**PAPER**

**Collection Development for Diversity and Inclusion: The Long Tail Challenge**

**Jeppe Nicolaisen**, University of Copenhagen, Denmark, jep.nic@hum.ku.dk

**Keywords: Collection development; Diversity; Inclusion; Skewed distributions**

# Introduction

Items of information are usually distributed among sources. Words in texts, authorships, and scientific articles in journals are just some examples. Such distributions tend to follow a power law where a small fraction of sources account for a large amount of the items (see figure 1).



Figure 1. The long tail

The three bibliometric laws (*Zipf’s law, Lotka’s law*, and *Bradford’s law*) adequately describe this phenomenon (de Bellis, 2009). In recent years, thanks to a bestselling book by Chris Anderson, former editor of WIRED magazine, the phenomenon is now widely known as *the long tail*. Anderson (2006) argues that products in low demand (i.e., products that individually have a low sales volume) can collectively build a better market share than products in high demand. If the store and distribution channel is large enough, the collective sales volume of products in low demand may even exceed the sales volume of bestsellers and blockbusters. That is actually how companies such as Amazon.com make most of their profit. They not only cover bestsellers and blockbusters (the head), but offer their customers more or less complete coverage of all products (including the long tail). Not only has this proven to be a strong business idea, it also solves the problem of developing an assortment for all needs. By simply offering (almost) everything, all needs may (in principle) be satisfied. Libraries and information systems (e.g., databases) face a similar problem: How to develop collections for diversity and inclusion. The same solution, “simply offer everything to the users”, do not apply here. The reason is that the “store” and “distribution channel” is not large enough. No library can afford to buy everything, and no database can afford to provide complete indexing of all documents. Some selection is required. The aim of this paper is to discuss the pitfalls and possibilities for developing collections for diversity and inclusion when dealing with skewed distributions. We will start by analyzing some typical proposals for selection based on quantitative methods. These include random selection and *Bradfordizing*. Next we will analyze typical proposals for selection based on qualitative methods – specifically expert assessments. We will see that the long tail of skewed distributions challenges the very idea of developing collections for diversity and inclusion. The paper ends with a discussion of this challenge, and a proposal for how to deal with it.

**Quantitative methods**

Statisticians have proved that, for small sample sizes, when drawing a random sample from a skewed population, the usual frequentist intervals for the population mean cover the true value less often than their stated frequency of coverage (Meeden 1999). Though it is hard to imagine that any library or database would ever want to select material or content purely at random, it is important to recognize that sampling from skewed populations is actually quite a challenge. *Bradfordizing* is one alternative to random sampling. Bradford’s law (Bradford 1934; 1948) concerns a regularity observed in scientific journals: Articles on a given subject are published unevenly by journals. A few journals publish a relatively high number of the articles whereas many journals publish only one or a few articles each. Burrell (1988) notes that although Bradford’s law strictly speaking is about articles and their concentration/dispersion in journals, it is customary to speak in terms of a population of sources producing items. Moreover, a number of studies have shown that Bradford’s law applies to other sources and items than just journals and articles (e.g., Worthen 1975; Kirby 1991; Tonta and Al 2006). Thus, by conducting a Bradford analysis of how items are distributed in sources, it is possible to identify the so-called *core* (i.e., the head of the long tail). According to Bradford’s law, about a third of the items are produced by just a few sources, and the so-called *Bradford multiplier* assesses that it takes four times as many sources to cover the next 1/3 of the items, and sixteen times as many sources to cover the last 1/3 of the items. Identifying the core, and thus limiting selection to the head of the long tail distribution has been proposed as an effective selection procedure. Bertram C. Brookes was among the first to address the potential. In a short note in *Nature* he wrote that it “seems to offer the only means discernible at present to reducing the present quantitative untidiness of scientific documentation, information systems and library services to a more orderly state of affairs capable of being rationally and economically planned and organized” (Brookes 1969, 953). The basic assumption underlying this idea is that Bradford’s law functions as a neutral and objective method. However, as demonstrated empirically by Nicolaisen and Hjørland (2007), operationalizing the concept of subject, when conducting Bradford analyses, influences on the results of the very same. Consequently, Bradford’s law does not automatically function as a neutral method. The results of utilizing Bradford analysis as a method for identifying the core information sources of any subject, field or discipline will depend in part on the way the subject is operationalized. Moreover, selecting just the core of information sources tends to favor dominant theories and views while suppressing views other than the mainstream at a given time (Nicolaisen and Hjørland 2007). Thus, collection development for diversity and inclusion will likely fail if Bradfordizing is employed as selection procedure.

**Qualitative methods**

Qualitative methods for collection development include a variety of methods (Evans 2000). Here, we shall limit the discussion to just one common approach: Expert assessment. This approach suffers from some of the same problems as the quantitative approaches mentioned above. Three examples: J. Bernal Gilmore points to Gottfredson (1978) and Mahoney (1985) as evidence for the claim that there is little reliability and validity in expert assessments. He therefore offers the suggestion that “nothing forces the conclusion that it would be the least bit foolish to make many of our choices by drawing lots” Gilmore (1991, 148). This would be a kind of random sampling, which, as we saw above, is actually quite delicate to perform without bias when the population is skewed. Cole, Cole and Simon (1981) reports the results of an experiment in which 150 proposals submitted to the National Science Foundation were evaluated independently by a new set of reviewers. Their results show that getting a research grant depends significantly on chance, or as the authors put it, “the luck of the reviewer draw” (Cole, Cole and Simon 1981, 885). To eliminate the element of chance, the authors suggest to increase the number of reviewers. This, however, is a bit similar to Bradfordizing, and consequently faces the same problem of tending to favor dominant theories and views while suppressing views other than the mainstream at a given time. To avoid this problem, experts are sometimes told to assume a neutral point of view. Wikipedia, for instance, operates with a neutral point of view policy (NPOV) “which means representing fairly, proportionately, and, as far as possible, without bias, all of the significant views that have been published by reliable sources on a topic"[[1]](#footnote-0). However, this policy has been criticized for being too naïve:

*No encyclopedia, no knowledge, no explanation can be said to be neutral in the sense that is free of its authors' position in the world. The NPOV policy's reliance on trusted, reliable sources to distinguish between facts and opinions, and between majority and minority views does not free Wikipedians from their "makeup and position in the world" (Nagel, 1986, 5) when they construct entries for Wikipedia* (Mai 2016, 21).

**Discussion**

Perhaps the late American R&B singer Bo Diddley said it best: “Don't let your mouth write a check that your tail can’t cash”[[2]](#footnote-1). Although diversity and inclusion are noble and worthy ideals for collection development, they are ideals, which in practice are impossible to honor completely. This, however, does not imply that we should give up these ideals in favor of less worthy but accessible goals. No, we should maintain them exactly as ideals. In practice, we should then be clear about the various constraints that limit us from completely honoring them. For instance, that costs makes it impossible to cover everything, that we have selected a large part of the head of the long tail distribution as this part generates most interest per item, and that we have used subject specialists as experts, which implies the usual selection biases. The more accurately we inform about the selection process and its limitations and biases, the better we assist the users in their search process. They will then know what to expect from our collection, what kind of material they may find, and consequently not find, and therefore need to search for elsewhere.

 *[All parts of this extended abstract will be developed further in the full paper]*

# REFERENCES

Andersen, Chris. 2006. *The Long Tail: Why the Future of Business is Selling Less of More.* New York, NY: Hyperion.

de Bellis, Nicola. 2009. *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics.* Lanham, MA: The Scarecrow Press.

Bradford, Samuel C. 1934. “Sources of Information on Specific Subjects.” *Engineering* 26: 85-86.

Bradford, Samuel C. 1948. *Documentation*. London, UK: Crosby Lockwood.

Brookes, Bertram C. 1969. “Bradford’s Law and the Bibliography of Science.” *Nature* 224: 953–956.

Burrell, Quentin L. 1988. “Modelling the Bradford Phenomenon.” *Journal of Documentation* 44, no. 1: 1-18.

Cole, Stephen, Cole, Jonathan R. and Simon, Gary A. 1981. “Chance and Consensus in Peer Review.” *Science* 214 (November 20): 881-886.

Evans, G. Edward. 2000. *Developing Library and Information Center Collections*. Englewood, CO: Libraries Unlimited.

Gilmore, J. Barnard. 1991. “On Forecasting Validity and Finessing Reliability*.” Behavioral and Brain Sciences* 14, no. 1: 148-149.

Gottfredson, Stephen D. 1978. “Evaluating Psychological Research Reports: Dimensions, Reliability, and Correlates of Quality Judgments.” *American Psychologist* 33, no. 10: 920-934.

Kirby, Steven R. 1991. “Reviewing United States History Monographs: A Bibliometric Survey.” Collection Building 11, no. 2: 13-18.

Mahoney, Michael J. 1985. “Open Exchange and Epistemic Progress.” *American Psychologist* 40, no. 1: 29-39.

Mai, Jens-Erik. 2016. “Wikipedian’s Knowledge and Moral Duties.” *Nordisk Tidsskrift for Informationsvidenskab og Kulturformidling* 5, no. 1: 15-22.

Meeden, Glen. 1999. “Interval Estimators for the Population Mean for Skewed Distributions with a Small Sample Size.” *Journal of Applied Statistics* 26, no. 1: 81-96.

Nagel, Thomas. 1986. *The View from Nowhere*. Oxford, UK: Oxford University Press.

Nicolaisen, Jeppe, and Hjørland, Birger. 2007. “Practical Potentials of Bradford’s Law: A Critical Examination of the Received View.” *Journal of Documentation* 63, no. 3: 359-377.

Tonta, Yaşar, and Al, Umut. 2006. “Scatter and Obsolescence of Journals Cited in Theses and Dissertations of Librarianship.” *Library & Information Science Research* 28, no. 2: 281-296.

Worthen, Dennis B. 1975. “The Application of Bradford’s Law to Monographs.” *Journal of Documentation* 31, no. 1: 19-25.

1. [https://en.wikipedia.org/wiki/Wikipedia:Neutral\_point\_of\_view](https://en.wikipedia.org/wiki/Wikipedia%3ANeutral_point_of_view) [↑](#footnote-ref-0)
2. <https://www.brainyquote.com/authors/bo_diddley> [↑](#footnote-ref-1)