TECHNICAL WRITING: A KEY TO COMPUTERIZED INFORMATION RETRIEVAL

PREPARED BY
S. O. JONES
CORPORATE LIBRARIAN
DOUGLAS AIRCRAFT COMPANY, INC.

PRESENTED TO
THE 14TH INTERNATIONAL TECHNICAL COMMUNICATIONS CONFERENCE OF THE SOCIETY OF TECHNICAL WRITERS AND PUBLISHERS
CHICAGO, ILLINOIS
24 - 27 MAY 1967

JUN 24 1967

DOUGLAS CORPORATE OFFICE

Best Available Copy
TECHNICAL WRITING:
A KEY TO COMPUTERIZED INFORMATION RETRIEVAL

by

Shirli O. Jones
Corporate Librarian
Douglas Aircraft Company, Inc.

ABSTRACT

The future success of mechanized retrieval of information from published material lies, in a large part, in the hands of the technical writer. The context is only as usable as its format when current retrieval methods are used. Examples are given of confusion arising through nondefinitive titles, particularly where permuted indexes, such as Key-Word-In-Context (KWIC) and Key-Word-Out-of-Context (KWOC), are used.

The appearance of an article in several different publications is another source of difficulty for the information retrieval system. How the creation of separate entries for the same item can damage the integrity of an information system is shown.

A plea is made to the technical writers for understanding the needs of the second generation of reader. Steps which can be taken without censoring the language by replacing all synonyms and near-synonyms with a single term are indicated. The use of meaningful titles, standardization of citations, and several other areas are explored from the viewpoint of the information specialist.
FIGURES

1. KWIC and KWOC Indexes 4
2. Title Repetition by Word 5
3. KWOC Index With Terms Repeated 6
4. List of Words Prevented From Indexing 7
5. Effects of Hyphen Usage Upon Computer Sort 8
6. Effects of Singular-Plural Usage Upon Computer Sort 9
7. Index Card With Multiple Corporate Authors 12
8. Variations of Single Corporate Author 14
9. Automatic Index 16

Preceding Page Blank
INTRODUCTION

Each year, more and more persons decry the proliferation of the printed word. Available statistics prove that unless restraint is applied, mankind will be submerged in published information by 1975. Concern over how to obtain essential information from this vast amount of material results in a rash of papers such as this. Information specialists involved with the retrieval of necessary data bemoan the unnecessary generation of papers, but of more importance is more careful consideration in both the preparation and the publication of all papers. Such a specialist is not in a position to determine the relative value of one document in comparison with another, but he is in a position to help establish guidelines suitable for advanced computer techniques that will provide the keys to necessary storage and retrieval capabilities for the future. Although information retrieval systems covering all degrees of complexity have been introduced, none can produce a product better than the original material. Therefore, authors and technical writers must cooperate with the information specialist to ensure the rapid and effective retrieval of valuable data.

Information retrieval, in general, relies on two forms of index—subject headings and concept coordination. The more familiar form of the subject heading index is the public library card. The concept coordination system uses "uniterms" or "descriptors"—a single term or concept for every major topic in a document. These terms can then be coordinated with other terms or concepts for flexible, specialized retrieval. Bibliographic information, such as personal author, corporate author, and title, is handled in much the same way, regardless of the type of retrieval system used. The equipment used in creating and manipulating the index can range all the way from a typist typing on 3 x 5 file cards to a fully computerized system for storage and automated retrieval in various formats. In every case, the technical writer's importance is evident, but it is never so obvious as in the computerized storage and retrieval system. Thus, the problems associated with this area require special emphasis.
DISCUSSION

Since 1959, when Douglas actively began to design its mechanized information retrieval program, many different systems have been developed. Government agencies, such as NASA's Scientific & Technical Information Division and DOD's Defense Documentation Center, have built massive systems utilizing the most sophisticated of computer equipment. Several private industries have expended vast quantities of time and money to provide their personnel with retrieval capabilities. Organizations such as the American Society for Metals and the American Chemical Society have called upon their professional talents to produce systems designed to cope with the complexities of their scientific disciplines. These are all major efforts, but the needs of the smaller systems are of equal importance. And either type of system must face much the same problems. Some of these problems, which will be described in detail, are the responsibility of the technical writers and/or publishers.

One of the most common forms of mechanized index is the Key-Word-In-Context or KWIC index, and a variation called KWOC, or Key-Word-Out-of-Context (Figure 1). As can be seen, these indexes consist of a sorting of the title with each word repeated in alphabetical order. The title will appear as many times as there are significant words in the title (Figure 2). The KWOC index provides a more readable format (Figure 3) but follows the same principle. To prevent listings of insignificant terms (such as "the" and "of"), a stop list is input to the computer. Chemical Titles employs a stop list of 1,335 "insignificant" words (Figure 4). Examination of this list indicates that titles such as "Laboratory Systems & Procedures" would be completely eliminated.

Another problem occurs with the inconsistent use of hyphens. In Figure 5, there are several interesting entries. Under "low," the first inconsistency occurs with "low energy"--two entries are unhyphenated, three hyphenated. Next is "low molecular weight"--two entries without hyphens, one with.
| IC LESIONS IN THE BROAD BREASTED BRONZE TURKEY/ THE INFLUENCE OF EXERCISE AND O2+ ON CREATIVE TREATMENT OF INFECTIOUS DISEASES | 7310 |
| OF THE COMMON SOLE/ ORBIT PRESSURE AND O2+ EFFECT OF O2+ | 5740 |
| SYSTEMIC EFFECTS AFTER BREATHING ON POSITIVE PRESSURE IN THE DISTRIBUTION OF VIABLE FRACTIONS OF RAIN | 4984 |
| AS RELATED TO SEASON, THE DISPERSION AND THE LEAST TERM/ THE GATINSON OF THE CHOICE OF SE FREE PATHOGEN FREQUENCY LICATION OF GENETICS TO CAL PROBLEMS IN LUCERNE E HYBRID FORAGE SORGHUM FLORIDA/ AUTUMNAL BANIC ACID FORMATION IN CAL COMPOSITION OF SAKU ADJUSTMENT OF BREEDING COLONY/ PROGRESS REPORT. DI BREEDING BIOTYPE/ THE STATE OF DEVEL 5745 |
| OXIDIZING/ STUDIES ON SABLE WATER/ STUDIES ON SABLE WATER/ STUDIES ON SABLE WATER/ STUDIES ON SABLE BREED/ BRIDING OF INTERRUPTED ATRIOVENTRICULAR BLOCK/ THE STATE OF DEVEL 5754 |
| BREEDING BEHAVIOR OF THE RAVEN WITH BREEDING BIOTYPE/ THE STATE OF DEVEL 7334 |
| BREEDING BIOLOGY AND ETHOLOGY OF THE BREEDING BIOTYPE/ THE STATE OF DEVEL 6544 |
| BREAIDING ON PULMONARY COMPLIANCE/ BREATHING POTENTIALLY MODIFIED AEROSOLS/ BREATHING/ DIAPHRAGM ACTIVITY AND TH 7234 |
| BREED, SEX, AND SEASON QUALITY/ THE T | 7593 |
| BREEDING SHORT/ THE INFLUENCE OF EXERCISE AND O2+ ON BREATHING ENERGETIC OUTPUT AND AEROB 7254 |
| BREATHEING MOVEMENTS IN ENTODONELLA-SO BREATHING AND PATIENTS WITH ADAMS STO | 8526 |
| BREATHING ON PULMONARY COMPLIANCE/ BREATHING POTENI MODIFIED AEROSOLS/ BREATHING/ DIAPHRAGM ACTIVITY AND TH | 5754 |
| BREEDING BEHAVIOR OF THE RAVEN WITH BREEDING BIOLOGY AND ETHOLOGY OF THE BREEDING BIOTYPE/ THE STATE OF DEVEL 7334 |
| BREAIDING ON PULMONARY COMPLIANCE/ BREATHING POTENI MODIFIED AEROSOLS/ BREATHING/ DIAPHRAGM ACTIVITY AND TH | 7593 |
| BREAIDING BEHAVIOR OF THE RAVEN WITH BREEDING BIOLOGY AND ETHOLOGY OF THE BREEDING BIOTYPE/ THE STATE OF DEVEL 7334 |

Figure 2. Title Repetition by Word
Figure 4. List of Words Prevented from Indexing
One of each appears under "low pressure." Following this is one of the most common: "low temperature," with eight unhyphenated entries to three hyphenated—an overwhelming verdict for abandoning the hyphen in this case.

An additional factor for consideration is that most computer sorting treats a hyphenated word as one word—that is, where "low temperature" appears as two separate words, the citation will appear under both "low" and "temperature." Where the hyphen is used, the citation will appear under "low," but not under "temperature."

Such details as the use of singular and plural word forms also can cause difficulty (Figure 6). "Line" and "lines" have been separated by seven entries of "linear" and "linearity." This may seem insignificant in the example, as shown, and it is, because the eye can scan enough of the page at one glance...
Figure 6. Effects of Singular-Plural Usage Upon Computer Sort

to see both entries even though they are separated. If this index covered
5,000 to 6,000 items, these terms could appear so far apart that the searcher
would not see them both, and thus would miss a portion of the indexed
information.

Many of the titles used today have meaningful terminology because the authors
were aware that the titles were likely to appear in KWIC or KWOC index for-

mats. But what becomes of a title such "A Look Ahead," or "Structures and
Economics," or "Pilot Studies and System Components Evaluation"? Each of
these papers is concerned with means of handling information, and each was
presented at a national symposium on Engineering Information, the theme of
which was "A Coordinated Engineering Information System." The difficulty
in indexing titles such as these shows that even information specialists are not making their own jobs easier. This is why librarians must enlist the help of societies and associations such as STWP in reminding others, and remembering themselves, that each piece of documentation is not a unique bit of printing destined to exist forever in its own little cubicle. The present store of documentation is so vast that it is almost incomprehensible, but every new publication must be added to it after the first use. To provide recall from this mass, the major subject or subjects of individual documents must be indicated to the indexer, whether this indexer is a machine or a man.

Examination of the growth of the application of the key-word indexes reveals their importance and indicates the stress that must be placed upon them. In 1958, H. P. Luhn proposed the KWIC concept. Later in that year, Chemical Abstracts Service received a National Science Foundation grant to apply Luhn's concept to chemical information. By 1962, over 30 different permuted indexes were in use, and by 1965 the number had grown to several thousand. The current indexes include such products as "Chemical Titles," published by the American Chemical Society, "Index to Legal Theses & Research Projects," published by the American Bar Foundation, "Current Medical Terminology," published by the American Medical Association, and "Meteorological and Geostrophical Titles," published by the American Meteorological Society. An announcement received recently by the Douglas Library listed permuted indexes available to technical publications such as the Military Electronics Conferences, National Telemetering Conferences, Institute of Radio Engineers, and Institute of Electrical and Electronic Engineers' Transactions, Environmental Sciences Proceedings, and others. Many more of these indexes have been developed for use in industry. Many are produced for internal use only, as, for instance, at Douglas. The extent of the influence of these indexes must be recognized, whether or not the form is accepted as a proper index.

Although the indexing requirement of a title should not take precedence over its main purpose—to describe the contents of the document—nor be considered the only important feature to an indexing system, its value in automatic indexing cannot be overstressed. The Abstract, the Table of Contents,
the Introduction, and the Conclusion also are relied upon to provide guidance for the indexer under almost all circumstances. However, under conventional cataloging conditions, the human indexer is able to exercise some judgment with regard to the relative merits of terms, descriptors, subjects, and so forth. The major emphasis here is directed towards the mechanically created index, where human judgment cannot be applied once the indexing program is established. So even those who question the value of permuted indexes should provide titles that can be used.

Another serious problem arises in relation to the publishing of all these papers and documents. Every author enjoys having his creation appear in print, and without qualification, "the more the better." Certainly, each publication is directed toward a different audience and justification for duplication is easy. But what does duplication do to the cataloger, the reference librarian, and the researcher? Imagine the frustration of finding five citations to a subject, only to discover that the five citations cover two articles published in five different sources! Such multiple publication of the same article causes problems for the scientist and engineer as well. Also, the engineer's sources or references may be credited to a corporate author or publication different from that used by the library for the same article.

Figure 7 shows an example of the many corporate authors possible for a single article. For this particular item, the work was done by a member of the staff of the California Institute of Technology, under funding from the Air Force Aerospace Research Labs, and submitted for publication to the "American Journal of Physics." The report was transmitted per contractual agreement to the Defense Documentation Center for dissemination throughout the defense community. Because of its application to aerospace endeavors, the National Aeronautics and Space Administration also acquired the document for dissemination. Thus, there are five sources established, three of which are primary and two (DDC and NASA) are secondary.

Are not all of these sources entitled to claim this document as their own? Certainly they are. The Aerospace Research Lab paid for the research, the California Institute of Technology prints its staff's output, the "American
Journal of Physics agreed to publish the article for circulation among the scientific community. Both DDC and NASA are obligated under their charters to announce and disseminate all work performed through government funding.

Those responsible for the publication of material cannot solve this problem by limiting the number of times a report is published, but they can ensure that each publication makes note of the other sources in which it appears. When previous or prospective publications are known, why should not this information be included in a footnote? If the specific report number or issue number were not available, it would be helpful if a statement such as this were used: "Also published as Jet Propulsion Lab Technical Report and Aerospace Research Lab Technical Report," or "Announcement is anticipated in the Defense Documentation Center's Technical Abstract Bulletin and the National Aeronautics and Space Administration's Scientific and Technical Aerospace Reports Bulletin." Either of these statements would provide both cataloger and scientist with a warning that the document may appear under any one of those various corporate authors.
Often library users do not recognize the same material in different forms. Having once seen a journal article, a user may not be aware that the grey-cover ARL Report contains the same material—at least, not until he has requested the document and has begun re-reading it. This can also be true of the cataloger and librarian. Yet, with the tremendous quantity of information published today, information specialists must do all they can to save the reader from such circumstances. Libraries are given generous budgets, professional staffs, and expensive computer programs to inform their clients of material of interest to them. Informing an individual over and over again of the same material is not pleasing to either the individual or the responsible information specialist. The integrity of the entire information handling system is downgraded by these occurrences. To help avoid this, the author and the technical writer should give notice of the multiple publication of any documents he handles.

With the widespread use of mechanized systems has come another problem—the variations used in individual corporate names. Some mechanized programs have avoided the problem to some degree by coding the corporate authors. However, this imposes another set of problems. The concern of the information specialist is with the use of company names, agency names, universities, and so forth, as they are input to a computer. Take, for example, "IBM." In most communications, the use of "IBM" conveys the desired message. But the name of the company is International Business Machines, so corporate source notations should appear with the full name, "International Business Machines." This is a simplified example, because most catalogers would know that the citation should use the full name. It is not always so obvious, however, whether initials can constitute an acceptable corporate entry. IBM is not acceptable, because this is not the legal name of the organization. IIT Research Institute is acceptable and correct, since this organization officially discontinued the use of the title "Illinois Institute of Technology Research Center" and adopted IIT Research Institute as its official name. These examples illustrate how confusing to new and
Inexperienced catalogers abbreviations in corporate names can be. Such abbreviations should be avoided unless they constitute a legal designation.

Even such variations as abbreviating "company" one time and spelling it out the next time can cause confusion. When a computer sorts, it sorts according to standard computer rules. Douglas could appear at least four different ways (Figure 8), without counting eliminating the periods or including divisional entities. The computer does not recognize all these names for the same company—it recognizes only a change in the characters in each.

Authors, technical writers, and information specialists must all be certain to use the corporate name in the same manner each time, so that the items will all appear together in a computer printout. It is important for indexers and catalogers to be precise about entries, but the tendency is to accept an

DOUGLAS AIRCRAFT COMPANY
DOUGLAS AIRCRAFT CO.
DOUGLAS AIRCRAFT COMPANY, INC.
DOUGLAS AIRCRAFT CO., INC.

Figure 8. Variations of Single Corporate Author
entry as it appears on the document. For this reason, the careful use of the corporate name can be of great help in facilitating the retrieval of valuable information.

Projections into the future indicate that, without doubt, the next major breakthrough in information systems will be total text input of some sort. Paper tapes and optical scanning methods have been tried with varying degrees of success but they reveal that there are several obvious questions to be answered in relation to full text input. These are as follows:

1. What about the extent to which the "stop list" would have to be enlarged beyond that of the KWIC stop list?

2. What about synonyms and near-synonyms? In the interest of variety, the writer will strive for a variation in terms, perhaps using words such as automobile, car, vehicle, sedan, all in reference to the same subject. When retrieving from total text input, care must be taken to ensure that each of these terms in used for recall.

3. What about antonyms? When a request arises for information on safety in handling chemicals, will there be a reference also to "hazards in handling chemicals"?

4. What about adjectives and adverbs that lose meaning when taken out of context? These words create problems, not so much in retrieval as in clutter (Figure 9). They are a vital part of all writing for logical communications, yet they are rarely used for retrieval. One interesting exception is the term "orbital." When first entered into the Douglas dictionary, the only term used was "orbital," although the usual procedure was to enter both the plural and singular forms of all terms. It seemed obvious that no one would ever use "orbital" by itself, much less in plural form. Yet within no more than 6 months, Douglas librarians were not only coping with "orbital paths," as expected, but also with "orbitals." The plural form has now become an accepted term in the aerospace industry.
Figure 9. Automatic Index
CONCLUSIONS

The technical writer's key to computerized information retrieval is not the restriction of all writing to a 10,000-word vocabulary. Rather, the key is the thoughtful selection and use of truly pertinent and descriptive terms, especially in those portions of a report which serve as key tools for effective storage and retrieval of valuable information, and in fully documented publication notations. Those who feverishly embrace the new word, or the new concept described by an old word, must consider the problems of the information retrieval specialists, and must realize that the kind of vocabulary used today must be retrievable tomorrow.


