Philip Hider, Charles Sturt University

The Search Value Added by Professional Indexing to a Bibliographic Database

Abstract

Gross et al. (2015) have demonstrated that about a quarter of hits would typically be lost to keyword searchers if contemporary academic library catalogs dropped their controlled subject headings. This paper reports on an analysis of the loss levels that would result if a bibliographic database, namely the *Australian Education Index* (AEI), were missing the subject descriptors and identifiers assigned by its professional indexers, employing the methodology developed by Gross and Taylor (2005), and later by Gross et al. (2015). The results indicate that AEI users would lose a similar proportion of hits per query to that experienced by library catalog users: on average, 27% of the resources found by a sample of keyword queries on the AEI database would not have been found without the subject indexing, based on the *Australian Thesaurus of Education Descriptors* (ATED). The paper also discusses the methodological limitations of these studies, pointing out that real-life users might still find some of the resources missed by a particular query through follow-up searches, while additional resources might also be found through iterative searching on the subject vocabulary. The paper goes on to describe a new research design, based on a before-and-after experiment, which addresses some of these limitations. It is argued that this alternative design will provide a more realistic picture of the value that professionally assigned subject indexing and controlled subject vocabularies can add to literature searching of a more scholarly and thorough kind.

Introduction

While indexers and catalogers might complain that empirical evidence pointing to the value their work adds to databases and catalogs is not always noted, or given much weight, by their employers, it is important that this evidence continues to be collected and reported, just as it is important that any evidence that suggests a decline in the value of professional indexing and cataloging is likewise reported. Catalogers and other metadata professionals may need to consider a range of survival strategies in a 'post-truth world', including those suggested by Gross (2015) and Borie et al. (2015), but they first need to convince themselves of the continued value of their work, and this is best done through a thorough, and open, engagement with the data.

The research described in this paper follows up on the studies conducted by Gross and Taylor (2005) and Gross et al. (2015), which provided evidence for the ongoing value of subject headings in a contemporary academic library catalog, i.e. that of the University of Pittsburgh. They found that, on average, about a quarter of "hits" in real-life keyword searches would not have been retrieved were it not for one or more subject headings, even after the catalog had been enriched with tables of contents and other derived indexing. The subject headings, of course, would been assigned by catalogers. The findings therefore suggest that this key component of catalogers' work, i.e. subject indexing, continues to significantly assist library patrons, at least to the extent that they still use the library catalog to perform subject searches.

Although the two studies by Gross et al. make a number of assumptions (including the use of the library catalog and/or its bibliographic data), as will be discussed shortly, they are based on a relatively straightforward methodology that can be readily replicated, and this paper reports on the findings of a similar study that examined the impact on retrieval of another branch of professional indexing, namely, that carried out for periodical and bibliographic databases. Again, it focuses on the value specifically of assigned subject indexing. If subject indexing is generally regarded as one of the most important and "professional" component activities of cataloging, it is typically an even more central activity in database indexing: if it was found to add little value, then the case for the professional database indexer would surely be weak. Conversely, if database searches are much assisted by professionally assigned subject indexing that could not be readily assigned by authors or other non-professionals, then the case for professional intervention would be intrinsically strong.

However, the limitations of the methodology employed by Gross et al. and in this study do cast some doubts on the resulting evidence. The paper discusses these limitations and subsequently proposes another research design that aims to address them. A second study based on this design and that is currently in progress, is briefly described; its results should provide a fuller picture of the extent to which indexers improve subject searching on a particular bibliographic database, namely the *Australian Education Index.*

Literature Review

The value of assigned indexing, and in particular assigned indexing using controlled language, was first called into question with the publication of results from the "Cranfield" experiments, which found that, for topical document retrieval, certain forms of derived indexing could achieve higher recall and precision ratios than those achieved by the various controlled vocabularies tested (Cleverdon 1967). Numerous studies and discussions of the relative merits of controlled and derived indexing since have pointed to the "received wisdom" of the two approaches' complementarity, each with strengths and weaknesses more or less exposed in different retrieval contexts (Rowley 1994; Bawden and Robinson 2012). The question remains, however, as to whether the value that controlled indexing (particularly of the sophisticated kind undertaken by information professionals) adds to a given search context is sufficiently large to justify its costs. This has recently been addressed by Gross and Taylor (2005) and Gross et al. (2015) in the context of the academic library catalog. The reality of this environment is not yet one of comprehensive "full text" retrieval (that is, retrieval based on full-text indexing), but rather of retrieval based, predominantly, on titles, tables of contents, summaries, and limited amounts of other "content", along with cataloger-assigned subject headings. Gross et al. (2015) found that the number of records retrieved in the University of Pittsburgh's library catalog by keyword searches, for topics, that were *only* retrieved because of the inclusion of one or more subject headings, represented, on average, about a quarter of total hits. Such a proportion might be considered insufficiently large to warrant the expense of professional subject indexing in the case of "casual" searching, but proponents of detailed cataloging argue that 'scholarly' searching requires more comprehensive results (Gross et al. 2015; Mann 2008).

While many experiments have been carried out to evaluate the effect of controlled subject vocabularies in bibliographic databases (indeed, more than in library catalogs),

the methodology employed by Gross et al. (2015) does not appear to have been replicated in this particular environment. Bibliographic databases are defined here as the products of the various journal indexing services, which sometimes also support direct access to full texts, online. It is unclear whether the subject indexing provided by these services, often based on a thesaurus, enhances retrieval to a similar extent, to that of LCSH in library catalogs. There are a number of differences between the two kinds of environment that might affect the indexing's relative impact, as indeed there are across individual bibliographic databases, such as the nature and quantity of other data elements (including abstracts) present in the keyword index, and the breadth and depth of the controlled vocabulary (if used) relative to the breadth and depth of the topics searched for by the database's users.

While studies of retrieval loss in specific databases would therefore be instructive, it should also be noted that in the modern environment, databases, including the library catalog itself, tend to be searched *within* a federated search: thus a more complete picture of retrieval loss caused by a lack of professional indexing would also involve replicating the methodology of Gross et al. on the kind of "discovery tool" that is now typically provided by academic libraries.

The methodology of Gross et al. has some limitations, however. As they themselves point out, it allows for a measure of "hits" lost, but is silent on whether or not these hits are *relevant* (Gross et al. 2015). Gross et al. speculate that the proportion of 'hits' that are in fact 'misses' is likely to be less on catalogs with subject headings, as precision tends to be a strength of controlled vocabularies, although this has yet to be demonstrated. Moreover, the measure provided by the methodology does not necessarily reflect *actual* retrieval loss, because it is based on *individual* search results (i.e. from a single query), whereas in real life users may perform *follow-up* searches, on the same topic, which might reduce, or otherwise affect, the proportion of misses.

There is no doubt that iterative searching takes place and is a significant factor in document retrieval (Hider 2006; Rieh and Xie 2006; Zhang 2013; Zhang and Soergel 2014; Pontis and Blandford 2015). On the other hand, supporters of controlled indexing have repeatedly stressed the challenges of the "synonym problem", even for the more committed searchers (Weber et al. 2006). One wonders how often topics are systematically searched for using *all* possible word forms of *all* synonyms and near-synonyms, in *all* languages. Subject headings and thesauri not only limit this problem, but also suggest search terms for related concepts that might well unearth other relevant resources. This can happen either *pre-hoc*, e.g. through preliminary thesaurus consultation, or *post-hoc*, e.g. through links in records and subject facet displays. Thus there are a number of ways in which the various possible elements of a *whole* search session can affect the actual level of retrieval loss, potentially both upwards and downwards. As Hider (2017) has recently pointed out, professional cataloging, including the assigning of controlled subject terms, may add value across a range of catalog user tasks, not limited to the retrieval of bibliographic records via the generic search box.

There are also issues to be considered, as mentioned earlier, around the *interpretation* of the measure produced by the methodology of Gross et al. (as opposed to its validity). That is, at what level does retrieval loss become "bad"? In some search contexts, there may be little need for a high recall ratio: relatively few, reasonably relevant resources may suffice. In other search contexts, on the other hand, the objective might be *full* recall,

or the user may be significantly disadvantaged if, say, one out of four relevant resources was missed. The *need* for resources, and particular recall levels, are themselves very difficult things to measure; indeed, they have yet to be convincingly measured, which is one of the reasons why there is no definitive answer to the relative values of controlled and derived indexing. Instead, we shall assume, for the purposes of the alternative research design described later in the paper, that a thorough search is, at least sometimes, required.

Design of First Study

The first study of the reported research project applies the methodology developed by Gross et al. (2015) to a particular bibliographic database, namely, the *Australian Education Index* (AEI), which "provides a complex and sophisticated subscription database consisting of more than 200 000 entries relating to educational research, policy and practice" (ACER Cunningham Library, 2017a). The database covers predominantly English-language material. The professional indexers who maintain AEI assign subject terms from the *Australian Thesaurus of Education Descriptors* (ATED), along with identifiers and geographic names where applicable. ATED includes "over 5,000 main entry descriptors", along with many cross-references, and "reflects terminology used to describe research and practice in Australian education" (ACER Cunningham Library, 2017b). AEI records also include the titles and subtitles, abstracts and journal names of the articles covered by the database, all of which may provide an indication of subject. However, it does not include author assigned "keywords" (although such terms are sometimes used by the indexers to assist their subject analysis).

Whereas in the studies by Gross et al. the proportion of resources missed was estimated by analyzing, in some cases manually, the content of the records retrieved from searches on the full database that included the LCSH, it was possible to calculate the loss level in the case of the AEI database by running the same search queries twice: firstly on all the basic keyword indexes, and secondly on all the basic keyword indexes except for those with the assigned subject terms.

The sample of queries used in the AEI study was derived in a similar, though not identical, way to that of the studies by Gross et al. (2015). In the latter case, a set of search terms was derived from a file taken from the catalog system's transaction log: after duplicate terms were removed, every (presumably chronologically) tenth term was taken for the sample. However, those terms that resulted in no hits or more than 10,000 hits were excluded from the sample. The AEI study did not have access to any search logs on the AEI database itself, but was provided with a recent transaction log of (general) keyword searches on *EdResearch Online*, which is based on AEI and provides access to "over 56 000 articles from more than 500 Australian education journals" (ACER Cunningham Library, 2017c). An inspection of the de-duplicated log suggested that taking every fifth (chronological) search query would reduce the number of interdependent queries—that is, queries from the same series of searches on a topic—in the sample to a reasonably small proportion. The resulting set of queries was found to include a large number that were clearly not *topics*, but instead represented searches for known articles, journals, authors, etc. After these were identified and eliminated, there was the additional issue of queries that had produced no hits or a very large number of hits. While Gross et al. (2015) had excluded those resulting in more than 10,000 hits for

practical reasons, the author decided that there were also theoretical grounds for excluding overly vast result sets from analysis: it was thought unlikely that researchers and scholars, or their assistants, would typically wade through quite so many records, even for "thorough" literature searches and even if they had immediate access to the full texts, and that they would likely limit the result set to a more manageable size, or conduct a different search.

The *EdResearch Online* log recorded the queries' hit numbers, and these were used as a guide to the number of hits one might expect, for a given query, on AEI (i.e. up to about 5 times as many). It was decided to exclude those queries with more than 100 hits in the log, so that only those queries yielding substantially fewer than 1,000 hits on AEI would be included. That is, it was felt that a very thorough research assistant may be prepared to look through entire result sets if they numbered in the hundreds, but not in the thousands.

Although queries with zero hits in the *EdResearch Online* log might have yielded some hits on the AEI database, it was decided to exclude these as well, along with those with more than 100 hits, so that the final sample size numbered 63. This made it considerably smaller than the 191 search terms analysed in the later study by Gross et al. (2015), but it was considered adequate for the purposes of providing indicative results. It should be noted that the queries were left in their natural (i.e. original) state, which meant that a few incorporated the Boolean logical operator "AND" or truncation. The sample queries are listed in Appendix A.

Results of First Study

The effect of the omission of the subject indexes on the 63 keyword searches is detailed in Appendix A. The percentage of lost hits across the sample ranges from zero to 78.1%, with a mean of 27.0% and a median of 23.3%. Interestingly, the mean *matches* that produced by later study by Gross et al. (2015) for all-language materials; the corresponding median was 17.6%. Overall, the sample of queries retrieved 5,256 hits with the subject indexes and 3,898 without them, representing a percentage loss of 25.8%. This compares with a loss of 27.7% in the later study by Gross et al. (2015) for all-language materials. Nine of the 63 queries lost 50% or more of their hits without the subject indexes: thus, for one in every seven "successful" searches, half or more hits would be lost. This compares with one in every five searches in the University of Pittsburgh catalog (Gross et al. 2015). In summary, the analysis indicates that similar loss levels, with respect to subject searching, might be expected if the AEI database and the University of Pittsburgh library catalog were not supported by professional indexing.

Design of Second Study

While other studies applying the same methodology as described above could be usefully carried out on other databases, for the purposes of comparison, whether users of the AEI really do miss out on about a quarter of relevant resources when subject searching remains something of an open question, given the methodological limitations outlined earlier. An alternative research design was developed to address those limitations. Specifically, the second study was intended to examine the proposition that professional indexing significantly improves the outcomes of scholarly literature searching. Assumptions are made here that comprehensive searches are necessary for

exhaustive accounts of the literature on a given topic, and that such accounts are necessary for convincing and high-quality scholarship.

A before-and-after experiment was constructed in which a research assistant, with experience in the field of Education as well as in reference librarianship, was provided with a list of topics that an academic might wish to engage an assistant to search for on the AEI. In the first stage of the study, the assistant was asked to conduct their literature searches using a version of the AEI stripped of its assigned subject terms (as well as its subject search option on the advanced interface), and to find as many relevant, or potentially relevant, articles as possible, with no limit placed on the number of searches she could try (for practical purposes, a time limit of 45 minutes per topic was imposed), and to compile a bibliography for each topic. The assistant could make use of all search functionality available, including links to full text, as she saw fit; she was not advised, at this stage, that the database had been stripped of its subject indexing.

The same research assistant was then asked, in the second stage of the study, to find any *additional* resources that she deemed relevant, or potentially relevant, for each of the topics previously searched for, on another version of the AEI database, this time with the assigned subject indexing, and subject search option, reinserted. She was advised that the database had been enhanced accordingly. The research assistant was asked to re-enter all the general keyword queries she had performed earlier, and could also enter other queries, or click on links she encountered, based on the assigned subject indexing in retrieved records (including those from new searches). Again, for practical purposes, the research assistant was given a maximum of 45 minutes per topic; she could make use of all search functionality available, although this did not include any facet displays or thesaurus look-up. She was asked to add entries for the new items (if she found any) to the bibliographies.

Twenty topics were derived from the sample of real-life queries used in the first study. Those topics, as expressed in the queries, which were thought likely to be clearer to the research assistant were selected. Although the sample size was small, it was considered large enough to yield an indicative measure of retrieval loss, given the exhaustive nature of the searching.

The results of the second study will be reported elsewhere. It should be noted that the search interface for the AEI database used in the study does not include all the features that might increase the effect of the assigned subject indexing, such as thesaurus look-up (a subject facet display might also significantly increase retrieval, particularly perhaps in less exhaustive searches). On the other hand, it should likewise be noted that the database also does not include any author-assigned keywords, which are present (and indexed), at least for some resources, in some of the other bibliographic databases.

Whether various databases register significantly different levels of retrieval when applying the methodology outlined above is a question inviting much further research. It would be interesting to compare results across disciplines, languages, different search interfaces, different controlled vocabularies, etc. Perhaps most tellingly, the research design could be replicated on a database that incorporated author-assigned keywords. It might also be possible to modify the design in the case of a database that searches on full text.

It was noted earlier that this alternative methodology does not address the question of whether a lack of retrieval from a particular database is likely to lead to its omission from

the eventual literature review. As well as the reality of researchers, and their assistants, tending to search on multiple databases, often using a federating discovery tool, there is also the possibility that they will encounter references to resources missed in the literature searching in the resources they do retrieve, or in citation indexes, or, perhaps, in follow-up author and journal searches. They might also try their luck on Google Scholar, or indeed on Google in general. The results of studies such as the one described here therefore have to be considered in light of all elements of the typical practices involve in modern scholarship.

Conclusions

It would appear that the levels of retrieval loss incurred by a lack of subject headings in catalog records are matched by similar loss levels in the case of bibliographic databases missing their professionally assigned subject indexing based on thesauri. It could be debated as to whether losing 27% of hits on a given subject search *matters*, but for more 'serious' information seeking, it is easy to imagine scenarios in which it would. Of course, this still does not mean that the costs associated with the indexing are justified, in comparison with other products and services that funds might be spent on, but it does suggest that professional indexing, and cataloging, should, at least in some cases, be considered as *candidates* for funding, and probably quite strong ones.

A fuller picture of the value of professional indexing, in terms of subject retrieval loss that its omission might cause, is possible to construct through the application of the before-and-after experimental design outlined in this paper. Although this still does not paint a *complete* picture, even of professional indexing's value to scholarship, it represents another piece in the jigsaw of evidence with which indexers and catalogers must now engage.

References

ACER Cunningham Library. 2017a. *Australian Education Index (AEI).* https://www.acer.org/library/australian-education-index-aei.

ACER Cunningham Library. 2017b. *Australian Thesaurus of Education Descriptors.* http://cunningham.acer.edu.au/multites2007/index.html.

ACER Cunningham Library. 2017c. *EdResearch Online.* http://opac.acer.edu.au/edresearch.

Bawden, David and Lyn Robinson. 2012. *Introduction to Information Science.* London: Facet.

Borie, Juliya, Katie MacDonald and Elise Sze. 2015. "Asserting Catalogers' Place in the "Value of Libraries" Conversation." *Cataloging and Classification Quarterly* 53, no. 3-4: 352-367.

Cleverdon, Cyril W. 1967. "The Cranfield Tests on Index Language Devices." *Aslib Proceedings* 19, no. 6: 173-194.

Gross, Tina and Arlene G. Taylor. 2005. "What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results." *College & Research Libraries* 66, no. 3: 212-230.

Gross, Tina, Arlene G. Taylor and Daniel N. Joudrey 2015. "Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching." *Cataloging & Classification Quarterly* 53, no. 1: 1-39.

Gross, Tina. 2015. “Naming and Reframing: A Taxonomy of Attacks on Knowledge Organization.” Knowledge Organization 42, no. 5: 263-268.
Hider, Philip. 2006. “Search Goal Revision in Models of Information Retrieval.” *Journal of Information Science* 32, no. 4: 352-361.
Hider, Philip. 2017. “A Critique of the FRBR User Tasks and their Modifications.” *Cataloging and Classification Quarterly* 55, no. 2: 55-74.
Mann, Thomas. 2008. “The Peloponnesian War and the Future of Reference, Cataloging, and Scholarship in Research Libraries.” *Journal of Library Metadata* 8, no. 1: 53-100.
Pontis, Sheila and Ann Blandford. 2015. “Understanding ‘Influence’: An Exploratory Study of Academics’ Processes of Knowledge Construction through Iterative and Interactive Information Seeking.” *Journal of the Association for Information Science and Technology 66*, no. 8: 1576-1593.
Rieh, Soo Young and Hong (Iris) Xie. 2006. “Analysis of Multiple Query Reformulations on the Web: the Interactive Information Retrieval Context.” *Information Processing and Management 42*, no. 3: 751-768.
Rowley, Jennifer. 1994. “The Controlled Versus Natural Indexing Languages Debate Revisited: A Perspective on Information Retrieval Practice and Research.” *Journal of Information Science* 20, no. 2: 108-119.
Weber, Michael A., Stephanie A. Steely and Marilou Z. Hinchcliff. 2006. “A Consortial Authority Control Project by the Keystone Library Network.” *Cataloging and Classification Quarterly* 43, no. 1: 77-98.
Zhang, Pengyi and Dagobert Soergel, D. 2014. “Towards a Comprehensive Model of the Cognitive Process and Mechanisms of Individual Sensemaking.” *Journal of the Association for Information Science and Technology 65*, no. 9: 1733-1756.
Zhang, Yan. 2013. “The Development of Users' Mental Models of MedlinePlus in Information Searching.” *Library and Information Science Research 35*, no. 2: 159-170.

Appendix A: Retrieval Loss in the AEI Database

Search query	**Hits in full search (*n*)**	**Hits in search excluding subject indexing (*n*)**	**Retrieval loss (%)**
lesson & planning	270	176	34.8
digital & storytelling	70	51	27.1
concept map	56	56	0.0
giftedness & music	11	9	18.2
saturday & school	49	38	22.4
astronomy	119	87	26.9
Middle & school & structure	163	117	28.2

free & online & articles & about & learning	1	1	0.0
physical & activity & academic & performance & children	10	9	10.0
boys & girls & learn	64	63	1.6
gender & balance	114	93	18.4
differentiated & instruction	71	62	12.7
nurture & students & development & through & communication & in & classroom	3	3	0.0
Writing. & Learning & to & teach & english & in & secondary & school	23	6	73.9
play-based & effectiveness	8	2	75.0
Angry & 'and' & aggressive & children	4	4	0.0
language & cueing & systems	4	4	0.0
reading & comprehension & importance	61	44	27.9
libraries & non & english	46	23	50.0
segregation	195	148	24.1
ecosystems	81	80	1.2
training & 'and' & crisis	44	38	13.6
positive & youth & development	158	92	41.8
intelligence & classroom	181	128	29.3
assessment & large & online & distance	29	15	48.3
assessment & large & online	189	129	31.7
Listening & relations & education	20	13	35.0
learning disabilities' & 'AND' & 'brain research'	8	3	62.5
neuromyths & in & education	5	4	20.0
learning & styles & 'and' & pedagogy	28	21	25.0
youth participation'	59	54	8.5
cloud & computing	32	16	50.0
parenting & skills	271	111	59.0
sensory & play	28	22	21.4
exploratory & play	89	84	5.6
Group & work & with & children	427	379	11.2
home-education	44	43	2.3
foundation & style	68	43	36.8
teacher & review & pedagogy	172	132	23.3
whiteboard & video	16	13	18.8
direct & instruction	320	272	15.0

Cyberbullying	87	78	10.3
transgender	54	42	22.2
flipped & learning	34	32	5.9
animal & assisted & therapy	5	4	20.0
importance & of & science & in & primary & school	8	3	62.5
first & generation & college & student	14	12	14.3
misconceptions, & primary, & science	73	16	78.1
Guided & Reading & Learning & Difficulties	11	6	45.5
sexual & assault & on & campus	3	1	66.7
positive & discipline	308	238	22.7
Out & of & school & care	145	123	15.2
gender & segregation	54	34	37.0
school & based & intervention & social & work	62	42	32.3
heavy & work	81	75	7.4
theology	250	217	13.2
authentic & student & engagement	229	142	38.0
reading & for & pleasure	6	5	16.7
art & therapy	60	33	45.0
year & 9 & selective	25	16	36.0
personalized & learning	26	18	30.8
new & arrival & program*	43	28	34.9
individualized & learning	67	45	32.8