been given to assisting foreign scientists who are planning trips to America, so that visits to appropriate individuals and institutions may be included on their itineraries. Foreign requests for information on everything from where to study cardiology in the U. S. to possible sources of ultra-pure rare earth samples are handled as a matter of routine. The same assistance is willingly given, of course, to American scientists, all within the limits of available time.

In summation, probably the most effective means of communication is putting the right people in direct touch with each other. We believe that the Science Attaches of the State Department have an important role in this effort, and one which will increase in importance and effectiveness as their numbers grow and as understanding of their work becomes more widespread.

---

# Development and Production of *Chemical Titles*, a Current Awareness Index Publication Prepared with the Aid of a Computer

By ROBERT R. FREEMAN and G. MALCOLM DYSON
The Chemical Abstracts Service, The Ohio State University, Columbus 10, Ohio
Received March 8, 1962

The introduction of *Chemical Titles* in 1961 marked the first publication produced almost entirely by computers and other data-processing equipment. The success of this innovation has generated many requests for more information about it. With this in mind, we hope to encourage other organizations to make use of this technique for dissemination of information by presenting here a history of *Chemical Titles*' development coupled with a description of its production.

**Reasons for *Chemical Titles*.**—With the great expansion of research activities in the chemical and allied disciplines after World War II came the now familiar but dramatic rise in the amount of published research information. It was not long before the traditional means of acquiring information, primary journals, reports, and abstract journals, grew so bulky that individuals were overwhelmed. Despite the desirability, both for intellectual and commercial reasons, of keeping up with current developments in one's field, one heard more and more that it was a hopeless task.

During the early 1950's many people began to look to the newly developing capabilities of computers as tools for organizing information. It was early suggested that indexes should be prepared by computers. While realizing the complexity of the process of indexing scientific reports, the newly formed Research Department at the Chemical Abstracts Service began to investigate the possibilities of using information-handling equipment to prepare an acceptable index to the newest research papers.

In the fall of 1958, we became convinced that the KWIC system of keyword indexing as designed by H. P. Luhn of IBM could be developed suitably as a scheme for indexing the titles of chemical communications by computers.[1]

Indexing by key words, with meaning clarified by context, was not new. Scholarly concordances have been known for centuries. The Central Intelligence Agency has prepared a permuted title word index using keypunch, reproducing punch, sorter, and tabulator since 1953.[2] However, Luhn's KWIC method offered easy and extremely rapid handling of large volumes of information, relatively simple preparation of the input to the computer, and output of a product which is readily reproduced by photographic offset methods and easy to use.

The Chemical Abstracts Service had long taken pride in its Subject Indexes, emphasizing standardized chemical nomenclature and indexing by concepts rather than words. At this point, however, the management recognized that to cope with the problems of timeliness, to the extent of producing an index to literature which has just appeared, required a departure from tradition. In the fall of 1959, the National Science Foundation's Office of Science Information granted $150,000 to the Chemical Abstracts Service to study for a year the feasibility and acceptability to chemists of keyword-in-context computer-produced indexes.

**Early Stages of the Research Program.**—The initial stage of the research program consisted of selection of a limited number of journals to be covered by the new service from among the over 9000 covered by *Chemical Abstracts*. It was decided initially to limit coverage to periodical publications which appear quarterly or more frequently. A number of staff members of the Chemical Abstracts Service cooperated in a survey of abstracts from presently published journals. They reached the conclusion that some 600–700 journals publish about 60 per cent. of the world's chemical research results. Several books dealing with the literature of chemistry were consulted in an effort to be certain that the major journals of each country and of each area of chemistry were included. The initial group selected, which was to be expanded later, consisted of 550 journals.

A second preliminary step consisted of selecting a group of words to be rejected as index entries. The machine program allows that a pre-selected set of words, the number of which depends on the memory capacity of the computer used, may be rejected as keywords because of

lack of value as retrieval terms in an index. Several staff members of the Chemical Abstracts Service made lists of words to be included in this list. After discussion, the lists were edited and combined. A total of 750 words was selected. After some early experience the list was expanded to 950 words. Later, as more statistics were collected, we decided that words which occurred less than once in 10,000 words should be allowed to index even if they are of no value, since they consume a rather insignificant amount of space. This decision caused the size of the list to fall to 328 words, the present composition.

Having taken these preparatory steps, we felt that a test of the program was in order. By using tables of contents of a sample of important chemical journals, some 2000 titles with their corresponding authors and journal references were keypunched. After the resulting deck had been processed by an IBM 704 computer, which had been given the list of words to be rejected and the instructions for the KWIC program, the results were studied carefully. This experiment proved that, with a few improvements, there should be no difficulty in producing a publication based on the keyword-in-context program. The rules for editing titles in preparation for keypunching began to take shape at this point.

**Pre-Editing for a Better Index.**—The experiment brought home the fact that some new editing concepts had to be formed to make a machine-created index amenable to human searching. In this program the computer regards as a word any set of characters (*i.e.*, their representation as punches in cards or magnetic dots on tape) which is unbroken by a space. Chemical terminology, particularly that of organic chemistry, is characteristically made up of words (in the definition given above) which often include several units of meaning. It is often the case that those segments which are useful to a chemist who is searching an index are buried behind several prefixes which tell only where (locants), how many, or in what configuration something else is.

Having grasped this idea, we realized that chemists would seek information on both specific and generic topics in the proposed index. If we allowed each word to index, the index would be highly specific. Therefore, to make a degree of generic searching possible, we began to separate editorially chemical words into the segments from which they were formed before keypunching the words as input to the computer. Thus, 2,3-dichlorobutane is split so that indexing takes place at "chloro" and "butane." Words beginning with numbers are automatically rejected as index entries by the machine program. As a corollary to the rule, we found it to be possible to prevent the indexing of some terms by connecting them to a previously occurring, related term by a hyphen, making a single word from several. For example, "van der Waals" will index at three locations. By inserting hyphens to create a single string of characters, one can editorially prevent the indexing of two useless terms. A word may also be prevented from indexing by allowing it to begin with a non-alphabetic character, such as a parenthesis.

It had occurred to us since early in the project that the tedious editorial step of indicating where words should be divided might be mechanized. Decomposition of words into various segments has been an integral part of programs for machine translation, particularly in highly inflected languages such as Russian. Recently published work in which a portion of the system of chemical nomenclature was analyzed linguistically has given us further confidence that the step will become feasible.[3] We are collecting many data on the occurrences of words in titles with automatic editing as a partial objective.

In present-day machines, one must also learn to do without lower case letters and the great variety of symbols which chemists are so prone to use. While there are definite indications now that a greater number of characters will be available in the future,[4] we had to develop pre-editing conventions which were compatible with existing machines. These conventions are:

A. Symbols and Abbreviations.—In general, abbreviated expressions are fully spelled out, although in the few cases in which the abbreviation is used almost exclusively (such as DDT), it may be acceptable to allow it to remain. The names of all chemical elements and compounds are spelled out. If the name of a compound cannot be established, an empirical formula in spelled-out, form is given, following the rules given below for sub- and superscripts. Names of symbols are spelled out. For example, $n_D^{20}$ would be written "refractive index."

B. Greek letters.—The letters $\alpha$ and $\beta$, when used as locants in organic chemical nomenclature, are so common that the letters A and B are used to represent them. In all other cases, the name of the letter is spelled out.

C. Subscripts and Superscripts.—1. Isotope mass numbers are connected to the element name by a hyphen.

2. The number of atoms of an element in a compound is written in parentheses immediately to the right of the element name.

3. Valence is indicated in Roman numerals in parentheses immediately to the right of the name.

4. Other subscripts and superscripts, in general, are placed in parentheses. In the event the entire expression in question is already within parentheses, it is often possible to use a hyphen in place of the parentheses which normally would enclose the sub- or superscript. For example, $\Delta^{1,4}$-ketosteroids would be keypunched (delta-1,4)- keto steroids.

**Market Research.**—The technical feasibility of keyword-in-context indexing having been proved, we began a survey of the market potential. A sample of 2700 titles was keypunched and, with the cooperation of H. P. Luhn, was processed by an IBM 704 computer at the IBM Research Center in Poughkeepsie, N. Y. The new publication was named *Chemical Titles*.

Seven thousand sample copies of *Chemical Titles* were distributed to registrants at the Cleveland Meeting of the American Chemical Society in April, 1960. An additional 18,000 copies were mailed to chemists and librarians around the world. A second issue similar to the first was distributed to another 16,000 potential users of an express index service in May.

During 1960, a total of about 100,000 copies of *Chemical Titles* was distributed free of charge, in order to give chemists and librarians throughout the world the opportunity to evaluate and comment on the new service. By July, 1960, enough data were available for the American Chemical Society's Publications Committee to decide to authorize *Chemical Titles* as a regular publication of the Society beginning in 1961. Promotion of

*Chemical Titles* during the latter half of 1960 resulted in 2500 subscriptions, and enabled the publication to be self-supporting when it went into production. Often new publications do not break even during the first three years of operation. One of the key factors enabling the *Chemical Titles* service to get a start was the generous support and confidence of the National Science Foundation. This support enabled a large, widely dispersed population to become aware of the existence of the unusual service and to gain some familiarity with it.

The way in which *Chemical Titles* is used is difficult to explain verbally. Visual inspection, however, quickly reveals the simplicity and effectiveness of use. We believe that sampling was the key to the success of the promotional campaign and that without sampling, it is questionable whether success for *Chemical Titles* could have been achieved at all.

Initially, the ACS used a survey-order type of promotion without committing either the Society or the prospective purchaser. As the campaign progressed and some idea of the potential market was obtained, firm subcription prices were set and intensive solicitation of orders was undertaken.

Those responsible for promotion believe they reached the vast majority of potential prospects. The following mailing lists were used:

All subscribers to *Chemical Abstracts*, *i.e.*, base rate, colleges and universities, and ACS member subscribers. (The *CA* subscribers were solicited repeatedly in all six promotions where sample copies were sent and, in a separate promotion, with a reprint of a *Chemistry and Industry* editorial praising *Chemical Titles*); 5,000 names from the *American Library Directory*; 5,000 names of Industrial Research Laboratories; ACS Division of Chemical Literature, 1,000 names; expired *CA* subscribers, 1,300; Colleges and Universities (ACS applied publications list, 1,100; American Rocket Society names (ACS applied publications list, 2,500; Industrial Research Laboratories (ACS applied publications list), 4,000.

**Feedback and Further Experimentation.**—One of the benefits of the lengthy experimental and promotion period was the feedback from those who had had a chance to use the new product. Their suggestions resulted in better coverage, fuller use of space in the keyword index, the addition of an author index, and an experiment in classification of the keyword index. Experience enabled us to improve the machining procedures.

Initially, we did not feel the need to include an author index as a separate feature. H. P. Luhn had developed a machine-derived code which consisted of the first four letters of the last name of the first author listed with a title, his two initials, the last two digits of the year of publication, and the first letter of each of the first three words, excepting fifteen prepositions and conjunctions. The dual purpose of this code was to provide a unique identifier for each paper, according to which the bibliographic listing of references could be ordered, and to provide a means for getting from the keyword index to the complete title and reference from which the keyword is taken.

Because of the nature of this code, we felt that the ordered bibliography itself would serve as an adequate author index. Many users of the sample issues pointed out, however, that the senior author of a paper may not necessarily sign his name first, thereby rendering our form of author index only partially useful. We found that, with very little extra effort, a complete author index could be added to *Chemical Titles* as a separate section.

Many users suggested that we cover journals not included in the original list. A number of important journals we had overlooked were thus added, bringing the coverage to slightly over 600 journals.

The program for producing the keyword-in-context index, as originally conceived by H. P. Luhn provided that each keyword be placed in the center of the column with its context around it. If a keyword happened to be the first or last in a title, nearly half of the line was left blank. By an addition to the program, we were able to instruct the computer to fill in this space with any remaining parts of the title which did not appear on another part of the line. The new feature considerably enhanced the value of the index by allowing the printing of additional context on which the user may base his interpretation of and interest in the keyword.

Improved machine processing was achieved through a compromise based on the experience of the first issues. According to the original plan for producing *Chemical Titles*, the machine-created keyword index as well as the original data with a code reference added by the computer, were to be transferred from magnetic tape to IBM cards, manually sorted and edited, then printed out from the cards. By using the equipment available to us, this process took six days. By accepting the existence of a relatively small percentage of useless index entries and the slightly higher cost of computer sorting, it was possible to reduce the processing time to one day by printing out directly from magnetic tape.

After the first two issues of *Chemical Titles*, there was some sentiment toward dividing the keyword index into more than one section. It has been a matter of controversy whether the chemist who is searching the index can define his interests as falling into only one of several areas. If he can do this, he can save time by having a shorter index available. If his interests overlap he probably must spend longer searching because he must look in several places.

As an initial experiment in classification, the keyword index was divided into biochemical and general sections. At first, a certain group of journals was listed as biochemical, the rest general. However, a significant amount of biochemical material appears in the so-called general journals. Titles were then classified individually in the following two promotional issues. This classification scheme drew a mixed reaction. Finally, an issue in which the keyword index was divided into twenty subject areas was produced. The opinion of future subscribers to *Chemical Titles* was solicited both in the issue and by direct mail. After consideration of the higher cost of producing a classified index and the two-to-one preference of subscribers for a single index, we decided to return to the original form.

**Chemical Patents-Another Experiment.**—Realizing the particular need for a current awareness indexing service for patents, we attempted to apply the keyword-in-context indexing principle to this area. Patent titles, it is well known are often written with the intent of concealing the

content of the patent by being as vague or general as possible. In order to remedy this situation, we rewrote over one half of the titles, using keywords derived from the text. *Chemical Patents*, as the publication was called, provided keyword, inventor, and assignee indexes as well as a bibliographic listing of patents by number within each country. Despite our attempts at producing better titles, *Chemical Patents* was not well received. Users commented that so many patents are issued for products or processes which differ only in some minute point from one another that it is impossible to discriminate on the basis of titles alone.

**Production of *Chemical Titles* During 1961.**—Production of *Chemical Titles* is organized as a research project, directed by a project leader who is responsible to the Director of the Research and Development Division of the Chemical Abstracts Service. The entire staff required for the production of *Chemical Titles* consists of the project leader (half-time), one clerical assistant, two keypunch operators, and approximately one day per week of the time of an IBM supervisor. We have purchased computer time outside our organization for operation of the KWIC and allied programs.

During 1961, the preparation of *Chemical Titles* proceeded as follows. Workers in the Chemical Abstracts Service library were instructed to send each issue of the more than 600 journals being covered to the *Chemical Titles* office. Many publishers cooperate with *Chemical Titles* by supplying tables of contents in proof before publication.

If a table of contents is published in a language other than English, it is next sent to a translator. Translators for *Chemical Titles* are all chemists who are associated with the Chemical Abstracts Service in one way or another. Each translator works at a fixed rate per title. Thus, *Chemical Titles* has been able to benefit from the chemical and language skills of a greater number of people than would have been possible had a full-time translator been employed. Some 18,640 titles were translated in this manner in 1961. This figure reflects the fact that many publishers of non-English language journals include an English table of contents in their journals.

At this point all of the titles are in English and are ready for editorial preparation. The editor, who is the project leader, first selects, by making a red line at the margin, each title which is to be indexed, according to the same selection rules employed for *Chemical Abstracts*. Briefly, these rules state that a paper must either present new, previously unpublished chemical information, or be a comprehensive review, and must not be anonymous. Translators also follow the selection rules.

Next, for reasons discussed earlier in this paper, the editor indicates with red lines the points at which the keypunch operators should leave spaces where they do not already exist, insert words or hyphens, delete unwanted expressions, *etc.* It was also necessary in 1961 for the editor to write the accepted abbreviation of the journal, along with its volume number and year, at the top of the page.

The clerical assistant then writes in the number of the last page of each article selected next to that of the first page. If the journal is available, the pagination can be given with accuracy. When only a table of contents is available, the assumption is made that the last page of an article is one number less than the first page of the succeeding article. Using a sequential numbering device, the assistant numbers each title.

The titles which have been selected, translated, and edited, are now keypunched in the form which will allow a computer to handle them. Each title may be thought of now as consisting of three parts, each of which is coded separately during the keypunching. The parts are the authors (code 1), the title itself (code 2), and the journal reference (code 3). Because more than one card frequently is needed for the information in each class, the cards are numbered sequentially within the class. Each card also is punched with the number given to that title.

All titles are machine verified and corrected. After this step, the titles are listed on an IBM 407 tabulator and the listing is sight-verified by the editor as a further check for accuracy. This step eliminates about twenty-five to thirty errors per issue of *Chemical Titles*, or about one per hundred titles, many of which are editorial rather than keypunch errors.

Despite the apparently great difficulty of the material being keypunched, our operators are able to achieve the above degree of accuracy while keypunching at the rate of nearly 10,000 key-strokes per hour, which is average for any keypunching operation. This rate of work yields about 1500 cards per day, or over 300 complete titles. Verification, together with correction of errors, proceeds at about the same rate as keypunching.

After having been sorted into order and sequence-checked according to the code described above, the cards are sent to the computer installation. There the information punched in them is transferred to magnetic tape and processed by the computer. About 2.5 to 3 hours for computing on the IBM 704 and about 2 hours for card to tape and tape to printer operations on the IBM 1401 are required for an average 125-page issue of *Chemical Titles*.

The major steps are : (1) derive the identification code described earlier. (2) Examine each title-word, passing by those which are in the dictionary of non-indexable words, and creating a KWIC index entry for every other word. (3) Sort of KWIC index into order by keywords and all other words to the right of them. (4) Sort the Bibliography into order by identification code. (5) Create and sort the Author Index into order. (6) Print out all three sections.

Upon receiving the printed copy from the computer installation, the clerical assistant pastes the sheets loosely on light cardboard and ships the resulting pack to the printer for offset reproduction.

We have obtained the tabulated data from the first full year of production.

**Further Developments and Changes.**—Early in 1960, independently of the *Chemical Titles* operation, we began to explore the possibilities for development of a unique four-letter mnemonic code to represent each of the periodicals covered by *Chemical Abstracts*. Machine handling of large volumes of chemical information seemed to call for highly condensed journal citations.

At that time we began to receive occasional criticisms of the reference code employed in *Chemical Titles* and

| | Total | Average per issue |
|---|---|---|
| 1. General Data | | |
| Titles | 68,400 | 2,850 |
| Pages | 2,997 | 125 |
| Words | 4,923,000 | 205,125 |
| Words | 1,641 | |
| Words required to completely list and index a title | | 72.4 |
| 2. KWIC Index | | |
| Total keywords indexed | 410,800 | |
| Keywords per title | 6.00 | |
| 3. Bibliography | | |
| Average words per title | 11.5 | |
| Keywords/Total ratio | 0.52 | |
| 4. Author Index | | |
| Total lines | 147,000 | |
| Average authors per paper | 2.15 | |

the resulting bibliographic arrangement of titles from users of the early experimental issues. The criticisms reflected two points: The apparent waste of space resulting from the year digits in the code being identical for most papers and the desire of users of *Chemical Titles* to be able to see a table of contents of some favorite journals. Re-examining the code, we realized that Luhn had originally conceived it for files which (1) are historical by nature and (2) do not provide a separate author index. Neither of these points is characteristic of *Chemical Titles*.

It became apparent to us that a large portion of *Chemical Titles* could be prepared by the use of punched card equipment alone if we were not dependent upon the computer for the derivation of the reference code. Finally, it occurred to us early in 1961 that all of these problems could be resolved at once by changing to an identification code based on the journal citation rather than the author. The Vari-Typer Headliner, a simple letter-by-letter photocomposing machine, has enabled us to print the heading of each table of contents.

An additional advantage of the new arrangement lies in saving paper and printing costs. With a journal reference line included with every title, approximately one quarter of the Bibliography Section in the format used during 1961 was devoted to journal citations. With the new system, a space equivalent to three lines is required for the headline. Since there are ten titles in the average journal, a net saving of seven lines per journal is apparent. The over-all saving in space required to print *Chemical Titles* should amount to about 7 per cent. The decision was therefore taken in mid-1961 to adopt the new format and journal-oriented code for 1962.

Some procedural differences between the former and present methods are now evident in the preparation of *Chemical Titles*. We are no longer dependent on a computer for the derivation of the code. The editor indicates at the top of each table of contents the code for the journal and its volume number. The keypunch operators complete the code by adding the first page number of each individual article. The title cards only are reproduced and sent to the computer for the KWIC Index preparation. At the same time, the Bibliography is listed on an IBM 407 tabulator and the author index is prepared by reproducing and sorting these cards into alphabetical order and listing them. An important factor to be considered is the balance of cost and time consumed in preparation. Under other circumstances, and at other times, it may indeed be more desirable to perform the entire routine with the aid of a computer.

**KWIC Indexing In Other Areas.**—Our experiments and the interest generated by *Chemical Titles* have led us to prepare several indexes to subject material in fields other than chemistry. We have applied KWIC indexing to papers in medicine,[5] meteorology,[6] and general scientific words in all fields.[7] It is our feeling that the method is even more easily applied to fields other than chemistry, where the peculiar problems of chemical nomenclature are not present. A possible exception is mathematics because of the prevalence of symbols, the expression of whose meaning requires many words. As might be expected a large proportion of the list of words rejected for indexing in chemistry applies also to other fields. On the other hand, the general words belonging to one field, such as "chemistry" itself, may be important keywords to another field.

The present keyword-in-context concept has proved to be of great utility for the rapid dissemination of large quantities of information in indexed form. As a pioneering effort in machine indexing it does not match, nor is it designed to match, the subject indexes produced by human effort.

Acknowledgment and appreciation is expressed to the Science Information Office of the National Science Foundation for the support of the research and development involved in this project.

## REFERENCES

(1) H. P. Luhn, "Keyword-in-Context Index for Technical Literature (KWIC Index)," IBM Advanced Systems Development Division, 1959, Yorktown Heights, N. Y.
(2) M. P. Veilleux, "Permuted Title Word Indexing Procedures for a Man-Machine System," paper presented at the Third Institute of Information Storage and Retrieval, American University, Washington, D. C., February 14, 1961.
(3) Eugene Garfield, "An Algorithm for Translating Chemical Names to Molecular Formulas," Institute for Scientific Information, 1961, Philadelphia, Pa.
(4) The development of better printing capabilities is taking two directions. These are the addition of more characters to the print chains of high speed printers and the use of computers to control automatic photocomposing machines.
(5) "Medical Titles" (Sample Issue), IBM Advanced Systems Development Division, Yorktown Heights, N. Y., 1960.
(6) "Meteorological and Geoastrophysical Titles," Sample Issue, Vol. 1, No. 1, January, 1961, American Meteorological Society, Washington, D. C., 1961.
(7) "KWIC Index to the *Science Abstracts of China*," Massachusetts Institute of Technology Libraries, Cambridge, Mass., 1960.