CONCLUSION

There are some substantial cooperative efforts involved in the processing of bibliographic information into NTIS. These cooperative efforts are not without serious problems and further cooperative efforts will be required to provide solutions. The particularly difficult problem of vocabulary control and retrieval in the NTIS multiple thesauri system will require further analysis and experimentation.

LITERATURE CITED

(1) Committee on Scientific and Technical Information, "Guidelines to Format Standards for Scientific and Technical Reports Prepared By or For the Federal Government," 1968, **PB 180 600.**
(2) Heald, J. H., "The Making of TEST, Final Report of Project LEX," Office of Naval Research, November 1967, **AD 661 001.**

CHEMTRAN and the Interconversion of Chemical Substructure Systems*

CHARLES E. GRANITO**
Institute for Scientific Information, 325 Chestnut St., Philadelphia, Pa. 19106

Received January 18, 1973

The need for the interconversion of chemical substructure systems is discussed and CHEMTRAN, a new service, designed especially for creating interconversion programs, is introduced.

The organic chemist studies the three-dimensional world of chemical compounds but, as with researchers in other fields, he communicates his results in one or at best, two-dimensional media. The principle of least effort[1] guides him in most of his efforts; organic chemistry is no exception. The organic chemist has adopted the structural diagram as the means of employing "least effort" in communicating his findings to students and colleagues.

NOMENCLATURE

Unfortunately, chemists have for too long been trying to use nomenclature to serve two basically divergent needs:

1. The need for a short descriptor to replace the entire chemical structure in speech and in writing
2. The need for systematic indexing names that describe the molecular makeup of a compound

Thus, we find, for example, both adamantane and tricyclo (3,3,1,$1^{3,7}$)-decane representing the same structure (Figure 1).

The problem has traditionally been that chemists usually prefer the shorter, less descriptive names where they have a choice. As the result of continual compromise, we find many "trivial" names adopted as standard, "systematic" names. For over a hundred years there has been a mixing of trivial and descriptive names. Not surprisingly, the situation has become chaotic. Organic compounds reported in the literature are not really indexed by any "systematic" nomenclature, but rather, by a nomenclatural system made up of many descriptive and many nondescriptive words. Compromise has, in this case, often reduced the system to a point of very limited utility. Furthermore, the resulting "nomenclature" has led to a dependence on whole molecules in using subject indexes, despite the fact that research chemists are usually interested in classes of compounds—i.e., compounds which have some common structural feature. Nevertheless, there is no substructure index to the 4 million compounds reportedly recorded within *Chemical Abstracts!*

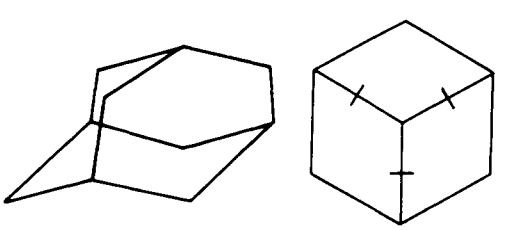
SUBSTRUCTURE SYSTEMS

The failure of nomenclature for substructure searching has been recognized for some time but has not diminished the many organic chemists' desire to do substructure searching. Instead, it has led them to develop other structure representations.

Over the past 30 years, a considerable effort has gone into the development of substructure search systems. There are today, hundreds of such systems in use around the world. These systems may be generally classified as: fragment codes, line notations, and connectivity tables.

The most widely used fragment code is the Ring Code.[2] The most widely used line notation is the Wiswesser Line Notation (WLN).[3] Connectivity tables (CTs) have also received attention, but are not presently used by many organizations. However, the decision by *Chemical Abstracts* to use a connectivity table in its registry system[4] has led to increased interest in this technique in the past several years.

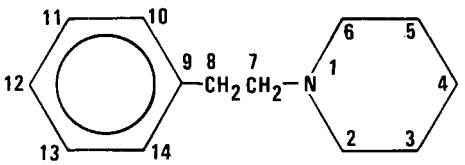
Figure 2 shows an example from each major system. We will not attempt to contrast the various systems in this paper.



ADAMANTANE TRICYCLO (3, 3, 1, $1^{3, 7}$) DECANE

Figure 1. Adamantane

*Presented before the Division of Chemical Literature, 164th Meeting, ACS, New York, N. Y., Aug. 30, 1972.
**Present address: C G Associates, 1948 Cardinal Lake Drive, Cherry Hill, N.J. 08003



WLN: T6NTJ A2R

Ring Code

2/12, 2/11
3/12, 3/2
6/12, 6/Φ
7/1
10/11, 10/Φ, 10/2
11/11, 11/6
12/3

CONNECTIVITY TABLE

Atom. No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Conn.		1	2	3	4	5	1	7	8	9	10	11	12	13
Element	N	C	C	C	C	C	C	C	C	C	C	C	C	C
Bond		*1	*1	*1	*1	*1	-1	-1	-1	*5	*5	*5	*5	*5
H-Count														

Ring Closure Pairs 001*1006 009*5014

Figure 2. N-Phenethylpiperidine coded in three systems: Wiswesser Line Notation, Ring Code, and Connectivity Table

INTERCONVERSIONS

Investigators at several organizations—e.g., Imperial Chemical Industries and NIH—have studied the interconversion of WLN and the CT used by CA. The Institute for Scientific Information has developed programs for converting WLN to the Ring Code fragmentation code.[5] The Internationale Dokumentionsgesellschaft für Chemie (IDC) has been developing programs for generating the GREMAS[6] (**G**enealogisches **R**echerchieren durch **Ma**gnet-band **S**peicherung) code from a topological (connectivity table) record.

There are, of course, many interested bystanders who cannot afford to work on interconversion programs but who, nevertheless, are very concerned with the results of such studies and what they mean to "their" system.

THE NEED FOR INTERCONVERSION OF STRUCTURE REPRESENTATIONS

There is no one "best" system! Depending on individual needs and resources, any one of the hundreds of systems that have been developed for representing structures may be "best" for a particular organization. "Best" depends, in part, on the specific application being considered.

The need for interconversion arises for many reasons. First, there is the question of *compatibility*. There are many private internal (single organization) files, and many external (available to the public) files. A given company may find it economical to continue using a fragment code it developed for internal use but at the same time, it may wish to compare its file against some external file—e.g., the ISI or CA registry systems or some supplier's file.

Second, there is the consideration of *costs*. There may be considerable cost attached to continuing a system by methods developed when a file was small, and/or growing at a slower rate. Retrieval using a given system may still be adequate but input costs may be getting out of hand. If this is the case, it would be wise to find a cheaper form of input that could be economically converted, via computer, to the older system.

Third, there is user *sophistication*. A system which may have been adequate for many years may now be moving towards obsolescence because the users are starting to ask more sophisticated questions (based, perhaps, on answers obtained using a more "primitive" system).

Fourth, there is the question of *flexibility*. A connectivity table user may wish to print out line notations as the end result of an analysis of some CT file. Or the WLN user may wish to use a fragment code for some statistical studies.

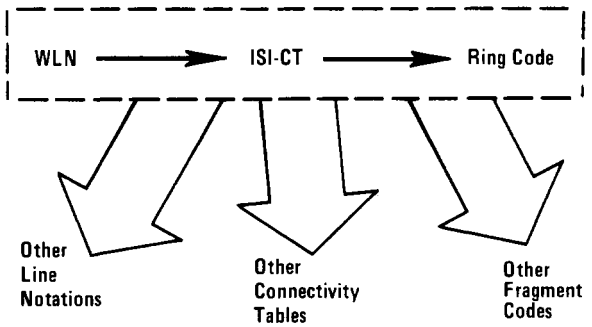


Figure 3. Interconversion of substructure systems

Today, we find *Chemical Abstracts* with over two million compounds coded in a connectivity table system and ISI with close to a million compounds coded in WLN. The U.S. Patent Office has large files coded in the Hayward notation; the IDC has large numbers of compounds in its CT and GREMAS Code. Derwent has a sizable patent file coded in one fragment code, and many journal literature compounds coded in the Ring Code fragment code. There are a number of individual companies and government agencies with over 100,000 compounds coded in "a" system. And almost all companies synthesizing new compounds have some internal system for their compounds. Finally, there are many universities with a wide variety of coded structure files.

There is, of course, overlap in the compounds covered by various systems but no one file contains all of the data elements. For example, *Chemical Abstracts* may have registered a particular compound and referenced articles which mention it, but the same compound may also have appeared in a patent not covered by CA, and be in 10 different organizations' files as well—probably in each case with the results from different tests.

If there is access to multiple files, it might be argued that it doesn't matter that each is coded in a different system. For example, a company can maintain a fragment code for internal purposes and use WLN when addressing the ISI file. However, there are legitimate reasons for wanting to compare whole files. Yet, logically, users do not wish to learn the rules for too many different systems. As a result, there is a definite need for programs that will allow automatic interconversion of systems. Unfortunately, some codes—e.g., the major fragment codes, cannot be converted to a connectivity table or line notation because all structural information is not contained in the fragment code. In such cases, only a one-way conversion can be effected but even this is useful.

COOPERATION

In order for there to be cooperation, two (or more) parties must reach a point where each believes that some benefit is to be derived as a result. In the case of chemical structure systems this point apparently has not been reached as there has been little (if any) evidence of inter-system cooperation to date. There has been considerable cooperation between users of a single system. And, as noted earlier, individual organizations have explored the development of interconversion programs. However, such studies have been rare and unilateral—i.e., one organization has done the work.

Cooperation in the form of making details of a system available does already exist, and this is an important and necessary first step for interconversion. However, necessary cooperation in effecting the interconversion of systems is still missing.

ECONOMICS AND KNOWHOW

To generate the necessary algorithms and computer programs for converting from one system to another, one needs both money and knowhow. Few organizations have both in the area of chemical structure representation.

The Institute for Scientific Information is one organization that has the necessary knowhow for creating such programs and has already undertaken the conversion of WLN to the Ring Code[5] fragment scheme. This was done for an economic reason, namely, to sell more subscriptions to the *Index Chemicus Registry System* (ICRS).[7] However, in developing these programs, another project became feasible—i.e., conversion of WLN to still other systems.

ISI believes that the Wiswesser Line Notation is the most economical form of structural input for a chemical retrieval system and that WLN has many additional advantages as well, which make it the system of choice for many organizations. However, because it is inexpensive, unique, and unambiguous it also serves as an ideal starting place for developing programs for the automatic generation of other codes used by individual organizations. In fact, ISI plans to introduce in 1973, a new service called CHEMTRAN for just this purpose. CHEMTRAN is a practical way of bringing money and knowhow together to improve both communication and cooperation in the chemical structure handling field.

CHEMTRAN

Simply stated, CHEMTRAN is a service that will develop programs to be used in converting (or translating) from one structure representation to another. At present, programs have been developed for converting WLNs to a unique connectivity table. Programs for converting from WLN to Ring Code fragmentation code are nearly complete. The programs developed for these projects, as well as the intermediate records that are generated, can successfully serve as the foundation for all subsequent interconversion programs. Consequently, an organization does not have to reinvest the effort ISI has already expended in reaching this level.

For example, any organization wishing to have its fragment code computer generated (with all the quality control this permits) could switch to WLN, take advantage of the great flexibility WLN offers, and still maintain the fragment code its users have become accustomed to for searching. Or it could use WLN as a less expensive form of input to its present system. In either case, the company investment made in programs developed for processing and searching a particular fragment code can be preserved. Furthermore, users can move towards what we feel is the ideal system—one which includes a fragment code, WLN, and a connectivity table. One CT is already deliverable under CHEMTRAN. Others can be. This planned new service is announced in the belief that it will expedite the development of programs for the interconversion of chemical structure systems.

LITERATURE CITED

(1) Zipf, G. K., "Human Behavior and the Principle of Least Effort," Hafner, New York, 1949.
(2) Steidle, W., "Possibilities of Mechanical Documentation in Organic Chemistry," *Pharm. Ind.* **19**, 88-93 (1957).
(3) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw Hill, New York, 1968.
(4) Morgan, H. L., "The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service," *J. Chem. Doc.* **5**, 107-13 (1965).
(5) Granito, C. E., Roberts, S., and Gibson, G. W., "Wiswesser Line Notations to Ring Codes. Part I," *J. Chem. Doc.* **12**, 190-6 (1972).
(6) Fugman, R., Braun, W., and Vaupel, W., "GREMAS—A New Method of Classification and Documentation in Organic Chemistry," *Nachrichten fur Dokumentation* **14**, 179-90 (1963).
(7) Garfield, E., Revesz, G. S., Granito, C. E., Dorr, H. A., Calderon, M. M., and Warner, A., "Index Chemicus Registry System: Pragmatic Approach to Substructure Chemical Retrieval" *J. Chem. Doc.* **10**, 54-8 (1970).

Some Chemical Notation Cooperative Activities*

WILLIAM J. WISWESSER,** CHARLES L. CRUM, KURT J. WINDLINX and RICHARD A. CREAGER
Fort Detrick, Frederick, Md. 21701

Received January 18, 1973

Cooperative efforts between Fort Detrick and various organizations over the past eight years and the consequences of these efforts are discussed.

Fort Detrick's first cooperative activity in chemical information management—with the J. T. Baker Chemical Co. in 1964—was one of the most stimulating and productive of a continuing series of such ventures. J. T. Baker's representative, C. T. Kleppinger, had asked Raymond R. Myers, then at Lehigh University, for suggestions having innovative value in their corporate plan to introduce a large new line of organic laboratory chemicals. Myers recommended management of the chemical structure information with a chemical notation, in cooperation with Wiswesser at Fort Detrick. A. J. Barnard, Jr., in charge of chemical information management at J. T. Baker, approved the plan and a punched-card deck was started with the mutual agreement that the information also would become experimental material for eventual development of a Fort Detrick chemical-biological data base. All agreed that this was a premium opportunity to

* Presented before the Division of Chemical Literature, 164th Meeting, ACS, New York, N. Y., Aug. 30, 1972.
** To whom correspondence should be addressed.