Pure & Appl. Chem., Vol. 49, pp.1803 - 1805. Pergamon Press, 1977. Printed in Great Britain.

THE INDEX CHEMICUS REGISTRY SYSTEM® - PAST, PRESENT AND FUTURE

Eugene Garfield
Institute for Scientific Information 325 Chestnut St., Philadelphia, Pa. 19106 USA

Michael Sim
Institute for Scientific Information 132 High Street, Uxbridge, Middlesex, U.K.

Abstract - ISI's chemical database -- the Index Chemicus Registry System --(ICRS) provides information on new organic compounds, new reactions, and new syntheses. ICRS, geared to the synthetic organic chemist, has covered over 2 million compounds since 1960, and has flexibility of access--computer tapes, microforms, printed copy --dependent on the user's needs. Structural diagrams in the form of machine-readable Wiswesser Line Notations (WLN) are an essential part of the database, allowing practical and economic Substructure Searching. This is augmented by fragment coding using the pharmaceutical "Ring Code". Future developments include structural diagram display, a Chemistry Citation Index, and conversion from WLN CT's to other fragment codes, or connectivity tables.

In the brief time available I shall try to give you a short overview of how our database came about, how it is presently used, and our plans for the future.

We are primarily setting out to serve the synthetic organic chemist, and it is both his curse and his blessing that he tends to think and communicate in terms of structural diagrams. It is his blessing because the structural diagram provides an accurate topological description of a molecule which is entirely independent of language; in fact it constitutes an international language of its own. It was with the intention of talking to the organic chemist in his own language that we started our chemical database in 1960 with a graphic abstract publication, which we now call Current Abstracts of Chemistry™ (CAC™). (1)

Its most distinctive feature is the liberal use of structural diagrams and reaction flow sheets, for the purpose of aiding visual scanning. The two other main characteristics of this publication (and they are linked) are:- (1) that it produces a manageable volume of material each week, by restricting itself to synthetic organic chemistry, and furthermore selecting only those articles which report new compounds, new reactions or new syntheses; it is compact enough to be browseable on a weekly basis; (2) that it is fast: this is helped by the fact that it is compact, and indeed we have maintained a median time-lag of only 6 weeks between the appearance of an article in a journal and its appearance in CAC. I think it is safe to say that it has become a standard tool in most chemical libraries.

The chemist's curse arises when he tries to express a structural diagram in linear form, either for verbal communication or for the creation of indexes and computer-based retrieval systems. I will not attempt to discuss the vagaries of systematic nomenclature or the merits of the various other methods of representing structure, since Messrs. Hyde and Grunewald will be dealing with this tomorrow. (However, as a small digression, I may say that in a recent newsletter (2), I read an exhortation to chemists that they should, quote: "actively encourage the death of nomenclature as the technique for retrieving published information on chemical structures".)

Suffice it to say that in 1968, when we decided to make our database available in machine-readable form, with the prime intention of giving the user a practical, economical method of searching for chemical structures and substructures, we chose Wiswesser Line Notation (WLN) (3,4,5). We felt that WLN was the most successful, widely-used method available at that time, and one which was capable of further development to cater for any of the forseeable demands which chemists might make. Eight years later we see nothing to contradict that view; if anything, it has been strengthened by events. This magnetic tape version of the database is known as the Index Chemicus Registry System® (ICRS®) (6); it is presently available from 1966 to date, and we are in the process of rendering the years 1960-65 into machine-readable form too. The years 1960-76 contain 2.25 million compounds, and already in the years 1966-76, we have encoded into WLN ca. 1.7 million compounds; the file is currently growing at a rate of 150,000 new compounds per year. However, the most significant aspect of this database for the user is not the large numbers, important though they are, but the fact that it is practically and economically searchable right now in response to substructure questions,

either for current-awareness purposes, or for retrospective search. As many organisations know to their cost, the mere existence of a system with a large compound file in machine-readable form does not automatically mean that economic substructure searches are possible or available.

We provide free to ICRS subscribers a series of computer programs called RADIICAL™ , which search both WLN and bibliographic records, but the database has also been successfully run by companies using their own in-house software, for example by ICI using their CROSSBOW programs (7,8,9). For those users who do not have the means to search the database on a computer, but who nevertheless wish to ask substructure questions, we provide a KWIC index to WLN portion of the file, called the Chemical Substructure Index® (10), and a computerised current-awareness service, called Automatic New Structure Alert® (ANSA®), run in Philadelphia.

This flexibility of access is important, and it has meant that a wide variety of organisations, from national institutions and multi-national companies down to comparatively small institutes, have been able to make use of the database in whatever form is best suited to their circumstances.

To illustrate what we mean by economic searching, allow me to quote some figures given in a recent paper by Mrs. Diane Eakin of ICI (11). Because they use the ICRS tapes both for current-awareness and retrospective search, they allocate half the subscription cost to each activity. At the moment, they run about 125 profiles against each monthly ICRS tape, 85% involving substructure questions, 10% bibliographic, and 5% involving both. The total cost per month, including all computer and staff costs (plus half the subscription cost), varies from $5-9 per profile. An average substructure search with all structures displayed costs $7, or $84 per year. To run a retrospective search on, for example, the period January 1973 to June 1976, a file of about 600,000 compounds, costs $45. ICI are currently running about 25 such retrospective searches per month, presently going back to 1969, and probably even further shortly.

The database has not, however, stood still since 1968. One of the first suggestions for improvement came from the Pharmadokumentationsring group of companies, who requested us to code the compounds in ICRS into the so-called Ring Code (12,13), as well as WLN. To accomplish this, while avoiding the expense of manually coding each compound twice, we have written computer programs to automatically convert the WLN for each compound into Ring Code (14), and the entire file is now available in both codes. (The RADIICAL programs have, of course, also been modified to search the new Ring Code record.) This brings with it two immediate benefits, and one long term.

The first benefit is that WLN and Ring Code are in many ways complementary, in the sense that a substructure question that is difficult to answer using one system is often well-suited to the other. Depending on the nature of the question, the user may choose to express the substructure in WLN, or Ring Code, or both; the latter often produces a synergistic effect on search time, since one code effectively acts as a screen for the other (15).

The second benefit is that a company which, for historical reasons, has always used Ring Code for its internal file, but now sees the advantages of adopting WLN, may purchase the conversion programs for internal use; this has, in fact, been done by the Ring group as a whole. In this way, they can "have their cake and eat it".

The long term benefit to ICRS and its users arises from the fact that, in order to derive Ring Code from WLN, our first task was to create a special connectivity table from each WLN, the corresponding Ring Codes were then derived from this connectivity table. At the moment, the connectivity table information is not retained in the machine once it has served its purpose, but we now have the potential to do several interesting things. In the first place, as described in Granito's 1973 paper entitled "CHEMTRAN and the Interconversion of Chemical Substructure Systems", (16), we can undertake to write programs which will convert from WLN, via our own connectivity table, to other fragment codes, other notations, or other connectivity tables, as dictated by the requirements of a particular user. Secondly, once users begin to demand more sophisticated techniques such as atom-by-atom searches on connectivity records, structure display of retrieved compounds, or advanced structure-activity relationship studies, we shall be in a position to provide them. Furthermore, such future extensions will share the pragmatic, economic characteristics of the present system, which are due in large measure to the fact that it is based on WLN (rather than topological) input and storage (17).

To sum up, we have developed the ICRS database according to what we perceived to be the organic chemist's most pressing needs. The methods we adopted have been heavily based on the pioneering work which has been done in the chemical and pharmaceutical industries, and they, in turn, have been among our most enthusiastic supporters. I am confident that this fruitful cooperation will continue, to the ultimate benefit of the chemical community as a whole.

As a postscript to the discussion of ISI's chemical information system, it is important to note that ISI also produces the Science Citation Index® (SCI®). At a minimum, at least one third of the annual source coverage of 450,000 articles in SCI is chemical. Probably an even higher percentage of the material in the citation index section of SCI is related to chemistry in all of its applications. Consider that several chemical journals, like the Journal of American Chemical Society (JACS) and Journal of Chemical Society (JCS) are amongst the most-cited science journals published (18).

SCI should be regarded not only as a primary source of chemical information, but also as a valuable adjunct to Chemical Abstracts and Current Abstracts of Chemistry. Indeed, it has been assumed in the design of CAC that once the primordial reference for a new compound or synthesis has been located, subsequent papers on biological testing, clinical use, or whatever could be located easily through SCI. This has been confirmed by over 15 years' experience with a file that now extends from 1961 to the present.

Perhaps it is an indictment of the insularity of the chemical profession that so few chemical libraries as such possess the SCI. The discipline oriented approach of the chemist, however, must be reckoned with. For this reason, we are now developing a Chemistry Citation Index, initially covering the period 1961-76. Such a large scale cumulation would represent an important complement to Beilstein, Chemical Abstracts and Current Abstracts of Chemistry and all other bibliographic tools used by the chemical profession.

REFERENCES

1. G. Revesz and A. Warner, J. Chem. Doc. 9, 106-109 (1969).
2. E. Ward, CNA (UK) Newsletter, Issue No. 5, 1-7 October (1976).
3. N. H. Wiswesser, A Line-Formula Chemical Notation, Thomas Y. Crowell, New York (1954).
4. E. G. Smith, The Wiswesser Line-Formula Chemical Notation, McGraw-Hill, New York (1968).
5. E. G. Smith and P. A. Baker, The Wiswesser Line-Formula Chemical Notation (WLN) Third Edition, Chemical Information Management, New Jersey (1976).
6. E. Garfield et al., J. Chem. Doc. 10, 54-58 (1970).
7. L. H. Thomson et al., J. Chem. Doc. 7, 204-209 (1967).
8. E. Hyde and L. H. Thomson, J. Chem. Doc. 8, 138-146 (1968).
9. J. E. Ash and E. Hyde eds., Chemical Information Systems, p. 73-85, Ellis Horwood, Chichester (1975).
10. C. E. Granito and M. D. Rosenberg, J. Chem. Doc. 11, 251-256 (1971).
11. D. R. Eakin, Chemical Structural Information - ICRS as Part of a Larger System vs. CROSSBOW Technology. Paper presented at ICRS Users' Meeting, Philadelphia, June (1976).
12. W. Steidle, Pharm. Ind. 19, 88-93 (1957).
13. W. Nubling and W. Steidle, Angew. Chem. Internat. Edit. 9, 596-598 (1970).
14. C. E. Granito et al., J. Chem. Doc. 12, 190-196 (1972).
15. H. Deforeit, Paper presented at ICRS Users' Meeting, p. 1-30, London, October (1976).
16. C. E. Granito, J. Chem. Doc. 13, 72-74 (1973).
17. C. E. Granito and E. Garfield, Naturwissenschaften 60, 189-197 (1973).
18. E. Garfield, Nature 264, 609-615 (1976).