Tamara Marnell
SLIS S503
Lala Hajibayova
8/11/2010

Final Exam

1. Systematic Representation

Define and discuss enumerative and faceted systems of representation. Provide comparative analysis of both systems from the perspective of knowledge organization by the system and information retrieval by the user. Provide specific [hypothetical] examples to illustrate your analysis and discussion.

In the effort to organize the vast amount of published literature on library shelves, there are two basic approaches to classification one could take: top-down enumerative classification, or bottom-up faceted classification. According to Hunter (2002), the former is "a process of division producing a series of subject classes in successive subordination." In a purely enumerative system, each of these classes is embedded in a hierarchy, mutually exclusive, and each resource can only exist in one. The indexer establishes a "place" for each based on personal judgment of the intellectual content of the document, but not necessarily the patrons' use of it (Mai 2004). From an information retrieval standpoint, this sort of classification is extremely rigid. While browsing or conducting a subject search, the searcher must be aware of not only the controlled vocabulary used to describe each class, but the order of the system-imposed hierarchy. For example, in a hypothetical collection of cookbooks, an indexer could organize each volume hierarchically like this:

A member of the public looking for the imaginary title *One Hundred and One Holiday Cookies* by browsing subject headings would need to look under Cooking: American: Baking: Desserts: Cookies. This appears straightforward enough, but what if that particular title did not consist of recipes for the modern baker with a convection oven, but of treats from the 19th century baked over hearths? The searcher would need to know to deviate after "American" to "Traditional" instead of "Baking." The searcher is also necessarily constrained only to American recipes, unless she goes out of her way to check the other cuisines for French Christmas cookies etc. One of the difficulties of enumerative classifications is that the searcher predetermines the importance of each set of classes, when the resources might be more accessible in a different order.

Faceted classification systems address this problem. In Hunter's (2002) description of a faceted system, "constituent parts of subjects [are] used as nuts and bolts to produce whatever subject classes are required." Instead of determining that the nationality of cuisine is more important than the cooking method, an indexer utilizing a faceted system would simply list each of the properties:

Nationality facet:
1. French
2. Italian
3. American
4. East Asian…etc.

Method facet
1. Traditional (SN: wood fire, Franklin stove, etc.)
2. Baking
3. Grilling
4. Dehydrating…etc.

Use facet
1. Entrees
2. Appetizers

3. Desserts

Type facet

1. Breads
2. Meats
3. Dairy
4. Vegetables…etc.

The faceted classification system gives the indexer a greater degree of flexibility when organizing items. As seen in the example above, it also provides the opportunity to expand the information contained in classes by listing more properties than could be contained in an enumerative class. Specialty diets, for instance, were not included in the original hierarchy, so books of vegan/vegetarian/raw recipes, low-sugar recipes for diabetics, or low-fat recipes for dieters would need to be allocated by an astute indexer to the closest fitting category. Thanks to facets, another "Specialty" branch could be added if deemed necessary for the collection when constructing a system of classification.

It is important to note, however, that faceted classification systems are not as radically different from enumerative systems as they first appear. Though the former affords indexers greater flexibility, and the result may be slightly easier to search and conceptualize from a retrieval perspective, both use controlled vocabulary and create mutually exclusive classes. In a faceted classification system, though no property below the primary facet is deemed necessarily more *important* than the others, a rational citation order must be established. Faceted classification is not a free-form "tagging"-like system or group of floating keywords that supports post-coordinate indexing. Essentially, patrons must approach it the same way they would an enumerative system. As Jacob (2004) points out, "Because a faceted classification scheme adheres to a fixed citation order during the construction of individual classes, the resulting structure, like an enumerative scheme, is necessarily hierarchical." Since the point of both systems is to fit resources into specific locations on library shelves, the difference is

primarily philosophical, affecting the methodology of the librarians but not necessarily the searching behavior of the patrons.

2. Indexing Languages

Define the concept of an "indexing language". Compare and contrast pre-coordinate indexing languages and post-coordinate indexing languages. Discuss the advantages and disadvantages of pre-coordinate and post-coordinate systems from the perspective of both the indexer and the searcher using an electronic retrieval system. Provide specific [hypothetical] examples to illustrate your analysis and discussion.

As defined by Jacob (2010), an indexing language is “The complete set of terms or descriptors (words/phrases) that is used to represent the conceptual content of the material being indexed.” Indexing languages consist of the natural language or controlled vocabulary used to describe the “is” properties (author, publication date, title, ISBN, call number etc.) and “about” properties (subject headings, keywords, abstracts, etc.) of resources. Since the “is” properties are more or less universally established, discussions of indexing language generally center on the difficulties of accurately describing the resources’ “about” properties.

There are two primary types of indexing languages: pre- and post-coordinate. The difference between the two is the balance of authority given to professional indexers and searchers in describing the use or intent of a document. In pre-coordinate indexing, the indexer establishes the precise class used to describe a resource in its entirety. Post-coordinate indexing, on the other hand, allows users to combine terms as needed to define their own conceptual categories of resources. For example, in a pre-coordinate system, a book may be placed under the hypothetical subject heading of “Celtic history in the middle ages.” The book will appear only under that heading, and not under any variant or blended into neighboring categories, like “Celtic religion” or “Celtic history in the Renaissance.” In a post-coordinate system, a user can define a group of documents appropriate to his needs by using Boolean logic. Instead of being confined to

the categories predefined by indexers, he could search for “Celtic history AND religion AND middle ages” to produce the hybrid category, “The history of Celtic religion in the middle ages.”

From the perspective of the indexer, pre-coordinate systems are more streamlined and easier to organize than post-coordinate systems. Post-coordinate systems necessitate a lot of time and effort to describe every potential aspect and use of a document for user retrieval. Pre-coordinate systems require only a summary classification, instead of an exhaustive list of keywords. However, pre-coordinate systems can be restrictive even for the indexer. A cataloger working with system that allows adult fiction to be described as either “Romance” or “Westerns” may be confronted with a cowboy love story that fits neatly into neither one. As emphasized by Mai (2004), standards for allotting resources to specific classes are ill-defined and personal; one indexer with certain experiences and favorite subject headings may assign a resource to a completely different class than another.

Post-coordinate indexing compensates for this by limiting the indexer’s authority over the resource to less complex categories. A member of the public interested in romantic novels set in the Old West can create that category through the OPAC herself, instead of attempting to guess whether the library shelved them under “Romance” or “Westerns.” On the downside, post-coordinate indexing requires skill on the part of the searcher, as well as trial-and-error. The aforementioned search for “Celtic history AND religion AND middle ages” may return no results because it is too specific, or a barrage of results the searcher did not intend, like books on the influence of Christianity in Celtic art in the middle ages. Repeated rewordings may be necessary to elicit the results the searcher wants, where as a pre-coordinate subject heading would have guaranteed a body of appropriate materials on the first try. Users who are

unaccustomed to modern electronic interfaces may become frustrated with these systems or require the assistance of a librarian to learn to conceptualize materials this way.

Today's libraries mostly take a hybrid approach to pre- and post-coordinate indexing systems, supplying the collection with both specific subject headings for browsing and descriptive keywords for searches tailored to user needs. This approach obviously requires more work than either would alone, but is necessary to compensate for the pitfalls of each and provide overall effective access to users.

3. Folksonomies

Drawing on the materials provided across the semester, provide your own definition of a "folksonomy." Compare and contrast folksonomies and controlled vocabularies. Discuss the advantages and disadvantages of folksonomies and controlled vocabularies both as tools for retrieval and as systems of organization. Would you characterize a folksonomy as an example of free-text searching, a post-coordinate system for retrieval, or something else entirely?

The term "folksonomy" was coined in 2004 by Vanderwal to mean "the user-created bottom-up categorical structure development with an emergent thesaurus," or "tagging that works." In practice, the "taxonomy" section of "folksonomy" is usually lacking, as users as a whole do not classify their descriptors hierarchically. Folksonomy, to me, is any system in which the users define the descriptors applied to resources. In a folksonomy, there is no standardization and little quality control. Only the "votes" of the masses over time determine whether a natural language term is appropriate for a body of documents.

To contrast, a controlled vocabulary, used by most libraries and academic databases, collocates documents of similar description under a single official term. These terms are arbitrary; selected by indexers who attempt to predict the needs and behaviors of the users. According to Buckland (1999), controlled vocabularies are problematic primarily because these

predictions are not entirely accurate or even practical. He cites examples like the LCSH "God – Knowableness" or the common difficulty of "Vietnam War" vs. "Vietnam Conflict," as well as the difficulties of representing titles or subjects in foreign languages. Shirky (2005) objects to controlled vocabularies because classification systems can be biased (like Dewey's WASP-centered, anti-Semitic structure of religious texts), and because indexers are neither fortune tellers nor mind readers. It is impossible for librarians to predict precisely how users will naturally conceptualize and describe materials. Unless the public is educated extensively in the use of thesauri, or the many vocabularies of different institutions are standardized, controlled vocabularies will be unnatural to use and produce less than ideal search results.

Folksonomies avoid this last problem to a certain extent, but here's the rub: though professional indexers cannot read minds, users do not even attempt to. User application of terms in systems like Del.icio.us and Flickr are often useless for retrieval by other users. Golder and Huberman (2006) described different types of terms users apply to their own bookmarks in Del.icio.us which have little meaning out of personal context: task-oriented tags (e.g. toread, jobsearch), ownership assertions (e.g. mystuff, mycomments), and expressing personal opinions (e.g. funny, stupid). A look through the Library of Congress photographs uploaded to Flickr during a pilot study shows tags that have little or marginal bearing on the content of photos, like "red" in a photograph of a working woman who happens to be wearing red lipstick or "four hats" in a photo of people driving a horse and buggy down a country road.

Some other problems evident in folksonomies are listed by Golder and Huberman (2006) as "polysemy, synonymy, and basic level variation." Polysemy results in inappropriate search results; synonymy results in the exclusion of many potential resources; and basic level variation results in large disagreement over the appropriate level of information necessary to describe a

document. Synonymy is the most obvious difficulty in the Library of Congress' Flickr effort. For example, many photos from the 1940s are tagged with "World War II, "World War 2, "WWII," and "WW2." A quick search for "wwii" would produce only a handful of many relevant results. A related problem is that of spelling and punctuation; a careless user could easily tag a photo "Wolrd War II" and render it irretrievable until someone else compensates for the mistake, or use "world_war_II" and create an entirely different class of images with no direct link to the other variants. Basic level variation also comes into play, as someone with a photography background would apply terms like "large format" and "transparency," whereas an amateur might leave it at "color photo."

Folksonomies are not free-text searching, because tags are applied deliberately by users as descriptors, and are not part of the original content of a document. They can only be marginally considered post-coordinate indexing during retrieval, because in principle the users are still retrieving documents by combining others' descriptors, just as they would in structured library collections. However, in true post-coordinate indexing systems, users do not have the authority to alter the document beyond organizing search results for personal use. The ability to tag after retrieval, therefore influencing subsequent search results by other users, completely upsets the responsibility of description. The system also obliterates any relationships between descriptors; though Vanderwal's definition included "an emergent thesaurus," the only practical way this occurs is if users deliberately copy other descriptors or "hop" from one tag to another in search of new resources. The only emergent pattern can be seen in "tag clouds," which define only the relationship between terms and popularity, and not between related or super/subordinate terms. In the spectrum of increasing formal structure, from user-controlled free-text searching to

indexer-determined pre-coordinate indexing, folksonomies lie somewhere between post-coordinate and free-text systems.

References

Golder, S. A. & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2). 198-208.

Hunter, E. J. (2002). *Classification made simple*, 2nd ed. (pp. 40-58, 70-81, 86-88). Aldershot: Ashgate.

Jacob, E. K. (2004). Classification and categorization: A difference that makes a difference. *Library Trends, 52*(3), 515-540.

Jacob, E.K. (2010). Session 6 Lecture Notes: Indexing. Retrieved August 10, 2010 from http://oncourse.iu.edu.

Mai, J. (2004). Analysis in indexing: Document and domain centered approaches. *Journal of Information processing and management 41,* 599-611.

Shirky, C. (2005). Ontology is overrated: categories, links, and tags. Clay Shirky's Writings About the Internet: Economics & Culture, Media & Community. Retrieved August 31, 2008, from http://www.shirky.com/writings/ontology_overrated.html

Vander Wal, T. (2007). Folksonomy coinage and definition. vanderwal.net. Retrieved August 31, 2008, from http://vanderwal.net/folksonomy.html