GAS CHROMATOGRAPHIC INFORMATION STORAGE AND RETRIEVAL BY THE UNITERM SYSTEM

M. ROGOZINSKI AND M. MEDWIED
Organic Chemistry Department, Israel Mining Industries Laboratories, Haifa (Israel)

(Received November 29th, 1962)

Within the last decade gas chromatography has become one of the most popular techniques of organic chemical analysis. This is clearly shown by the number of papers which have appeared on this subject. *Gas Chromatography Abstracts* have published close to 4000 abstracts for the period 1952–1961 and even this figure probably does not represent a complete coverage. The number of publications in this field is growing at such a rate that it is becoming increasingly difficult to keep up with the literature. It is clear from the size of the literature on gas chromatography that collecting it will have little utility without a suitable technique of information retrieval. However, little has been published concerning the suitability of the various information systems for gas chromatography. SPENCER AND JOHNSON[1] have reported details of a punched card system for the storage of gas chromatographic data for hydrocarbons. LEWIS *et al.*[2] expanded this method by using key punched IBM cards and including more data. However, both these systems, as described, have only been applied to the storage of a limited number of concepts.

Of the various systems of information storage and retrieval, we chose the Uniterm system[3,4] for our card file on gas chromatography which at present consists of close to 2500 literature items. Information is being stored under more than three hundred Uniterms. The Uniterm system involves no codes, is highly flexible and simple to operate and both storage capacity and depth of indexing are almost unlimited. However, the direct retrieval of retention data, as described by both SPENCER AND JOHNSON and LEWIS *et al.*, has not been attempted. Although this is undoubtedly possible with the Uniterm system, the punched card system would appear to be the more suitable method for this particular application.

The Uniterm system is essentially a system of co-ordinate indexing consisting of an information store and a concept (Uniterm) store. In the information store the information cards (abstract cards, reprints etc.) are given consecutive integral numbers (non-significant accession numbers) in the order in which they are received, so that each information card is characterised by an arbritary storage number which serves as its address. The information that these cards contain is broken down into a number of simple concepts or bits of information called Uniterms.

The Uniterms, which are subject headings for unit concepts, all of equal hierarchical stature, are stored independently in the form of Uniterm cards, each of which represents a single concept (Uniterm). This store of Uniterm cards makes up the information retrieval system.

Examples of Uniterms are the names of stationary phases, the names of materials

analysed, operating conditions such as temperature, flow rate etc., application of technique such as surface area determination, relative volatility determination etc. As there is no limit to the number of terms which can be chosen there is no limit to the depth of indexing. For example, a generic term such as HYDROCARBONS can be broken down into subgroups such as PARAFFINS, NAPHTHENES, AROMATICS, OLEFINS etc. and the subgroups can be broken down further into specific molecular weight ranges or even individual compounds. Parametric concepts such as COLUMN PROPERTIES can be more rigorously specified by qualifying this Uniterm with the addition of nouns, prepositions, adjectives etc. For example: COLUMNS–CAPILLARY; COLUMNS–PACKED; COLUMN–EFFICIENCY, EFFECT ON; COLUMN–EFFICIENCY, NUMERICAL VALUES; COLUMN–GEOMETRY, EFFECTS OF, etc. The efficiency of the retrieval system depends directly on the specificity of the Uniterms used. The more qualified the concept involved, and consequently the larger the number of Uniterms, the more rapid and efficient will be the information retrieval. However, such systems would require a greater amount of indexing so that the depth of indexing, which is a matter of choice, depends on the individual requirements. Translation of the information into Uniterms is of necessity performed by personnel experienced in the field.

The freedom of choice of terms and the fact that all terms are independent and of equal hierarchical stature makes this system highly flexible, permitting the introduction of new terms as necessary. This is one of the great advantages of the Uniterm system.

Indexing consists of entering the accession number of the information card on the relevant Uniterm card, the column being determined by the terminal digit of the accession number. The accession number 351 is entered in column 1, the number 659, in column 9 etc. The numbers are arranged in ascending order in the columns to facilitate comparison and searching (see Fig. 1).

OLEFINS

0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
100	121	62	93	44	85	406	27	78	99
140	351	182	373	84	555	436	167	98	129
160	371	302	703	174	635	476	187	208	369
200	561	332	783	234	645	636	227	218	439
350	631	372	863	634	665	646	297	278	539
390	691	432	923	644	815	686	347	408	589
410	701	592	973	734	825	736	387	428	599
430	731	602	1053	794	895	766	397	558	679
440	851	642	1133	824		806	407	628	699
540	911	682	1413	1004		816	437	638	729
660	1281	702	1483	1134		976	477	728	779

Fig. 1. A typical Uniterm card.

If card 740 discusses the relative merits of dinonyl phthalate and squalane for the separation of a mixture of paraffins and aromatic hydrocarbons in terms of π bond–polar substrate interaction and vapour pressure effect, the number 740 is written in the 0 column of the following Uniterm cards: DINONYL PHTHALATE–STATI-

ONARY PHASE, SQUALANE–STATIONARY PHASE, PARAFFIN HYDROCARBONS, AROMATIC HYDROCARBONS, SOLUTE–SOLVENT INTERACTION and VAPOUR PRESSURE EFFECTS. Depending on the requirements, the depth of indexing can be increased and Uniterm cards can also be made out for the individual paraffins and aromatics mentioned as well as for the Uniterms, POLAR STATIONARY PHASES and NON-POLAR STATIONARY PHASES.

As the information retrieval consists of a collection of independent concepts it is possible to retrieve information in the form of combinations of specific qualified concepts, which when co-ordinated, represent the sought for idea. This is done by withdrawing the relevant Uniterm cards (representing the simple concepts) and scanning them vertically, columnwise, for common numbers. The common numbers will then represent information cards which contain the information sought, that is, the complex concept represented by the combination of simple concepts. For example, if information is required on the relative merits of polar and non-polar stationary phases for the separation of paraffin–aromatic hydrocarbon systems then the four Uniterm cards POLAR STATIONARY PHASES, NON-POLAR STATIONARY PHASES, PARAFFIN HYDROCARBONS and AROMATIC HYDROCARBONS are scanned for common numbers. A card whose number is found on all four Uniterm cards will contain information on all four concepts. However, it is not always the case that the concepts involved are related in the desired manner. A card may contain information on all four concepts without actually discussing the relative merits of the two types of stationary phases for the separation of paraffin–aromatic systems. The information might concern the separation of paraffins on non-polar stationary phases and the separation of aromatic–olefin systems on polar stationary phases. This does not invalidate the indexing but indicates a lack of depth of indexing or qualification. It can be avoided by increasing the number of Uniterm cards by the addition of the Uniterm PARAFFIN–AROMATIC HYDROCARBON SEPARATION. The desired information would then be sought on the three Uniterm cards (instead of four): POLAR STATIONARY PHASES, NON-POLAR STATIONARY PHASES and PARAFFIN–AROMATIC HYDROCARBON SEPARATION.

Certain Uniterm cards tend to become overloaded. Examples are the Uniterms PARAFFIN HYDROCARBONS and SILICONE STATIONARY PHASES. The greatest amount of work on any single molecular type in gas chromatography has been the separation of the enormous number of paraffin isomers present in natural hydrocarbon fractions. Several hundred articles have been published on this subject. It is clear that a single Uniterm covering this concept will be inadequate. However, making Uniterm cards for each individual isomer will be advisable only for laboratories specifically interested in this subject. A good way to solve this problem is to make Uniterm cards for specific molecular weight ranges, the number of Uniterm cards can thus be limited according to requirements.

Silicones have always been one of the most popular stationary phases and in this case too it is also desirable to increase the depth of indexing. This is readily done by making Uniterm cards for each type of silicone.

Indiscriminate information storage can also result in overloading. For example, an article discussing the separation of ketones and esters might also mention benzene. The storage of this information, which has little informational value relative to the large amount of work published in far greater detail on the separation of aromatics,

will cause overloading which is not compensated for by the informational value involved. This can be avoided by screening (that is, not storing this information) or by the use of the proposed colour coding system mentioned below. The great majority of published papers supply operating conditions and, unless unusual conditions are involved, the storage of information on numerical values of such operating conditions as flow rate and temperature will result in gross overloading and seriously reduce the efficiency of the method.

Some minor refinements to reduce noise and increase retrieval efficiency are planned in these laboratories. As all terminal digits in any column of the Uniterm card are identical there is actually no necessity for their presence. If these terminal digits were eliminated it should make the search for common numbers on several Uniterm cards simpler. For example the numbers 12, 562 and 2532 could be written in column 2 as 1, 56 and 253 respectively with the understanding that the number 2 be added as a terminal digit to any number in column 2 when the abstract card number is sought. Numbers represented by units only could be entered into the corresponding column as zeros to avoid their elimination.

Noise could be reduced by indexing address numbers of cards containing information of particular interest in a different colour. Colour coding could also be used to reduce the number of Uniterm cards by combining two similar Uniterms on one card, for example RETENTION TIME DATA and RETENTION TIME, EFFECT OF VARIOUS PARAMETERS ON.

The compilation of a thesaurus of the Uniterms in use is advisable to avoid duplication. Uniterms seldom consist of single words and multiword Uniterms can frequently be written in several ways, for example TEMPERATURE PROGRAMMING and PROGRAMMED TEMPERATURE. Synonyms also occur such as injection block, flash heater and flash evaporator. A very useful list of Uniterms for gas chromatography and a ready made thesaurus is the index of Gas Chromatography Abstracts. In fact, it should be highly profitable to combine the Uniterm system with Gas Chromatography Abstracts by writing the Gas Chromatography Abstract numbers on the appropiate Uniterm cards (colour coded).

In these laboratories the Uniterm cards have been classified into related groups. For example all Uniterm cards concerning stationary phases are stored together in alphabetic order. Uniterm cards of stationary phase functions such as STATIONARY PHASE COMPARISON, TEMPERATURE LIMIT, CONDITIONING, PRETREATMENT, etc. are similarly stored adjacent to the stationary phase cards. This system greatly facilitates the search for these cards and simplifies Uniterm combinations to build more complex concepts. The subdivision of the Gas Chromatography Abstracts index is a good example of such a classification.

As many of the abstract cards in these laboratories were prepared from references and abstracts, an author index has been found necessary to avoid duplication. This system has been in use three years and has been found to work very satisfactorily.

ACKNOWLEDGEMENT

The authors are indebted to the Israel Mining Industries for permission to publish this work.

SUMMARY

The organization of gas chromatography literature for storage and retrieval by the Uniterm system of coordinate indexing is discussed. In this system the contents of the separate literature items are broken down into simple concepts which are stored independently. By combining these simple concepts into a more complex one retrieval of the sought idea is achieved.

REFERENCES

1 C. F. SPENCER AND J. F. JOHNSON, *Anal. Chem.*, 30 (1958) 893.
2 J. S. LEWIS, G. T. MCCLOUD AND W. SCHIRMER, *J. Chromatog.*, 5 (1961) 541.
3 W. E. BATTEN, *Specialized Files for Patent Searching*, in R. S. CASEY AND J. W. PERRY (Editors), *Punched Cards*, Reinhold, New York, 1951.
4 M. TAUBE and Associates, *Studies in Coordinate Indexing*, Vol. 1–5, Documentation Inc., Washington D.C., 1953–1959.