and where $H = 2 + 2(13) + 2 - 0 - 2(2) = 26$. Final m.f. = $C_{13}H_{26}O_4$.

As a third example consider 1,4-bis[3-bis-(diethylamino)propylamino]butane. By morphemic analysis, it becomes

$$\begin{aligned}
&2[2(2[2C] + N) + 3C + N] + 4C + 0\\
&2[2(4C + N) + 3C + N] + 4C\\
&2(8C + 2N + 3C + N) + 4C\\
&16C + 4N + 6C + 2N + 4C = 26C + 6N = C_{26}N_6\\
&H = 2 + 2(26) + 6 - 0 - 0 = 60 \text{ and the m.f.} = C_{26}H_{60}N_6
\end{aligned}$$

Finally, consider the example of 1,2,3,4,5,6-hexanitrohexatriene. By morphemic analysis, it becomes

$$\begin{aligned}
&6(N + 2\phi + DB) + 6C + 3DB\\
&6N + 12\phi + 6DB + 6C + 3DB = 6C + 6N + 12\phi + 9DB\\
&\qquad\qquad\qquad\qquad\qquad\quad = C_6N_6O_{12} + 9DB\\
&H = 2 + 2(6) + 6 - 0 - 2(9) = 2 \text{ and m.f.} = C_6H_2N_6O_{12}
\end{aligned}$$

In this particular case the morphemic analysis is not as straightforward since there are several potentially ambiguous morpheme combinations.

Consider the chemical 2,3,4-tris-[3-bis-(dibutylamino)-propylamino]-pentadiene-1,4. Off the computer, this compound results simply in $3[2(2\,C_4 + N) + C_3 + N] + C_5 + 2DB$. Carrying out the simple multiplications and additions gives a partial molecular formula of $C_{62}N_9 + DB_2$ and $H = 2 + 2(62) + 9 - 2(2) = 131$. m. f. = $C_{62}H_{131}N_9$. The structural diagram of this chemical is shown (see Fig. 1) to indicate how time-consuming it can be to go through the procedure of drawing such a diagram in order to calculate the molecular formula.

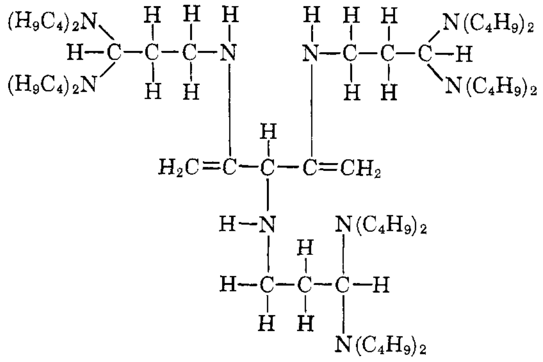

Fig. 1.

With a little practice, one quickly memorizes the common morphemic values and is able to get the basic notion of how to identify them quickly. Obviously, if you want to calculate such names as 17B-amino-3B-androstanol, your dictionary (or your memory) must tell you that androstane contains nineteen carbon atoms and four rings (double bonds). Most steroid chemists would have this morpheme memorized. However, even a clerk can look it up in the dictionary. Using the algorithm one quickly finds the molecular formula directly from the chemical name

$$\begin{aligned}
&N + 19C + \phi + 4DB\\
&H = 2 + 2(19) + 1 - 2(4) = 33
\end{aligned}$$

The formula is $C_{19}H_{33}NO$

(1) E. Garfield, "An Algorithm for Translating Chemical Names to Molecular Formulas," Institute for Scientific Information, 1961. See also E. Garfield, *Nature*, **192**, 192 (1961).

The Data Compilation as Part of the Information Cycle*

By N. B. GOVE and K. WAY
Nuclear Data Project, National Research Council, Washington, D. C.

Received March 22, 1961

The majority of scientists are hard at work doing their very best to increase the information explosion. A few scientists spend a portion of their time trying to make sense out of the data produced by their colleagues. The first step in such a process is to collect the data, convert to consistent units where necessary, choose best values where possible, and arrange the results in some useful and clearly stated way. This activity may be called compiling and the resultant work a data compilation. In this respect a data compilation is an information retrieval tool—not adding to the available information but rather trying to arrange it in a more convenient form.

Many people think of information as a kind of nutrient. If you feed the right scientist with the right information he will grow. While this view is sometimes useful it has a serious drawback.' It ignores the fact that the scientist may not remember the information. Now the fact that the scientist may forget the information may indicate that he is the wrong scientist for the job. But that is a shortsighted view. If a scientist reads of an experimental result now and next year he reads of another result couched in different units, with different correction factors, referred to a different standard, in a different journal in a different language, he may not recognize that here is a discrepancy that suggests a new experiment, or that here a trend is shown which suggests a new theory. While many information specialists are concerned only with supplying a complete list of references, current or retrospective, this should be regarded as an intermediate goal. To be successful, information systems should provide the right information to the right person *at the right time, in the right units, in the right language,* with discrepancies or new trends clearly visible, and with a minimum amount of chaff. To design systems that will do this we need, among other things, better data compilations.

Data compilations are most useful in active fields of science, fields in which new results are coming in rapidly

* Presented before the Division of Chemical Literature, American Chemical Society National Meeting, Washington, D. C., March 24, 1962.

and in which there is some degree of confusion concerning how best to display and interpret these results. In these fields compilations can and should play an important role in the information cycle—that is, the process in which information is published, received, stored, retrieved, studied and finally used in the production of new information. As more and more documents pour in it becomes increasingly unpleasant to think of reading all the papers in one's own specialty plus all those in related specialties and still have time for research. Even if present retrieval methods are developed to the limit of their capabilities and retrieve all, and only, the relevant articles there is still the job of reading them. In this regard the compilations can be used to ease the job of the researcher, at least in the borderline specialties. Compilations, if properly made and properly used, can save many hours of library searching.

The researcher who wants to use a compilation is severely limited by three factors: (1) he may not know if it exists (2) he may not know the exact reference and (3) he may know the exact reference and still be unable to find the compilation.

Compilations can be very difficult to find. Many are unpublished; some are privately circulated; some are in progress reports. The absence of any standard publication procedure has resulted in an array of strange documents. Some examples are shown in Table I. Consider the one that reads: Coefficients of Internal Conversion of Gamma Radiation, Part II: L-Shell—USSR Press; Moscow, Leningrad; issued in USA as 58ICCL1. Many, perhaps most, librarians could eventually obtain a copy if given this information. But things may not be that simple. Suppose that you are a librarian or information officer and some scientist says to you, "I seem to remember a table of conversion coefficients put out two years ago in Russia. Can you find it for me? It had a light blue cover with a piece of dark blue tape on the binding." Now the problem has become more difficult. But let us assume that you manage to find this compilation. Do you think the scientist will go away and leave you alone? Not at all. He will say, "Thank you very much, I'm sorry I gave you the wrong year and the wrong country. I hope that didn't make things any harder for you... But at least I got the color right... By the way, is this the latest and best collection of conversion coefficients?"

Table I. Examples of Unusual Compilations

Tables of Clebsch-Gordan Coefficients	NAA-SR-2123
Relative Isotopic Abundances	1959 Nuclear Data Tables
Radionuclides Arranged by Gamma Ray Energy	Nucleonics Data Sheets
Comprehensive List of Nuclides with Atomic Mass, Half-Life and Specific Activity	AHSB-44
Cross Sections for Fast Neutron Reactions	Texas Nuclear Corp.
Neutron Cross Section Trends Around 14 MeV	1961 Progress Report, Dept. of Chem., U. of Ark.
Coefficients of Internal Conversion of Gamma Radiation Part II: L-Shell	USSR-Press; Moscow, Leningrad; issued in USA as 58ICCL1
Radiations from Radioactive Atoms in Frequent Use	USAEC-Feb. 1959
Effective Cross Section Values for Well-Moderated Thermal Reactor Spectra	CRRP-960 EANDC-4 TNCC-30 AECL-1101

It is in trying to answer this last question that many information systems and information officers break down. However, this example is intended to illustrate only a small part of the over-all problem with respect to compilations, namely, that compilations are not realizing their full potential usefulness in the information cycle.

It would seem worthwhile for some group containing both documentalists and scientists to examine the over-all situation with respect to data compilations and compilation systems in order to make recommendations for improving the quality, coverage, and availability of data compilations. Some ideas that might be considered are: (1) a co-ordinated directory giving descriptions and locations of compilations, (2) establishment of journals specializing in compilations, (3) asking editors of abstract journals to mark compilations in a distinctive way so that they can be found more readily, (4) coding retrieval machines to distinguish compilations, (5) encouraging publication of compilations in areas where most needed, (6) establishing standards of style and format, and (7) establishing standards of quality for compilations.

We don't claim that all of these things are simultaneously desirable but we do claim that some improvements can and should be made and that a cooperating group of documentalists and scientists together could find these improvements.

Of course some work is already being done along these lines. A directory of continuing numerical data projects and their publications has been issued by the Office of Critical Tables.[1] A directory to Nuclear Data Tabulations has been published by the Nuclear Data Project,[2] showing that some 83 compilations were published in 1958 and 1959 relating to low energy nuclear physics, only 17 in regular periodicals. The American Chemical Society has a journal which accepts compilations.[3] The American Institute of Physics is seriously considering a proposal for a journal to be devoted primarily to compilations. With respect to item (5) above, the Sub-committee on Nuclear Constants of the National Academy of Sciences has encouraged a number of compilation projects.[4] Improvements in areas (5), (6), and (7) above have been brought about by the Office of Critical Tables of the National Academy of Sciences. Still a great deal remains to be done.

One sometimes hears the statement that cooperation between scientists and documentalists has not been what it should be.[5] We wish to point out that in a study of data compilations such as that proposed here both scientists and documentalists are essential for any progress. Both groups would realize this and so it seems a good place to begin improving the rapport.

In conclusion, another well-known statement is that most of the information retrieval we have today is merely document retrieval. It's also obvious that we can never catch up with the information explosion with document retrieval alone, no matter how good. The idea of a compilation system is not profound, or new, or revolutionary, but at least the data compilation gets inside the document and works with the information directly. Any improvement that can be made in the availability, quality, and completeness of data compilations will be a positive step in the direction of true information retrieval.

BIBLIOGRAPHY

(1) A Directory of Continuing Numerical Data Projects, Office of Critical Tables, National Academy of Sciences-National Research Council, Washington, D. C., August, 1960.

(2) Directory to Nuclear Data Tabulations, R. C. Gibbs and K. Way, 185 pp., 1958; Supplement, 1959 Nuclear Data Tables, pp. 1-38; both U. S. Government Printing Office, Washington, D. C.

(3) *Journal of Chemical and Engineering Data.*

(4) See *Physics Today*, **14**, No. 10, p. 70 (August, 1961).

(5) K. Way, N. B. Gove, and R. van Lieshout, "Waiting for Mr. Know-It-All (or Scientific Information Tools We Could Have Now)," *Physics Today*, **15**, No. 2, p. 22 (February, 1962).

Library Applications of Permutation Indexing

By R. A. KENNEDY
Bell Telephone Laboratories, Incorporated, Murray Hill, N. J.

Received May 24, 1962

Within recent years interest has been growing in the possibilities of delegating to computers not only functionally simple information storage and retrieval operations, but also certain complex intellectual tasks such as determining the content of a document and labelling it in some systematic maner for future recall. In addition to automatic indexing, machine preparation of abstracts also has aroused close attention. The technical feasibility of performing at least some part of these human tasks with the aid of data processing equipment has been demonstrated. However, certain conceptual and economic problems remain to be worked out before the library administrator can count on his cataloging section becoming a unit of a computation center.

While automatic indexing in any interpretative and analytical sense is therefore not yet a practical matter, a simpler mode of machine indexing is coming into wide use. This is the technique known variously as permutation indexing, permuted title indexing, subject-in-context indexing or, as H. P. Luhn calls it, keyword-in-context indexing (KWIC). Whatever the name for the process, current practice was primarily stimulated by the publication in 1958 and 1959 of reports by Ohlman, Hart and Citron of the System Development Corporation[1] and Luhn of IBM.[2] Differing in method and product, both of these approaches pointed up the practicality of compiling subject indexes automatically, or largely so, from key-punched titles or other parts of documents. To oversimplify, merely by mechanically shifting a significant word to a fixed filing position, surrounding the word with some of its original context, and listing the whole in an alphabetical array, a highly useful index could be got.

Since that time well over thirty applications of the fundamental technique appear to have been made. In a number of cases it has been taken up as a very satisfactory all-round way of combining mechanized listing and swift indexing in a current announcement package. While certain observers have fastened upon the obvious limitations of the approach and dismissed it out of hand, the method is being increasingly adopted either for single uses or regularly scheduled purposes. Among the uses made in the last several years are: *Chemical Titles*, the American Chemical Society's semi-monthly publication covering some 600 journals of pure and applied chemistry; *Bibliography of Chemical Reviews*, a selection of abstracts from *Chemical Abstracts; BASIC*—a title index published in each semi-monthly issue of *Biological Abstracts*; the *KWIC Index* to the *Science Abstracts of China*, issued December 1960 by the MIT Libraries and covering some 3300 Communist Chinese scientific papers; the *KWIC Index to Neurochemistry*, prepared in 1961 by the Mimosa Frenk Foundation for Applied Neurochemistry in cooperation with IBM and covering some 2100 papers; *Dissertations in Physics*, an indexed bibliography of the 8418 doctoral theses accepted by American universities from 1861 to 1959, compiled by the IBM San Jose Laboratory and published by Stanford University Press, 1961; and at the Bell Telephone Laboratories, a number of applications which are the particular concern of this paper.

It may be of interest to note some of the considerations entering into the decision, in 1959, to try permutation indexing. First, the Libraries had been directed to provide Laboratories personnel with more adequate means of access to internal reports than available through the existing non-library facilities. Secondly, we wanted to get a first-hand, working appreciation of the methods and problems of producing relatively simple indexes by punched card and computer processes. Permutation indexing was considered worth experiment under the circumstances because, among other reasons:

1. The report titles to be processed were well endowed with words of indexing and retrieval significance to their author-users. An average of 5 to 6 subject tags designating physical and chemical processes, elements, materials, systems, equipments and so on would be provided for each report indexed.

2. The additional essential search tools—author index, case or project number index, and a listing of reports by the issuing department—would be provided automatically, as adjuncts to the permuted title index.

3. User requirements for both a current awareness bulletin and a retrospective, multi-aspect search facility, could be handled by the one system.

4. An index in book form would be produced. Apart