A Workshop in Chemical Substructure Information Handling Systems[a]

ALAN GELBERG[b] and ISAAC D. WELT[c]
Center for Technology and Administration, American University, Washington, D.C.

Received January 4, 1973

The Workshop in Chemical Substructure Information Handling Systems was offered as an experimental course at American University during the Fall semester of 1969. This was a graduate level, 3-credit course that met weekly for 2½ hours, for 14 weeks. Topics discussed were: conventional systems (chemical nomenclature, molecular formula indexes, and ring systems); nonconventional systems (fragment codes, chemical notations, topological codes, and connection tables); and automated literature services.

Although the chemical information specialist has been recognized as a bona fide member of the chemical profession for several decades, his education has been sadly neglected by the academic community. In most instances, chemists have simply left the bench for the library or information center without the benefit of special training. Over the years, they have amassed the necessary know-how or expertise which would enable them to function efficiently in their milieu. This "on-the-job" training has produced a number of highly-qualified individuals who manage the many information centers within the chemical industry and government. However, it is not the most efficient way of developing a new discipline or profession. Meetings, such as those of the Division of Chemical Literature have been of great importance in this regard, but they cannot replace the classroom entirely.

TRAINING PROGRAM FOR CHEMICAL INFORMATION SPECIALISTS

During the past five years, the American University has been successful in developing a graduate program for the training of chemical information specialists, with particular emphasis upon the needs of the Washington, D.C., metropolitan area for such individuals. The program evolved as a result of the close collaboration between the University's Department of Chemistry (ACS-accredited) and the Center for Technology and Administration.

The Center (CTA), a unit within the College of Continuing Education, is concerned with the training of individuals in the broadly-based areas of Technology of Management, which may be defined as follows:

> "It evolved from the acceptance of the electronic digital computer with its associated communications capabilities, as well as from the emergence of a broad and substantial quantitative theory of the properties, structure, and flow of information. It approaches problems in all fields involving information transfer, transformation, and storage. The discipline should be included in the education of those who will be involved in the management, or design, or development of large, complex systems involving people, resources, and objectives."

Three of the Center's areas of concentration are of interest to chemists: the Management Science option, the Computer Systems, and the Scientific and Technical Information Systems. The latter provides the information storage and retrieval component in the education of literature and information chemists.

Although courses in information handling and in chemistry are available to all graduates with a Bachelors degree in chemistry on a nondegree basis, registration for a Masters degree in chemistry (information science) is preferred. A minimum of 33 hours of graduate level courses are required, of which roughly half are in the field of chemistry while the remainder deal with information handling.

It should be emphasized that the graduate level courses in chemistry are designed to round-out the students' backgrounds in chemistry and are taken in the company of laboratory-oriented graduate students. Laboratory work, however, is not required. The information courses are those offered to all other CTA students (except for the Workshop in Technical Information: Chemical Substructure Handling Systems).

In this program, there is, first of all, an introductory survey course which covers the information science field quite broadly. This is followed by a choice of two further courses from among: (a) a detailed course in "Abstracting and Indexing" during which the students abstract and index sample documents from their own field of interest; (b) "Technical Information Machine Systems" stressing the role of the computer in information storage and retrieval and the design of systems for the handling of scholarly information; (c) "Natural Language Data Processing" which represents an introduction to the field of computational linguistics and designed for those interested in automatic abstracting and indexing and mechanical translation; and (d) "Publication Techniques" an introduction to computer-aided publication and involving the economics of journal publication (Joseph H. Kuney, formerly Director of Business Operations, ACS Books & Journals Division, was the instructor for this course); and (e) Workshops on new developments such as the "Chemical Substructure Handling Systems." Students must then enroll in a Seminar and present a detailed study before the class. After adequate "feedback" the report is written up as for publication and represents the M.S. "thesis." A Comprehensive examination must be successfully completed in each of the two fields, that is, in chemistry, as well as in information science before the Masters degree can be conferred. It must be emphasized that this is a bona fide degree in chemistry.

[a] Presented in part in the Joint Symposium on "The University and Chemical Information," ACS Division of Chemical Education and Chemical Literature, 160th Meeting, ACS, Chicago, Ill., Sept. 14, 1970.

[b] To whom inquiries should be addressed concerning the Workshop at BD-240, Food & Drug Administration, 5600 Fishers Lane, Rockville, Md. 20852.

[c] To whom inquiries should be addressed concerning the academic program at American University.

Since the beginning of this program, some five students have been awarded this Masters degree. They were all part-time students, taking the course offerings during the weekday evenings and on Saturday morning, and all were employed by either the Federal government or by the information industry in the area. In this respect, they constituted a cross-section of the several thousand CTA students from other disciplines ranging from mathematics to business administration.

Within the last year, an experimental Ph.D. program in chemical information has been established. In addition to the above mentioned courses, students will be required to take additional courses in the computer area as well as in other allied fields. The dissertation requirements will involve original research of a high caliber dealing with the use of computers in the science of chemistry. Several highly qualified candidates have indicated their interest in this program, and are currently enrolled.

As a result of these experiences, it should no longer be necessary for prospective chemical information specialists to take their graduate work in laboratory-oriented research and then transfer to the information center where laboratory experience may not be as important to their professional responsibilities as information handling know-how. Furthermore, the management orientation of the program tends to produce an individual who is more flexible in his training and can more easily move about in an organization. As the demand for chemists changes with the times, so must our educational system, and this program has been designed to keep pace with the times.

CHEMICAL SUBSTRUCTURE INFORMATION HANDLING SYSTEMS

The course entitled "Workshops in Technical Information Handling" is a graduate 3-credit elective course, and as described in the University catalog is "for specialization and practice in the common process subsystems of technical information machine systems—e.g., acquisition, indexing, abstracting, and film media." Specific subject matter coverage varies and depends on the interests and capabilities of the instructor and the students. During the Fall semester of 1969, within the aegis of the Workshop, an experimental course was offered, subtitled, "Chemical Substructure Information Handling Systems." The suggestion of offering this course emerged from informal discussions between the authors. The experiment was adjudged successful, and the course was reoffered in 1970, and again scheduled for the Fall semesters of 1971, 1972, and 1973.

A large number of governmental agencies in the Washington, D.C., area are actively involved in chemical information storage and retrieval projects: These agencies include the National Institutes of Health, the U. S. Patent Office, the National Library of Medicine, the National Bureau of Standards, the Food and Drug Administration, the National Agricultural Library, and the Walter Reed Army Institute of Research. In addition, local universities and commercial organizations share an interest in this area. Many employees continue their education for either advanced degrees of for the maintenance of their professional proficiency. It was estimated that a sufficient number of students could be enrolled to meet the University minimum requirement of six students. In 1969, 12 students enrolled and stayed with the course for the entire semester.

The large number of chemical information programs in the Washington, D.C., area not only provided potential students, but also provided a number of experts in different systems that could (and were) called in as guest lecturers. In addition, the students themselves were called on to discuss the programs with which they were associated.

Of the 12 students enrolled in the course, in 1969, nine were actively involved in chemical information programs, one student was a full-time graduate student at the University, and the remaining two enrollees expressed an interest in chemical information science, but had no experience or related qualifications beyond undergraduate degrees in chemistry.

The Syllabus. Each session's discussion was preceded by the question: "What is meant by a chemical substructure search?" Each subject covered was evaluated as to how it analyzed chemical structures for input, storage, search, and retrieval, and how it answered the question.

The course followed the general outline:

History of Chemical Information Systems and a survey of the topics that were to be discussed during the semester, and a description of the related rationale.

Conventional Methods. Use of Chemical Abstracts, Beilstein, and the Ring Index. Uses of nomenclature, molecular formulas, and definitions of rings.

Nonconventional Methods. Development and use of the Fragment Codes: Wiselogle, Frear, Chemical-Biological Coordination Center (CBCC), Ringdoc, the U. S. Patent Office, and the Smith, Kline and French codes.

Chemical Notations as developed by Dyson/IUPAC, Gordon-Kendall-Davison, Wiswesser, Silk, and Hayward (and later Skolnik). Polish strings.

Connection Tables as proposed by Ray and Kirsch, Walter Reed Army Institute of Research, Chemical Abstracts, and "CROSSBOW."

The Mechanical Chemical Code proposed by Lefkovitz.

Automated literature chemical searches from *Chemical Abstracts Condensates* and *Chemical-Biological Activities* (CBAC); and the Institute of Scientific Information *Index Chemicus Registry System* (ICRS), ASCA, the *Science Citation Index*, and recently ANSA.

The U. S. Army *Chemical Information and Data System* (CIDS).

The National Registry System.

In 1970, 1971, the contributions of Corey and Wipke (at Harvard and Princeton); Chauvenet, Farrell, Koniver, Feldman, and Heller from the Division of Computer Research and Technology, NIH, in computerized chemical graphics were added.

Presentations. The following experts graciously donated their time to present lectures to the class:

Winston Hayward, U. S. Patent Office (Washington, D. C.), who described the development of the Hayward Notation to the 1969 class

Wm. J. Wiswesser, Fort Detrick (Frederick, Md.), who discussed "Computer Applications of the WLN" to the 1969, 1970, 1971 and 1972 classes

Chas. E. Granito, Institute for Scientific Information (Philadelphia, Pa.), presented "Computerized Information Handling Services from ISI" to both the 1969 and 1970 classes; and in 1971 and 1972 the class visited the Institute for Scientific Information in Philadelphia.

Paul Olejar, formerly of the National Science Foundation (Washington, D. C.), projected the "Future of the National Chemical Registry System" to the 1969 class.

In addition, although unable to come in from Columbus, Ohio, John Gibson, formerly of Chemical Abstracts Service, provided extensive literature describing the CAS systems for distribution and discussion for the 1969 class, which was also used for the 1970 class.

Site Visits. As could be arranged, the class met at facilities for demonstrations. For example, the class visited:

a. The National Cancer Institute to observe the Micrositе System as explained and demonstrated by G. F. Hazard

b. The Division of Computer Research and Technology, NIH, to view the use of the Rand Tablet for diagram input and programmed conversion to the Wiswesser Line Nota-

tion (WLN), search and retrieval, and high speed printer two-dimensional drawings, as explained by D. Koniver in 1970; revisited in 1971 where Richard Feldman described his experimental further development of input methods and file structuring based on nested structural characteristics

c. The Institute for Scientific Information, Philadelphia, to examine a chemical information publishing operation and obtain knowledge of the products as well as how they are prepared.

Reference Materials. Since there is no basic text for this type of a course, the following materials were used as a guideline and for source materials:

"Survey of Chemical Notation Systems," NRC Publication 1150; and the companion "Survey of European Non-Conventional Notation Systems," NCR Publication 1278. During the term, the National Research Council Committee on Chemical Information published "Chemical Structure Information Handling. A Review of the Literature 1962-1968," NRC Publication 1733. These three sources provided the basic reference materials. Future courses will include "Computer Handling of Chemical Structure Information," by M. F. Lynch *et. al.*, published by MacDonald/American Elsevier. The author and guest speakers provided a number of reprints which were distributed in advance of the related discussions. Selective readings, mostly from the *Journal of Chemical Documentation, Information Storage and Retrieval,* and *Nachrichten für Dokumentation* were assigned.

Student Participation. As earlier indicated, nine of the students in 1969, were actively engaged in chemical information projects. Each was an expert in his respective system, and provided a lively commentary on not only his own system, but a critique of other systems with which he was familiar. The students were called on to give informal presentations describing the programs with which they were involved.

Course Guidelines. The development of chemical structure information systems was traced with the parallel development of available equipment. In addition, organization needs, the level and type of personnel required, and availability of equipment was discussed. The equipment included 3 × 5 card files, edge-notch cards, tab cards, Termatrex, microfilm, and finally computers. Input/Output devices reviewed included: the Army Chemical Typewriter, the Mohawk keytape, the MTST, the Dura Mach, the Shell chemical "golf-ball" for the Selectric typewriter, and the use of drumprinters or special print chains—e.g., that developed by Smith, Kline and French. Where possible, economics of chemical information systems were discussed—i.e., the level of training required for manual coding *vs.* machine coding. Emphasis was placed on the effectiveness of a chemical search for each system under review, and the role of a chemist/information scientist in responding to a chemical search request. The transition of a model system was followed as demands on the system expanded, and the file size grew. In the true spirit of a Workshop, open discussions prevailed. The class significantly contributed ideas and problem areas consistent with their level of experience. Assignments were made to gain understanding of subject matter and in preparation for the invited speakers. In addition to being a survey of "whom was or is doing what?," details of methods were analyzed so that the participants had more than a shallow understanding of the varying approaches that were discussed.

The aforementioned assignments were not collected, but, when specific problems were noted, they were discussed, either in class or in private student-instructor conferences. Rather than try to assemble a final exam, a term paper was required dealing with subjects that either the student was familiar with from his own experience, or that mentioned or described in class. The success of the class may in part be judged from the fact that three of the papers were recommended for publication in the *Journal of Chemical Documentation.* Two of these three papers were presented at Middle Atlantic Regional Meetings, and were subsequently revised and published. These were:

Spann, M. L. and Willis, D. D., "A Comparative Study of a Fragementation vs. a Topological Coding System in Chemical Substructure Searching," *J. Chem. Doc.* **11,** 43-7 (1971).

Pick, R. O., Eckermann, E. H., Schafer, J. A., and Waters, J. F., "Strategy of Data Retrieval and Analysis from Large Biological and Chemical Files," *J. Chem. Doc.* **12,** 35-7 (1972).

CONCLUSIONS

In summary, the need and value of a specialized course of this type seems to have been well proven. The course was judged to be successful in view of the number of students that enrolled, and the course was reoffered in 1970, 1971, and 1972.