which could be considered, is related to the documentation of physical and chemical properties of materials. Our principal interest is in the properties of chemicals under a wide variety of conditions.

The "International Critical Tables," our "old-standby" will not be revised. Instead, the Office of Critical Tables (of the National Academy of Sciences) will coordinate decentralized collections of critical data. For the user, the problem is first to become aware of all sources of potentially useful data, then to obtain publications which might be useful in a given field, and then to be able to use the various systems in which the data are presented in order to locate the value you need.

One group has decided to do something about this. It is the Thermophysical Properties Research Center at Purdue University. Picking up in time where the "International Critical Tables" left off, they intend to collect, evaluate and disseminate all recorded information on the thermophysical properties of all substances. Incidentally, in the research phase of their work they also intend to measure new properties and incorporate them into the literature. Since its beginning in January, 1957, the Center has proceeded to do that which some said couldn't be done. Their financial support has come from industry, which would indicate that some companies are willing to pay large sums for necessary information about properties of materials. More recently, Chemical Abstracts Service has begun plans for a physical properties data service.

The chemical industry has supported, directly or indirectly, several research projects which have generated a large amount of data. The *Journal of Chemical and Engineering Data* was developed by the ACS to publish in a single place papers that formerly went into their other journals. What we need now is better control and dissemination of properties scattered in many primary publications, possibly by beginning some kind of a specialized secondary publication. Perhaps with industry-wide support a physical properties serial publication could be established. It would issue simple tables of new data collected from primary journals, manufacturers' technical bulletins, research center compilations, *etc.*, organized in serial number order. A separate coordinate index would enable users to relate compound, conditions and property and turn to the correct table for the value. Maybe this wouldn't be such an impossible job?

CONCLUSION

Most of you will agree that there is a need for these services, and probably you could add several to the list. There is no lack of equipment or know-how. All of these services could be performed today accurately, promptly and in an acceptable format. The remaining questions are of economics and initiative. The ACS and other organizations have shown some initiative, yet there is still plenty of room for the private profit seeker. If the vacuum is left too long, however, the Federal Government may have to fill this gap. Perhaps someone else is willing to sell these services and do it now!

(1) Plans to publish *Chemical Patents* have been abandoned.
(2) This service has now been developed as the Chemical Specifications Microfilm File, an affiliate of the *Oil and Gas Journal*. The first file and indexes will be issued about August, 1962.

An Algorithm for Translating Chemical Names to Molecular Formulas*

By EUGENE GARFIELD
Institute for Scientific Information, 33 South Seventeenth St., Philadelphia 3, Penna.,
Received March 13, 1962

To calculate a molecular formula, a human or a machine computer must first be able to recognize the chemical name or the structural diagram on which the molecular formula calculation must be based. Prior to the publication of my book [1] on this subject, there has never been a serious consideration of the possibility of computing molecular formulas directly from chemical names. Chemists have always assumed that it is first necessary to draw a structural diagram before the molecular formula of a chemical can be calculated. Furthermore, the vagaries of chemical nomenclature have created the psychological climate that this step must be necessary. It has been axiomatic that in order to obtain the same molecular formula the chemist must work from the same structural diagram. Naturally, when you give it a second thought, you know this is not true. For example, if I say butane, the average chemist knows its formula to be C_4H_{10}. It is not necessary for him to draw the two dimensional ideograph $CH_3CH_2CH_2CH_3$ or the linear notation to arrive at the correct molecular formula. Once you accept the idea that the structural diagram is not necessary, then you can proceed to the question of how one "recognizes" a chemical name.

The chemist reads a chemical name and has a built-in mental dictionary that tells him certain combinations of letters have a particular referential meaning. For example, butane is a string of four carbon atoms. However, a computer is a far less sophisticated "reader" and must be instructed in a very precise fashion how to "recognize" the occurrence of meaningful strings of letters. However, as chemical names get more complicated, the chemist also has difficulty in identifying meaningful segments of chemical names. It is, therefore, important and very

* Presented before the American Chemical Society, Division of Chemical Literature, March 22, 1962, Washington, D. C.

useful to develop simple schemes for identifying these name segments. Unfortunately the rule books for such systematic nomenclature as I.U.P.A.C. and *Chemical Abstracts* are written not to help in the comprehension of chemical names, not to help recognize them, but as aids in generating names from diagrams. To simplify discussion we can speak of one scheme as a "recognition grammar" and the other as a "generation grammar." Modern structural linguists hope that the time is approaching when a grammar will consist of a series of algorithms, with the difference between the two being reduced to a matter of precision. An algorithm is a set of operations reduced to a uniform procedure. A grammar is a loose set of rules. Even in the domain of relatively precise chemical nomenclature, existing chemical grammars are so loose that a chemist may arrive at several different names for the same chemical.

How then do we go about "mechanically" recognizing chemical names? The naive young chemist might inquire, "Isn't it simply a matter of dictionary look-up? Why don't you simply put all existing chemical names on a magnetic tape along with the molecular formula?" These might be easy solutions, although the second may be quite expensive, if our problem were confined to previously reported chemicals. However, the problem of translating chemical names involves the large influx of new chemicals as well as the old, each of which must be separately calculated.

The less naive chemist might then ask, "Well, isn't it simply a question of adding up the values of the various syllables in the name, each of which is to be found in a dictionary of chemical syllables?" This would-be solution is what I call the syllabic approach. Syllables are useful for teaching spelling and preparing manuscripts, but they are almost useless for calculating molecular formulas. One need only cite the case of benzene and benzoic acid to illustrate quickly that the syllable "*benz*" does not always "mean" the same thing. And, more than this, neither does "*ene*" or "*oic*" acid. Each of these syllables is a homonymic expression and each conveys a different meaning depending upon the linguistic environment in which it is found. "*Oic*" acid means one thing in benzoic acid and another in pentanoic acid. It is true that their meanings are related but the acids are not the same. The syllabic approach for translating chemical names to molecular formulas is completely hopeless.

How then is it possible to recognize the meanings of words if we cannot work from syllables? In a natural language, the same problems are faced but they are resolved intuitively. When you try to recognize a word mechanically, you face the same problems. The linguist resolves this difficulty by breaking a word into meaningful units of language called morphemes. While there are many definitions of a morpheme, each is perfectly satisfactory for a particular grammatical theory. There are also different techniques for deriving a list of morphemes for a particular language. Each technique results in the compilation of a dictionary or an inventory of morphemes. This list may vary from dialect to dialect. For example, in I.U.P.A.C. nomenclature, ethanol consists of three morphemes, *eth, an* and *ol*. However, in *C.A.* nomenclature, ethanol must be treated as a single morpheme in order to properly distinguish the meaning of a prefix in such cases as *di* in diethyl and diethanol.

Before illustrating how the algorithm or recognition procedure works, let me summarize what it must be capable of doing. The algorithm must perform a syntactic analysis of the chemical name such that each morpheme in it is correctly identified both with respect to its referential meaning (calculation value) and its relationship to the other morphemes in the name. With chemical nomenclature, prior morphological analysis produces a dictionary, which not only gives meanings, but also syntactic rules for otherwise ambiguous expressions. The procedure must be able to distinguish between the meaning of *penta* in pentadiene, pentane, and pentachloropentane.

I think that it can be seen readily that the complexities of programming a machine for such a recognition procedure are much greater than programming a chemist or non-chemist. Machines have a long way to go in matching man's capabilities for learning. Those of you who are interested in the computer procedures for analyzing chemical names can refer to the examples given in my book.

The manual algorithm consists of eight basic steps. (1) Ignore all locants. Locants do not enter into the calculation of molecular formulas. They would be important for an algorithm which attempted to produce a structural diagram or a unique cipher. (2) Retain all parentheses. If you are dealing with *C.A.* nomenclature, you will have to add "parens" in cases like *di* in *di(ethanol)*. (3) Replace all morphemes by their calculational values. For example, *eth* equals two carbon atoms while *nitro* equals one nitrogen atom, two oxygen atoms and one double bond. (4) Resolve the ambiguity of any occurrences such as *penta* in pentadiene. Remember: (a) you cannot have two multipliers in a row unless separated by a paren; (b) if either of the next two morphemes (after an ambiguous morpheme) is alkyl ending, it is not a multiplier, as in pentadienoic acid; (c) if either of next two morphemes is not an alkyl ending it is a multiplier. (5) Place a plus sign after all morphemes except multipliers. (6) If there is a plus sign at the far right of a parenthesized term, place it outside right paren. If at far right of name, always drop it. (7) Carry out multiplications. (8) Calculate hydrogen value using the formula: $H = 2 + 2n_C + n_N - n_X - 2n_{DB}$, where n_{DB} is the number of double bonds.

Let's consider several examples of increasing complexity. In the chemical *methylaminoethane*, there are no parenthesized expressions, no locants, and no multiplier morphemes. The morphemic analysis is meth, yl, amin, o, eth, an, e. Each morpheme is assigned the following meanings which can be memorized quickly: Meth = C, yl = +, amin = N, o = +, eth = 2C, e = +. By simple addition of the equation C + N + 2C + you obtain the partial formula 3C + N. In conventional notation, this is C_3N. To calculate hydrogen (step 8): H = 2 + 2(3) + 1 − 0 − 2 (0) = 9. The complete formula is C_3H_9N.

As a second example let us consider the chemical, N-[3-(diethylamino)propyl]-N-ethyl-2-amino-1,4-butanedioic acid. By a similar morphemic analysis it becomes:

$$(0 - [2(2C) + N] + 3C) + 2C + 0 + N$$
$$+ 0 + 4C + 0 + 2(2\phi + DB) \qquad \phi = \text{oxygen}$$

$$(7C + N) + 6C + N + 4\phi + 2DB = 13C + 2N$$
$$+4\phi + 2DB = C_{13}N_2O_4 + 2DB$$

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

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

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

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

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.

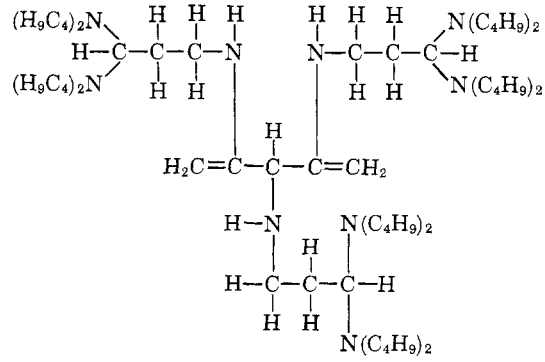

Fig. 1.

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

$$N + 19C + \phi + 4DB$$
$$H = 2 + 2(19) + 1 - 2(4) = 33$$

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.