Essays of an Information Scientist, Vol:1, p.386-388, 1962-73 Current Contents, #50, December 13, 1972

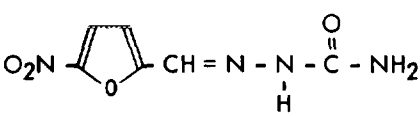
"Current Comments"

Are You Ready for Chemical Linguistics? Chemical Semantics? Chemical Semiotics? Or, Why WLN?

December 13, 1972

In a recent editorial,[1] I discussed the Wiswesser Line Notation (WLN) as one of many chemical "languages". We chose WLN for ISI®'s *Chemical Substructure Index*™ [2] because WLN had proved to be the most economical of available systems in terms of human input (encoding) and of computerized output for printed indexes and computer manipulation. Furthermore, WLN offered the most flexibility to meet unforseen future requirements.

Although the use of WLN is of fairly recent vintage, the principle of "line" notation is an old and familiar one. The molecular formula $C_6H_6N_4O_4$, is in fact ambiguous and may represent several different isomers or compounds. If this formula is expanded, a less ambiguous line or linear notation is obtained: $O_2NC_4H_2OCH{=}NNHCONH_2$. The syntax of this chemical sentence is further clarified by inserting periods, as in $O_2N.C_4H_2O.CH{=}N.NH.CO.NH_2$. To make the compound more immediately recognizable, one can go to the structural formula or diagram:



The presumed value of the structural diagram is that it expresses the topological nature of the molecule in a way that is immediately intelligible to most chemists. Certainly it is not because it is appreciably more "accurate", but it is a bit more realistic than $C_6H_6N_4O_4$.

Like Chinese characters, structural diagrams are ideograms. The spoken sounds for different Chinese dialects used to express the concept represented by a particular ideogram or combination of ideograms are frequently totally dissimilar. The dialect speakers may not understand one another, but they can all "read" the same ideogram. The same is true of the chemical structural diagram. An English chemist may not understand the spoken words used by a Japanese chemist, but they will both be able to "read" the structural diagram.

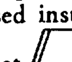
Like the Chinese ideogram, the structural diagram also has an *intra*national utility. One English chemist may "speak" pure Geneva nomenclature. Others may choose to use the slang of chemistry, that is, "generic" and trivial names. In other words, one scientist may speak a chemical Mandarin and vocalize the ideogram shown above as "5-nitro-2-furfural semicarbazone", while another makes do with the slang of "Furacin".

The disadvantage of the structural diagram and Chinese ideogram alike is that they can only "suggest" phonemes associated with the symbol. They cannot *record* in perpetuity any particular phoneme or morpheme, much less any absolutely accurate representation of all that a three-dimensional, indeed multidimensional concept may imply. Two-dimensional diagrams have another disadvantage, as far as Western cultures are concerned.[3] The structural diagram cannot be "drawn" on Gutenberg's printed line using the conventions of its particular ideographic style. Certainly the *standard* alphanumeric keyboard can't handle anything except acyclic or relatively simple cyclic compounds. (Many chemical typewriters have been invented.)

Line notations like WLN can be dealt with on the standard typewriter or the standard computer printer. For purposes of listing, these linear structural diagrams can be "alphabetized" and permuted. As a by-product, chemical Mandarins and chemical hipsters may communicate a little easier. The use of WLN is preferred to the structural formula for indexing because it can be processed like any other piece of linear alphanumeric data. It is also superior to a "structural diagram" because the canonical notation for any particular compound is unique and invariable, if the rules are followed. There can be only one notation for a given compound. It can be "drawn" or written in only one way. The same is not true of a structural diagram. The structural diagram may be unique, but it is not invariable. Like the Chinese ideogram, it is more or less subject to a calligraphy permitting stress on certain elements for visual or merely esthetic emphasis. The "best" way of drawing structures like strychnine or adamantane has not yet been decided.

Despite its advantages for representing compound structures, WLN suffers from a lack of familiarity. Chemists, like most human beings, resist change. In spite of its simplicity, the newness of WLN is excuse enough to prefer and to tolerate the demonstrable disadvantages of what may be inadequate but nevertheless comfortingly familiar. As with other new ideas, citation indexing for example, it is a necessary task to educate in its use and also to make the transition as simple as possible.

Any chemist can learn to "read" or to decode WLN expertly in a few hours. It would certainly take him longer to learn to write it correctly enough to produce canonical notations. Since many chemists simply won't spend the time necessary to learn to write WLN, we have developed some dictionaries that can be used instead.

Thus if you insist that  is "furan", our *Dictionary of Frequently Found Substructures*[4] will tell you that in WLN a furan is T50J. The user already knows, but didn't think to ask for it as a heterocyclic (T), five-atomed ring (5), with one oxygen atom (0). (The final J merely indicates a stop.) The dictionary will also tell him that phenothiazine is T C666 BM ISJ. He really knew all along that it is heterocyclic (T), contains a linearly arranged series of consecutive six-atomed rings (C666), with an NH at the b-position (BM), and a sulfur at the i-position (IS) (the J indicates end of the ring system,) and you have T C666 BM ISJ.

It's not my purpose here to teach readers WLN, but to show that it is merely another method of writing or drawing the structural formulas. WLN is highly translatable, because it represents structures rather than names.

This translatability of WLN has already produced *CHEMTRAN,* ISI's computer system for conversion of linear notations into atom-by-atom connectivity tables, and from there into the fragment codes used by the Dokumentationsring system.[5] Since there are many chemical "dialects", the capability of translating WLN into other dialects will be necessary. Some of these other dialects may prove to be "artificial languages."

The easy machine-processing of linear notations is already having its effect on chemical nomenclature. ISI is looking into the generation of "systematic" names from line notations. It is not surprising that CA is now making major changes in its system of nomenclature. Heavily influenced by the interest shown in systems for searching by chemical substructure, CA has adopted WLN for its dictionaries. Nevertheless, try as they may to standardize or systematize nomenclature, the alphabetic arrangement will not put in the same place two chemicals whose structural similarities are evident to the student of elementary organic chemistry--such compounds must be separated as long as they are called respectively thiazine and phenothiazine, or testosterone and estradiol.

Contrary to Orwell's warnings regarding "Newspeak" and any other controlled language, the chaos of completely "natural" chemical nomenclature would remain a hindrance to the development of chemical communication if we did not simultaneously develop new translation languages like WLN. Should the use of WLN ever hinder the process of chemical thinking, it will have outlived its usefulness. I cannot foresee that time for at least a few decades.

1. Garfield, E. ISI's *CHEMTRAN* "compatibilizes" files of encoded chemical structures.
Current Contents No. 46, p. 5-6, November 15, 1972.

2. ISI's *Chemical Substructure Index (CSI* ™ *)* to new compounds and new syntheses reported in the current chemical literature. Entries in the *CSI* refer the user to abstracts in ISI's *Current Abstracts of Chemistry* ™ . The *CSI* uses the Wiswesser Line Notation to encode chemical structures, then permutes (rotates), alphabetizes, and lists the notations to produce the printed index. The WLN is, thus, the language of the *CSI.*

3. In this connection, the producers of the *Ring Index* might have profited from consultation with Chinese lexicographers.

4. The *Dictionary of Frequently Found Substructures* is reproduced in the front of each monthly and annual issue of the *CSI.* A much more complete *Dictionary,* with four or five thousand entries, is now in preparation.

5. Steidle, W. Possibilities of mechanical documentation in organic chemistry. *Pharm. Ind.* 19:88-93, 1957.