Exercise 04 -- Controlled & Uncontrolled Vocabularies:

Perception and Relevance of Controlled Vocabulary Quality Issues

Rosanna M. Longenbaker

ILS 531-S70: Indexing and Abstracting

Spring 2013

Dr. Yan Quan Liu

February 19, 2013

Exercise 04 -- Controlled & Uncontrolled Vocabularies:

Perception and Relevance of Controlled Vocabulary Quality Issues

During research conducted by Fidel, many subjects expressed that their perception of controlled vocabularies guided their decision of whether to consult with a thesaurus or use free-text (1991). Quality issues in controlled vocabulary are very relevant because they can lead to inaccurate results in computerized or print searches.

Perception of Controlled Vocabulary Quality Issues

Fidel began his article by presenting one view that was held of controlled vocabularies. Fidel writes that, "the notion that controlled vocabularies are an unnecessary burden on information specialists, as well as on end users, has been the driving force behind much research in recent years" (1991, p. 501). During research reported by Fidel, professional "searchers" were asked to provide reasons for their choices of controlled or uncontrolled vocabulary. "The reasons provided by searchers to explain their selection of search keys reflect their perceptions." (Fidel, 19991, p. 502). Empirical research was conducted so the "perceptions" could be supported by "objective measurements" (Fidel, 1991, p. 502).

The searchers' perceptions of quality affected their choice of using "textwords" or descriptors. Fidel noted that, "they used *only* textwords when they did not trust the descriptors and/or the indexing" (1991, p. 506). "16% of the instances in which searchers decided to enter a term without consulting a thesaurus at all, they did so because they believed the indexing and/or the descriptors would not be useful" (Fidel, 1991, p. 513).

In the research reported by Fidel, "*Over half the time searchers neglected to con-sult a thesaurus, they did so either because they did not trust the quality of the thesaurus, because the*

thesaurus was not available, or because they had to search several databases for a request" (Fidel, 1991, p. 511).

This was also supported by the fact that "some specific thesauri were heavily consulted, while others were ignored most of the time" (Fidel, 1991, p. 513). Dubois also included information in his article that could help to argue that the perception of the quality of a thesaurus affects its use. He wrote that, "potential users familiar with, but not actually using directly, online databases in a research centre environment surveyed by Tatalias… expressed an overwhelming desire (91%) for thesauri as available search aids before embarking on direct searching" (1987, p. 245). These users must have had the perception that the controlled vocabulary would be of use to them otherwise they would not have asked for thesauri.

Some other perceptions affecting the choice to use controlled vocabulary have been disproved through research. The idea that "searching with textwords would result in high re-call, while descriptor searching secures high precision" was disproved. Fidel adds that, "This notion is prevalent even though thesauri were first introduced to improve recall" (1991, p. 505). The "searchers" who participated in the study did not cite this as a reason for their decision of using controlled vocabulary or free-text (Fidel, 1991). Other perceptions were also disproved. For example searchers stated that they used a "textword key" because "the use of textwords increases recall" while empirical evidenced showed that "textword keys are not used with significantly higher frequency to increase recall than descriptors" (Fidel, 1991, p. 506).

The perceptions that people have of natural language could also cause them to want to use a controlled vocabulary. Muddamalle wrote that, "There are serious apprehensions that natural language is full of ambiguities, and, as such, it is not suitable in the retrieval process (1998, p.881). Muddalle referenced a series of experiments conducted in Cranfield. "The

objective of the second of the Cranfield experiments was to test the effectiveness of vocabulary control. Most cited was the result that a minimum controlled index language, one in which only synonyms and word endings were normalized, also per-formed well, or better, in retrieval than any index lan-guage with full vocabulary control. Muddamalle adds that, "Many accepted the findings of Cranfield II and similar experiments, despite their flawed methodologies" (1998, p. 882). This could be an example when false perceptions affect the use of controlled vocabulary.

Nowick & Mering have the view that the process of adding new terms to the Library of Congress Subject Headings is slow, but they mentioned that a method exists to add new terms. They suggested that a "A free-text keyword field could be used for highly technical or new terms not yet incorporated into a CV" (2008, p. 29). In this case, their perception that the process is slow has led them to recommend a quick solution.

Relevance of Controlled Vocabulary Quality Issues

Quality issues in a controlled vocabulary are relevant to a person who is conducting a search because problems in the construction of a controlled vocabulary could cause useful articles not to be returned as the result of a search. "Terminological conditions, request characteristics, and the availability and quality of databases and thesauri, all interact to af-fect the selection of search keys" (Fidel, 1991, p. 506).

Dumais, Furnas, Gomez, and Landauer conducted research to find out how often experts or skilled individuals in certain fields would choose the same key word to describe something. The group wrote that, "The idea of an 'obvious,' self-evident,' or 'natural' term is a myth! Since even the best possible name is not very useful, it follows that there can exist no rules, guidelines or procedures for choosing a good name, in the sense of 'accessible to the unfamiliar user'" (1987). This highlights one quality issue with controlled vocabulary. That is that enough cross-

references are not always available for users to find the preferred search term. They continue by saying "Clearly the only hope for untutored vocabulary driven access is to provide many, many alternate entry terms. Thus aliases are, indeed, the answer, but only if used on a much larger scale than usually considered" (Dumais et al, 1987). Dumais et al still believe controlled vocabulary should be used to locate information even with more "aliases." Their idea being that researchers would enter their terms and receive a list of suggested terms from the computer. In this manner researchers would learn the controlled vocabulary over time.

"In tests designed to secure high to moderate relevance (9.5 in a scale of 10), it was found that the use of free text plus a classification term gave 73.3% recall in a search for semiconductors patents and 62.75% in a search for layered products" (Dubois, 1987, p. 246). These results show the relevance of quality to perception because they led online services to provide both key word and controlled searches. "The type of finding illustrated above has led on the whole to a recognition by providers of online services that both free text and controlled vocabulary search provision is a wise choice" (Dubois, 1987, p. 246). Muddamalle also concluded that both controlled and natural language searches could produce good results. However, Muddamalle noted that using both searches would find even more relevant material than either search alone (1998).

Conclusion

The perception that a person has about the quality of a thesaurus or indexing will impact the decisions that person makes about using controlled vocabulary or free-text. It was believed that users would choose free-text to increase the amount of material recalled, but that was not shown to be the case. The quality issues in controlled vocabulary are relevant because if users are unable to find the correct search words they will not have good results. It is recommended

that people use both controlled vocabulary and free-text to improve accuracy of the results returned.

References

Dubois, C. P. R. (1987). Free text vs. controlled vocabulary: A reassessment. *Online Review*, *11*(4): 243-253.

Dumais, S. T., Furnas, G. W., Landauer, T.K., & Gomez, L. M. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, *30*(11): 964-971. Retrieved from http://0-go.galegroup.com.www.consuls.org/ps/i.do?id=GALE%7CA6365073&v=2.1&u=a30sc&it=r&p=AONE&sw=w

Fidel, R. (1991). Searchers' selection of search keys: II. Controlled vocabulary or free-test searching. *Journal of the American Society for Information Science*, *42*(7): 501-514. Retrieved from: http://0-search.proquest.com.www.consuls.org/docview/216896855/13C58D55DD63A46F14B/8?accountid=13743

Muddamalle, M.R. (1998). Natural language versus controlled vocabulary in information retrieval: A case study in soil mechanics. *Journal of the American Society for Information Science, 49*(10): 881-887. Retrieved from: http://0-search.proquest.com.www.consuls.org/docview/216908500/13C58DF404FE59C7DB/9?accountid=13743

Nowick, E.A. & Mering, M. (2003). Comparisons between Internet users' free-text queries and controlled vocabularies: a case study in water quality. *Technical Services Quarterly*, *21*(2): 15-32. doi:10.1300/J124v21n02_02