What Have We Got to Lose?

The Effect of Controlled Vocabulary on Keyword Searching Results

Tina Gross and Arlene G. Taylor

Presentation for the LITA/ALCTS
Authority Control in the Online Environment IG
ALA Annual Conference – Orlando, Florida
June 27, 2004

 1

Subject Heading searches...

- were scoffed at in the 1830s:

  Strout: "classified catalogs and indexes were not needed because living librarians were better than subject catalogs...[and] any intelligent man who was sufficiently interested in a subject to want to consult material on it could just as well use author entries as subject, for he would, of course, know the names of all the authors who had written in his field."

 2

Subject searching in the 20th century ...

- continued to be denigrated, even though Cutter had convinced American librarians to use subject headings in dictionary catalogs.
- Catalog use studies (in academic libraries) showed that most searching was for known items, or at least for known authors. (Studies in public libraries showing high use of subject searching tended to be ignored.)

 3

Subject searching in the 1990s ...

- was shown by transaction logs to be the most popular kind of search in online catalogs (to the surprise of many)
- but such searches were difficult because users did not know how to construct exact subject heading strings
- so they compensated with title keyword searching

 4

Subject searching in 2004 ...

- is often ignored in favor of keyword searching, which can now be used to search every word in a record in many OPACs ...
- and so has resulted in a suggestion that subject headings should be stripped from records in order to save gigabytes of space and to save the time required to create them

 5

Remove subject headings?

- Some experienced librarians had observed that some keyword searches retrieved records that only had one or more sought-after word(s) in a subject string in a subject heading field.
- How often is this true?

 6

The Study

- Research Question:
  - What proportion of records retrieved by a keyword search has a keyword only in a subject heading field, and thus would not be retrieved if there were no subject headings?

 7

Methodology

- 3397 keyword searches were obtained from a transaction log of a small university library.
- De-duplicating of searches resulted in 2270 unique searches.
- A sample of 227 searches was selected after use of a common statistical formula for determining sample size.

 8

Methodology (cont.)

- The sample was used to replicate the searches in PittCat, the OPAC of the University of Pittsburgh
  - limited to holdings of the University Library System and the Law and Health Sciences libraries
  - limited to English language materials
  - could not omit provisional acquisitions records that lack subject headings

 9

Methodology (cont.)

- Collected the following data:
  - # hits with all keyword(s) anywhere
  - # hits with all keyword(s) in record with at least one in a subject, but not all in title
  - # records from the 2nd set (or # of first 50 records) with at least one keyword in subject only
- Example: *metal sculpture* – 19 hits with keyword anywhere

 10

Keywords in subject headings and also in title:

Gold, silver, and bronze : metal sculpture of the Roman baroque / Jennifer...

Title: Gold, silver, and bronze : metal sculpture of the Roman baroque / Jennifer Montagu.
Gold, silver & bronze
Author: Montagu, Jennifer.
Published: Princeton, N.J. : Princeton University Press, c1996.
Series: Bollingen series ; XXXV, 39. The A.W. Mellon lectures in the fine arts ; 1990
Bollingen series ; 35, 39.
A.W. Mellon lectures in the fine arts ; 1990.
ISBN: 0691027366 (CL : acid-free paper)
Physical Description: xvii, 262 p. : ill. (some col.) ; 29 cm.
LC Subject Heading(s) (limits do not apply): Metal sculpture, Baroque--Italy--Rome.
Metal sculpture, Italian--Italy--Rome.
Notes: Includes bibliographical references (p. [249]-255) and index.
Contents: 1. Introduction
2. Roman Bronzes around 1600
3. Adoration of the Host. Tabernacles of the Baroque
4. Medal. Some Drawings for the Hamerani
5. Silver Tribute from a Prince to a Grand Duke. The 'Piatti di San Giovanni'
6. Giardini and Giardoni. Silversmiths and Founders of the Eighteenth Century
7. From Rome to Lisbon. The Patronage of John V

App. A The Tabernacles of Jacopo del Duca

App. I The Ciborium of Sta Maria Maggiore

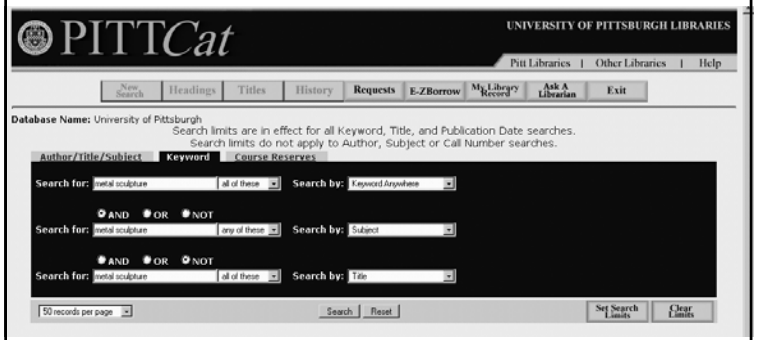
App. II The Ciborium of Sta Maria in Vallicella

App. III The Tabernacle of the Cappella Aldobrandini in Bologna

App. IV The Virgin and Child for Lisbon
App. V The Virgin of the Immaculate Conception for Lisbon.

 11

Second search to find at least one keyword in subject, but not all in title



 12

Record with all keywords in fields other than subject headings (3 hits):

Enduring memory, in stone, in metal, in beauty.

Title: Enduring memory, in stone, in metal, in beauty.
Author: National Sculpture Society (U.S.)
Published: [New York, 1946?]
Physical Description: [52] p. illus., plates. 35 cm.
LC Subject Heading(s) (limits do not apply): Sepulchral monuments.
Sculpture.

These were removed from the "lost" count by manual examination.

 13

Record with keywords only in subject headings:

Paul Wayland Bartlett and the art of patination / Carol P. Adil, Henry A....

Title: Paul Wayland Bartlett and the art of patination / Carol P. Adil, Henry A. DePhillips, Jr.
Author: Adil, Carol P.
Contributors: DePhillips, Henry A.
Paul Wayland Bartlett Society.
Published: Wethersfield, Conn. : Paul Wayland Bartlett Society, c1991.
Physical Description: xiii, 80 p., [1] p. of plates : ill., port. ; 23 cm.
LC Subject Heading(s) (limits do not apply): Bartlett, Paul Wayland, 1865-1925.
Patina of metals.
Metal castings.
Sculpture, American.
Notes: "Books of reference [used by Bartlett]": p. 80.
Includes bibliography (p. ix-x).

Seven records had at least one of the keywords only in a subject heading – 7 of 19 or 36.8%.
These would be lost if there were no subject headings.

 14

Methodology (cont.)

- When retrieved set was larger than 50:
  - viewed only 1st 50 and used % of those 50 to determine % for entire set
  - justified because display results are in reverse chronological order – presumably those records displayed first would often be considered more useful than older records

 15

Example – more than 50 hits

- Example: *crime policy*
  - 388 hits
  - 218 with at least one keyword in a subject heading, but not all in title
  - 42 of first 50 of the set of 218 had at least one keyword in a subject field only – 84%
  - 84% of 218 = 183.1
  - 183.1 is 47.2% of 388
  - we estimate that in a search for *crime policy*, 47.2% of hits would be lost without subject headings

 16

Searches rejected for analysis

- 41 of the 227 searches did not yield valid results
  - 9 retrieved more than 10,000 hits (PittCat's maximum) – e.g., *diseases*, *civilization*
  - 32 retrieved no hits
    - many typos or spelling errors – e.g., *hollecaust*, *vintriloquism*
    - others just not there – e.g., *pet doctors*, *capital punishment china*

 17

Findings

- Average proportion of hits lost in absence of subject headings – 35.9%
- Median – 30.2%
- Total % of all hits lost if subject headings were not present, combining all searches – 35.4% (36,319 out of 102,580 hits)

 18

Results by number of keywords in search

	all searches	1 keyword	2 keywords	3 keywords	4 or more keywords
# of searches	186	44	98	30	14
median # of hits	66	390	57.8	39.5	9
average % lost	35.9%	26.0%	37.3%	44.9%	38.0%
median % lost	30.2%	19.7%	36.6%	34.7%	26.5%

 19

High loss individual searches

keyword(s)	# hits	# hits with kw in subj. only	% hits lost if no subject headings
airplanes military parts	23	23	100%
businesswomen	173	171	98.8%
divorced people	55	51	92.7%
horror films	402	332.8	82.8%
mass media politics	372	292.5	78.6%
history slang	22	17	77.3%

 20

Findings (cont.)

- For 31.7% of the searches, the percentage of hits with a keyword only in a subject field was 50% or greater
- That is, for about 3 of every 10 successful keyword searches, half or more of the hits now retrieved would not be retrieved if there were no subject headings.

 21

Addition of TOC and Summaries

- Since completion of our data collection, the University of Pittsburgh has begun adding tables of contents and summaries to records in PittCat
  - Such enhancements can help users to assess relevance of retrievals.
  - Enhanced records could include highly specific search terms not typically present in a traditional MARC record.

 22

Inclusion of enhancements...

- increases number of hits
- decreases chances that a search will produce no hits at all
- reduces precision – i.e., increases the number of irrelevant hits

 23

Reduction in precision

- *metal sculpture* now yields 46 hits, but among the first 10 are:
  - Jazz modernism
  - Collected poems
  - Rapid prototyping casebook
  - Animaculture [book of poems]
- searching “as a phrase” eliminates some of these, but also eliminates records such as those in slides 13-14, which seem potentially relevant. (An “as a phrase” search now yields 13 hits, 5 of which have “metal sculpture” in a subject heading only, and 1 of which is not relevant [!])

 24

Reduction in precision (cont.)

- a keyword search for *morning after pill* had no hits in our study
- it now has 2 hits:
  - Paper trail: common sense in uncommon times (containing the essays: "Good morning spamerica," "After 20 years of cultivation," "A pill for what haunts you")
  - Fear of dreaming: the selected poems of Jim Carroll (containing "Morning," "After St. John of the Cross," and "Blue pill")
- it still has no hits "as a phrase"

 25

Reduction in precision (cont.)

- *geometric patterns* also retrieved no hits in our study, but now retrieves more than 50 records
  - such a specific topic is not well represented in subject headings
  - but only 2 of the first 10 results appear relevant
- although a sophisticated user could make *geometric patterns* a phrase search and retrieve better results, the average user tends to use the default settings ("all of these" in PittCat) without understanding them.

 26

Reduction in precision (cont.)

- And performing a phrase search would not help with one-word searches (23.7% of our study)
  - in a search for *athletes*, only 3 of the first 10 hits appear relevant to a general search on *athletes*.
  - but opening the first of these yields a linked subject heading: "Athletes—Biography."
  - clicking on this linked heading retrieves a list of subject headings in which the user may scroll forward or backward (possible, of course, only if subject headings are added and maintained)

 27

Conclusion

- The study found that if subject headings were to be removed from catalog records (or no longer added to them), users performing keyword searches would lose more than one third of the hits they currently retrieve (35.9% on average).
- The loss of hits would be in addition to the loss of other functions and advantages provided by controlled vocabulary in general.

 28

- Without subject headings, a keyword user whose search retrieves an overwhelming number of hits with a high proportion of “false drops” would have few options in trying to find a smaller, more relevant set of hits.
- And a large proportion of the lost one-third of hits would be likely to be relevant to the user because the lost records would actually have at least one keyword in subject headings, and not just in any random place.

 29

Thank you.

Further questions may be addressed to:

ataylor@mail.sis.pitt.edu
and/or
tinag@pitt.edu

 30