. From this perspective, most of the facets on the web today are not regarded as faceted classifications, as these primarily describe objects. Schwartz (2008) points out what seems to be the crux of the problem: the term facet is used differently "in information architecture (IA) and guided navigation, where 'topic' is one among many facets, (than it is in library science and) thesaurus development, where 'topic' is the primary object of facet analysis." The research illustrated in this paper assumes that the concept of facet can be applied to aspects of any universe, regardless of whether the universe is subject-based or object-based. This is consistent with Ranganathan's faceted classification theory (Ranganathan, 1967, p. 567).

The Postulational Approach to Facet Analysis

The postulational approach to facet analysis refers to a methodology used for both the creation (by a classificationist) and subsequent usage (by classifiers)

of a faceted classification scheme. It was introduced by the Indian librarian and mathematician Shiyali Ramamitra Ranganathan, widely considered to be the father of faceted classification, in the second edition of his *Prolegomena to Library Classification* (1957) and further developed in the third edition (1967). The Classification Research Group (CRG), a UK-based group of information organization professionals established to "discuss the principles and practice of bibliographic classification" (Vickery, 1966, p.10) continued to develop the approach throughout the 1960s and 1970s.

The approach has informed much of the work on faceted classification during the twentieth century through today, and it provides the theoretical underpinnings of the research presented in this paper. It is based on a set of normative rules in the *Prolegomena* consisting of 43 Canons of Classification, 12 Postulates and 22 Principles. Of these, seven Canons and three Postulates pertain to facet choice, while the rest concern facet sequence and are not relevant for this research.

In the postulational approach to facet analysis, a classificationist facet analyzes a given universe under the guidance of Ranganathan's Canons of Classification[8], yielding a faceted classification representing the universe. The classificationist then proposes postulates to guide classifiers in the identification of corresponding facets in the objects to be classified according to the scheme. In the facet analysis of a folksonomy performed in the research illustrated in this paper, the classificationist is absent. Instead, the facet analysis is performed directly on users' descriptions of objects within a universe. This is essentially the reversal of the classificationist's process of facet analysis, the implication of which is that the facets found in the folksonomy are truly inductive, representing an aggregated users perspective of the universe.

Ranganathan proposed three postulates to guide in the choice and identification of facets. These were analyzed in light of relevant theoretical discussions they elicited by members of the CRG and used as the basis upon which to form the methodology used in this research.

1. The Postulate of Fundamental Categories

Ranganathan postulated that five conceptual categories exist in every universe: Personality, Matter, Energy, Space and Time (PMEST). The CRG argued for a more flexible use of conceptual categories, stating that the number and type of categories may vary from universe to universe and that "any such list of fundamental categories should not be used mechanically and imposed upon the subject, but to use it as a provisional guide in approaching a new field can be helpful" (Vickery, 1960, p. 24).

2. The Postulate of Basic Facet

[8] The Canons of Classification are the crux of faceted classification theory. They provide strict rules for the division of any universe into its core facets. They are thus responsible for the ontology representing any given universe. There are seven Canons that concern the choice of facets: 1. The Canon of Differentiation; 2. The Canon of Relevance; 3. The Canon of Ascertainability; 4. The Canon of Permanence; 5. The Canon of Concomitance; 6. The Canon of Exhaustiveness; 7. The Canon of Exclusiveness.

With this postulate, Ranganathan proposed that “every compound subject has a basic facet” and that “to identify the Basic Facet of a compound subject, a general knowledge of the schedules of Basic Subjects is necessary” (Ranganathan, 1967, p. 402). These constitute the main classes and the main subdivisions of each class in a classification scheme of subjects. In essence, this postulate seals the facets of the objects to be classified to the classification scheme; each facet is really a facet of a class within the schedule. The postulate can be extrapolated to apply to objects as well: every complex object has a basic facet, which is represented as a class in the classification scheme of objects. In the absence of a classification scheme, one of the major tasks of the facet analysis in this research was to discern the basic facets representing classes.

3. The Postulate of Isolate Facet

Here, Ranganathan posits that “each isolate facet of a compound subject can be deemed to be a manifestation of one and only one of the five fundamental categories” (Ranganathan, 1967, p. 403). This is a fairly straightforward postulate concerning the relationship between facets and conceptual categories. Although a category can be represented by several different facets in an object, each facet represents one and only one category.

Dataset and Methodology

Tags from LibraryThing

The dataset consists of tags from LibraryThing[9], a social networking website where users can catalog, tag and share their book collections, thus enabling people with similar tastes in books to connect. To date, the website has more than 940,000 members who have cataloged over 45 million books representing nearly 5 million individual works, to which over 58 million tags are affixed[10]. The dataset of tags was constricted to those depicting non-fiction books about history. This was accomplished by creating a TagMash[11] with the tags *history* and *non-fiction*.

A TagMash created with *history* and *non-fiction* yielded 45 of the most popular tags for each of the 250 most popular books tagged with both the two tags. The dataset was further constricted to include only tags representing those books that had also been indexed with the subject heading ‘history’ by the Library of Congress. Only 76 of the original 250 books (30.4%) filled this criterium. The final dataset consisted of 107,375[12] instances of 1,288 unique tags depicting 76 non-fiction history books.

[9] http://www.librarything.com

[10] http://librarything.com/zeitgeist. Retrieved November 27, 2009.

[11] http://www.librarything.com/blog/2007/07/tagmash.php. Retrieved November 27, 2009.

[12] This number includes 34 instances of 13 tags that were deemed by me to be too ambiguous to classify. These tags were taken out of the dataset and are not used when calculating percentages of the total dataset, making the total: 107,341 instances of 1,275 distinct tags.

The Facet Analysis

The method followed in this research was a non-linear and highly iterative process aimed at placing each tag in a mutually exclusive facet. Based on the in-depth study of Ranganathan's postulational approach discussed above, the following postulates were proposed to serve as guidelines throughout the process of facet analysis:

1. Look for conceptual categories to which all the facets in the universes to be classified belong. Use PMEST as a starting point.
2. Look for explicit or implicit basic facets. These represent classes in the universes to be classified.
3. All the explicit or implicit facets found will belong to one and only one of the conceptual categories found. By extension, each tag in the user-generated metadata will belong to one and only one facet.

An algorithm was developed to use in the initial analysis of each tag. The main reason for using the algorithm was to make the dataset more manageable by sorting the tags into smaller groupings. This would presumably facilitate in the identification of the facets, basic facets and the remaining conceptual categories by providing a systematic overview of the types of tags present in the dataset. The algorithm applies Ranganathan's Method of Residues, which is a technique intended to aid the classifier in figuring out the conceptual category to which identified facets belong. According to the Method of Residues, "if a certain manifestation is easily determined not to be one of 'Time', 'Space', or 'Energy', or 'Matter', it is taken to be a manifestation of the fundamental category 'Personality'" (Ranganathan, 1967, p. 401).

The completion of the initial analysis of the tags resulted in a rough division of the original dataset into eight categories: Time, Space, Energy and Other in both the universe of books and the universe of subjects. Each of these was then concurrently examined for facets and basic facets, and the search for more conceptual categories continued within the two "Other" categories. The identification of facets, basic facets and conceptual categories in this stage of the analysis was an ad hoc process in which tags were grouped together based on linguistic or operational similarities and then tested for the following criteria based on Ranganathan's faceted classification theories:

To ascertain that the grouping represented a facet, the following criteria had to be fulfilled:

- Facets are the results of a single principle of division
- A facet is a facet of a class, which is represented by a basic facet
- Every facet belongs to a conceptual category

To ascertain that the grouping represented a basic facet, the following criteria had to be fulfilled:

- Basic facets represent classes in the universe

- Classes are differentiated by facets

To ascertain that the grouping represented a conceptual category, it had to differentiate the entire universe and contain at least one facet.

The grouping together of tags was thus a highly iterative process in which the above criteria were checked, and adjustments and readjustments to the groupings were made accordingly. In this sense, the facet analytical process can be compared to puzzle-solving; the verified identification of facets often led to the identification of either implicit or explicit basic facets and the identification of conceptual categories often led to the identification of facets therein.

According to the rule based on the Postulate of Isolate Facet used in this research, each tag was placed in one and only one facet, and each facet was placed in one and only one category. This led to difficulties when compound tags were encountered. Into which facet should the compound tag be placed? Compound tags were initially sorted out of the dataset. Upon completion of the facet analysis, if facets had already been identified for each aspect of the compound tags, then the compound tag was placed in the facet that was deemed to be the least concrete. For example, *medieval Europe* is a compound tag made up of *medieval ages*, which represents the facet "by Time" and *Europe*, which represents the facet "by Place". After ascertaining that both of the facets had already been identified in the dataset, the tag was placed in the "by Time" facet. If a facet of a compound tag had not already been identified, the tag was placed in the new facet.

Results

Over 107,000 instances of 1,275 unique tags representing 76 history books make up the folksonomy analyzed in this research. Subjecting them to a facet analysis resulted in the discernment of two distinct implicit universes: the universe of books and the universe of subjects contained within the universe of books. Basic facets, conceptual categories and facets were identified in the tags representing each of the universes (see Table 1, Table 2, and Figure 2). Basic facets were identified implicitly in the universe of books (books as physical objects and books as works[13]) and explicitly in the universe of subjects (subjects as disciplines). These represent here the top-level classes in each of the universes.

The initial conceptual categories discerned in both the universes were based on those postulated by Ranganathan: Personality, Matter, Energy, Space and Time. All of these were identified in the universe of books, while only Personality, Energy, Space and Time were identified in the universe of subjects. An additional two conceptual categories were found that apply solely to the universe of books: Agent and External Reception. As will be shown, while it was fully possible to facet analyze the metadata representing

[13] Work is here defined as encompassing everything about a book that doesn't pertain to its physicality. This is a much broader definition than that used in the FRBR model, where aspects such as translation and ISBN would be considered facets of manifestation and expression rather than work.

the universe of books, results of the facet analysis of the metadata representing the universe of subjects remain incomplete.

In Table 1 and Table 2, each universe is presented with the conceptual categories and facets identified in each. The total number of tags representing each category and facet are shown in parentheses next to the category or facet name as a percentage of the total number of tags in the dataset. For example, 75,858 of the 107,341 instances of tags in the dataset (70.67%) belong to facets in the Personality category in the universe of books; 75,713 (70.54%) of these belong to the facet “by Subject”, 130 (0.12%) to the facet “by Type” and 15 (0.01%) to the facet “by Title”. The basic facet in the universe of subjects was identified as being implicit, thus accounting for 0% of the dataset.

The Universe of Books

Tags that describe the universe of books account for 29.46% of the dataset. Of these, 27.93% describe books as works, while the remaining 1.53% describes books as physical objects.

Universe	Category	Facet	Examples of tags
Universe of Books	Basic Facet (0%)	By aspect (0%)	Work (implicit), Physical Object (implicit)
	Personality (70.67%)	By subject (70.54%) (facet of Work)	See Table 2: Universe of Subjects
		By type (0.12%) (facet of Physical Object)	audiobook, library book
		By title (0.01%) (facet of Work)	the histories
		By isbn (0.003%) (facet of Work)	Isbn
	Matter (22.4%)	By genre (21.89%) (facet of Work)	historical, mystery, non-fiction
		By binding (0.26%) (facet of Physical Object)	hardcover, paperbook
		By version (0.12%) (facet of Work)	Translation

		By format (0.07%) (facet of Work)	Audio, mp3
		By edition (0.06%) (facet of Work)	first edition
		By series (0.01%) (facet of Work)	Hinges of History
	Energy (4.95%)	By activity (work) (3.74%) (facet of Work)	read, tbr, unread
		By activity (object) (1.17%) (facet of Physical Object)	borrowed, own, owned, wishlist
		By process (0.03%) (facet of Work)	illustrated, made into movie, translated
	Agent (1.11%)	By author (0.76%) (facet of Work)	gibbon, Albert Manguel, Australian author
		By publisher (0.28%) (facet of Work)	folio, folio society, penguin classics
		By user (0.07%) (facet of Work)	Book club, adult, teen, ya
	Space (0.17%)	By place (0.17%) (facet of Physical Object)	library, box 2, storage
	Time (0.36%)	By year written or published or by year read (0.36%) (facet of Work)	100s, 1984, 2006, 2007
	External Reception (0.33%)	By source (0.15%) (facet of Work)	Comedy Central, daily show, npr, This American Life

		By award (0.11%) (facet of Work)	pulitzer prize, national book award
		By new expression (0.05%) (facet of Work)	film, movie, Hinges of History
		By rating (0.02%) (facet of Work)	favorite, staff pick

Table 1: Universe-level presentation of results of the facet analysis of the Universe of Books

A number of the tags in the folksonomy can be identified as "task-oriented" tags. These include *read* and *tbr* from the "by activity" facet of books as works, and *borrowed* and *wishlist* from the "by activity" facet of books as physical objects. Both of these facets belong to the energy conceptual category. The identification of task-oriented tags is consistent with Kipp & Campbell (2007), who found "tags relating to time and task which suggest the presence of an extra dimension in classification and organization." While they propose that conventional two-dimensional classification systems are unable to facilitate these types of tags, their unproblematic inclusion here suggests that faceted classifications, which allow for multi-dimensional representations, are ideally suited for the task. Additionally, the user is represented in the faceted classification ontology, primarily implicitly as the agent of task-oriented tags, but also explicitly, as in the tag *Book club*. As will be discussed, the inclusion of users in a faceted classification may provide novel ways to personalize faceted navigation.

The Universe of Subjects

The "by Subject" facet in the universe of books accounts for 70.54% of the dataset. It represents the universe of subjects and has been subjected to a facet analysis of its own (see Table 2). The results of the facet analysis of the universe of subjects are inconclusive. Percentages are therefore only given for the conceptual categories.

Universe	Category	Facet	Examples of tags
Universe of Subjects	Basic Facet (52.82%)	By discipline (52.82%)	biology, history, literature, religion
	Personality (16.16%)	By person	Aaron burr, Rasputin, sickert, us president, serial killer

		By group	American Indians, secret societies, marine corps, merovingians
		By entity	animals, Mayflower, theory, codes, map, television, culture, christianity, books
	Energy (12.72%)	By energy (find facets?)	cultural diffusion, crime, evolution, murder, politics
	Space (14.94%)	By place	america, boston, college, sea, the west, world
	Time (3.36%)	By time	19th century, 1990s, antiquity, dark ages, renaissance

Table 2: Universe-level presentation of results of the facet analysis of the Universe of Subjects

The universe of subjects is the very object of universal library faceted classifications. There are at least two different ways to model a universal faceted classification: one can either divide the universe of subjects into smaller units of knowledge, such as the traditional division of the universe of knowledge into disciplines, or one can base it on the concepts in the universe and divide it by some other means, as was attempted in the CRG's unsuccessful quest for a New General Classification in the 1970s and as is currently being attempted in the Integrative Levels Classification (Gnoli, 2008).

Due to the high occurrence of tags indicating disciplines in the tag set, it was here assumed that users acknowledge disciplines to be a natural initial division of the universe of subjects. As will be discussed, however, intra-facet relationships are not exposed during the course of facet analysis and it was therefore neither possible to reveal what types of phase relationships the disciplines had with one another nor to which disciplines the remaining tags indicating subjects differentiated.

The Induced Faceted Classification Ontology

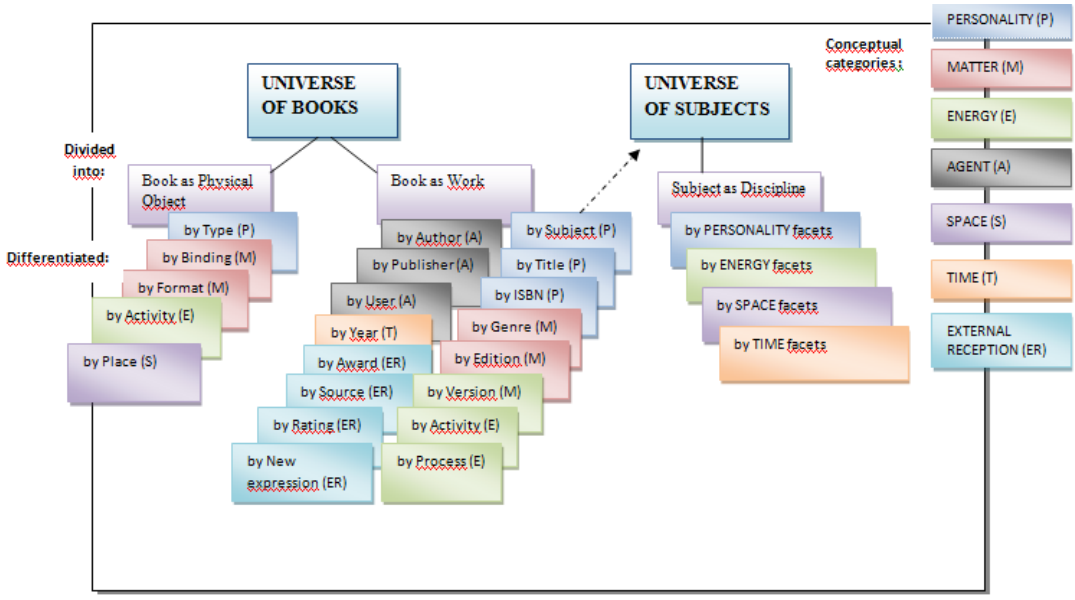

Figure 2: Ontological model of the facet analyzed universes of books and subjects

The ontological model of the facet analyzed universes of books and subjects in Figure 2 and the exemplified ontological relationships shown in Figure 3 illustrate some of the relationships discerned in the folksonomy. Each facet evokes a specific "differentiated-by" relationship to the object of which it is a facet via the basic class to which the object belongs. For example, there is a "differentiated-by-author" relationship between the book (as Work): *A History of Reading* and the tag: *Albert Manguel*, and a "differentiated-by-format" relationship between the book (as Physical Object): 1776 and the tag: *hardcover*.

The identification of facets in the tags space is thus significant because facets represent a new way of grouping tags. The most common grouping of tags is the tag cloud, which clusters tags together based on the frequency of tag co-occurrences. Here, tags are grouped together based on shared common characteristics that distinguish them from other tags in the tagspace in relation to aspects of the object they represent.

Unfortunately, intra-facet relationships are not explicit. In faceted classifications, intra-facet relationships are semantic relationships, like synonyms and hierarchical relationships (Broughton, 2006). Thus, the hierarchical relationships between *massachussetts*, *new england* and *united states*, which all belong to the "by Place" facet of the subject of the book, *Mayflower: A Story of Courage, Community, and War*, remain implicit, as do the synonyms *united states*, *us* and *usa* from the same facet. These can be inferred by those with knowledge of the domain, but they are not directly discernible in the model. This is consistent with Kwasnik's analysis of the role of classification in knowledge structures. She notes that one of the major disadvantages of faceted classifications lies in their lack of explicit intra-facet

relationships, such that, “in terms of theorizing and model building, the faceted classification serves as a useful and multidimensional description but does not explicitly connect this description in an explanatory framework” (Kwasnik, 1999, p. 42). The consequences of this failure are seen most clearly in the lack of explicit intra-facet relationships between the basic facets representing classes in the universe of subjects. Since the tag space was unstructured prior to the facet analysis, this resulted in the inability to determine how each discipline related to one another and which facets belonged to which discipline.

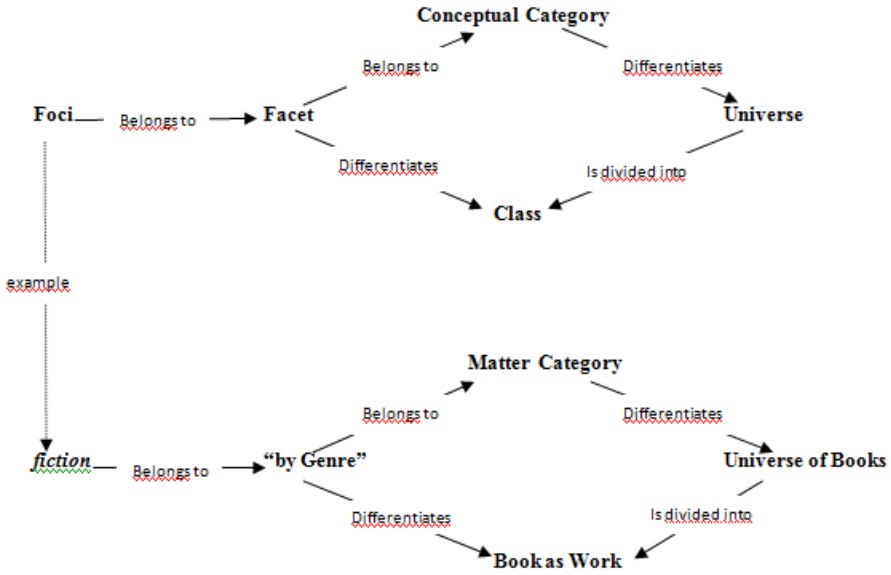

Figure 3: Ontological relationships between foci, facet, category, universe and class (top) with example from the folksonomy depicting the tag “fiction” (bottom)

Inter-facet relationships, on the other hand, are explicit in the model. Inter-facet relationships are syntactic relationships, “the number and variety (of which) seem unique to faceted classification” (Broughton, 2006). Since all tags belong to facets, they inherit both inter-facet and inter-category relationships. Examples of inter-facet relationships in the dataset would include (author)-(activity)-(subject) relationships, as in *orwell*-writes about (implicit)-*history*; and (user)-(activity)-(place) relationships, as in *book club-borrowed-library*. Additionally, since all facets belong to categories, they inherit inter-category relationships. Broughton (2009) notes that the use of “categories allows general rules to be proposed about the relationships in the domain (as opposed to the relationships between particular pairs of concepts).” Although the conceptual categories borrowed from Ranganathan are by definition diffuse, it has been found that the smaller the universe, the more specified the relationships can be defined. This is consistent with Vickery’s work on special faceted classifications (Vickery, 1960).

Discussion and Suggestions for Further Research

Despite the fact that the nature of this research was primarily theoretical, there are a number of practical implications of the results. First of all, it has been shown that facet analysis techniques may be used for the identification and characterization of facets and conceptual categories in tags representing books. This indicates new knowledge: the most populated facets and categories are not the same as the most popular tags. In this light, it is possible that the "wisdom of the crowd" may be invoked to alleviate one of the major challenges involved in the creation of faceted structures, namely the selection of facets.

In 1960, Vickery noted that, "theoretically, an unlimited number of facets could represent the various perspectives contained in each topic" (Vickery, 1960, p. 20) and that the first step in a facet analysis is to "define the domain and interests of domain participants" (ibid, pp. 12-13). Kwasnik (1999, p. 44) cites the difficulty of choosing the right facets as one of the leading problems in the creation of faceted classifications. The challenge of selecting facets is particularly relevant when constructing faceted classifications for websites, where usually only a select number of facets in a predefined order are displayed on the user interface. Unearthing which facets users choose most frequently to describe a given domain may provide valuable clues as to which facets should provide most prominent placement in user interfaces.

The facets and categories discerned in this research indicate the following ranking in popularity:

1. by Discipline (basic facet of subjects) (ca 37%)
2. by Genre (facet of book as work) (ca 22%)
3. by facets in the Personality category (category of subjects) (ca 16%)
4. by Place (facet of subjects) (ca 15%)
5. by facets in the Energy category (category of subjects) (ca 13%)
6. by Activity (facet of book as work) (ca 4%)
7. by Time (facet of subjects) (ca 3%)
8. by Activity (facet of book as physical object) (ca 1%)

Even when taking into consideration the possible distortions of the results due to bibliographic information being readily available at LibraryThing and the disproportionate number of history and non-fiction tags in the dataset, it is quite clear here that facets indicating the subjects are more popular than facets indicating other aspects of books. Of the three facets listed above that indicate the universe of books, genre is clearly the most popular, followed by the two task-oriented facets indicating what users do to books as works and to books as physical objects.

Conversely, the selection of facets for use on websites is often limited to the metadata available at the time of facet construction. The results of this research imply that facets could be culled from user-generated metadata, either in the form of tags or otherwise . In an online public access catalog, for example, this could perhaps manifest itself as facets indicating awards associated with books in the collection.

This research has also shown the possibility of including a user dimension in a faceted classification. This presents many possibilities for designers of faceted structures on websites to allow for user interaction without bothering the basic structure of the classification system. It implies, however, that steps be taken to take into account the overall purpose of the classification so that relevance for all users is preserved. One way this may be accomplished is through efforts to personalize the faceted classification, such that users are only presented with task-oriented facets that are relevant to them.

One can, for example, imagine its use in a faceted online public access catalog where a user could log on and mark objects in the catalog based on specified criteria, like whether or not the user has read or enjoyed them. Logged on users could then be presented with a facet on the search page indicating task-oriented tags in addition to all the other facets of documents normally presented there. This personalized version of the catalog would allow for the user to narrow search results by, for example, books that haven't yet been read by the user ("by User Activity") that are by such and such author ("by Author") and about such and such subject ("by Subject").

In the presentation of the results of the facet analysis of tags representing the universe of subjects, it was seen that inherent difficulties were encountered involving in the identification of what the facets were of. This implies that, in order to successfully expose a faceted classification ontology in a tag space in which some of the tags indicate aboutness, some prior form of initial division of the universe of subjects appears necessary. This is consistent with Schmitz' (2006) examination of Flickr tags, which found that the use of "domain-specific upper model ontologies" is necessary for the inducement of a faceted ontology in the tags.

There are several ways this may be accomplished. One is a tagging system that requires users to choose a core category in which to place each resource before tagging it. This requirement, however, seems counterintuitive to the freedom and ease of tagging. Furthermore, it would not answer to the contention made by the León manifesto that the distinction between disciplines is becoming less and less rigid (Gnoli & Szostak, 2007). Schmitz (2006) examines the automatic correlation of gazetteers and common taxonomies to tags.

More research is recommended based on the examination of different ways with which to initially divide the universe of subjects in order to facilitate the facet analysis of tags. In this light, the Open Shelves Classification (OSC) project at LibraryThing is interesting. The OSC aims to create top-level, statistically-tested classes in the universe of subjects through a bottom-up collaboration of LibraryThing users (Public Library Association, 2009). An analysis of the use of disciplines in tags compared to the OSC top-level classes is recommended, in particular vis-á-vis differences in fiction and non-fiction books.

Fu et al. (2009) have developed a model for the prediction of tag choices based on a cognitive study of the imitation effect in tagging. It was suggested here that LibraryThing folksonomies are largely comprised of "uninfluenced" tags, meaning that users choose tags removed from prior tags of the resource in question. Further research on the validity of this suggestion and an

eventual comparison of folksonomies based on “imitated” tags and those based on “uninfluenced” tags would be very interesting.

Finally, the methodology applied manually to the dataset in this research was a highly laborious and time-consuming effort. While the facet analysis of the universe of subjects remained incomplete, the analysis of the universe of books was successful. It is likely that algorithms (based, for example, on the Method of Residues and on likely existent conceptual categories) could be developed to partially automate the facet analysis of smaller domains. Further research into automating this process is recommended. Interested readers are referred to Stoica et al. (2007) for their inspiring work on the automatic extraction of faceted hierarchical metadata in texts, Marchetti et al. (2007) for their work on the extraction of tags expressing semantic relationships. and to FLOR, a FoLksonomy Ontology enRichment tool developed to automatically assign semantic relationships to tags based on existing Semantic Web ontologies (Angeletou et al., 2008).

References

Abrera, J. (1974). Traditional Classification: Characteristics, Uses and Problems. Drexel Library Quarterly, 10 (4), 21-36.

Angeletou, S., Sabou, M., and Motta, E. (2009). Improving Folksonomies using Formal Knowledge: A Case Study on Search. Paper presented at the 4th Asian Semantic Web Conference, Shanghai, China, December 7-9, 2009.

Broughton, V. (2006). The need for a faceted classification as the basis of all methods of 92 information retrieval. Aslib Proceedings: New Information Perspectives, 58 (1/2), 49-72.

Broughton, V. (2009). Facet analysis as the theoretical basis of vocabulary tool construction: Fundamental Purposes in Knowledge Organization. PowerPoint Lecture presented at the ISCO UK Conference, 2009. Retrieved November 22, 2009, from http://www.iskouk.org/conf2009/presentations/broughton_ISKOUK2009_presentation.pdf

Dahlberg, I. (1992). Cognitive Paradigms in Knowledge Organization. International Classification, 19(3), 125 and 145.

Ellis, D. & Vasconcelos, A. (1999). Ranganathan and the net: Using facet analysis to search and organize the World Wide Web. Aslib Proceedings, 51 (1), 3-10.

Fu, W-T., Kannampallil, T. & Kang, R. (2009). A Semantic Imitation Model of Social Tag Choices. Proceedings of the 2009 International Conference on Computational Science and Engineering, 4, 66-73.

Gnoli, C. & Szostak, R. (2007). The León manifesto. Retrieved March 13, 2009, from http://www.iskoi.org/ilc/leon.htm

Gnoli, C. (2008). Categories and Facets in Integrative Levels. Axiomathes: an international journal in ontology and cognitive systems, 18 (2), 177-192.

Halpin, H., Robu, V. & Shepherd, H. (2007). The Complex Dynamics of Collaborative Tagging. In WWW 2007, May 8-12, 2007, Banff, Alberta, Canada. Retrieved November 10, 2009, from http://www2007.org/papers/paper635.pdf

Ingwersen, P. & Wormell, I. (1992). Ranganathan in the perspective of advanced information retrieval. Libri: international library review, 42 (3), 184-201.

Kipp, M. & Campbell, D. (2006). Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices. In Proceedings of the American Society for Information Science and Technology, Austin, Texas.

Kwasnik, B. (1999). The Role of Classification in Knowledge Representation and Discovery. Library Trends, 48(1), 22-47.

Lichtblau, D., Trice, A., & Wartik, S. (2006). Taxonomic and Faceted Classification for Intelligent Tagging and Discovery in Net-Centric Command and Control. Paper presented at 2006 CCRTS, Command and Control Research and Technology Symposium, June 20–22, 2006, San Diego, California

Marchetti, A., Tesconi, M., Ronzano, F., Rosella, M. & Minutoli, S. (2007). SemKey: A Semantic Collaborative Tagging System. In WWW 2007, May 8-12, 2007, Banff, Alberta, Canada. Retrieved November 10, 2009, from http://www.ibiblio.org/www_tagging/2007/paper_45.pdf

Mills, J. (1960). A Modern Outline of Library Classification. London: Chapman & Hall.

Morville, P. & Rosenfield, L. (2007). Information Architecture for the World Wide Web (3rd ed.). California: O'Reilly.

Petersen, E. (2006). Beneath the Metadata: Some Philosophical Problems with Folksonomies. D-Lib Magazine, 12(11). Retrieved September 3, 2009, from http://www.dlib.org/dlib/november06/peterson/11peterson.html

Public Library Association. (2009). New Classification System for Public Libraries?

PLA Blog: The official blog of the Public Library Association. Retrieved November 4, 2009, from http://plablog.org/2009/01/new-classification-system-for-public-libraries.html

Quintarelli, E. (2005). Folksonomies: Power to the People. Paper presented at the ISKO Italy-UniMIB meeting, June 24, 2005, Milan (IT). Retrieved March 22, 2009, from http://www.iskoi.org/doc/folksonomies.htm

Quintarelli, E., Resmini, A. & Rosati, L. (2006). Facetag: Integrating Bottom-up and Top-down Classification in a Social Tagging System. Paper presented at the EuroIA Conference 2006, Berlin (DE).

Rabourn, T. (2003, August 10). Faceted Movable Type. Pixelcharmer. Retrieved November 27, 2009 from http://www.pixelcharmer.com/fieldnotes/2003/faceted-movable-type/

Ranganathan, S.R. (1967). Prolegomena to Library Classification (3rd ed.). Bombay: Asia Publishing House.

Redmond-Neal, A. (n.d.). Facets Help Move from Search to Found (White paper). Retrieved November 27, 2009, from http://www.dataharmony.com/library/whitePapers/facets_help_move_from_search_to_found.html

Rosenfield, L. (2005). Folksonomies? How about Metadata Ecologies? Retrieved September 3, 2009, from http://www.louisrosenfeld.com/home/bloug_archive/000330.html

Schmitz, P. (2006). Inducing ontology from flickr tags. In WWW 2006, May 22–26, 2006, Edinburgh, UK. IW3C2.

Schwartz, C. (2008). Thesauri and facets and tags, oh my! A look at three decades in subject analysis. Library Trends, 56 (4), 830-842.

Spiteri, L. (2010). Incorporating Facets into Social Tagging Applications: An Analysis of Current Trends. Cataloging & Classification Quarterly, 48(1), 94-109.

Steckel, M. (2002). Ranganathan for IAs: An Introduction to the Thought of S.R. Ranganathan for Information Architects. Retrieved September 3, 2009, from http://www.boxesandarrows.com/view/ranganathan_for_ias

Stoica, E., Hearst, M. & Richardson, M. (2007). Automating Creation of Hierarchical Faceted Metadata Structures. In Proceedings of NAACL-HLT, Rochester NY, April 2007.

Uddin, M. & Janecek, P. (2007). Faceted classification in web information architecture: A framework for using semantic web tools. The Electronic Library, 25(2), 219-233.

Vanderwal, T. (2005, February 21). Explaining and Showing Broad and Narrow Folksonomies. Vanderwal.net. Retrieved September 1, 2009, from http://www.vanderwal.net/random/entrysel.php?blog=1635

Vickery, B.C. (1960). Faceted Classification: A guide to construction and use of special schemes. London: Aslib.

Vickery, B.C. (1966). Faceted Classification Schemes. New Brunswick, N.J.: The Graduate School of Library Services, Rutgers.

Weaver, M. (2007). Contextual metadata: Faceted schemas in virtual library communities. Library Hi Tech, 25 (4), 579-594.

Weinberger, D. (2006).Taxonomies and Tags: From Trees to Piles of Leaves. Retrieved April 28, 2009, from http://www.hyperorg.com/blogger/misc/taxonomies_and_tags.html

Xiao, Y. (1994). Faceted Classification: A Consideration of its Features as a Paradigm of Knowledge Organization. Knowledge Organization, 21(2), 64-68.

[Intentionally left blank]