The Retrieval power added by subject indexing to bibliographic databases

Abstract
Earlier studies have found that topic queries would lose, on average, about a quarter of hits if bibliographic databases did not incorporate the subject indexing based on controlled vocabularies and typically carried out by information professionals. Topic queries were run on three more databases, namely ERIC, PsycINFO and SocIndex, to investigate whether this loss level might be generalizable. Results indicate that the amount of “retrieval power” lost varies across databases, probably due to a range of factors. In two of the three databases studied here, over a third of hits were lost on average, whereas in the other case about a sixth were lost. These results point, if anything, to an even stronger case for retaining the assigned indexing included in these databases, as well as to the need for further research into the value of professional indexing across contemporary search environments.

Introduction
In a recent study (Hider 2018) the methodology previously employed by Gross et al. (2015) was used by the author to gauge the “retrieval power” of the subject descriptors and identifiers assigned by professional indexers for a particular bibliographic database. Gross et al. (2015) had found that about a quarter of hits would be lost, on average, if keyword searches on a university library catalog were not supported by Library of Congress Subject Headings (LCSH). In the author’s study the mean percentage of hits that would be lost to keyword queries on the Australian Education Index (AEI) was found to match exactly that of the later Gross study, i.e. 27.0%, despite the different type of database and different controlled subject vocabulary (Australian Thesaurus of Education Descriptors). Hider (2018) also reported a second study in which a searcher was asked to use the AEI stripped of its assigned vocabulary, and then to search, for the same topics, on the full AEI, with the subject indexing reinserted. Again, the indexing was found to support the retrieval of a similar proportion, on average across topics, of additional, relevant resources.

The objective of the follow-up study reported here was to gauge the retrieval power added by subject indexing based on controlled vocabularies to other bibliographic databases, as part of a larger program of research examining the search value of professional indexing across a range of search contexts. In order to facilitate comparison with the findings in the initial AEI study, the same measure of retrieval power, and similar methodology, was applied. The remarkably consistent results in the earlier studies, for a university library catalog and a bibliographic database with quite different content, raised the question of whether the “retrieval power” of subject indexing in bibliographic databases and catalogs might be generalizable, to some
extent, in the same way that Lotka’s Law has been found to apply, fairly accurately, to the frequency distribution of name headings in different library catalogs (Smiraglia 2002).

The measure of “retrieval power” can be associated with the experimental design conducted in classical “information retrieval” (IR), in which retrieval effectiveness is measured by calculating recall and precision ratios based on the outputs from individual queries entered into the system. These indexes, and their derivatives, have been used to assess the effectiveness of subject indexing on numerous occasions since they were first developed for the “Cranfield” experiments of the 1960s (Cleverdon 1967), despite the various methodological issues that have been identified their application to real-life search contexts (see e.g. Schamber 1994). Within this positivist framework, the relative value of controlled indexing has been called into question, although the received wisdom is that both derived and controlled indexing have their strengths and weakness, and that their relative effectiveness depends very much on particular search environments (Rowley 1994; Bawden and Robinson 2012). Moreover, they are not necessarily mutually exclusive: controlled indexing, typically created by information professionals, generally adds some value to a database. The questions are how much value and is this sufficient to justify its cost?

Some of the added value that a controlled vocabulary can bring to search is based on what is often termed the “synonym problem” (Beall and Kafadar 2008). Human searchers cannot always be relied on to think of (or spend time thinking of and entering as queries) all the synonyms and related words that might be included in derived indexing on a given topic. If the database is not supported by a search thesaurus – and most are not – this can become a major barrier to recall. Even when searchers do not make use of a controlled vocabulary’s reference structure, and simply enter “keywords”, they may retrieve resources that they would not otherwise retrieve because of the additional subject indexing, without losing any of the resources they retrieve via the derived indexing. These extra resources are likely to be relevant, assuming that searcher and controlled vocabulary use a given term for the same, or similar, concept. The proportion of resources that assigned indexing (typically based on a controlled vocabulary) adds to result sets, across a range of topics, is what this paper labels the indexing’s “retrieval power”.

This measure of the indexing’s value is not quite as refined as the classical measure of recall, as it is not based on “relevance”, but instead serves as a proxy for this measure. Ultimately the number of relevant resources in any result set is a subjective matter, to be determined by the person doing the searching; the subjective nature of “relevance” has been one of the main criticisms levelled at classical “information retrieval” (IR) experimentation, based on recall and precision measures that apply “standardized” relevance judgements (Schamber 1994). (Recall itself is also not a very practical
measure in naturalistic experiments, where the number of relevant resources in a given collection cannot be easily established). Within the positivistic frame of classical IR, however, the “retrieval power” measure is considered, by the author, to provide a reasonable approximation for “recall”, for the purposes of this research: it is likely that the additional resources retrieved via the assigned indexing will, if anything, tend to be more relevant than those retrieved via the derived indexing, as controlled vocabularies are thought to generally provide for more precision than natural language indexing (Gross et al. 2015). Nevertheless, it should be noted that the methodological criticisms that apply to classical IR experimentation also apply to this measure, and are more fully discussed by Hider (2018).

The author is not aware of studies of “retrieval power” in naturalistic settings prior to the research reported by Gross and Taylor (2005) and Gross et al. (2015). Gross et al. (2015) repeated their earlier study on the retrieval power provided by the LCSH on the newer version of the University of Pittsburgh’s library catalog, in which records had been enhanced by tables of contents, summaries, and limited amounts of other “content”. They found that the retrieval power added by LCSH had not, in fact, diminished all that much. Whether similar levels of added retrieval power are provided by assigned indexing across a range of bibliographic databases is not clear, and answering this question would be a first in comparing the added value that professional indexers bring to different databases. The measure may not be so applicable to databases that search on full text, where precision in the higher ranked hits is a more practical consideration, given the very large numbers of resources that are likely to be retrieved, but the reality still is that most bibliographic databases and library catalogs do not retrieve on full text for the most part (as opposed to provide full text, post-search). There are thus many contemporary databases for which the measure is relevant.

Methods

The three new bibliographic databases chosen for this follow-up study were: ERIC (https://eric.ed.gov), PsycINFO (www.apa.org/pubs/databases/psycinfo; specifically, the 1987-2017 version) and SocIndex (www.ebsco.com/products/research-databases/socindex). They support different disciplines, and use different controlled subject vocabularies (ERIC Thesaurus, the Thesaurus of Psychological Index Terms, and Sociology Thesaurus). It should be noted that, in addition, the search interfaces used to interrogate the three databases differed in their functionality and indexing rules: for example, words were retrieved on separately in the case of ERIC, but as phrases in the case of PsycINFO and SocIndex. Perhaps most importantly, the SocIndex system provided an “apply related words” option, which resulted in queries being automatically expanded to cover additional words with variant spellings and word forms. The hosts of the databases used in the study were ProQuest, Ovid and EBSCOhost, respectively.

A range of sampling sources was also employed, to best represent what terms might
be searched on these different databases. For ERIC, 55 of the keyword search queries used for the earlier AEI study were re-employed (ERIC covers the same field as AEI, and is international in scope and outlook; the AEI queries might in some cases comprise Australian terminology). For PsycINFO, the “sixty most-frequent multiword terms” identified in a search log from PsycINFO (Yi et al. 2006) were tested, while for SocIndex, the sample of 96 terms was based on lists of “Research Topic Ideas” for anthropology, criminal justice and sociology students found on a webpage of the Frances Willson Thompson Library at the University of Michigan-Flint (http://libguides.umflint.edu/topics/socialscience). Each source has both methodological pros and cons: the sample for ERIC represented actual search queries, but on a different database; the sample for PsycINFO also represented actual search terms, entered on the same database, but was modal rather than random; the sample for SocIndex constituted terms not derived from actual searching, but was thus independent of the database’s controlled vocabulary. The sample sizes were not large, but allowed for indicative findings (with an estimated confidence level of at least 70% for a 4% margin of error).

The terms in the samples were searched for on the relevant database thrice: first, on the full set of keyword indexes; then on all the indexes except for the index with the controlled subject vocabulary; and finally on all the indexes except for the indexes with the controlled vocabulary and the identifiers (additional, uncontrolled terms assigned by the indexers for concepts not represented in the controlled vocabulary). It was hypothesized that the loss of the identifiers would not reduce retrieval power all that much, as they would tend to be terms present in the abstracts, and thus covered in the full search.

**Results**

Table 1 shows the mean percentage of hits that would be lost per topic if the subject descriptors and identifiers were missing. The percentages for SocIndex are relatively low, but would probably be considerably higher if either the “apply related words” function was turned off, or its thesaurus also utilized the reference structure of the controlled vocabulary. The other two databases, ERIC and PsycINFO, both lost considerably more retrieval power than did AEI, without their assigned indexing. The results, combined with that of the AEI study, suggest that the retrieval power added by assigned indexing varies significantly across databases, and that this is the case even with databases covering the same discipline (such as ERIC and AEI). An interesting finding shown in table 1 is that, contrary to the author’s hypothesis, the loss specifically of identifiers can sometimes make a significant difference to retrieval power. In such cases, the identifiers must have been created more independently of the other bibliographic elements, such as the abstracts. The relevance of the additional resources that the identifiers yield merits investigation.
Table 1: Mean % loss of hits

<table>
<thead>
<tr>
<th></th>
<th>Without descriptors</th>
<th>Without descriptors &amp; identifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERIC</td>
<td>35.0</td>
<td>36.5</td>
</tr>
<tr>
<td>PsycINFO</td>
<td>26.6</td>
<td>35.2</td>
</tr>
<tr>
<td>SocIndex</td>
<td>11.0</td>
<td>17.2</td>
</tr>
</tbody>
</table>

Table 2 shows the median percentage of hits that would be lost per topic, for the three databases. Again, the variation across the databases is considerable, but in all cases the median is less than the mean, indicating losses of much larger magnitude for some topics. Indeed, there are topics with over 50% losses on all three databases.

Table 2: Median percentage loss of hits

<table>
<thead>
<tr>
<th></th>
<th>Without descriptors</th>
<th>Without descriptors &amp; identifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERIC</td>
<td>28.3</td>
<td>30.9</td>
</tr>
<tr>
<td>PsycINFO</td>
<td>18.0</td>
<td>26.8</td>
</tr>
<tr>
<td>SocIndex</td>
<td>9.8</td>
<td>16.4</td>
</tr>
</tbody>
</table>

Finally, the overall percentage loss, combining all the searches, was calculated for each database, and is shown in table 3. There is even more variation here, with the percentages dependent on the relationship between the individual topics’ loss levels and their overall retrieval power.

Table 3: Overall percentage loss

<table>
<thead>
<tr>
<th></th>
<th>Without descriptors</th>
<th>Without descriptors &amp; identifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERIC</td>
<td>85.7</td>
<td>86.1</td>
</tr>
<tr>
<td>PsycINFO</td>
<td>28.9</td>
<td>38.5</td>
</tr>
<tr>
<td>SocIndex</td>
<td>19.5</td>
<td>26.7</td>
</tr>
</tbody>
</table>

Conclusions

Given that the experiments on the ERIC and AEI databases used the same search queries and retrieved on their respective indexes in the same way, the differences in their results suggest that variations in content affect the extent to which assigned indexing increases retrieval power. Differences in both the derived and assigned indexing may well be factors. Nevertheless, the results across the three databases suggest that functionality and the keywords’ source may affect retrieval loss as well. Since discipline influences content and users’ choice of keywords, it can be postulated that discipline is likely to also influence retrieval power. The various possible factors do not preclude an underlying “law” that applies across bibliographic databases, but no evidence for it was found in this study. Rather, the results suggest that any such law would make, at best, a weak impression on retrieval power in the face of other factors that real-life settings involve, both within and without the database. These might
include the size of the database, its data elements, and the amount of indexing, both assigned and derived, the scope of the controlled vocabulary, as well as the disciplinary nature of the indexing language, and the functionality of the system.

Although additional retrieval power produced by assigned indexing seems to vary across databases, even the lowest yields, resulting from systems that poorly utilize controlled vocabulary, appear potentially significant. If searchers are looking for reliability, it is worth noting that some of the losses on SocIndex were much greater than the mean of 16.4%. In fact, 25% or more of hits were lost in the case of almost 30% of the topics. For other databases, retrieval loss would appear to be considerably higher, sometimes more than the loss estimated in the AEI and library catalog studies. Further research, across more databases and in more depth, is needed to assess more fully the extent that subject indexing based on controlled vocabularies adds value to bibliographic databases. The evidence garnered thus far suggests that it is of a magnitude that database providers and subscribers need to consider carefully, before making any decision to cut costs by jettisoning it.

References