By C. D. GULL

Alphabetic Subject Indexes and Coordinate Indexes: An Experimental Comparison*

Mr. Gull is a member of the staff, Documentation, Inc.

ONE OF THE OBJECTS of this contract is to make an experimental study of classification systems,[1] alphabetic subject indexes and coordinate indexes. Because the existing catalogs present difficulties of size, location and security restrictions, a sampling technique was employed. The first comparison was made between the alphabetic subject indexes of the Technical Information Division (TID) of the Library of Congress and of the Document Service Center (DSC) in Dayton and coordinate indexes developed by Documentation Incorporated. Cards were obtained from TID representing 1207 reports cataloged under its Office of Naval Research contract, and cards were obtained from DSC, representing 543 reports cataloged for the Air Force.

All cards found under headings beginning with certain words, such as Antennas, Electric, Electronic and Microwaves, were chosen as part of the sample, because the headings incorporating these words were thought to be the most heavily used headings in the list for unclassified reports and would thus illustrate the maximum concentration of numbers on coordinate index cards that could be obtained for the sample. These cards represented 707 TID reports. The remaining 500 cards were in numerical order, from U20400 through U20899, making the sample 1207 out of 21,000 unclassified TID reports. The subject matter of these cards is so diverse that they are considered representative of the complete subject catalog. The 543 DSC cards were chosen at random and not in consecutive numerical order, thus making the samples equally representative for both catalogs.

Preparation of Sample Subject Heading Catalogs

Sample subject heading catalogs were set up from the two groups of cards, and the following figures were determined from them:

* Technical Report No. 5, Prepared under Contract No. AF 18(600)-376, for The Armed Services Technical Information Agency, by Documentation, Inc., Washington, D.C.

[1] A comparison of classifications and coordinate indexes will be described in a later Technical Report.

TABLE I

Catalog Cards From	Reports	Different Headings Used	Subject Heading Assignments	Average Subject Headings Per Report	Cross References Needed				Total Cards	Total Cross References	Cross References Per Heading
					To Headings		To Subdiv.				
					See	See Also	See	See Also			
TID	1207	1110	1950	1.61	630	208	212	25	3025	1075	0.97
DSC	543	899	1357	2.50	460	67	229	5	2118	761	0.85
Totals or Averages	1750	2009	3307	1.89	1090	275	441	30	5143	1836	0.93

It was not difficult to underline the subject headings on the cards and arrange them in alphabetic order, but it took a great deal of time to establish the cross reference structures for the two samples. The cross references had to be included to make the samples correspond to the original catalogs and to permit comparison of the subject catalogs and the coordinate indexes for reference purposes, as well as to determine the relative difficulty of preparation. The second edition of the Navy Research Section (NRS) *List of Subject Headings*[2] was followed in creating the *see* and *see also* references for the TID catalog; for the few headings found on the cards but lacking in the list, the references were made according to the policies used for the *List*. The *ASTIA Document Service Center Subject Heading List** lacks the type of cross reference structure of the NRS list, and it provides only seven *see* references and 28 *see also* references for the 899 subject headings and none to subdivisions. It was there fore necessary to supply the cross reference structure for the DSC sample catalog, and this was done according to the policies followed in the NRS list. Thus both samples were supplied with all cross references required by the permutations of words in the various subject headings. While it was possible to provide all of the necessary *see* references, no attempt was made to supply any but the most obvious *see also* references for the DSC sample catalog, since it is extremely difficult to guess the relationship of headings in a list lacking in its own *see also* references. The situation accounts for the small number of *see also* references in the DSC sample compared to the greater number for the TID sample.

The form of the NRS list also indicated the production of 220 *see also* references which could not be used in the sample TID catalog as such, but 186 of these are changed to *see* references and added, leaving only 34 to be discarded. A similar situation prevailed for the sample DSC catalog.

It is particularly noteworthy from these samples that 97 cross references are needed for every 100 TID headings and at least 85 cross references for every 100 DSC headings. These references are required by the inherent difficulties of alphabetic indexes: they must be included for synonyms, relations between headings, and the permutations of words in multiple-word headings.

Preparation of Sample Coordinate Indexes

Certain assumptions about coordinate indexes were current in our thinking when we undertook to prepare the first coordinate index from the TID cards:

1. Coordinate index terms should be simple.
2. A coordinate index is used by coordinating two or more terms to discover the original materials providing the desired coordination. Coordination is accomplished by any of the logical operations of conjunction, alternation, and negation, or any combination of them.
3. In order to make it unnecessary to search the entire index, the record of the original materials should be posted on the coordinate index term cards. Since numbers are very convenient for such posting, the original materials should be arranged in numerical order.
4. Since we lacked the original materials to put in numerical order, a numerical or accessions catalog was essential. The numbers already on the cards were ideal for this purpose.
5. Coordinate indexing can be accomplished by manual and mechanical means. The samples described here were made on cards for manual coordination, divided into ten vertical columns according to the terminal digits of the numbers, a device expected to facilitate the coordination of numbers.[3]

[2] U.S. Library of Congress, Navy Research Section. *List of Subject Headings*, 2d ed. Washington, 1950.

* U.S. Dept. of Defense. Armed Services Technical Information Agency. Document Service Center. ASTIA Document Service Center Subject Heading List (Alphabetically), Dayton, 1952. The subject catalog maintained by the Document Service Center has been provided with cross references, even though they are lacking from the printed list.

[3] A sixth assumption was this: The distribution of coordinate index terms into categories will facilitate both the cataloging operations and reference use. Although an attempt was made to categorize the terms, this phase of the investigation is yet to be completed.

After the TID file was set up in order by TIP number, the coordinate index was started by considering the card with the lowest number: in this sample, U 23, bearing the two subject headings—*Power meters* and *Microwaves—Absorption.* At this stage, only the subject headings were considered in preparing the coordinate index, and no attention was paid to the titles and abstracts included on the cards. Clearly, the term *Microwaves* could be used on one coordinate index card and *Absorption* on another card, and 23 entered on each card in the column headed 3 (for the final digit), but what about the phrase *Power meters*? If used as a phrase, it is not as simple as if broken into two words, and it requires in an alphabetic file a cross reference from the permutation, *Meters, Power.* If broken into two words, the specialized meaning of the phrase becomes lost in the general character of the single words, but is recovered when the two words are coordinated, showing 23 to be common to both words. In an attempt to test the assumptions and with a keen realization of the costly, time-consuming character of a cross reference structure, the phrase was broken up into two words, and the next cards were considered. As the work progressed, it soon became a goal to create the coordinate index without any cross references, if possible. However, not all phrases seemed as easy to break into single words, and the progress of the work was marked by indecision and inconsistency. A chronicle of the efforts to solve the problem, and the solution itself, are found in our Technical Report No. 3, November, 1952[4] The rule for the solution is repeated here because of its importance to coordinate indexing:

"Enter every word in a coordinate index system as a filing word on a single coordinate index card. Whenever in a particular system a word is used in one, and only one, descriptive phrase, enter that word as the filing word on a card, followed by the remaining word or words in the phrase. The word or words following the filing word on any card will themselves be filing words on other cards."

An example shows the practical application of this rule. Given a report dealing with digital computers, two cards are made, one headed *Computers* and the other *Digital.* If there are no computers in the system except digital computers, the *Computers* card is modified to read *Computers, Digital.* If there is nothing digital in the system except computers, the *Digital* card is modified to read *Digital computers.* If later a report is received on analog computers, the card for *Computers, Digital* is shortened to read *Computers* and a new card is made reading *Analog computers,* providing, of course, there is nothing analog in the system but analog computers. The *Digital computers* card is not affected until a report is received on some other digital device, when the term is shortened to *Digital* alone.

With this rule for a guide in choosing unit terms, the coordinate index for the TID cards was rapidly completed, with no further problems, and the cards were arranged in alphabetic order. The sample coordinate index possesses these characteristics:

1. Every term in the system is a filing term.
2. Since there are no subdivisions, every term is on equal footing with every other and can be the subject of a complete search.
3. All "see" references required in a standard system, by virtue of the order of words in index-headings, are eliminated.
4. All "see also" references from general to specific subjects are eliminated.
5. The subjective choice of the indexer between possible permutations of multiple-term descriptions is eliminated.
6. Since every word in the system is a filing word and each word in the system appears only once as a filing word, searching for

[4] Also published as "Unit Terms in Coordinate Indexing" by Taube, Mortimer, Gull, C. D., and Wachtel, Irma S., in *American Documentation,* 3:213-218, October 1952.

the "proper subdivision" in the proper phrase is unnecessary.

7. Since serial numbers do not reveal the security classification of reports, a single coordinate index can be used for all classifications without compromising the security requirements based on the "need-to-know."

1214 unit terms for the combined coordinate index, since 372 terms were common to both indexes.

The merged coordinate index provided a marked contrast to the sample subject heading catalogs, as shown in these figures:

TABLE 2

Catalog Cards From	Reports	Cards in Sample Subject Heading Catalog	Cards in Separate Coordinate Indexes	Cards in Merged Coordinate Index	Subject Heading Assignments	Subject Headings Per Report	Unit Terms Assigned	Converted to Unit Terms Per Report
TID	1207	3025	815	443	1950	1.61	4249	3.52
DSC	543	2118	771	399	1357	2.50	2317	4.26
Common to TID & DSC	—	—	—	372	—	—	—	—
Totals or Averages	1750	5143	1586	1214	3307	1.89	6566	3.75

The new rule made it easy to prepare a coordinate index from the headings on the DSC cards, and both coordinate indexes possessed the same characteristics.

At this stage of the work, the two sample alphabetic subject heading catalogs were of approximately the same quality for reference purposes because of their full cross-reference structures; but they could not be combined easily into one catalog because the headings are uninverted for TID (i.e., Digital computers) and inverted for DSC (i.e., Computers, Digital). Any attempt at combination would require extensive changes on at least one set of cards, as well as new cross references.

Merging the Coordinate Indexes

It was soon perceived that the contrary was true of the two coordinate indexes. Because the terms in each coordinate index were unit terms, and predominantly single words, it was entirely feasible and easy to merge the two indexes into one. Before the merger there were 815 unit terms for the TID coordinate index and 771 for the DSC coordinate index, a total of 1586 terms; but after the merger there were only

The merged coordinate index requires less than one-fourth the number of cards in the two subject heading samples, yet the average number of indexing assignments was doubled, even though the assignment of unit terms was restricted by the policy of creating unit terms from the subject headings only.

Improving the Quality of Coordinate Indexing

It was recognized that an improvement in quality could be obtained for the merged coordinate index by

1. Assigning additional unit terms based on information obtained from the titles and abstracts on the cards, or
2. Assigning additional unit terms based on titles and abstracts on the cards *plus* a review of the original documents.

The second alternative was not tested, under the assumption that it would be too expensive an undertaking for any large collection of documents, but an investigation of the first alternative was undertaken. A new merged coordinate index was created from a sample of 200 cards, comprised of 100 DSC cards (the lowest numbers in our

non-consecutively numbered sample) and 100 TID cards (U20200-U20299). Since all unit term assignments required by the subject headings were retained, the base of the new coordinate index was identical with the old coordinate index for these 200 cards. The preparation of the new index revealed that 388 unit terms were used for the 200 cards in the old merged coordinate index. The review of titles and abstracts resulted in the use of 90 terms already in the old merged coordinate index and in the addition of 117 new terms, bringing the total to 595 unit terms. The average number of unit term assignments was increased from 3.52 to 6.88 for each TID report and from 4.26 to 5.57 for each DSC report, or an average of 6.24 unit terms per report.

This last average is three and a quarter times the average number of subject headings per report, and it indicates that the depth of indexing is much greater for coordinate indexes as we assume from this test they would be prepared than for the conventional subject heading catalogs as they are now prepared. This difference is not as great as indicated here, since the conventional subject heading catalogs provide access to reports by means of cross references in addition to entry under the headings. Lacking an accepted terminology for the sum of subject headings plus cross-references for any report, we are calling this sum "access points." The following Table has been developed to show per report the comparison between subject headings, access points, converted unit terms, and unit terms resulting from improved coordinate indexing.

It is interesting to note that while the DSC reports have more subject headings and more converted unit terms per report than do TID reports (2.50 to 1.61 and 4.26 to 3.28, respectively), they have fewer unit terms after a review of titles and abstracts (5.64 to 6.88). As an explanation of this situation, we conjecture that it is probable that the DSC policy of assigning headings liberally assures a better conversion to a coordinate index than does the TID policy of restricting the assignment of headings, but that the TID abstracts are more informative for indexing purposes than the DSC abstracts.

If a search of a subject catalog is considered from the viewpoint of an average TID report, it will be found entered under 1.61 subject headings and access to it will be provided under 1.58 cross references, or a total of 3.19 access points, compared to 3.52 entries when the same subject headings are converted to unit terms.

A similar comparison for DSC cards shows 2.50 subject headings plus 2.12 cross references per report, or a total of 4.62 access points per report compared to 4.26 entries per report when the same subject

TABLE 3
PER REPORT

	1	2	3	4	5	6	7	8
Catalog Cards from	Subject Headings	Times Cross References Per Heading	Equals Cross References	Access Points (1+3)	Column 1 Converted to Unit Terms	New Assignments of Existing Unit Terms	Assignments of New Unit Terms	Unit Term Assignments (5+6+7)
TID	1.61	0.97	1.58	3.19	3.52	2.69	0.91	6.88
DSC	2.50	0.85	2.12	4.62	4.26	0.95	0.43	5.64
Averages	1.89	0.91	1.72	3.61	3.75	1.82	0.67	6.24

headings are converted to unit terms. If the pattern of the TID cards had been repeated, these should have been five or more unit terms per DSC report, rather than 4.26. An examination of the subject headings on the DSC cards reveals why there are fewer unit terms than access points, for many of the reports are assigned overlapping headings with certain words in common which are used only once in converting to unit terms, for example, *Meteorological equipment* and *Meteorology—Research,* in which four words (or four access points when cross references are included) reduce to three unit terms:

1. Meteorology; Meteorological (on one card)
2. Equipment
3. Research

Since the number of access points is equal for all practical purposes to the number of unit terms for both catalogs, it might be assumed that a coordinate index whose terms are converted directly from subject headings offers no advantage in reference use over a subject heading catalog, but this assumption is incorrect for these reasons:

Coordinate Index	Subject Heading Catalog
1. Reports are listed on all cards consulted, for there are no cross references and no subordination of words.	Reports are listed under subject headings only—just over half of the access points—and not on the cross references—the remainder.
2. Unit terms can be freely combined in the searching process, thus providing combinations to meet each searcher's need, i.e., more generic or more specific searches.	Combinations of words are frozen because of the use of multiple term subject headings and cross references.
3. The searcher is certain that he has access to all reports listed under a single word.	Because no cross reference system includes all permutations of words in the headings, the searcher is never certain he has access to all reports to which a word applies.

The searcher is interested in how many reports can be provided to meet his particular need with the least effort and time, rather than in the number of access points or unit terms per report. An extensive comparison of subject catalogs and coordinate indexes for reference use is planned, but until the statistics are available, the value of converting subject headings to unit terms can be measured only as shown above, although demonstrations performed with the samples indicate that the reference advantage of this level of coordinate indexing is considerable.

The value of improving the level of coordinate indexing by considering titles and abstracts of reports in addition to converting subject headings has been demonstrated, however, in unit terms per report. If it is assumed that the average of 6.88 unit terms per report is the optimum for coordinate indexing of the 100 TID reports (and here we recognize that all cataloging and indexing are subjective accomplishments), then 2.69 terms per report of this total are assignments of unit terms already used in the previous sample—in other words, unit terms under which the searcher would expect to find reports but under which he would not find them in a subject heading catalog or in a coordinate index prepared by converting subject heading assignments. The same condition applies to 0.95 unit terms out of the average total of 5.64 unit terms for the 100 DSC reports. New unit terms were needed for both sets of reports: 0.91 unit terms per TID report and 0.43 unit terms per DSC report. Thus the review doubled the unit terms used for the TID reports (from 3.28 to 6.88) and increased those for the DSC reports by one-third (4.26 to 5.64), and these figures are a measure of the superiority of this level of coordinate indexing over subject headings.