Katherine Thornton. 2011. Contentious categories. In Smiraglia, Richard P., ed. *Proceedings from North American Symposium on Knowledge Organization*, Vol. 3. Toronto, Canada, p. 217-226.
https://digital.lib.washington.edu/ojs/index.php/nasko/issue/view/862

Contentious Categories: Discussions of the Design of the Category System in Wikipedia

Katherine Thornton

The Information School
Box 352840
Mary Gates Hall, Ste 370
Seattle, WA 98195-2840

thornt@uw.edu

Katherine Thornton. 2011. Contentious categories. In Smiraglia, Richard P., ed. *Proceedings from North American Symposium on Knowledge Organization*, Vol. 3. Toronto, Canada, p. 217-226.
https://digital.lib.washington.edu/ojs/index.php/nasko/issue/view/862

Contentious Categories: Discussions of the Design of the Category System in Wikipedia

Introduction

Wikipedia is an online encyclopedia created entirely of user-generated content. Roughly two years after Wikipedia began, the community decided to create a category system to organize the content of the site. The category system has changed over time, as have conceptualizations of what role it should serve in Wikipedia. In this paper, I consider six months of discussion about the design and implementation of the category system in Wikipedia. I analyze the comments editors shared and attempt to understand how these early decisions shaped the category system as it currently exists.

Related Work

The category system in Wikipedia has been considered by scholars from inside the computer-supported collaborative work (CSCW) community as well as scholars from the knowledge organization (KO) community. Kittur et al. (2009) quantitatively mapped the categories used in Wikipedia into seven topical areas. This work did not account for the fact that sub-categories are often associated with multiple levels of parent categories, and thus obscures the significant amount of work that the editors of Wikipedia have undertaken to systematize category hierarchy and encourage editors to apply category labels according to community convention. In this paper I will highlight evidence of early concern for these issues of convention and consistency in the category system.

In their study of the types of work valued in the Wikipedia community, Kriplean et al. (2008) point out that both category creation and category link application are types of work acknowledged to be valuable. This work will support their argument by providing rationales for elaborating content organization advanced by the community in the early stages of the development of the category system.

Forte and Bruckman (2008) outline the social roles that exist in the Wikipedia community. They describe the negotiation that takes place in the wiki, and on related mailing lists between individuals who are trying to create community-wide policies. Through looking at the discussions related to the creation and implementation of the category system, I will describe the strategies that the Wikipedia community employed to attempt to make the implementation of category designations consistent throughout the site.

From the perspective of information organization, Voss (2006) described the category system as a thesaurus built through collaborative tagging. While Voss notes that the creation of hierarchy is facilitated by adding categories to other categories, he does not outline the implications of what happens in the system when different types of relationships that are used in the system between categories. I will outline the types of relationships present among super- and sub-categories in Wikipedia. I will also provide examples of how there is little evidence of the community

prioritizing designs which would support users' ability to distinguish between these different types of relationships in the early period of discussion about the category system in Wikipedia.

Method

Because of the wiki format of Wikipedia, all changes to the site are preserved. I collected the transcripts of the text on a single discussion page for a period of six months from June 2004 through January 2005.[i] I chose this time period due to the fact that it was immediately after the community agreed to implement a category system in Wikipedia, when the final design of the system had not yet been finalized, and changes were still being discussed and codified. I then coded the data thematically in order to undertake a content analysis of that section of the archived talk page for categorization. I developed a code book of 31 codes.

Several themes emerged from my coding of the data. The editors who were participating in the discussions of how to design and implement the category system frequently returned to issues of hierarchy, scope, and navigation. By contrasting the editors' discussions of hierarchy with theory about constructing hierarchy in indexing languages, it is clear that many of the same conclusions reached through trial and error in Wikipedia have been previously described in the KO literature. The editors' discussions of how to scope the number of categories and the type of categories in the system are especially interesting when considered in the context of the recurring appeals to the community to consider extant schema for information organization as potential models. Discussions of navigation through Wikipedia via the category system reflect a very clear understanding of the category system as a navigational tool, something that no longer seems to be the case in Wikipedia.

Hierarchy in the category system

The community of editors who participated in the discussions in this data set were very concerned with ensuring that hierarchy would be a feature of the category system. The fact that the community felt hierarchy to be such an important element for organization is consistent with how the KO literature describes the advantages of hierarchical structures. Elaine Svenonius (2000) said of hierarchical relationships "They are a powerful means for optimizing recall and precision, and at the same time, they are the quintessential means for navigating a knowledge domain". This is evident in the Wikipedia system, in that when looking at the categories displayed at the bottom of any page it is possible to click on any of them (they are all hyperlinks) and see related categories and all subcategories. The editors who contributed to the design of the category system share a vision of this feature that it would allow users to more quickly understand the context of any individual article by making relationships between content visible.

Aitchison et al. (2000) define hierarchical relationships as defining larger units and their subunits "The superordinate term represents a class or whole, and the subordinate terms refer to its members or parts. This relationship is used in locating broader and narrower concepts in a logically progressive sequence". Hierarchical structures allow users to see how different concepts relate to one another in a given system. There are four types of hierarchical relationships defined by Aitchison et al. (2000): generic, whole-part, instance, and polyhierarchical. Much of the discussion in the data set illustrated a failure to discriminate

between these different types of hierarchy, and multiple, sometimes contrasting, assumptions of what type of hierarchy was being proposed. All four types of hierarchy are in use in the category system of Wikipedia. In an ideal hierarchical structure, "terms are hierarchically related only if both are members of the same fundamental category" (Aitchison et al. 2000). This recommendation for the construction of consistent hierarchical systems is not followed in the category system of Wikipedia, and is a source of unaddressed, perhaps unrecognized conflict in discussions of how categories should be managed.

According to Aitchinson et al. (2000), the generic hierarchial relationship "has the mathematical property of inheritance, whereby what is true of a given class is also true of all the classes subsumed under it". There are very few examples of this in the category system. Other indexing tools that are used in Wikipedia are organized in this way, for example see the Wikipedia page for 'List of birds'.[ii] That is, if we take the class to be 'birds', all of the pages for which links are supplied in the list are pages for birds.

Aitchinson et al. (2000) define the whole-part relationship as a single concept or entity as the class and parts of that concept or entity as the subclasses. An example of this type of hierarchical relationship in the category system of Wikipedia would be the pages listed under the category 'States of the United States'.[iii] Other than the page 'U.S. state', the other pages under this category are all parts of the category itself.

Instance relationships are defined by Aitchinson et al. (2000) as a general concept or entity as a class and then specific instantiations of the concept or entity as the subclasses. It is difficult to find examples of categories for which the subcategories are all instances of the category. This is more often achieved through lists in Wikipedia. An example of this type of hierarchical relationship in Wikipedia would be the 'List of cathedrals' page.[iv] If we take 'cathedrals' to be the class, all of the cathedrals listed on that page are instances of the class.

The final type of hierarchical relationship as outlined by Aitchinson et al. (2000) is polyhierarchial. This type of relationship describes instances when one term is located underneath more than one category. Many of the categories in the category system of Wikipedia are located in more than one parent category. Polyhierarchy is very common in the category system of Wikipedia.

Browsing through the category system in order to examine the types of relationships that exist between superclasses and subclasses, the predominant relationship is one of association. The Associative relationship is not a hierarchical relationship. The category system of Wikipedia, although perhaps envisioned to be a hierarchical system by some in 2004, is now full of categories related to other categories by the associative relationship. This is significant because although the designers felt that they were creating a hierarchical category system, many of the relationships in the category system are not hierarchical. Aitchinson et al. (2000) define this type of relationship, saying: "Put simply, the associative relationship is found between terms that are closely related conceptually but not hierarchically and are not members of an equivalence set". For example, the Wikipedia page for the category 'Plants' provides links to 82 pages within the category.[v]

The fact that the relationships between supercategories and subcategories in Wikipedia include both hierarchical relationships as well as associative relationships is due to the fact that the category system emerged from a community in which there were divergent views of what the system should look like. One of the editors who contributed to the discussions in the dataset stated:

> "So I think we need a way of distinguishing between a category where (a) you are asserting that everything in the category is an example of the thing it is in (ie list categories), and (b) categories where you are just providing hierarchical links for convenience" (editor 43).

The first type of category they describe encompasses the first three types of hierarchical relationships described by Aitchinson et al. (2000), generic, whole-part and instance. The second type of category this editor describes would make use of the non-hierarchical associative type of relationship. This editor is highlighting the need to be clear about the different types of relationships in the category system as it is being designed and created, and the differences this editor points to are elaborated in the knowledge organization literature.

Another editor provided the following example of why one page might need multiple category designations:

> "I'm thinking about some of the dog topics. For example, dog is a member of pets; dog is also a member of mammals; both mammals and pets are members of animals but neither is a subcategory of the other. Now, how about dog agility? It needs to go under the dog sports category, which needs to be under the dog category, because it's related to dogs. It also needs to go under the sports category, because it's a sport. It probably also needs to go under the hobby category. But dog and sports do not at any higher point in the hierarchy have a common parent; possibly hobby and sports might fall together again under leisure activities (?), but not all sports are hobbies and not all hobbies are sports. Just wanted to give another example of why something might need to be in multiple categories" (editor 11).

This statement is a clear articulation for the need for polyhierarchy. The editor would like their to be hierarchy, but would also like a single category to be able to belong to more than one supercategory. This is an excellent example of the community working through the issues until they come to a point of recognizing what they need. This type of hierarchical relationship is clearly defined in the literature of knowledge organization, and members of the Wikipedia community recognized the same issues when planning out the category system.

Scope of the category system

Another theme that the editors of Wikipedia who participated in the discussions about the category system in 2004 emphasized was the scope of the category system. The editors were very concerned about the number of category labels that would be applied to each page. One editor commented:

> "Even if each of these categories is relevant (which can be doubted) the original page starts to clutter up rapidly. Logically, there is no almost limit to the extent to which categories can be applied to any page for the imaginative editor" (editor 54).

This editor was worried that categories would be assigned unevenly. Some individuals would choose to apply a large number of category labels, while others would apply few.

Others shared the concern that some editors would apply a large number of category labels to any given page, whiles others would apply few. One editor argued:

> "I suggest that there should be a Guideline for categorisation by which editors (1) exercise caution and err on the side of not ascribing a category unless the text of the page justifies it (2)limit the size of the categorisation link text so that it remains small in relation to the size of the page. It seems to me that without these minimal requirements we will end up with pages exhibiting content confusion. Pages will go through a long gestation period looking rather weird - perhaps even looking like they derive from a banner supported commercial website. Is there any solution to the problem of trying to create hierarchies in a more logical order?" (editor 29).

While this editor clearly wanted to create hierarchical relationships between categories, they had observed how inconsistently these hierarchies were constructed.

Other editors felt that the work of scoping the category system of Wikipedia was such a large task that it should be modeled on existing structures for information organization. One editor brought up the challenge of making relationship types explicit and suggested modeling the category system on the Resource Description Framework (RDF)[vi]:

> "The fix is to label the arrows: describe the relations. This is, in my limited understanding, what RDF does. That uses the terms subject, predicate, and object. The subject is the thing you're categorizing. The object is the category you're adding it to. And the predicate describes the relation. Predicates allow you to make semantic inferences programmatically. So far in the wiki I've seen two predicates, which I would summarise as Is an example of (John Lennon is an example of a vocalist) and Is, er, related in some way to (Musical groups are, er, related in some way to Music; 251 Menlove Avenue is, er, related in some way to John Lennon)" (editor 15).

In the examples provided, this editor is pointing out an example of an instance relationship (hierarchical) and an associative relationship (non-hierarchical). The community did not discuss the challenges of mixing hierarchal and non-hierarchical relationships with no disambiguation.

Other editors suggested modeling the category system on extant directories created to index the World Wide Web:

> "Using other sites as a guide to category structure
> To minimise reinvention of wheels, consider the category structures of Web directories such as www.zeal.com, which have been painstakingly thought out over long periods. Some of them cater well for the "converging path" problem, eg "Country and Western Dancing" should be locatable under "Music" and under "Folklore" and under "Dance Styles". See, as a starter, Zeal's main top-level categories (Entertainment, Work & Money, Computing, Shopping, People & Chat, Sports, Lifestyle, Travel, Library, and Personal), and follow a few down: http://www.zeal.com/category/preview.jhtml?cid=302562 - and/or see the three paths that lead to the "Country and Western Dancing" category at http://www.zeal.com/category/profile.jhtml?cid=225504 (two of them are not the main line and are listed at "Symlinks to this category"). Wikipedia can improve on those Web directories with its upward links (Zeal doesn't let you easily browse up in any line except the main path; others are probably the same)" (editor 46).

This editor is pointing out how much effort could be saved if the category system were modeled on an extant structure.

The design, creation, monitoring and evaluation of the category system require significant community effort. In order to create the most effective system, many discussions were based around how best to scope the category system. Members of the community experessed concern over inconsistency in the average number of categories that might get applied to a given page, made appeals to modeling the syntax of the system on RDF and suggested web directories as other potential models for the category system.

Navigation via the category system

At the time the discussions in this data set were taking place (June, 2004 through January, 2005) the hyperlinked category labels for each Wikipedia page were displayed at the top of each page. All category information is currently displayed in a box at the bottom of each page. There are many pathways to any individual page in Wikipedia. Users arrive to specific pages via a link from a results page from a search engine, from a link in the text of another page, from a link from the infoboxes located in the upper-left or upper-right-hand corners of many pages with a large amount of related content, from a list of links, or from the list of pages provided on the page for any category, etc. Additional work remains to be done to quantify how often the category system is used for navigation in Wikipedia. Regardless of the reality of 2011, with regard to the use of the category system for navigation, many of the contributing editors, who participated in the discussions about the design and implementation of the category system in 2004, felt that navigation was one of the primary ways in which the category system would be used.

One contributing editor articulated the following vision of navigating through the category system:

> "We have to think from the encyclopedia user's point of view. He/she is starting at the top level of the hierarchy with a subject in mind, and they need to know which blind path to go down to find an article on that subject. It might help to think of the problem as a game of twenty questions. The first question we may ask is, "Is your subject a Category:Persons, Category:Places, or Category:Things?" If they choose Category:Persons, then ALL the articles from then on should be about persons. Why? Because we may someday be able to click a link to collapse the hierarchy, and display all the articles below that level in one alphabetical order. If they wanted to know about Stephen King's books, they might choose Category:Things, and have a choice of Category:Animals, Category:Vegetables, Category:Minerals, Category:Ideas, etc., and go down one of those paths. My point is, Categories link only as a hierarchy; Wikipedia articles link as a network to every related article. So as long as the user reaches the article on Steven King (the person), or the articles on Steven King's books using the categories, the articles themselves link to each other" (user 27)."

This statement clearly indicates that the editor felt people would begin their search by looking at the category system from top to bottom for desired content.

Lee and Olson (2005) compared the hierarchical navigation structure of Yahoo! directories with information retrieval via keyword searching in a search engine. One of the factors that they consider in their study is the location of the hierarchical browsing tool on the Yahoo! main webpage. "Quite a few students were new to the Yahoo! directories, which are placed well down the busy Yahoo! homepage. Google has gone even further to take directories off of the homepage (a searcher must click on "more" search options to find a set of icons that includes

"directories" represented rather cryptically by a graphic of an open book)" (Lee and Olson 2005). From looking at this dataset (a subset of all discussions about the category system at this time) it is clear that the decision to move the category box to the very bottom of each page predated community understanding of how the category system would be utilized.

Another editor presented a contrasting vision for how the category system would be used:

> "On the other hand, I think there's a good case to be made for a more bottom-up approach; let's take a look at how things are being categorized, and try to find the patterns in that. It's more the Wikipedia way, too. For example, I've noticed that there are a lot of categories that are non-plural, such as Category:Medicine, Category:Biology and Category:Law. In those cases, rather than being categories containing only one article (Medicine, Biology and Law, respectively) they are instead full of articles and subcategories that are about the indicated topic. It's like there's an implied "Topics relating to -" in front of the categories with singular titles. I think that'd be a good approach to consider as a standard, personally, since it seems to be natural and it would also cut down on the wordiness of many category titles (there'd be a heck of a lot of categories starting "Topics relating to" otherwise)" (editor 3).

This editor is articulating a need for specific guidelines for term construction to facilitate vocabulary control, another example of the Wikipedia community echoing principles frequently discussed in the KO literature. This editor's comments suggest that some members of the community felt that the amount of effort that was being expended in the design of the category system could be reduced if the purpose of the category system were explicitly articulated.

Conclusion

The discussions around the creation of the category system in Wikipedia were very fast-paced in the first six months after it was instituted. Many of the discussions coalesced around a few themes. Although the content of Wikipedia is connected as a graph structure, many members of the community argued for the creation of a category system in which hierarchical relationships between categories would be created. There were many discussions related to hierarchy, but the community did not distinguish between different types of hierarchical relationships, nor did they discuss the challenges users would face when trying to interact with a system in which so many different types of relationships existed between categories without being explicitly described. Members of the community were also conscious of the issue of appropriately scoping the category system. They worried about consistency in terms of the number of categories and super-categories that might be applied to a page. They also suggested external models for the structure and syntax of the category system. In the early days of the existence of the category system it was conceived to be a navigational tool. Many recommendations were made as to how to facilitate navigation through Wikipedia using the category system. The three themes of hierarchy, scope, and navigation, encompass the issues that were most salient to the editors of Wikipedia who contributed to the design and implementation of the category system in 2004.

Acknowledgements

I would like to thank David W. McDonald for his feedback throughout the development of this research project.

REFERENCES

Aitchison, Jean, Gilchrist, Alan, and Bawden, David. 2000. Thesaurus Construction and Use: A Practical Manual. Chicago: Fitzroy Dearborn Publishers.

Forte, Andrea, and Bruckman, Amy. 2008. Scaling Consensus: Increasing Decentralization in Wikipedia Governance. In Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS '08).

Kriplean, Travis, Beschastnikh, Ivan, and David W. McDonald. 2008. Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. In Proceedings of the 2008 ACM conference on Computer supported cooperative work (CSCW '08). ACM, New York, NY, USA, 47-56.

Kittur, Aniket, Chi, Ed H., and Bongwon Suh. 2009. What's in Wikipedia?: mapping topics and conflict using socially annotated category structure. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI '09).

Lee, Hur-Li. & Olson, Hope A. 2005. Hierarchical navigation: An exploration of Yahoo! directories. Knowledge Organization, 32: 10-24.

Svenonius, Elaine. 2000. *The intellectual foundation of information organization*. MIT Press.

Voss, Jakob (2006). Collaborative Thesaurus Tagging the Wikipedia Way, http://arxiv.org/abs/cs/0-604036. Accessed Jan 21, 2011.

[i] http://en.wikipedia.org/wiki/Wikipedia_talk:Categorization
[ii] http://en.wikipedia.org/wiki/List_of_birds. Accessed March 12, 2011.
[iii] http://en.wikipedia.org/wiki/Category:States_of_the_United_States. Accessed March 12, 2011.
[iv] http://en.wikipedia.org/wiki/List_of_cathedrals. Accessed March 12, 2011.
[v] http://en.wikipedia.org/wiki/Category:Plants. Accessed March 12, 2011.
[vi] http://www.w3.org/RDF/. Accessed March 15, 2011.