Rolf G. Henzler
Deutsches Krebsforschungszentrum, Heidelberg

Free or Controlled Vocabularies

Some Statistical User-Oriented Evaluations of Biomedical Information Systems

Henzler, R. G.: **Free or controlled vocabularies.** Some statistical user-oriented evaluations of biomedical information systems.
In: Intern. Classificat. 5 (1978) No. 1, p. 21–26
Based on CANCERNET (a literature information system, formerly called SABIR-C) a quantitative comparison of free text and controlled vocabulary used in indexing and retrieval is given. The advantages of free text search are compared with the convenience of searches using controlled vocabulary. An indexing consistency text of 683 double indexed articles is commented on. The results are evaluated with regard to loss of information from the author to the user. They show the necessity of both free text and controlled vocabulary in an optimum combination. Author

1. Indexing typology of information systems

An evaluation of information systems indexing deserves special attention. Table 1 shows a typology of some computerized biomedical information systems available in Germany. It shows two coordinates: the further to the right a system is located, the more detailed the classification structure of its vocabulary (1). The further down the system is the more complex are the means of describing the documents and of selecting key words from the documents. The two main factors are (down) which and what kind of tool for condensation of the content in the documents is used (e.g. free or controlled vocabulary) and (right) how detailed and well differentiated the terms are arranged for a later retrieval. One could assume that the more complex a vocabulary structure is, the more expensive the information service. Indeed only information systems with little structure can be viable without subsidy. The best example of this is proved by the information systems of ISI in Philadelphia. They have other advantages however: they provide for rapid performance, are up to date, make use of the citation approach, and exclude subjective factors; among the disadvantages is a retrieval with relatively low recall and no weighting of the value of the article.

It is surprising that there are so many US information systems without controlled vocabulary (one exception is MEDLARS). Only one system (the European CANCERNET) has controlled vocabulary *and* free key words being sufficiently open to changes. It can be assumed that users' needs are not only recall-oriented; particularly in medicine, with its difficulties in nomenclature, there seems to be a trend towards free text. This explains why the two systems MEDLARS and CANCERNET have experienced a policy change during the last two years. They now take up more free text for document analysis. Of course, this was favored by the further development of search technology. Especially in Europe the main approach to document storage was to analyze the contents of documents semantically; the hope was to provide the optimum basis for a later information retrieval by indexing the documents in as much detail as possible. Because the role of indexing was overstressed the following points have not been given adequate attention.

ÜBERSICHT UND KLASSIFIZIERUNG
der wichtigsten biomedizinischen Informationssysteme

		Begriffsordnung / Titel	ohne Begriffsordnung	Klassifikation oder Syntax	Hierarchischer Thesaurus	Hierarchischer Thesaurus und Klassifikation oder Syntax
Gemeinsame Kommunikationsbasis ist	natürliche Fachsprache	ohne Zusatz	(ASCA)	–	–	–
		mit Klartext	CANCERLINE (Carc.Abstr.)	–	–	–
		mit freien Schlagwörtern	–	BIOSIS	–	–
		mit freien Schlagwörtern und Klartext	CANCERLINE (Cancer Ch. Ab.)	BIOSIS (Biolog. Abstracts)	–	–
	normierte Systemsprache	mit normierten Schlagwörtern	–	–		MEDLARS (bis 1975)
		mit normierten und freien Schlagwörtern	–	CANCERNET (bis 1977)	–	–
	normierte Systemsprache natürliche Fachsprache	mit normierten Schlagwörtern und Klartext	–	–	(INIS)	MEDLARS (ab 1976)
		mit normierten und freien Schlagwörtern und Klartext	–	EXCERPTA MEDICA		CANCERNET (ab 1978)

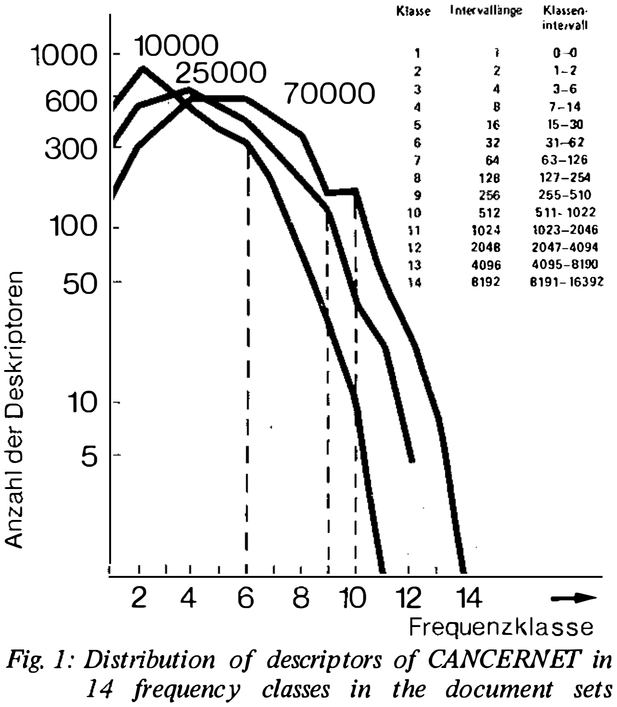
Table 1: Survey and classification of the most important biomedical information systems.

* We gratefully acknowledge the permission of the Deutsche Gesellschaft für Dokumentation e.V. to print this revised English version of a paper presented at the "Deutscher Dokumentartag 1977" in Saarbrücken. German version is contained in its proceedings volume.

1. Is it economically justifiable to devote a great deal of time and money on input if retrieval is the factor which limits information output: is the great effort involved in storing articles worthwile if only a small percentage of them are likely to be recalled?
2. Does the construction of a thesaurus enable retrieval without the use of free text information?
3. Is not condensing the data of a document into key words (and even more into controlled vocabulary) even more inefficient on recall than a later free text search?
4. Is a thesaurus up to date with present knowledge? Is it not necessary that the topical facts expressed in words appearing for the first time in the documents are available immediately for information retrieval? Is free text the only means of fulfilling this task in an ideal way?
5. Can high precision always be achieved? Is it not inevitable that one can have *either* high recall *or* high precision?

2. Frequency **distribution of** free and **controlled** vocabulary

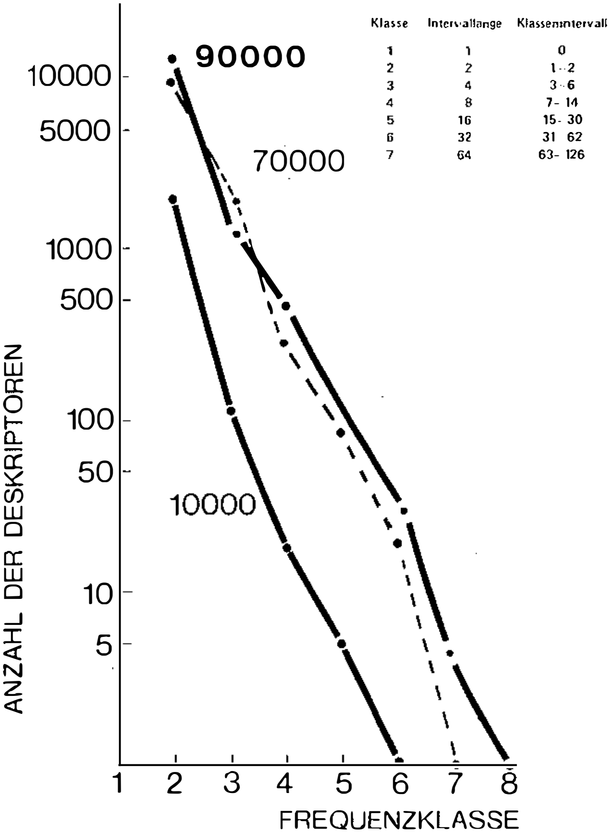
Bhattacharyya (2) reported in length on the role of natural language in indexing and in retrieval in the various scientific disciplines. The above mentioned questions are to be answered with reference to statistical material based on CANCERNET. In order to prove or contradict the above theses the structure and the vocabulary of thesaurus, indexing, and search were examined with regard to the frequency of use of the key words. The distribution of the indexed vocabulary is shown in Fig.1: the greater the number of documents stored in the data bank, the higher the shift to the right; i.e. the fewer keywords in the lower frequency classes. Fig.2 shows the frequency of the free key words not belonging to the controlled vocabulary. When plotting on a logarithmic scale it is remarkable that the lines which should be parallel (according to Zipf's law) are actually intersecting.



Klasse	Intervallänge	Klassen-intervall
1	1	0–0
2	2	1–2
3	4	3–6
4	8	7–14
5	16	15–30
6	32	31–62
7	64	63–126
8	128	127–254
9	256	255–510
10	512	511- 1022
11	1024	1023–2046
12	2048	2047–4094
13	4096	4095–8190
14	8192	8191- 16392

Fig. 1: Distribution of descriptors of CANCERNET in 14 frequency classes in the document sets 10 000, 25 000, 70 000 (logarithmic scale)

It follows that the vocabulary collections shown have not developed "freely". The free key words are indeed regularly examined for incorporation into the thesaurus. High frequency is not the only determining factor since certain free key words are meaningful only in their proper context and when read in the key word category of the document. Using 6000 evaluations, Sicker (3) found that many key words are rarely or never used as search terms despite high information value in documents. The division into two kinds of key words (search terms and non-search terms as context description), which is also employed in CANCERNET, is recommended especially because the size of the controlled vocabulary in the thesaurus can be kept optimum and manageable, thus meeting both aims of the user:



Klasse	Intervallange	Klassenintervall
1	1	0
2	2	1-2
3	4	3-6
4	8	7- 14
5	16	15- 30
6	32	31 62
7	64	63- 126

Fig. 2: Distribution of "free" key words of CANCERNET in 8 frequency classes in the document sets 10 000, 70000, 90 000 (logarithmic scale).

High recall by using thesaurus key words (for common terms)

High precision by using free key words and also free text (for topical concepts).

The frequency distributions of title words, free key words and descriptors are shown in Fig. 3. Again the advantage of a controlled vocabulary for better recall becomes apparent: There are few different types, especially in lower frequency classes. This entails fewer key words for an average search formulation since several synonyms and various derivations are combined into one descriptor of the controlled vocabulary.

The data bank can be subdivided into segments of arbitrary length. Thus indexing frequencies can be collected at intervals as illustrated by the following results obtained over a period of six months. By means of our automatic search evaluation program we were able to

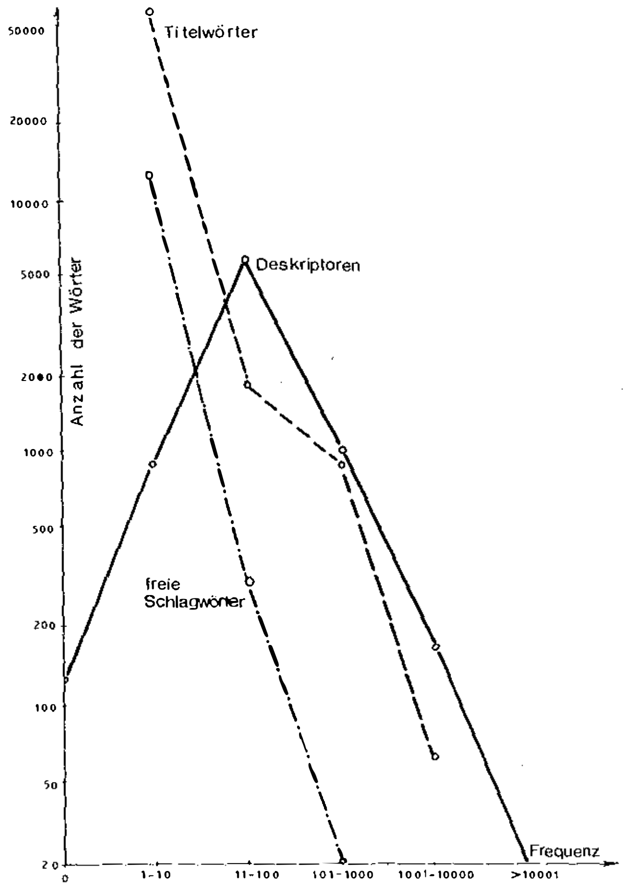


Fig. 3: Comparison of frequency distributions for descriptors, "free" key words, and title words in the document set 90 000 (logarithmic scale).

evaluate the requested searches with reference to various criteria. For this purpose we have selected the following categories:

- Entering article numbers of references found into the memory
- Entering key words, free terms and other search characteristics used during the search into the memory.

Evaluation of the first category will prove useful in the following discussion. From Fig. 4 it becomes apparent that the articles most frequently retrieved (with search frequencies of 13 and 12) were also the most recent (articles from 1977 are naturally represented less since the evaluation began only in the beginning of 1977). In this evaluation it was most interesting that more than 20% of the total information was extracted by the searches. For this reason there is no doubt that the effort on the input side is worthwile, since these 20% (19459 different citations) were retrieved after only 6 months. The average quote per search was 47 citations. The distribution of the article frequencies is shown in Fig. 5.

3. The value of indexing

To determine whether indexing is worth the effort or whether free text searches would be more economic, the following aspects must be considered:

- is indexing better than adequate free text incorporation if ideal conditions are assumed in both cases?
- how good is indexing in practice?

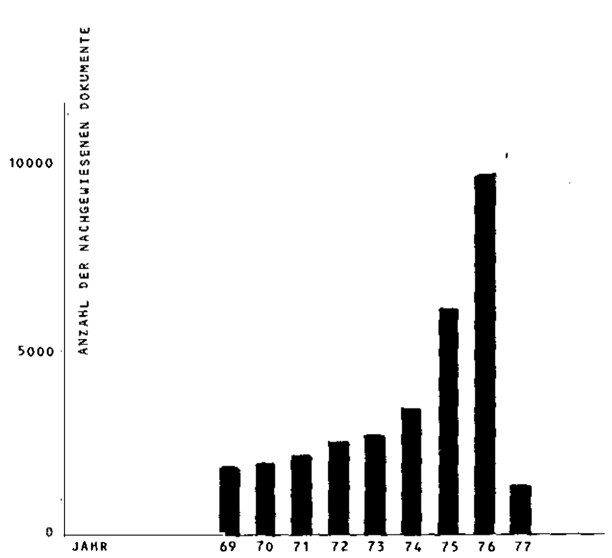


Fig. 4: Age distribution of documents retrieved in 666 requests during January–June 1977 in CANCERNET (Heidelberg).

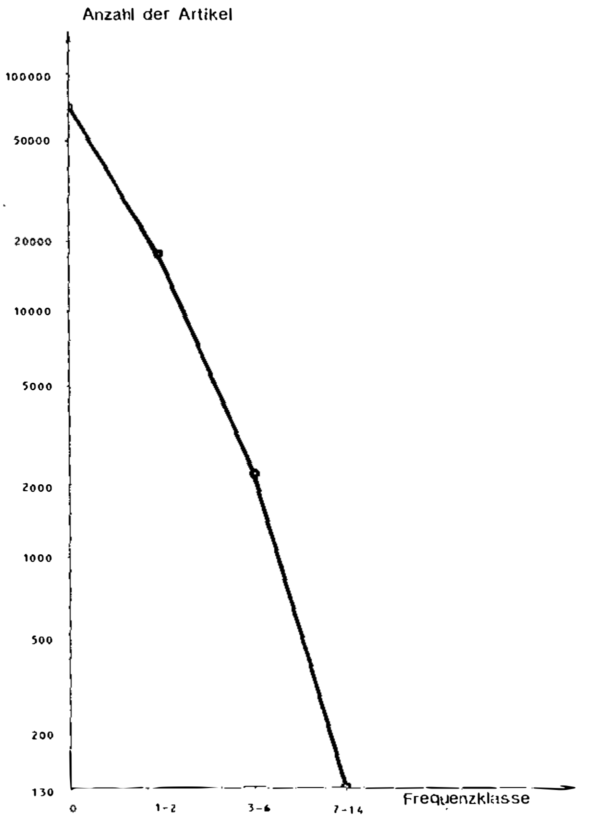


Fig. 5: Frequency distribution of number of references retrieved in 666 requests during January–June 1977 in CANCERNET (Heidelberg).

The first question can be answered by showing that either free text search or searching with standardized key words yields higher precision. In some special disciplines one cannot be sure that the user means the same as the author with his free text term, often because the author has used the same word years ago with a totally different connotation from the one the user would use today. Therefore it is questionable whether the special vocabulary in free text is really a better means of communication than a controlled vocabulary. Otherwise there would not be several committees on standardizing and systematizing vocabularies for each discipline. However a thesaurus enables any special vocabulary to be arranged in such a manner that orientation is easy.

It is well known that not all important facts in a publication find proper representation, either in titles or in abstracts: e.g. for cancer research the very important descriptor "survival time *more* than five years" can hardly be found in free text circumscription, because a standardization of free text input cannot be assumed as in the case of controlled vocabulary.

A sample of 100 out of 3900 descriptors of CANCERNET shows that only about 50% of the descriptors are represented properly in free text; when one uses some kind of classification (as in CANCERNET with 41 document-oriented categories) a further 20% can be retrieved by a combined free text and classification approach. A further 10% can be retrieved by means of sophisticated free text retrieval with operators as "adj"[1]. If there is no classification at all, some problems are not well solved as, for example, poor associativity of the "adj" function. When searching abstracts, this case without associative adjacent is especially critical. These investigations verify the experiences of Rothkirch–Trach (4), who has stated that 70% of all key words are terms consisting of more than one word and hence only retrievable with a more complex retrieval tool.

An evaluation of our own shows further that when the key words are correctly represented by free text expressions, barely 50% of its key word frequency can be found as title word frequency. Of course abstract word frequencies are much higher but give less precision in retrieval.

The converse evaluation of two samples of 100 titles words, one out of 2800 most recent title words within the last half year and the other out of 61 000 different title words (of the whole CANCERNET data base) has shown that 35% of all title words could not be mapped into the thesaurus without loss of information. 5% were trivial words which were not included in the stop word list, and among the remaining 30 terms were 25 special words and 5 names or numbers (Fig. 6). This proves that free text search cannot be abandoned. This is especially important in cases of descriptors with low indexing frequency (often concepts newly implemented into the thesaurus).

4. The value of free text searching

By using free text searching not only precision but also recall can be improved considerably. The reason for this is probably that indexers do not see the same importance of some concepts for the user as the authors did in writing their titles and abstracts. Because of the occurrence of some hierarchical "brothers" the indexer might have thought it more reasonable to apply broader terms. In this case, free text searching would mean indeed just a further retrieval aid to ensure a compensation of probable indexing weaknesses. The evaluation of the searched terms in CANCERNET (Heidelberg) within the first six months of 1977 shows that the number of "tokens" (amount of words used irrespective of repetitions) consists of 86% controlled vocabulary and 14% free text vocabulary; 2% of the free text elements used could not be mapped into the set of the controlled vocabulary. The corresponding articles could be retrieved only by means of free text searching (Fig. 7).

The surprisingly high number of free text terms is understandable because of the additional usage of title words, title word stems (truncated in the beginning and at the end of the word) and key word stems thus allowing convenient string search using the most important semantic units without controlled vocabulary, e.g. the stem THERAP allows a retrieval concerning all therapeutic terms as radiotherapy, hormone therapy, immunotherapy, therapeutic complications (except surgery),... (5,6,7).

New ideas in the latest articles are not always retrievable by using only new combinations of concepts already known. Especially in cancer chemotherapy there are always new chemical substances which do not have even the word stem in common with known concepts. In order to retrieve these articles the new names have to be used, but they occur only in free text because they are so new that no indexing rule and no thesaurus has incorporated them yet. There will always be new names as long as we have active research. Certainly the vocabulary does not converge to a limit (8). The older the scientific topic the more interesting is a search for a survey or collection of most relevant articles for a specific aspect of that topic: Such a search can best be done using the vocabulary of a thesaurus!

Economic considerations make it obvious that it is hardly possible to integrate *all* new concepts of the stored articles immediately into the thesaurus. However, the searcher should have all new concepts available derived from free text in assorted lists. As soon as the terms have been used for searches they will be more important for the thesaurus and should be incorporated in it.

[1] Acute "adj" Leukemia means that both terms should occur connected. This operator is necessary for further specification in searches using inverted files.

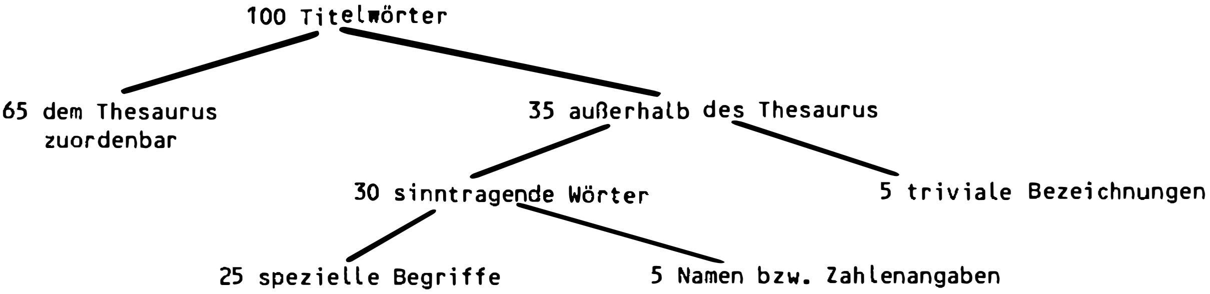


Fig. 6: Sample evaluation of 100 title words out of 2800 newest title words: qualitative classification.

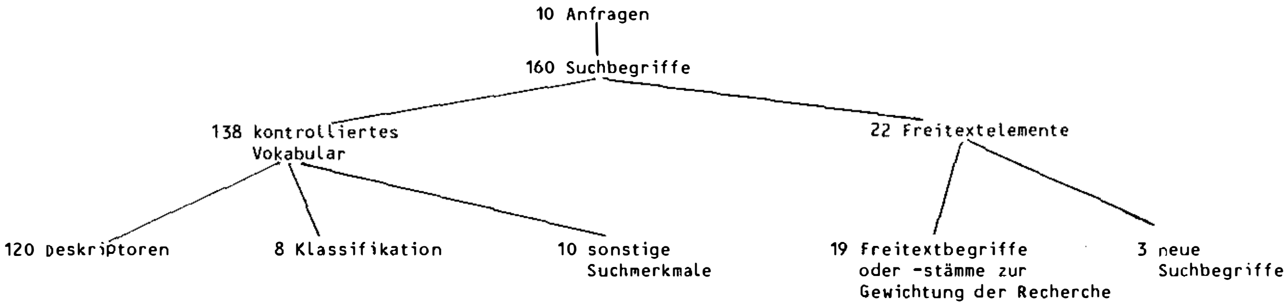


Fig. 7: Average distribution of search vocabulary per 10 requests (based upon 666 search requests).

Retrieval with controlled vocabulary only has disadvantages, as does free text search only: they have relatively low recall. Searches with controlled vocabulary do not always have the highest precision especially for topical subjects.

5. Indexing consistency tests

The question "How good is indexing in practice" requires an evaluation of indexing consistency studies. The tests already known were performed with small samples only so the results were of little statistical significance (9). They also differed quite considerably. Thus Hooper (10) found indexing consistencies of 10%–80%, Jacoby (11) reports an indexing consistency of 10% in chemical patents, Slamecka and Jacoby (12) 35%–45%, Jacoby and Slamecka (13) 16% for experienced and 12% for inexperienced indexers without indexing guide-lines. According to the quality of indexing guide-lines, classifications, and thesauri, a better indexing consistency was found. There are no plans for an evaluation of these studies nor for a further development of the consistency measures used (14).

Häufigkeitsverteilung der Indexing-Konsistenz in CANCERNET anhand von 683 doppelt geindexten Arbeiten

% der Übereinstimmung	Anzahl der Arbeiten
0 - 10	24
11 - 20	97
21 - 30	153
31 - 40	140
41 - 50	127
51 - 60	58
61 - 70	34
71 - 80	26
81 - 90	14
91 -100	10

Table 2: Frequency distribution of indexing consistency tests in CANCERNET based upon 683 twice indexed articles

However, an evaluation with more indexed articles involved will be presented using the so-called TANIMOTO[2] measure which was used in most of the literature. According to this measure, an indexing consistency of 38% within confident limits from 35%–40% on average was found in CANCERNET (table 2). This result is rather precise because it is based on 683 articles which happened to be double indexed and were found during the process of a system's refinement (and thus at random). The above percentage is a minimum value because 32 articles which were 100% identical were excluded from the evaluation since double input was assumed in these cases. So even the 38% may increase.

6. Some comments on indexing consistency tests

On presenting these values the question arises whether indexing at all is worth the effort. The following example (Fig. 8) of two descriptor chains with an indexing consistency of 40% illustrates the manner in which the numerical value has to be judged. The above value neither says anything about the similiarity of the exchanged descriptors nor about the probability of their usage in retrieval. If one would define indexing consistency from the searcher's viewpoint instead of the indexer's viewpoint one would attain far better results.

I. MAUS BALB C. MAMMA (T. EXP.), *T. CHEMISCH (INDUZIERT)*
CHEMOTHERAPIE EXP., CYCLOPHOSPHAMID

II. MAUS BALB C. MAMMA (T. EXP.), *KARZINOGENESE CHEMISCH EXP.*, OESTROGENE.
CHEMOTHERAPIE EXP., CYCLOPHOSPHAMID,
ADRIAMYCIN, METHOTREXAT, PREDNISON

Fig. 8: Comparison of two indexed descriptor chains with 40% consistency according to TANIMOTO measure

The measure $\frac{2c}{a+b}$, which is also common in literature, would yield for the above example an indexing consistency of 57%.

The only consequence of these studies is to keep the overall number of descriptors as low as possible in order to avoid errors by interchanging similiar descriptors. Gilchrist (15) recommends only 500 descriptors for an information system of medium size and is thus within the range of classification experts. Indeed one should remember that when using only 500 classification concepts corresponding to the descriptors, indexers will achieve higher consistency, assuming an exact definition of the classification areas. Condensing of facts in descriptors or classification makes sense, as long as one adds some free key words and some free text to the descrip-

[2] The TANIMOTO measure is: $\frac{c}{a+b-c}$ with a = number of descriptors assigned by first indexer;– b = number of descriptors assigned by second indexer;– c = number of descriptors assigned by both.

tors assigned by experts. The combined method enables a retrieval with higher recall than each of the exclusive approaches: using free text only, one cannot always be sure of high precision, and using thesaurus descriptors only one runs the risk of having a non-topical thesaurus.

Missing content analysis not only yields high connect times during searches but also a true lack of recall in retrieval;

missing free text is not only inaccurate but also a basic defect in retrieval of newly implemented ideas.

The three extreme approaches "title only", "free text only" and "thesaurus only" find an ideal synthesis which enables document acquisition to be as precise and still as condensed as possible: a content analysis of documents by controlled *and* non-controlled vocabulary, allows indexing with a few simple rules.

Thus one may find an economic compromise between the alternatives "high effort in input, low effort in search" and "low effort in input and high effort in search" with "medium effort in input and search".

Despite the above observations concerning the common vocabulary of indexers and searchers other recall lowering factors in information systems should not be disregarded: the scope of an information system and poorly selected journal lists.

These factors often decrease recall times more than the existence or non-existence of the thesaurus. Thus the cancer literature information systems, CANCERLIT and CANCERNET have quite different journal lists, CANCERLIT stores more articles from US journals and is more specialized in cancer chemotherapy and chemical carcinogenesis, whilst CANCERNET stores more articles from European journals and covers a broader cancer spectrum, including radiology, virology, and surgery with respect to experimental and clinical cancer research. So ist was no surprise that 30 typical search requests processed in CANCERLIT and CANCERNET had an overlap of 5% on the average. It was remarkable that 11 of the requests had no overlap at all. Expressed in positive terms, this means the two information systems are complementary to each other.

It was interesting to compare the frequencies of the search results (Fig. 9), which show that in frequency class "15–30 citations" CANCERNET (9 searches) was better than CANCERLIT (5 searches) – probably the advantage of the controlled vocabulary – but CANCERLIT is ahead in frequency class "7–14" with 11 as opposed to 3 in CANCERNET – probably the advantage of the greater amount of free text (abstracts). The second maximum in "3–6" may be caused by the successful free text searches in titles. A Wilcoxon sign test verifies for the above 30 searches that the two information systems are complementary. Of couse, the distributions should be investigated in further tests.

Thus the alternative "free text or controlled vocabulary" is no longer an alternative: There should always be both free text *and* controlled vocabulary in an ideal combination.

References

(1) Dahlberg, I.: Zur Theorie des Begriffs. In: Intern. Classificat. 1 (1974 (No. 1, p. 12–19.

(2) Bhattacharyya, K.: The effectiveness of natural language in science indexing and retrieval. In: J. Doc. 30 (1974) p. 235–254.

(3) Sicker, W.: Zur Untersuchung zwischen Such- und Informationswert bei der Schlagwortgebung. In: Nachr. Dok. 21 (1970) p. 206–207.

(4) Rothkirch-Trach, Ch. v.: Prinzipien der Thesauruserstellung dargestellt am Beispiel 1. des Aufbaus eines internen Informationssystems für die Studiengruppe für Systemforschung, 2. der Erstellung von Fachthesauri für das Projekt Bundestagsverwaltung der Studiengruppe für Systemforschung. München: Verlag Dokumentation 1970.

(5) Byrne, J. R.: Relative effectiveness of titles, abstracts, and

For the following see page 32

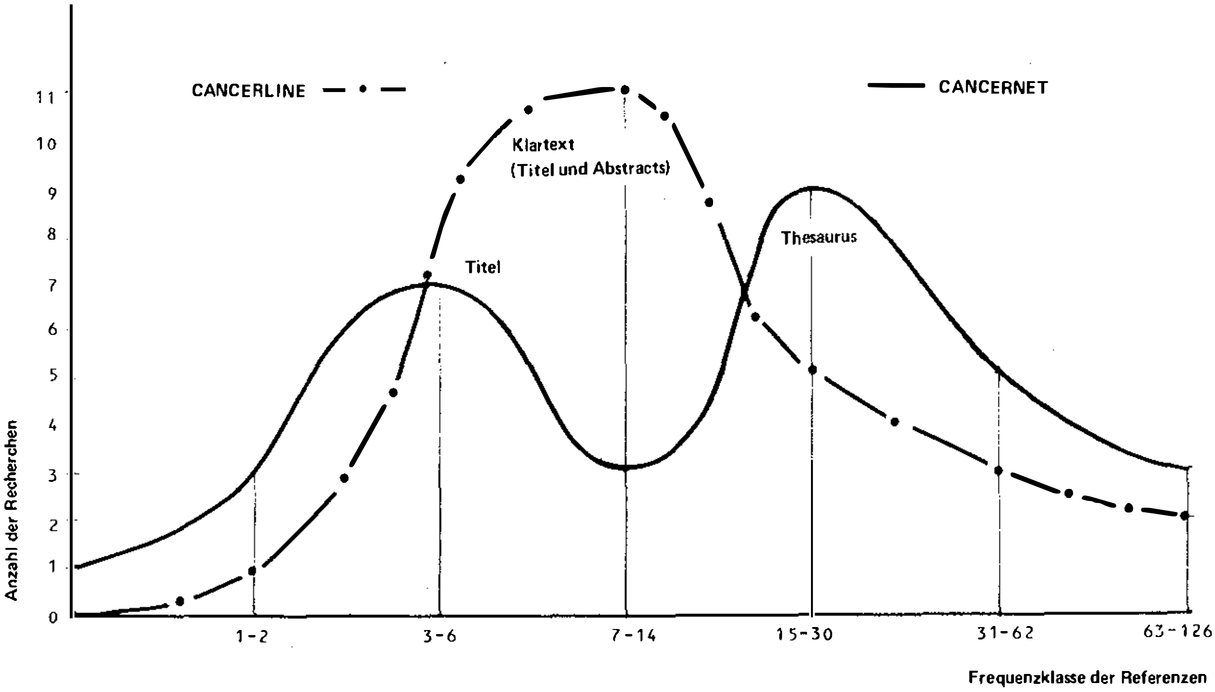


Fig. 9: Comparison of the frequency distributions of found citations based upon 30 typical search requests in CANCERLINE (CANCERLIT) and CANCERNET.