

<http://people.ischool.berkeley.edu/~buckland/naminglib.pdf>

Preprint of “Naming in the Library: Marks, Meaning and Machines.” In: *Nominalization, Nomination and Naming in Texts*. Christian Todenhagen & Wolfgang Thiele, eds. Tübingen, Germany: Stauffenburg Verlag, 2007, pp. 249-260. Published version may differ slightly.

Naming in the Library: Marks, Meaning, and Machines

Michael K. Buckland

Introduction

In a library there is a lot of naming: Marking documents with descriptive names and assigning documents to named categories. This necessary naming activity is, however, the site of tensions between the procedural need for stable marks and the inherent multiplicity and instability of linguistic expressions used to represent topics. Here we provide a brief introduction to the issues, tensions, and compromises involved.

Documents, Collections, and Topic Descriptions

Bibliographers list and librarians collect documents in whatever media or genres (books, journals, data sets, movies, etc.) are expected to be most useful for the communities and the purposes to be served. But once included, documents have 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 needs 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 achieved through very concise descriptions. Svenonius (2000) and Taylor (2004) provide introductions.

Librarians make descriptions of documents in their catalogs and through classified arrangements on their shelves. Assigning topic names to documents and assigning documents to named topical categories is central. 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 library 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, 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 category names in the library's catalog, classification, and bibliographies. Describing is inherently a language activity, even if restricted or artificial languages are used, since they too are culturally grounded and so partake of the character of natural language.

Bibliographic descriptions follow rules. For more than a century there has been gradual international standardization of rules for representing the imprint (where and by whom published), collation (physical features of a document), proper names (authors, institutions, and places), and other attributes of documents. The real difficulty, however, for both librarians and library users is in describing what a document is about, in naming its topic, which is usually presented as a two-stage process: First, the cataloger examines a document to determine what concepts it is about; then, second, assigns terms (linguistic expressions) from a vocabulary to denote those concepts. The literature of librarianship has very little to say about the first stage and concentrates on the second. Research has revealed that different indexers will commonly assign different index terms to the same document, as will a single indexer at different times.

Documentary Languages for Naming Topics

There are a variety of methods for representing what documents are about: subject classifications, lists of subject headings, thesauri, and so on. Recently developed varieties include "ontologies" and "folksonomies." A traditional collective term for all of them is "documentary languages" (or, sometimes, "bibliographic languages"). We need not examine each type, but will note four dimensions along which they vary.

Notation

Verbal approaches, using natural language words, are a simple and popular way to create descriptions. However, using ordinary vocabulary has disadvantages, and ease of creation does not lead to easy effective use. 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 (“controlled”) vocabulary.

Natural language words do not arrange themselves in a helpful way. Alphabetical filing order is determined by accidents of spelling rather than meaningful semantic relationships. “If the names of the classes, in a natural language, are used to arrange them, we do not get a helpful order. In fact names scatter classes in a most unhelpful chaotic order. It will give us an order like algebra, anger, apple, arrogance, asphalt, and astronomy,” wrote the famous Indian librarian, S. R. Ranganathan (1951, 34). Another limitation of using natural languages to create indexes is that they are ordinarily created only in a single language.

These problems can be addressed by using an artificial notation for the descriptive names (as in the Dewey Decimal Classification) designed to achieve some desired arrangement, 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 and 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.

Vocabulary Control

Language is characterized by multiplicity, such as singular and plural forms, variant spellings, synonyms, and antonyms. The same topic could be assigned any number of names, or represented in an indefinite number of ways (“unlimited semiosis”), so documents on the same topic could be scattered under any of several different headings. A searcher might find some not others. The librarians’ solution is “vocabulary control” whereby one form of name, e.g. *Violins*, is “preferred,” and used exclusively. Other commonly-used but “non-preferred” terms are listed, but only to re-direct the searcher to the preferred term: e.g. “*Fiddles* see *Violins*.” An “authority file,” a list of carefully differentiated preferred and non-preferred terms, is compiled and followed.

Vocabulary control can take care of synonyms, near-synonyms, antonyms, and variant spellings. Exact synonyms are quite rare. It is near-synonyms that are frequent. For example, *Birds* and *Ornithology* are very closely related but not quite the same. Near-synonyms require endless situational judgements concerning what to combine and what to differentiate.

In practice, vocabulary control also extends to hierarchical and other relationships (“See also”). Library vocabulary control extends beyond semantic to functional relationships, which differentiates this kind of thesaurus from a traditional lexicographic thesaurus. For example, *Biogas*, *Pig manure* and *Water hyacinths* are very different in etymology and denotation, but, since pig manure and water hyacinths are important ingredients in making biogas, anybody interested in one, might well be interested in the others, and so “see also” references in both directions

between each and *biogas* are justifiable in a library subject catalog.

Coordination

Many documents are concerned with complex topics, needing a phrase to express the scope. A simplistic approach ordinarily used in current search engines is to merely list the terms, in any order, needed to comprise the meaning. Documents about the “parents of handicapped children” would have three terms: the three keywords *children* and *handicapped* and *parents*. But there are also some documents on “the children of handicapped parents,” which would also be retrieved by the same keywords, but, being relatively few, would probably not be noticed in the retrieved set. Computers can easily handle keyword searches, but the earlier technology of catalog cards cannot: any such combination has to be “pre-coordinated” using some syntax at the time of cataloging to differentiate and to express relationships among the terms. The Library of Congress Subject Headings has two quite separate headings: *Children of Handicapped Parents* and *Parents of Handicapped Children*, and, because they constitute grammatical phrases, there is no confusion between them. This is a simple case. Syntactic rules are used to generate quite elaborate headings in which a primary term is progressively qualified, either as a complex phrase, such as *Hand-to-hand fighting, oriental, in motion pictures*, or with a chain of qualifying terms, as in *God--Knowableness--History of doctrines--Early church, ca. 30-600--Congresses*. The latter is a single subject heading in which the syntax is implicit from the positioning of the terms. For an English speaker accustomed to adjectives preceding the nouns they qualify, it sounds more natural if such headings are read in a reverse order with some conjunctions and prepositions added: “Congresses on the history of doctrines in the Early Church, ca 30-600, concerning the knowableness of God.” The artificial notation of library classification schemes allows elaborately coordinated topics to be expressed more concisely. In this way all documentary languages for naming topics, beyond the simplest use of keywords, have grammar and well as a vocabulary.

Fineness

A collection composed of one or very few documents needs no catalog. At the other extreme, distinguishing every little nicety in order to differentiate every document becomes cumbersome. Collections of millions do need very detailed description in order to achieve the fineness of sifting required to select a handful rather than a flood of records. In practice 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 tends to not to follow local needs.

Naming 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. It is not simply a matter of what the document is about, but of where it might be useful in an imagined future. Familiarity with the community and its purposes, ways of thinking, and terminology is an important requirement for the effective librarian.

Vesa Souminen (1997) asked the question “What is it that makes a good librarian?” Drawing on Saussure’s ideas, he answers that the task is one of “filling empty space.” The good librarian is one who is effective in arranging documents in relation to each need of each library user. 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.

Suzanne Briet (1954: 43) extended the idea of this forward-looking stance with her image of the librarian as a hunting dog, guided by the hunter (researcher), but prospecting ahead and pointing to prey invisible to the hunter in a dynamic partnership (*Comme le chien du chasseur – tout à fait en avant, guidé, guidant.*) (See also Briet (2006: esp. 50-51)).

Naming is Backward-Looking

The librarian’s effort to be forward-looking is, however, affected by the describing (naming) process. Topical description is a matter of naming what a document is about and describing is a matter of summarizing. Assigning subject headings is an extreme of summarizing what a document is about. 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. An explanation of what a subject heading (and, therefore, a document) is “about” must be derived from the discourse from which the name originates (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: a statement, a description, an explanation, or, at least, a discussion of whatever they are curious about. So a subject heading “about” a topic derives its importance 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 the reality of library naming practices is made much worse by time, by technology, by the nature of language, and by social change.

Naming, Time, and Instability

Time of Inscription

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 indexes and catalogs. 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 would evolve with those changes. Also, as the future becomes the present, new futures continue to be foreseen, and the forward-looking perspective would increasingly be related to changed 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 drift away from relevance to the perceptions of an advancing present. Assigned names are, therefore, inherently obsolescent with respect to both the past and the future. Discourses and the librarian flow forward with time, but the assigned names have been inscribed for, and fixed in, a receding past.

Figurative use of language

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. The instability of language is not of librarians' making, but they must follow. They take a conservative approach because changes in terminology call into question older terminology and the task of making retroactive alternations 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 or 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, the "semantic web", and the "virtual." 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 *information* and *retrieval* commonly co-occur in that order, then they are presumed to constitute a phrase. And if the phrase *information retrieval* and the phrase *vector space* 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 the same words are mentioned when the same or very similar language is in use. 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, are helpful.

The importance of language and of naming has not, however, engendered much mutual interest between librarianship and linguistics, despite some awareness (e.g., Sparck Jones & Kay 1973). Technical writing on information retrieval is 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 in the library literature. 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. 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. See also Blair (2003).

Research on the social practices of science has had an impact during the past decade on the understanding the use and role of documents and document description. *Sorting Things Out: Classification and its Consequences* by Bowker and Star (2000) is strongly recommended for its case studies revealing social agendas in the design of categorization systems.

Naming and Cultural Change

It is not simply that a new document has to be positioned in relation to both past discourse and that future needs. Additional complexity arises because there are, of course, not one but many simultaneous communities of discourse.

Language evolves within communities of discourse and produces and evokes those communities. 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 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 the 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.

Aside from these "dialect" differences, 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. Sixty years ago the word *computer* meant a human who performed calculations, but now always means a machine. More recently the word *printer* made the same transition.

Questionable naming practices can have non-linguistic causes. As one example, the *International Classification of Diseases*, widely used on death certificates to name causes of death, excluded some known causes of death. The explanation is that doctors thought that naming diseases for which there

was no known cure might draw attention to the inadequacies of the medical profession, so, instead of naming the actual cause of death, some other, broader or vaguer name was used.

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 the Far East, but it is now considered too offensive to use even though there is no convenient and acceptable replacement name and the phrase is needed in historical discussion.

Fighting Words

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 an excellent introduction and Joan Marshall's *On Equal Terms: A Thesaurus for Non-Sexist Indexing and Cataloging* (1977) is another classic treatment. (See also Olson (2002).

Berman picks out scores of subject headings, explains why 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* did not refer to the culture of Native Americans, but to progress in the eradication 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.

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, is 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.

Acknowledgments

The author benefited from the comments of Howard Greisdorf, Vivien Petras, and Julian Warner and from the research assistance, in 1992, of Janice Woo.

References

- Blair, David C. (1990): *Language and Representation in Information and Retrieval*. Amsterdam: Elsevier Science.
- Blair, David C. (2003): Information retrieval and the philosophy of language. *Annual Review of Information Science and Technology* 37:3-50.
- Bowker, Geoffrey & Susan Leigh Star (2000). *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA/London: MIT Press.
- Briet, Suzanne (1954): "Bibliothécaires et documentalistes". *Revue de la Documentation* 21, fasc. 2: 41-45.
- Briet, Suzanne. (2006): *What is Documentation?* Transl. and ed. by R. E. Day & L. Martinet. Lanham, MD: Scarecrow.
- Buckland, M. K., Hailing Jiang, Youngin Kim & Vivien Petras (2001). "Domain-Based Indexes: Indexing for Communities of Users." In: 3e Congrès du Chapitre français de L'ISKO, 5-6 juillet 2001. *Filtrage et résumé informatique de l'Information sur les réseaux*. Paris: Université Nanterre Paris X. 181-185. <http://metadata.sims.berkeley.edu/papers/ISKOBUCK.pdf>
- Fairthorne, Robert A. (1961): *Towards Information Retrieval*. London: Butterworths.
- Fairthorne, Robert A. (1974): "Temporal Structures in Bibliographic Classification". *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*, ed. by Jerzy A. Wojciechowski. Pullach, Germany: Verlag Dokumentation, 404-412.
- Flanzreich, Geri (1993): "The Role of the Library Bureau and Office Technology". *Libraries & Culture* 28, 403-429.
- Greisdorf, Howard/Brian O'Connor (2003): "Nodes of topicality: Modelling user notions of *On topic* documents". *Journal of the American Society for Information Science and Technology* 54, no. 14, 1296-1304.
- Krajewski, Markus (2002): *Zettelwirtschaft: Die Geburt der Kartei aus dem Geiste der Bibliothek*. Berlin: Kulturverlag Cadmos.
- Marshall, Joan, comp. (1977): *On Equal Terms: A Thesaurus for Non-Sexist Indexing and Cataloging*. New York: Neal-Schuman.
- Norgard, Barbara A. (2002): *Linguistic Expressions and Indexing Information Resources*. Ph.D dissertation in Library and Information Studies, University of California, Berkeley.
- Olson, Hope (2002): *The Power to Name: Locating the Limits of Subject*

- Representation in Libraries*, Dordrecht/Boston/London: Kluwer Academic Publishers.
- Petras, Vivien (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.
- Ranganathan, S. R. (1951): *Classification and Communication*. Delhi: University of Delhi.
- Souminen, Vesa (1997): *Filling Empty Space: A Treatise on Semiotic Structures in Information Retrieval, in Documentation, and in Related Research*. Oulu University Press. (Acta Universitatis Ouluensis, Humaniora B27).
- Sparck Jones, Karen & Martin Kay (1973): *Linguistics and information science*. New York, Academic Press. FID publ. 492.
- Svenonius, Elaine (2000): *The Intellectual Foundations of Information Organization*. Cambridge, MA: MIT.
- Taylor, Arlene (2004): *The Organization of Information*. 2nd ed. Westport, CT: Libraries Unlimited.
- Wilson, Patrick (1968): *Two Kinds of Power: An Essay on Bibliographic Control*. Berkeley: University of California Press.