Title

Obsolescence in subject description

Permalink

https://escholarship.org/uc/item/084792jk

Journal

Journal of Documentation, 68(2)

ISSN

0022-0418

Author

Buckland, Michael K

Publication Date

2012-03-02

DOI

10.1108/00220411211209168

Preprint. Published in the *Journal of Documentation* vol 68, no 2 (2012): 154-161. Published text may differ slightly.

Obsolescence in Subject Description

Michael K. Buckland

School of Information, University of California, Berkeley, USA

Introduction

Librarians collect and bibliographers list documents in whatever media and genres (books, journals, data sets, movies, etc.) are expected to be most useful for the communities and the purposes to be served. Once collected, documents need to be made accessible in an organized way. In part this is a matter of scale. A collection of one or very few documents can simply be placed in a list or on a shelf and need neither a catalog nor a systematic arrangement, but making each of a million different documents usefully accessible is a different matter. Effective bibliographical access is usually achieved through very concise descriptions. Svenonius (2000) and Taylor & Joudrey (2009) provide introductions. A current development is the conceptual model for the Functional Requirements for Subject Authority Records (International Federation of Library Associations and Institutions 2010).

Librarians assert the subject of a document in two different ways: assigning subject headings to documents' catalog records and also assigning documents to named subject categories through shelf classification. In Robert Fairthorne's colorful terms:

> . . . all retrieval systems demand marks of some kind . . . An object can be marked by changing it intrinsically in some recognizable way--as by painting it, punching a hole, or introducing it to a skunk. This I call 'inscribing.'
> Or it can be changed relative to its environment by putting it upside down, on one side, in an inscribed pigeon-hole, and so forth. This I call 'ordering' the item. Better terms, for less formal contexts are 'marking' and 'parking'. (Fairthorne 1961: 84-85).

Names (marks) are essential for systems to function, but they are, necessarily, linguistic expressions and, as we shall see, they create tensions and difficulties beyond librarians' effective control. Libraries are cultural institutions concerned with recorded knowledge and their mission is to support learning, both research (knowing more) and teaching (sharing understanding). Libraries exist to advance learning, knowledge, understanding, and belief. But what people know, what they would like to know, and what others have learned and written about, all resist mechanical treatment. If it were otherwise, education and knowledge management could be reduced to data processing.

Library users seeking documents relevant to their interests have to locate what they need in the library's terminology. There is, or should be, collaboration, with librarians seeking to anticipate their users' interests and vocabulary, and users trying to make sense of the topic names and codes in the library's catalog, classification, and bibliographies.

The task is exacerbated by increased scale. Collections of millions do need detailed description in order to achieve sufficient fineness of sifting to select a handful rather than a flood of records. In principle the level of detail in subject cataloging is situational, depending on how many different books are acquired in each topic. Since, as an economy, most libraries use whatever subject headings the Library of Congress has assigned, the fineness of detail commonly exceeds local needs.

Describing is inherently a language activity, even if restricted or artificial languages (e.g. classification numbers) are used since they too are culturally grounded and so partake of the character of natural language. All descriptive metadata arise in and reflect a cultural context.

Subject description is usually presented as a two-stage process: first the cataloger examines a document to determine what it is about and then, second, assigns terms (linguistic expressions) from a vocabulary or set of codes to denote those concepts. The literature of librarianship has very little to say about the first stage and concentrates on the second. Research has repeatedly revealed that different indexers will commonly assign different subject index terms to the same document, as will a single indexer at different times (Wolfram & Olson 2007).

Documentary Languages for Naming Topics

Verbal subject approaches, using natural language words, are a simple and popular way to create descriptions, but the multiplicity and fluidity of natural language vocabulary makes for unpredictable results: Should I look under *violin* or *fiddle* or both? The multiplicity of natural language terminology can be mitigated by adopting a restricted vocabulary, either a “controlled vocabulary” of natural language terms (e.g. “*Fiddles* see *Violins*”) or an artificial notation for the descriptive names (e.g. “787.1" in the Dewey Decimal Classification), with natural language indexes to the class numbers in as many different languages as desired. Having an artificial notation of letters, numerals, and other symbols does not mean that it is no longer a language. It is an artificial language–traditionally, a “documentary language”--and it is not immune to the problems of obsolescence and perspective discussed below. It is the same approach as the use of artificially constructed, restricted languages used, for example, in botanical and chemical nomenclature.

Subject Description is Forward-Looking

Patrick Wilson’s classic examination of the nature of bibliographic control, *Two Kinds of Power* (1968), formulates the task as a matter of fitting descriptions. The challenge is to create descriptions that will enable those to be served to identify and select the best documentary means to whatever their ends may be. By definition, the descriptions used by librarians are for future use. This requires the librarian to think about likely needs and to describe (name) in a forward-looking way. To do this the librarian constructs, consciously or not, some mental narrative about future use, some story in which the document in hand would be relevant to future needs. Will some future searcher consider this document “on topic” and better yet “relevant”? It is not simply a matter of what the document is about, but of how it might be viewed in an imagined future. (Greisdorf & O’Connor (2003) provide useful background and discussion). Familiarity with the community and its purposes, ways of thinking, and terminology is an important requirement for the effective librarian.

Vesa Souminen asked the question “What is it that makes a good librarian?” Drawing on Saussure’s ideas, he answers that a good librarian is one who is effective in the task of “filling empty space” with an arrangement of the most relevant documents in relation to each need of each library user (Souminen 1997, 18-19). That the populations of documents, of library users, and of needs are all very large and quite unstable makes the task more difficult, but does not undermine the principle.

This forward-looking stance is also reflected in Suzanne Briet's image of the librarian as a hunter's dog, guided by the hunter (researcher), but prospecting ahead and pointing to prey not visible to the hunter in a dynamic partnership "comme le chien du chasseur – tout à fait en avant, guidé, guidant" ("like a hunter's dog–out in front, guided, guiding" (Briet 1954: 43; 2006, 50-51)).

Subject Description is Backward-Looking

The librarian's effort to be forward-looking is, however, affected by the describing process. Topical description is a matter of naming what a document is about. In practice, descriptions summarize. Assigning subject headings is an extreme of summarization. But what, actually, is "aboutness" about? Stating that a subject heading represents a topic or a concept is valid, but unhelpful because saying that merely points to another name and does not explain. Saying that the subject heading *Dowsing* is 133.32322 in the Dewey Decimal Classification provides an alternative name but does not explain what dowsing is. An explanation of what a subject heading (and, therefore, a document) is "about" must be derived from the discourse from with which the name is associated (Fairthorne 1974). A subject description assigned to a document says that this discourse (document) relates to that discourse (literature, discussion, or dialogue), which means that the subject description is invariably based in the past. Similarly, library users don't want topics, they want discourse concerning a topic: a statement, a description, an explanation, or, at least, a discussion of whatever they are curious about. So a subject heading must derive its meaning from past discourse.

Meanings are established by usage, and so always draw on the past. The librarian, then, is creating descriptions by drawing on the past, but expressing them with an eye to the future. This Janus-like stance might seem difficult enough in a stable world, but reality is made much worse by time, by technology, by the nature of language, and by social change.

Subject Description, Time, and Instability

The librarian's formal act of naming, of recording the topical description of a document or of specifying a relationship between named topics, is necessarily performed at some point in time and inscribed into the apparatus of bibliographies, catalogs, and indexes. As time passes that act recedes from the present into the past. During the same flow of time the prior discourse, upon which the choice of name was derived, has continued, evolved, and changed, and naming practices can be expected to have evolved with those changes. Also, as the future becomes the present, new futures continue to be foreseen, and the forward-looking perspective of the subject cataloger increasingly comes to be related to new and different future discourses. However, an assigned name, once inscribed, is fixed. So, with the passing of time, its relationship with both the then-past discourses and also the then-future expected discourse needs both drift away from the perceptions of an advancing present. Assigned names are, therefore, inherently obsolescent with respect to both the past and the future. Discourses and the librarians flow forward with time, but the assigned names have been inscribed for, and fixed in, a receding past.

New Names for Subjects

New names arise, especially for new topics, through figurative use of language, especially through metaphor. Well-established terms are used figuratively, based on some perceived similarity, for emerging concepts, e.g. a computer *Mouse*. Then, through usage, the new meaning becomes fixed, at first within its context, then more widely. (For an excellent discussion see Norgard (2002)). The instability of language is not of librarians' making, but they must follow. They take a conservative approach because changes in terminology call older terminology into question and the task of making retroactive alterations to the marks in a catalog takes resources away from other worthy purposes.

Libraries and Technology

Libraries depend heavily on technology. Documents are physical objects on paper, film, magnetic disks, or other physical media. Libraries could not operate as they do if the tasks to be performed were not heavily routinized and, most of them, reduced to clerical procedures performed by support staff and, increasingly, delegated to machines. The modern library arose in the spirit of late nineteenth-century technological modernism as "library economy," imbued by Melvil Dewey and others with an emphasis on standards, system, efficiency, and collective progress that lives on in visions of digital libraries and the "semantic web." Detailed control is needed for effectiveness and for efficiency, and librarians, pioneers of new technology for filing and record-processing, inspired modern office management procedures (Flanzreich 1993, Krajewski 2002).

In a library, the machinic and the cultural collide like two tectonic plates and naming lies at the fault-line where librarians use "vocabulary control" to try to mitigate the linguistic ruptures and slidings they can neither prevent nor avoid. So, in the quiet bustle of the library there is an endemic battle between the incorrigibly cultural and aesthetic character of the underlying mission and the machinic tendencies essential for cost-effective performance. The central battle-line of these tensions is in the naming of documents and what they are about.

Mention and Meaning

The fact that the documents in libraries are overwhelmingly textual has allowed the heavy use of natural language processing techniques to infer semantic relationships between documents and between documents and queries. But this is a matter of lexical entities, of character strings, not of meanings. Fairthorne (1961) analyzed this difference by saying that these techniques deal with *mentions* not *meanings*. For example, if *real* and *estate* commonly co-occur in that order, then they are presumed to constitute a phrase. And if the phrase *real estate* and the phrase *property listing* tend to co-occur in the same texts, they are computed as being close in "document space," and a topical relationship is inferred from this "spatial" proximity. If relationships between marks are statistically significant, semantic affinities are implied but not explained. Machines can be programmed to detect regularities and inconsistencies among marks, even if they cannot distinguish sense from nonsense.

It is further evidence of the inherently linguistic character of bibliographical access that formulaic natural language processing techniques work quite well, but not always and not very reliably. It is the textual (lexical) similarity between documents that allows relatedness between discourses and/or descriptions to be inferred, since different discussions of the same topic will tend to use the same terms. From the method employed, homographs with different meanings (e.g. *host* (landlord) and *host* (crowd)) will dilute the precision of retrieval. The compelling economic attraction of this approach is, of course, that it is mechanical and so can be delegated to machines. The poverty of this approach arises when different vocabularies are used to refer to the same topic without using (mentioning) the same terms. For this and for cross-lingual search, formal structures, such as bilingual dictionaries or statistical associations, can be helpful.

Technical writing on information retrieval is, understandably, heavily engaged with natural language processing, especially named entity extraction, parsing to identify adjective-noun phrases, and all manner of frequency counts and statistical association. The name of George K. Zipf, the pioneer of word frequency analysis, is invoked rather than Peirce, Saussure, or Wittgenstein. It is only in recent years that the literature on the nature of language has received much attention. David Blair's explanation, in his *Language and Representation in Information and Retrieval* (1990), of the relevance of Wittgenstein's ideas to subject description and the insoluble problem of unlimited semiosis was a major milestone. (See also Blair 2003).

The relevance of the work of Eleanor Rosch and George Lakoff on categories and language (e.g. Lakoff 1987) is now widely recognized as important. Norgard (2002) provides a good overview of how linguistic expressions resist automatic indexing. Research on the social practices of science has had an impact during the past decade on the understanding of the use and role of documents and document description. *Sorting Things Out: Classification and its Consequences* by Bowker and Star (2000) is recommended for its case studies revealing social agendas in the design of categorization systems.

Naming and Multiplicity

It is not simply that a new document has to be positioned in relation to both past discourse and future needs. Additional complexity arises because there are, of course, not one but many simultaneous discourses. Language evolves within each community of discourse and produces and evokes that community. So every such community has its own more or less specialized, stylized practice of language. Attempts at controlled or stabilized vocabulary must deal with the multiple and dynamic discourses and the resultant multiplicity and instability of meanings. Most bibliographies and catalogs have a single topical index but cover material of interest to more than one community. Since each community has at least slightly different linguistic practices, no one index will be ideal for everyone and, perhaps, not for anyone. For example, in vernacular discussion of health, the terms *cancer* and *stroke* are commonly used, but in a professional medical discourse *neoplasm* and *cerebrovascular accident* are preferred names. So, in theory, multiple, dynamic indexes, one per community, would be ideal. It is not, however, only a matter of linguistic variation, but also of perspective. Different discourses discuss different issues or, when the same issue, from different perspectives. A *rabbit* can be discussed as a pet, as a pest, or as food. In medicine, specialists in anesthesiology, geriatrics, and surgery might all ask for recent literature on, say, *Cardiac arrest*, but because they are interested in different aspects they will not, in practice, want the same documents. (For a detailed examination see Petras (2006), also Buckland, Jiang, Kim & Petras, (2001)).

Naming and Cultural Changes

The vocabulary used by librarians to characterize their documents can become problematic for other reasons as the world changes. There are cognitive developments: New ideas and new inventions need new names. *Horseless carriages* were invented, then renamed *Automobiles*. Also, new referents emerge for existing names. Seventy years ago the word *computer* meant a human who performed calculations, but now always means a machine. More recently the word *printer* made a same transition.

The use of artificial notation, such as the Dewey Decimal Classification 330 for Economics, is intended to reduce dependence on natural language and is likely to prove more stable when new names replace old names for the same topic. However, such a notation also positions the topic in relation to other topics in a away that mere naming does not and the perspective implied by this positioning is likely to be contested and unstable and thereby obsolescent as will the verbal explanation of what the notation denotes.

There are also consequences for library naming from affective changes. Even when the denotation is stable, the connotation or attitudes to the connotation may change. Always, some linguistic expressions are socially unacceptable. That might not matter much, except that what is deemed acceptable or unacceptable not only differs from one cultural group to another, but changes over time, and, especially during changes, may be the site of contest. The phrase *Yellow peril* was widely used to denote what was seen as excessive immigration from East Asia, but it is now considered too offensive to use even though there is no convenient and acceptable replacement name and the phrase remains needed in historical discussion.

Much has been written concerning the social correctness of library subject headings, both the terms used and how they are related to each other. "*Sexual perversion* see also *Homosexuality*" was once, but is no longer acceptable. Sanford Berman's *Prejudices and Antipathies: A Tract on the LC Subject Heads Concerning People* (1971) is still an excellent introduction and Joan Marshall's *On Equal Terms: A Thesaurus for Non-Sexist Indexing and Cataloging* (1977) is another classic treatment. Berman picks out scores of subject headings, explains why he thinks each is offensive, and proposes more neutral alternative terminology. His examples and commentary show how naming always reflects a cultural perspective, that terminology acceptable to one group may be offensive to another, and that attitudes change. His examples are far too many and too interesting to summarize adequately here. *Jewish question* implies untenable assumptions; *Gypsies* are not from Egypt and prefer to be called *Roma*; the cross-reference "*Rogues and vagabonds* see also *Gypsies*" exhibits prejudice; the headings *Mammies* and *Negroes* are offensive to those so named; *Eskimos* are properly called *Inuit*. One's own behavior is reflected as superior to that of others: Rebellions by slaves are named "insurrections," rebellions by Whites are more positively named "revolutions." *Indians of North America, Civilization of* in the Library of Congress Subject Headings did not refer to the culture of Native Americans, but to progress in the eradication and replacement of their culture, as the Library's instruction made clear: "Here is entered literature dealing with efforts to civilize the Indians…" European powers have colonies; the U.S. has off-shore "territories and possessions" not called colonies. Many of Berman's examples reflect a male and Christian world view, the social attitudes of past times, and obsolete medical and psychological terminology (e.g. *Idiocy*). In some cases, counter-arguments can be made. For example, using *Roma* for Gypsies is counterproductive if the library's users are unfamiliar with that term.

Conclusion

Tracing shifts in library naming back through time is a highly educational form of cultural and linguistic archaeology. The Library of Congress Subject Headings, a hundred years old, with well over 100,000 different headings, and difficult to update, remains an easy target in spite of many reforms. It is a good example of a problem that is endemic in indexes and categorization systems: Linguistic expressions are necessarily culturally grounded, and, for that reason, in conflict with the need to have stable, unambiguous marks to enable library systems to perform efficiently. A static, effective subject indexing vocabulary is a contradiction in terms.

Acknowledgments

The author benefitted from the comments of Howard Greisdorf, Vivien Petras, and Julian Warner and from the research assistance of Janice Woo. This paper draws in part on Buckland (2007).

References

Berman, S. (1971). *Prejudices and antipathies: A tract on the LC Subject Heads concerning people*. Scarecrow, Metuchen, NJ.

Blair, D. C. (1990). *Language and representation in information and retrieval*. Elsevier Science, Amsterdam.

Blair, D. C. (2003). Information retrieval and the philosophy of language. *Annual Review of Information Science and Technology* 37, 3-50.

Bowker, G. & Star, S. (2000). *Sorting things out: Classification and its consequences*. MIT Press, Cambridge, MA.

Briet, S. (1954). “Bibliothécaires et documentalistes”. *Revue de la documentation* 21, fasc. 2: 41-45.

Briet, S. (2006). *What is documentation?* Transl. and ed. by R. E. Day & L. Martinet. Scarecrow, Lanham, MD.

Buckland, M. K. (2007). “Naming in the library: Marks, meaning, and machines”, In: Todenhagen, C. & Thiele, W. (Eds.), *Nominalization, nomination and naming in texts,* (pp. 249-260). Stauffenburg, Tübingen.

Buckland, M. K., Jiang, H., Kim, Y., & Petras, V. (2001). “Domain-based indexes: Indexing for communities of users”, In: Chaudiron, S. & Fluhr, C. (Eds.). *3e Congrès du Chapitre français de L’ISKO, 5-6 juillet 2001. Filtrage et résumé informatique de l’information sur les réseaux* (pp. 181-185). Paris: Université Nanterre Paris X. Preprint retrieved Nov 6, 2010, from http://metadata.sims.berkeley.edu/papers/ISKObuck.pdf

Fairthorne, R. A. (1961). *Towards information retrieval*, Butterworths, London.

Fairthorne, R. A. (1974). “Temporal Structures in Bibliographic Classification”, In Wojciechowski, J. A. (Ed.). *Conceptual basis of the classification of knowledge : proceedings of the Ottawa Conference on the Conceptual Basis of the Classification of Knowledge, Oct. 1-5, 1971* (pp. 404-412). Verlag Dokumentation, Pullach, Germany.

Flanzreich, G. (1993). “The role of the Library Bureau and office technology”, *Libraries & Culture* 28, 403-429.

Greisdorf, H. & O’Connor, B. (2003). “Nodes of topicality: Modelling user notions of *on topic* documents”, *Journal of the American Society for Information Science and Technology,* 54, 1296-1304.

International Federation of Library Associations and Institutions. Working Group on the Functional Requirements For Subject Authority Records. (2010). *Functional Requirements for Subject Authority Data (FRSAD). A Conceptual Model*. Eds Editors: M. L. Zeng, M. Žumer, A. Salaba. Retrieved Jne 3, 2011. http://www.ifla.org/files/classification-and-indexing/functional-requirements-for-subject-authority-data/frsad-final-report.pdf Visited 3 June 2011

Krajewski, M. (2002). *Zettelwirtschaft: Die Geburt der Kartei aus dem Geiste der Bibliothek.* Berlin: Kulturverlag Cadmos. (English edition forthcoming, MIT Press).

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press, Chicago.

Marshall, J., comp. (1977). *On equal terms: A thesaurus for non-sexist indexing and cataloging*. Neal-Schuman, New York, NY.

Norgard, B. A. (2002). *Linguistic expressions and indexing information resources*. Ph.D dissertation in Library and Information Studies, University of California, Berkeley.

Petras, V. (2006). *Translating dialects in search: Mapping between specialized languages of discourse and documentary languages*. Ph. D dissertation in Information Management and Systems, University of California, Berkeley.

Souminen, V. (1997). *Filling empty space: A treatise on semiotic structures in information retrieval, in documentation, and in related research*. (Acta Universitatis Ouluensis, Humaniora B27). Oulu University Press, Oulu, Finland.

Svenonius, E. (2000). *The intellectual foundations of information organization*. MIT Pr., Cambridge, MA.

Taylor, A. & D.N. Joudrey. (2009). *The organization of information*. 3nd ed. Libraries Unlimited, Westport, CT.

Wilson, P. (1968). *Two kinds of power: An essay on bibliographic control*. University of California Press, Berkeley, CA.

Wolfram, D. & H.A. Olson. (2007). A method for comparing large scale inter-indexer consistency Using IR modeling. In Proceedings of the 35th Annual Conference of the Canadian Association for Information Science. Visited June 4, 2011 at http://www.cais-acsi.ca/proceedings/2007/wolfram_2007.pdf

Michael K. Buckland can be contacted at buckland@ischool.berkeley.edu