French National Policy for Chemical Information and the DARC System as a Potential Tool of This Policy*

J. E. DUBOIS
The Paris VII University, 1, rue Guy de la Brosse, Paris, France

Received October 27, 1972

The incentive and the main lines of French policy on chemical information, as well as their implementation, are described. New governmental bodies have been created aiming at setting up a national network for scientific and technical information. Among these, the CNIC (Centre National d'Information Chimique) is in charge of the chemical field. The DARC system is being implemented as a tool of national policy for chemical information. An exhaustive chemical data processing system, it features topological encoding, input, and retrieval methods which are described in this paper.

This paper presents the main features of French policy for chemical information. More specifically, it discusses how this policy originated, the structures designed for its implementation, and the main features of the DARC topological system which is the major tool of this policy owing to the necessity of a compound registry system and of processes for substructure search.

NATIONAL POLICY ON CHEMICAL INFORMATION

Survey of the Past Situation. For many years, French research centers have been aware of the possibilities and the problems of technical and scientific information, and have instigated various achievements either in the national or the private field.

Thus the CNRS (National Center for Scientific Research), as soon as it was created in the 30's, set up a multidisciplinary documentation center in charge of the information provided by the various scientific publications; thus originated the *Bulletin Signalétique*. In the same way, several documentation centers were created by the main national research centers—CEA, INSERM, and INRA. Meanwhile, other documentation centers appeared, either within major firms for their internal use, or as a result of a joint decision taken by professionals eager to receive the information related to their specific activities—e.g., petroleum, textile, and metallurgy.

Numerous centers were therefore developed, but in a somewhat anarchical way which did not favor an optimization of activities. As a result, until recently, the main feature of the French situation was, as in many other countries, a lack of coordination which was felt:

at the national level, where there was no organization in charge of coordination between public and private activities

at the sectorial level, where there was no standardization of exchanges between the various private centers on the one hand, or between these centers and the CNRS and Universities on the other hand.

In recent years different trends in the development of international exchanges and the eventual implementation of a complex worldwide information network have aroused government interest in restructuring our means for achieving an integrated national network.

Present Trends. The need for structures better adapted to serve the interests of the scientific community led the French government to define a policy for scientific and technical information which is part of the national policy for scientific and industrial development.[1] This decision was based on a study of the problems arising from:

the difficulties (owing to existing structures) in coordinating all activities

the need to reinforce the French position to better contribute to international exchanges and to benefit in return from foreign achievements

The realization of this policy relies on the organization of sectorial activities within a national network. The latter is under the control of a body created by the Ministry of Industry and Scientific Development, the National Bureau for Scientific and Technical Information (BNIST) (Figure 1), which will start operating at the end of 1972. The BNIST includes:

a direction committee which is composed of representatives of the government, the public sector, and the private sector

a permanent secretariat

The BNIST has been allotted a budget which amounts to about 15% of the total governmental expenditure for scientific and technical information. The BNIST aims at defining and promoting the general policy of the national network and at coordinating all sectorial activities. The various information centers are grouped according to their specific activities—chemistry, electronics, nuclear science, etc.

The chemical field is managed by the National Center for Chemical Information (CNIC). The CNIC's aim is to provide the members of the scientific community with all of the chemical information they need wherever this information comes from, in the best conditions of speed, relevance, and convenience, and to promote relevant studies.

The CNIC comprises two organizations: the French Association for Machine Documentation in Chemistry (AFDAC) and the Association for Research and Development in Chemical Information (ARDIC).

*Presented in the Symposium, "Chemical Information Systems Abroad," 164th Meeting, ACS, New York, N. Y., August 29, 1972.

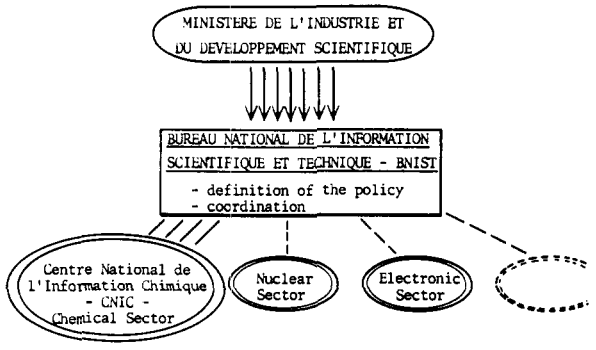

Figure 1. French National Network for scientific and technical information

To understand the objectives and relationships of the above organizations more clearly, we shall first try to define a model information network.

A Model Information Network (Levels and Centers). The organization of an information system implies a model network whose centers are located in a consistent order along the information flow, the acquisition taking place "up-stream" and the dissemination "down-stream."[2]

The network is based on three levels (Figure 2).

1st Level—Abstracting and Indexing Centers or "Signaletic Centers" (SC). Their function is to process primary information (books, reports, papers) into secondary information by means of light but adequate indexing. Each document is associated with indexing data elements (author names, title terms, keywords) which permit their retrieval.

2nd Level—Data Banks (DB). Their function is the compilation and a more or less thorough analysis of primary and secondary information in a given field. They produce data tables, monographs, etc.

3rd Level—Dissemination Centers (DC). Their function is the dissemination of processed information. They retrieve "on demand" information, either by selective dissemination—a profile is established according to the specific needs of each customer and documents are periodically sent to him—or by retrospective search of primary and secondary information provided by the centers at the first two levels.

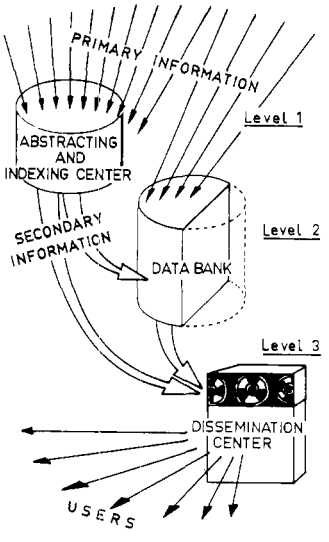

Figure 2. A model information network

The organization bodies that deal with information are "centers" which differ according to their activity level. The centers at levels 1 and 2 have different features:

at level 1, the Signaletic Center (SC) covers a wide range of information and gives a superficial set of indexing data—author names, journal references, title terms, keywords, which enable the documents to be located both geographically and according to content.

at level 2, the Data Bank (DB) covers a limited and specific field, and provides thorough information which includes numerical and experimental data.

Chemical information has one major characteristic which is based on the need of topological indexing data, as distinguished from "literary" indexing data.[3-6] The former are the main concepts of a chemical corpus because of its bulk, its fast growth, and its perenniality. It is therefore essential to the researcher or engineer to be able to retrieve a document by means of these data.

The retrieval of information related to chemical compounds rests on the organization of a Topological Data Base (TDB), which includes:

a topological code (TC) which is a unique and unambiguous code. The code is generated at the indexing stage, like the keywords, but through a specific process.

a Topological Screen System (TSS) which is a set of structural parameters aimed at optimizing the substructure search.[7]

a Registry Number (RN) which gives the link between the structure files and the information files.[8-10]

CNIC Policy within the Information Network. In the three-level network, the CNIC plans to be directly present at levels 1 (Signaletic) and 3 (Dissemination) and to promote the public and private centers at level 2 (Data Banks).

To implement its policy, consideration has been given to various factors (Figure 3). Specifically, the following are planned:

starting in 1972, dissemination of the secondary information provided by the Chemical Abstracts Service and development of a topological indexing system

in 1976, or thereabouts, setting up of a system which will cover French-language literature, from France and abroad, and will include a topological indexing center to implement the DARC system.

At level three, the CNIC aims at promoting new procedures to help communications between the user and the dissemination center. Furthermore, starting in 1972, the CNIC will take part in the development of dissemination procedures in the framework of a cooperation agreement with the Chemical Abstracts Service. The tasks implied by this policy will be shared between the AFDAC and the ARDIC (Figure 4).

The AFDAC will be in charge of the dissemination of chemical information (in the first stage, the information

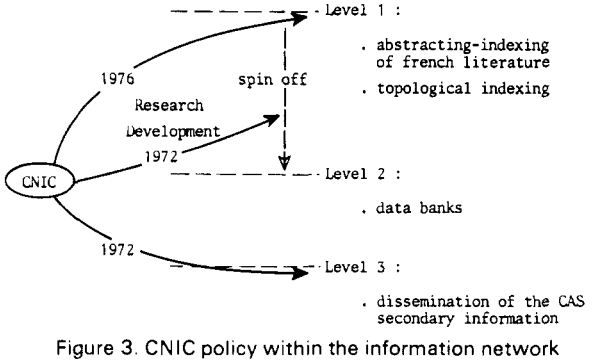

Figure 3. CNIC policy within the information network

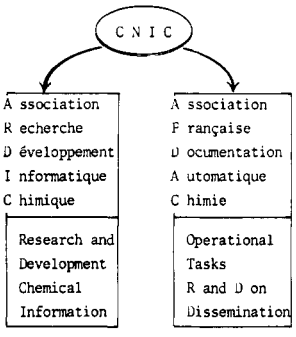

Figure 4. CNIC organization

provided by the CAS magnetic tapes), and of a topological pilot center based on the use of the DARC system, which will be operating in 1974 and will test acquisition and retrieval procedures. In a later stage, the AFDAC will cover the abstracting and indexing of French-written chemical literature.

The ARDIC will promote, carry out, or have carried out all research and development work aimed at creating or improving information systems applying to chemistry and all research operations using these systems: information acquisition and retrieval and correlation studies. The ARDIC will continue research on the DARC topological system, instrument of documentation policy whose main features we shall briefly describe here.

THE DARC SYSTEM

The DARC system (Description, Acquisition, Retrieval, Correlation), which has been developed since 1963,[11-18] provides an integrated system for data processing, which ranges from data representation to computer-aided design. [This work is supported by the following French government organizations: DGRST (Délégation Générale à la Recherche Scientifique et Technique) and D.I. (Dé légation à l'Informatique).] It can handle any chromatic graph-representable data, that is to say a graph where the nodes and edges are symbolically differentiated by various colors. Chromatic graphs may be used to represent electric networks, syntactic graphs, patterns, chemical compounds, reaction pathways of flow charts.

Within a chemical information network such as the one described above, the DARC system can be applied at every level (signaletic centers, data banks, dissemination centers) and also for correlations. This paper explains only the main features of the system for description, acquisition, and retrieval (DAR), and shows the applications to chemical information at the level of an abstracting and indexing center and the spin off at the level of data banks.

DESCRIPTION

The description of the chemical compounds in the DARC code or DEL (descriptor propagated by limited environment) meets the requirements of the above defined policy at the level of:

the Signaletic Center (SC) because of:

- its "*biunivocity*" (it is unique and unambiguous)
- its *ability to self generate* the retrieval elements especially the general, specific, and topological screens
- its *convertibility* with the other topological codes and its ability to generate fragment codes and a linear notation

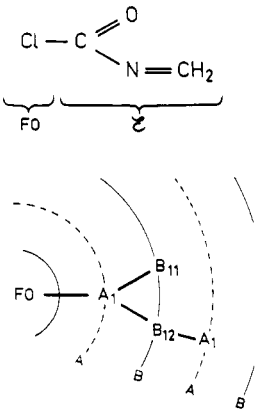

Figure 5. Focus and environment

Focus (FO), starting point for generating chemical graph Environment (ℰ) (atoms and bonds outside the focus), generated progressively in concentric layers (level A, level B) and ordered

the Data Banks (DB) because of:

- its *ability* to reach exhaustivity
- its *efficiency* as a design tool

The DARC code arises from the generation of a chromatic graph, which is the diagram of a chemical compound, and from the implementation of several concepts, a generation principle and a description method.

Basic Concepts. There are two, the focus and the environment.

The *Focus* (FO) is the starting point for generating the chromatic graph. For documentation, the choice of the focus is based on formal rules which provide the biunivocity of the description. For design, the choice results from an approach of the molecule which meets the chemist's specific requirements.

The *Environment* (ℰ) includes the chromatic sites (atoms and bonds outside the focus), and is organized in concentric layers,[19,20] centered on the focus. These layers are regrouped by pairs called limited environments E_B. Within a given E_B, there are subdivisions into limited environment segments e_B (Figure 5).

Generation Principle. The environment is generated progressively by propagating a limited environment E_B. Within a given segment e_B, the nodes of level A are generated before those of level B. The generation order of nodes on the same level is determined by applying and propagating an ordering function which requires assigning a specific index to each node. In this way, a chromatic, organized, and ordered graph is defined.

Description Method. After generation, each segment of the chromatic graph with its ordered nodes gives rise to a modular description according to the method shown in Figure 6. The elementary descriptors are ordered according to the generation principle. The ring-closing bonds are given in a specific descriptor. The full description provides the DARC code (Figure 7).

Features of the DARC Code. The DARC code is a *topological linear representation* of a molecular structure with an "open" logic, inasmuch as it can reach exhaustivity without altering its basic principles. It is particularly well adapted to the description of: stereochemistry (Figure 8), hydrogen atoms (Figure 9), and polymers.

ACQUISITION (A)

The acquisition is aimed at generating the DARC code in a unique and unambiguous form starting from an input structure which is either a diagram or a description in an-

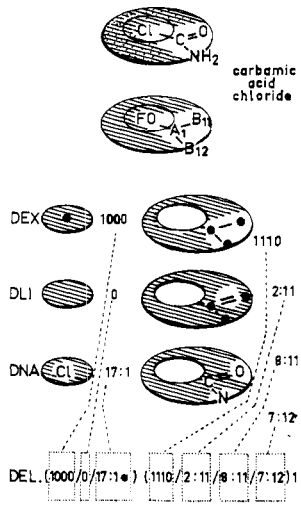

Figure 6. Building and DARC code

The DARC code results from the topological linking of modules, each of them consisting of three parts which describe respectively: the topology (DEX), the bond multiplicity (DLI), and the atom nature (DNA)

In the above example, there are two modules—the focus (chlorine atom) and the environment. The focus is described in the first three boxes, the environment in the following ones: 1000 (1st box) is the topological descriptor of the focus, 0 (2nd box) means no bond in the focus, 17 (3rd box) is the atomic number of chlorine, 1110 (4th box) is the topological descriptor of the environment, 2 (5th box) indicates a double bond in the environment, 8 and 7 (6th and 7th boxes) are the atomic numbers of the oxygen and nitrogen atoms of the environment

Slashes and parenthesis are separators. Numbers which follow the colons and the closing parenthesis are topological coordinates. The star allows for module linking

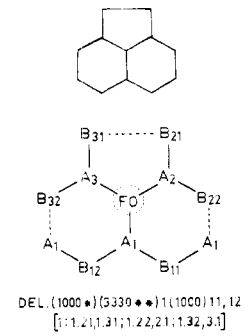

Figure 7. Code DARC tricyclo [6.3.1.0.$^{4-12}$] dodecane

The ring closure bonds are materialized by dotted lines which appear between square brackets in the descriptor

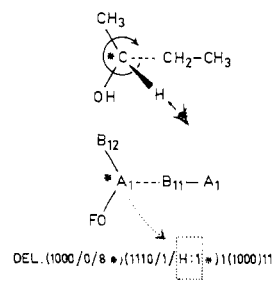

Figure 8. Assymetric carbon DARC representation

The DARC topological ordering rule allows us to class the substituents around the carbon and an observation rule allows us to specify its orientation. In this case (H)

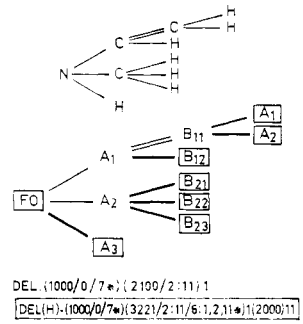

Figure 9. DARC codification of hydrogen atom

For specific studies (NMR) it is possible to include the H's codification in the DARC code

other code. Several sets of rules for the choice of a unique focus have been defined. They are selected with respect to the input procedure for the whole file:

Manual Input. The rules take into consideration those structural features which are easily and visually noticeable—for example, rings and complex ring structures.

Automatic Input. 1. By means of a specific device—the rules take into consideration features which are put forward by algorithms (connectivity degree, type of nodes, and edges) 2. By means of interconversion—in this case, if the input code leads to a unique representation starting from an origin (such as the CAS code), this origin is maintained. Now the DARC code is convertible with CAS code and Wiswesser Linear Notation.[21]

RETRIEVAL (R)

One of the major points in chemical information is to retrieve information related to a compound or to a set of compounds with one or several identical structural patterns. For this purpose, the DARC system includes an interrogative topological language close to the topological code and a retrieval strategy

This strategy is implemented on the DARC Topological Data Base (DARC/TDB), which includes three basic components (Figure 10):

DARC code, topological representation of each compound; DARC Topological Screens (DARC/TS); and a registry number which links the DARC Topological Data Base with the Information Data Base (IDB).

The main features of the DARC Screen system and the various possibilities for the organization of the DARC/TDB will be described.

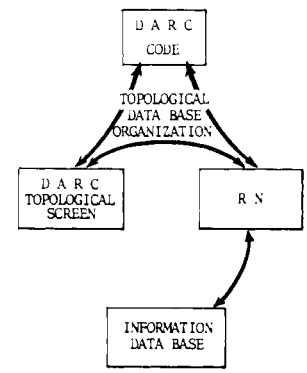

Figure 10. DARC topological data base (DARC/TDB)

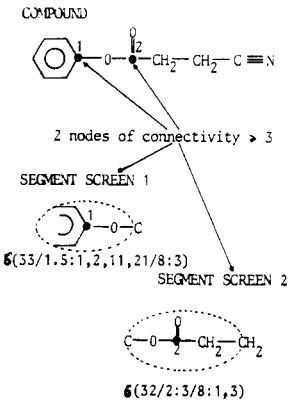

Figure 11. DARC segment screens (FREL subcode)
Segment screens are obtained by generating segments around each node whose connectivity is higher than or equal to three. The screens are computer-assignable starting from the DARC code. They follow the same rules as the code. Altogether, the segment screens constitute a fragment code (FREL subcode)

DARC Topological Screen System (DARC/TSS). There are three different types of screens, all of which can be directly generated from the code.

General Screens. These are related to generic properties of chemical compounds—molecular formula, cyclic features (cyclic nuclei count, ring count, ring size), acyclic features (chain length, heteroatoms, etc.)

Specific Screens. They are defined with respect to the trends of a given information center (particularly at level 2), and also result from experiments carried out with the users over a certain period of time. The specific screens are more extensive patterns or more complex cyclic systems than the general screens.

DARC Topological Screens (DARC/TS). They include two groups of screens:

DARC Canonical Screens (DARC/CS) which include segment screens and chain screens. The segment screens are fragments of the DARC code. They are determined by means of polyfocalization on each node having a connectivity higher than or equal to three, and generating the description of the E_B of each focus, called DENSE FOCUS. All the segment screens associated with a compound provide a unique though unambiguous fragmentary description of the compound, which is called FREL Subcode (limited environment fragment). The chain screens are the chains between two dense focuses or between a dense focus and a terminal node, provided they include more than two bonds (Figure 11).

DARC Dummy Screens (DARC/DS) which originate in the canonical screens and sharpen the discrimination ability of the latter by means of a statistical analysis of the corpus leading to a screen system providing the equipartition of the corpus (Figure 12).

The DARC procedure for accurate substructure search by means of selective polyfocalization (PS) is based on the description defined by segment screens and chain screens.

DARC Topological Data Base Organization (DARC/TDB). Four different types of organization are being currently evaluated, each corresponding to specific hindrances for volume and processing mode of the signaletic center or the data bank.

Organization 1—(Limited volume below 10^3 structures, batch processing). The screen file and the DARC code file are merged into a sequential organization. The record contains: the general screens, the canonical screens, and the DARC code.

Organization 2—(Volume below 10^4 structures, on line).

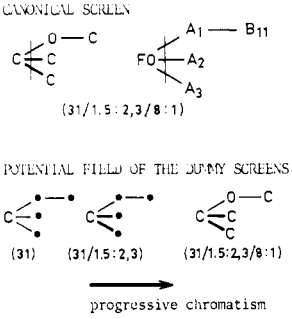

Figure 12. Derived descriptors—dummy screens
Dummy screens reflect the progressive chromatism and the dummy screens are selected in the potential field by applying statistical criteria

Screening is performed by two hierarchical inverted files (dummy screens and general screens), with direct access. The DARC code file has an indexed sequential organization, with direct access.

Organization 3—(Volume ranging between 5.10^4 and 3.10^5 structures, on line). Screening would be performed by 2 files: dummy screens inverted file and general and canonical screens sequential file. The DARC code file is the same as in organization 2.

Organization 4—(Volume above 3.10^5 structures). Screening could be performed by two hierarchical inverted files (canonical screens and general screens), with direct access. The DARC code file is the same as in organization 2.

The DARC System responds to the need for stocking a very large body of compounds while yet optimizing global data input through the DEL code, as well as to the need for rapid retrieval, not requiring an atom-by-atom search. For this purpose, use is made of the FREL topological code for developing screens which are substructures. Automatic generation of FREL screens is one characteristic of the canonical correspondance of the DEL and FREL sub-codes which are component parts of the DARC code. Furthermore, the corpus of screens is of a dynamic nature.

DARC System Current Development. A cooperative agreement has recently been established between the Chemical Abstracts Service and the CNIC/ARDIC, that part of the CNIC in charge of research and development studies. It is planned to establish an experimental program, based on use of a portion of the CAS Registry structure file which would call for the identification and examination of the differences in structural representation between the CAS and the DARC System. Use of the DARC System in accomplishing substructure search will also be examined.

In addition to these investigations, exchanges between the CAS and the ARDIC could lead to the definition of other areas of cooperation such as investigation of the problems of: stereochemical nomenclature, automatic nomenclature generation, and correlating measured data with structure.

Furthermore, the ARDIC is currently developing the use of the DARC Code, and various acquisition (A) and retrieval (R) methods and techniques in several explorative projects of data banks are implemented—pharmacodymanics, radioprotective properties, kinetics, and N.M.R.

CONCLUSION

In conclusion, French chemical documentation is gradually taking its place in a general organization of all scien-

tific disciplines. Paradoxically enough, the existing delay in both industry and state agencies with regard to automation, documentation, and data banks, and the almost total absence of any system of structural description have, in fact, created a situation favorable to the establishment of an up-to-date national network. Thanks to belated choices, this network, both complex and diversified, will enjoy the benefit of unified documentary techniques.

LITERATURE CITED

(1) Delorme, J., *Automatisme*, Vol. **XVII**, 6-7, 193 (1972).
(2) Dubois, J. E., and Cornier, M., *La Recherche* **19**, 35 (1972).
(3) Marden, E. C., and Koller, H. R., "A Survey of Computer Programs for Chemical Information Searching," National Bureau of Standards, Tech. Note **85**, Feb. 1961.
(4) Hunsberger, I. M., Ed., "Survey of Chemical Notation Systems," NAS, NRC Publ. **1150**, 1964.
(5) Frear, D. E. H., "Survey of European Non-Conventional Chemical Notation Systems," NAS, NRC Publ. 1278, 1965.
(6) Holm, B. E., Ed., "Chemical Structure Information Handling," NAS Publ. **1733**, 1969.
(7) Lefkovitz, D., Milne, M., Hill, H., and Powers, R., *J. Chem. Doc.* **12**, 183 (1972).
(8) "The Chemical-Biological Coordination Center," NAS, NRC, Washington, 1954.
(9) Chemical Abstracts Service, "Progress towards a Chemical Registry System at Chemical Abstracts Service," NSF, Contract **NSF-C414**, Task 1 (March 1969).
(10) O'Dette, R. E., and Terrant, S. W., Jr., *J. Chem. Doc.* **6**, 161, (1966).
(11) Dubois, J. E., Laurent, D., and Viellard, H., *C.R. Acad. Sci. Paris* **264**, C, 1019 (1966).
(12) Dubois, J. E., Viellard, H., *Bull. Soc. Chim.* 900 (1968); 905 (1968); and 913 (1968).
(13) Dubois, J. E., *Entropie* **27**, 1 (1969).
(14) Dubois, J. E., and Viellard, H., *Bull. Soc. Chim.* 839 (1971).
(15) Dubois, J. E., and Bonnet, J. C., *C.R. Acad. Sci. Ser. A* **270**, 1002 (1970).
(16) Dubois, J. E., Alliot, M. J., and Panaye, A., *C.R. Acad. Sci. Ser. C* **273**, 224 (1971).
(17) Dubois, J. E., and Herzog, H., *J. Chem. Soc. Chem. Commun.*, 932 (1972).
(18) Goldwaser, D., and Fontaine, J. C., *Automatisme*, Vol. **XVII**, 6-7, 229 (1972).
(19) Dubois, J. E., and Maroni, P., *C. R. Acad. Sci. Ser. C* **243**, 138 (1956).
(20) Penny, R. H., *J. Chem. Doc.* **5**, 113 (1965).
(21) Dubois, J. E., The Paris VII University, Paris, France, unpublished data.

Chemical Information in the United Kingdom*

A. K. KENT
United Kingdom Chemical Information Service (UKCIS),
University of Nottingham, Nottingham, England

Progress in the development of advanced chemical information systems in the United Kingdom is briefly described.

The major producers of primary chemical information in the United Kingdom are chemistry and chemistry-related departments of universities and similar academic institutions (Colleges of Advanced Technology, Polytechnics etc.) There are approximately 70 University Chemistry Departments with a large postgraduate, postdoctoral, and staff population. Several hundred other departments (Biochemistry, Pharmacology, and so on) in both Universities and other educational establishments swell the number of potential producers of primary chemical literature in the academic field very substantially.

Private industry and Government sponsored research institutions of one kind or another are certainly less significant as producers, but have a much more significant role as users of primary and, thus, secondary and tertiary information sources.

The Chemical Society is a major publisher of primary chemical information in the UK, but, apart from the activities of the United Kingdom Chemical Information Service (UKCIS), it has an insignificant role in the secondary information field. Indeed it is fair to say that compared with its input to the world's chemical literature, the UK makes a relatively insignificant contribution to the production of secondary literature sources. There are of course notable exceptions to this somewhat sweeping generalization—Derwent Publications for example, is internationally well-regarded for its work in the field of patents. Nevertheless the UK is making, and has made, some significant contributions in the area of research into methods of handling secondary sources of chemical literature, particularly computer-based ones, and in the area of international collaboration on both the collection and analysis of primary literature for secondary use, and its subsequent dissemination. The work of M. F. Lynch at Sheffield University,[1,2,3] of the Imperial Chemical Industries (ICI) team that developed the CROSSBOW system for handling Wiswesser Line Notations (WLN),[4,5] and of UKCIS itself[6,7] are examples.

Private chemical industry in the UK is dominated (except possibly in the pharmaceutical field) by a very few, very large national or multinational corporations. There is a very definite gulf between the few very large and the large number of rather small companies. The large companies have the technical and financial capability to develop internal programs for handling their chemical information problems, using external publicly available sources as well as their own internal sources; indeed some of the most notable contributions to the art of handling chemical information have come, in the past few years, from these major companies. The lack of a substantial number of middle-sized industrial users of chemical information has produced a situation in which there is a very marked gap between the sophisticated information-handling practices

* Presented at the Chemical Literature Division of the American Chemical Society, New York, August 29, 1972.