

Some Thoughts on Preserving Functions of Library Catalogs in Networked Environments

by Koraljka Golub

EDITOR'S SUMMARY

Classification and subject indexing systems have long been the mainstay of established information providers to deliver content precisely on topic. Logical semantic hierarchies and rich interconnections of related terms and synonyms enable accurate retrieval and browsing of similar resources and ideally should be available in online environments. But the cost of features may not be sustainable with massively growing resources. Efforts to merge databases and map disparate subject terminology require considerable human intervention. A possible solution combines controlled and uncontrolled terms from three sources: authoritative professional indexing, automated term suggestion and uncontrolled keywords proposed by authors or end users' social tags. Research is required to investigate the effectiveness, cost and applicability of combining controlled and uncontrolled terms for information retrieval..

KEYWORDS

subject indexing
controlled vocabularies
automatic classification
collaborative indexing
retrieval effectiveness
cost effectiveness
precision

Koraljka Golub is senior lecturer and associate professor in the Department of Cultural Sciences at Linneaus University. She can be reached at koraljka.golub@lnu.se

Searching for information resources by topic (subject searching) is a common but most challenging task, especially if we compare it with a search where we know which item we want to retrieve (known-item searching). The underlying reasons why subject searching is often difficult is that different people will use different terms to refer to the same concept (synonyms). In addition, many words have many meanings (polysems), some of which are entirely not related (homonyms – for example, a bank of a river or a bank as a financial institution). Let us then observe the fact that search algorithms largely rely on statistical, locative and linguistic data, such as considering that a subject is a string occurring above a certain frequency in a given location and is not a stop word (like an article or a preposition). These factors are some of the common reasons why search algorithms fail to retrieve relevant results, retrieve too many results or are not specific enough for the search query at hand.

In order to address these problems, libraries and other information institutions, including commercial ones like providers of journal article databases, have been using subject indexing and classification systems, such as subject headings, thesauri and classification schemes. These systems have been designed in a way that allows each concept to be represented by only one index term in the catalog; but with equivalent/related/narrower/broader terms cross-referenced to it. Identifying these relationships means that the user can use any of the cross-referenced terms in her query and retrieve all the relevant resources.

Furthermore, in many cases end users do not know which search term to use. To this purpose, libraries allow browsing of physical library shelves which are arranged topically, following hierarchical classification schemes. A similar option should be offered in an online environment, and some

library services do include it, such as LIBRIS, the Swedish union catalog (see <http://libris.kb.se/subjecttree.jsp> for the SAB, the Swedish classification system, and <http://deweysearchsv.pansoft.de/webdeweysearch/> for the DDC, Dewey Decimal Classification). However, many library software vendors do not seem to support this facility, in spite of the fact that research points to the need for browsing, as in many situations users actually prefer browsing to searching (see, for example, Koch, Golub & Ardö [1]), contrary to the popular belief reflected in search boxes of Web search engines.

The danger that established objectives of bibliographic systems (finding, collocating, choice, acquisition, navigation) may be left behind in light of the exponentially increasing numbers of resources has been recognized in the literature (see, for example Svenonius [2], p. 20-21). For example, when it comes to integrating subject access in library catalogs across collections such as repositories, archival and museum collections, and journal databases in order to preserve the established objectives of bibliographic systems, subject access fields need to be merged, mapped to one another or transformed to match the domain, language or granularity. This mapping is challenging to achieve and therefore quality-controlled subject access often lacks in such services today (see, e.g., Sondera, <http://sondera.kb.se>, a cross-search service of LIBRIS, the Swedish Media Database and the Swedish National Archives).

Moreover, as subject indexing and classification systems involve significant human resources, some argue for using only full-text indexing as is common in web search engines instead. However, because of problems mentioned above, the search results will not suffice, especially when it comes to subject searching and for purposes when high precision and recall are needed. The latter includes, for instance, search tasks by researchers, lawyers, medical doctors and pharmacists who often need to find all information available on a certain topic, and without a lot of noise.

Possible Solutions to Subject Searching

(Semi)-automated mapping/indexing as well as end-user and/or author indexing represent some potential solutions to retain the established objectives of library information systems in light of the exponentially

increasing numbers of digital documents and integrated databases, each utilizing various subject indexing and classification systems.

Software vendors and experimental researchers speak of the high potential of automatic indexing tools. While some claim to entirely replace manual indexing in certain subject areas, others recognize the need for both manual (human) and computer-assisted indexing, each with its (dis)advantages. Reported examples of operational information systems include NASA's machine-aided indexing which was shown to increase production and improve indexing quality ([3]); and the Medical Text Indexer at the U.S. National Library of Medicine, which by 2008 was consulted by indexers in about 40% of indexing throughput [4]. A more experimental stage of a mapping tool for Europeana subject terms recognized at least 50% of all possible assignments on a small sample [5].

However, hard evidence on the success of automated indexing tools in operating information environments is scarce; research is usually conducted in laboratory conditions, excluding the complexities of real-life systems and situations. Having reviewed a large number of automated indexing studies, Lancaster concluded that the research comparing automated versus manual indexing is problematic ([6], p. 334). For a suggested comprehensive framework for evaluation of automatic indexing see [7].

End-user or author indexing, closely related to social tagging, refers to indexing by end-users or authors which can take place in social web applications, resulting also in folksonomies (groups of social tags). Subject access points assigned in this manner are cheap to produce compared to resources needed for professional indexing. In addition, such terms can be more up-to-date and may also include more end-user terminology compared to the more traditional systems. At the same time, the challenges are not insignificant, which is why they cannot be used on their own. The language control provided by subject indexing and classification systems is missing. Because they do not follow any indexing policies, there is a lack of consistency. Not the least, there is a limited number of people who actually do contribute in existing services. Literature has suggested provision of an existing subject indexing or classification system from which the end users or authors could choose, in order to at least address the language-control

issue. For example, one study [8] explored the use of Dewey Decimal Classification (DDC) with mappings to Library of Congress Subject Headings (LCSH) and showed that the DDC and LCSH helped the taggers find focus for tagging, strengthened consistency and led to increase of access points in retrieval. Also, three times as many search terms were found in end-user index terms as in manually assigned controlled terms. Thus, both catalogers' index terms and end-users' index terms combined help improve retrieval.

Future Research

Future research is needed to address the above issues in relation to all types of information resources. The provision of combined controlled and uncontrolled index terms of three different origins – manual by professional indexers, (semi-)automatically suggested and end-user index terms like

social tags and author keywords – may both alleviate high costs associated with human-only indexing, as well as provide a more complete set of access points, thereby improving retrieval and library services in general. However, to what degree these three items may be combined and for what purposes, users and contexts needs to be investigated. A comprehensive methodology framework to evaluate these complex factors in the context of retrieval has been recently suggested [7].

Based on the results of such an evaluation, we should be able to judge more reliably to what degree automation is today possible and for which applications it is suitable. The complex ratio of cost/revenue of both automated and intellectual/manual approaches may then also be explored for a range of use cases. This research could eventually feed into the practical indexing guidelines for designated document collections as to which combination of automated or intellectual approaches to apply. ■

Resources Mentioned in the Article

- [1] Koch, T., Golub, K., & Ardö, A. (2006). Users browsing behavior in a DDC-based web service: A log analysis. *Cataloging & Classification Quarterly*, 42(3-4), 163-186.
- [2] Svenonius, E. (2000). *The intellectual foundations of information organization*. MIT Press, Cambridge, MA.
- [3] Silvester, J. P. (1997). Computer supported indexing: A history and evaluation of NASA's MAI system. *Encyclopedia of Library and Information Services*, 61(supplement 24), 76-90.
- [4] Ruiz, M. E., Aronson, A. R., & Hlava, M. (2008). Adoption and evaluation issues of automatic and computer aided indexing systems. In *Proceedings of the American Society for Information Science and Technology*, 46, 1–4.
- [5] Lancaster, F. W. (2003). *Indexing and abstracting in theory and practice*. (3rd ed.). London: Facet.
- [6] Manguinhas, H., Charles, V., Isaac, A., Miles, T., Lima, A., Néroutidis, A., ... Gordea, S. (2016). Linking subject labels in cultural heritage metadata to MIMO vocabulary using CultuurLink. Presentation at the 15th European Networked Knowledge Organization Systems (NKOS) Workshop, in collaboration with the German ISKO, TPD L 2016 Conference in Hannover, Germany, Friday 9th September 2016. Available at: <http://ceur-ws.org/Vol-1676/paper4.pdf>
- [7] Golub, K., Soergel, D., Buchanan, G., Tudhope, D., Lykke, M., & Hiom, D. (2016). A framework for evaluating automatic indexing or classification in the context of retrieval. *Journal of the Association for Information Science and Technology*, 679(1), 3-16.
- [8] Golub, K., Lykke, M., & Tudhope, D. (2014). Enhancing social tagging with automated keywords from the Dewey Decimal Classification. *Journal of Documentation*, 70(5), 801-828.