search as a means of substructure retrieval will depend on the type of requirements submitted to a search system. For example, if a majority of "real" requirements submitted by a representative user population are for fragments for which no molecular environment can be defined ((d) above), WLN will be an ineffective tool for substructure retrieval. Table III shows a classification of the subject requirements of 101 users according to the type of search strategy required. We were able to develop reasonably effective strategies for the majority of these users.

ACKNOWLEDGMENT

We wish to express our thanks for the help given by A. K. Kent and I. C. McCracken of the United Kingdom Chemical Information Service; by the Institute for Scientific Information and, in particular, A. E. Cawkell, C. E. Granito, and M. Rosenberg; and by the Oxford University Computing Laboratory. We also wish to thank the research workers who comprised our user population. The project was funded by a grant from the Office for Scientific and Technical Information (U.K.). The grant-holder was Sir Ewart Jones, F.R.S., Oxford University.

LITERATURE CITED

(1) National Academy of Sciences, Publ. No. 1733, "Chemical Structure Information Handling: A Review of the Literature: 1962-1968," Washington, D. C., 1968.
(2) Lynch, M. F., Harrison, J. M., Town, W. G., and Ash, J. E., "Computer Handling of Chemical Structure Information," Macdonald and Co. (Publishers) Ltd., London, and American Elsevier Publishing Company Inc., New York, 1971.
(3) Campey, L. H., Hyde, E., and Jackson, A. R. H., "Interconversion of Chemical Structure Systems," *Chem. Brit.* **6**, 427-30 (1970).
(4) Dammers, H. F., and Polton, D. J., "Use of the IUPAC Notation in Computer Processing of Information on Chemical Structures," *J. Chem. Doc.* **8**, 150-60 (1968).
(5) Bowman, C. M., Landee, F. A., Lee, N. W., Reslock, M. H., and Smith, B. P., "A Chemically Oriented Information Storage and Retrieval System. III. Searching a Wiswesser Line Notation File," *Ibid.*, **10**, 50-54 (1970); and earlier work by these authors cited therein.
(6) Leiter, D. P., Morgan, H. L., and Stobaugh, R. E., "Installation of a Registry System for Chemical Compounds," *Ibid.*, **5**, 238-42 (1965).
(7) Garfield, E., Revesz, G. S., Granito, C. E., Dorr, H. A., Calderon, M. M., and Warner, A., "*The Index Chemicus Registry System:* Pragmatic Approach to Substructure Chemical Retrieval," *Ibid.*, **10**, 54-8 (1970).
(8) Rössler, S., and Kolb, A., "The GREMAS System, an Integral Part of the IDC System for Chemical Documentation," *Ibid.*, **10**, 128-34 (1970).
(9) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, 1968.
(10) Palmer, G., "Wiswesser Line-Formula Notation," *Chem. Brit.* **6**, 422-6 (1970).
(11) Leggate, P., Rossiter, B. N., and Rowland, J. F. B., "The Evaluation of a Current-Awareness Service based on the *Index Chemicus Registry System,*" in preparation.
(12) Crowe, J. E., Leggate, P., Rossiter, B. N., and Rowland, J. F. B., "Development and Evaluation of a Current-Awareness Service based on the *Index Chemicus Registry System,*" The Experimental Information Unit, Oxford University. Report to the Office for Scientific and Technical Information (U.K.), June, 1973.
(13) Granito, C. E., Becker, G. T., Roberts, S., Wiswesser, W. J., and Windlinx, K. J., "Computer-Generated Substructure Codes (Bit Screens)," *J. Chem. Doc.* **11**, 106-10 (1971).

The *Integrated Subject File.* I. Data Base Characteristics

W. C. ZIPPERER, R. E. STEARNS, Jr., and M. K. PARK*
Computer Center, University of Georgia, Athens, Georgia 30602

Received October 11, 1972

Characteristics of Volume 71 of the *Integrated Subject File* (ISF), the computer-readable data base corresponding to the *Chemical Abstracts Subject* and *Formula Indexes,* are reported. Minimum, maximum, and average lengths and frequency counts for the data elements in the Chemical Substance and General Subject segments of the Standard Distribution Format (SDF) files distributed by Chemical Abstracts Service are presented. Similar data are tabulated for the same files as converted for use with the UGA Text Search System and for a merged data base created from the index entries from the ISF and the bibliographic information from *CA-Condensates*.

As a part of its continuing evaluation of computer-readable bibliographic data bases, the University of Georgia (UGA) has begun a research study on the *Integrated Subject File* (ISF), the computer-readable data base generated by Chemical Abstracts Service (CAS) which corresponds to the semiannual *CA Subject and Formula Indexes.* This study is designed to evaluate alternative file organizations, data element content, and search strategies for the ISF data base, as well as combinations and comparisons with *CA-Condensates,* the corresponding bibliographic data base.[1] The project description defines seven major tasks for investigation over an 18-month period. These tasks include: creation of bibliographic data base search files, collection and characterization of questions, comparison of biblio-

* To whom correspondence should be addressed.

graphic data bases, effect of iterative profile revision, characterization of bibliographic data bases, user evaluation of searches, and comparison of nomenclature and structure search. This paper, the first in a series on this study, reports the creation of the search files, presents statistics on data base content and size, and relates general and specific characteristics of the ISF.

The information reported for the versions of the ISF as distributed by Chemical Abstracts Service is, of course, pertinent for any use of these files. While the characteristics reported are not particularly difficult to obtain, they have not yet been reported nor are they available in the ISF data base documentation. The files converted for use in the Georgia center were created to facilitate comparison of alternative search strategies, file organizations, and data contents. The conversion effort was deliberately constrained to provide files suitable for use on an available bibliographic retrieval system, the UGA Text Search System, for the evaluation studies. Analysis of the evaluation results, especially the retrieval failures, will be used to determine desirable characteristics of an optimum retrieval system for the ISF. The rationale for some of the data base modifications made during conversion is stated briefly where appropriate. The detailed discussion on data representation and associated retrieval implications will be reported in a subsequent paper on the comparative retrieval results.

GENERAL CHARACTERISTICS

The *Integrated Subject File* is distributed as semiannual collections corresponding to the semiannual volumes of the printed publications. The data base is issued in two physically separate files, the Chemical Substance (CS) segment containing the specific chemical compound index entries and the General Subject (GS) segment containing the conceptual index entries. Each is organized in inverted order according to the alphabetical collating sequence of the printed *Subject Index*. The file format is CAS' Standard Distribution Format (SDF).[2] Each physical record corresponds to one complete index entry associated with the appropriate *Chemical Abstracts* (CA) abstract number. Redundancy between records is retained.

The Chemical Substance segment contains the index entries for the specific chemical substances selected for inclusion in the *CA Subject* and *Formula Indexes*. The data content consists of the CA systematic nomenclature entries, Registry Numbers, molecular formulas, associated text modifications (the conceptual statements which describe the context of documents with respect to index entry points), CA section numbers, CA abstract numbers, sortkeys, and various control and supplemental information. A typical annotated index entry for the major data elements of the Chemical Substance segment is shown in Figure 1.

The General Subject File contains the conceptual index entries of the printed *CA Subject Index*. Entries are characterized by concept headings and by concept modifiers which may occur as qualifiers or text modifications. An annotated index entry for the General Subject File containing the concept heading and qualifying descriptors, the text modification, and the CA abstract number is given in Figure 2.

More detailed descriptions of the file contents and format can be found in the CAS documentation for the data base.[3]

CHARACTERISTICS OF THE VOLUME 71 ISF FILES

The ISF data base corresponding to volume 71 of *Chemical Abstracts* (July–December 1969) was used for analysis of data base characteristics as this was the first volume to be distributed. The first task of this study was the determination of the size and frequency characteristics of the data elements for each of the data bases to be used as these data provide the foundation for this or any other study of the ISF. These characteristics have important uses in determining the requirements of the computer software—e.g., maximum field lengths and estimation of execution times—in formulating methods of constructing search profiles, and in calculating file sizes for alternative file organizations and structures. The characteristics are reported for the ISF files as distributed by Chemical Abstracts Service as well as the converted forms of the files to be used with the UGA Text Search System in later tasks.

The data base analysis tables which follow contain information on the length attributes of both individual data elements and logical records. All of the data bases investigated are in the same basic file format. The University of Georgia chose to adopt the CAS-developed file structure for use in its own data processing activities, and, while there are a few small differences in implementation, files in either the CAS or the UGA format can be processed in the same manner and with the same utility programs. Since the file structures are compatible, the data for the CAS and UGA files can generally be compared directly. In general, the file structure consists of a variable length directory of "n" eight byte segments (8-bit bytes), each of which defines the type of data by a hexidecimal identification number, its length, relative location in the record, and storage mode. The directory is followed by a variable length data field. To clarify the nomenclature, CAS generally uses Standard File Format (SFF) to refer to its internal files which may contain information in a wide range of character sets and storage modes. Standard Distribution Format (SDF) is a proper subset of SFF in that it is the same file structure but has predefined specifications for character sets (usually upper and lower case ASCII-8) and data recording conventions—e.g., density, block size, record size, etc. The UGA implementation should probably be considered a subset of SFF also since most of the internally created data bases are in EBCDIC and conform to other recording conventions. However, none of the UGA programs preclude the use of the full SFF definition. One additional difference between the CAS and the UGA files which should be taken into consideration by programmers when using the data tables is the convention for work alignment. Data in the CAS files are aligned on double

①Acetic acid
----, ②chlorofluoro-
③methyl ester, ④(-)-③[25197-79-9]
⑥hydrolysis of, kinetics of, ⑦70042e

①Heading Parent
②Substituent
③Name Modification
④Stereochemistry
⑤Registry Number
⑥Text Modification
⑦Abstract Number

Figure 1. Typical index entry for the Chemical Substance File

①Light, ②ultraviolet, biological effects
③on *Blastocladiella emersoni*, ④58113w

①Concept Heading
②Qualifier
③Text Modification
④Abstract Number

Figure 2. Typical index entry for the General Subject File

word boundaries (IBM 360), while full alignment is used in the UGA implementation.

Each of the analysis tables contains the following statistics for the number of data elements (DE_COUNT):

COUNT—the number of records in the file with one or more data elements
ZERO_CT—the number of records with no data elements
MIN—minimum number of data elements in any one record
MAX—maximum number of data elements in any one record
AVERAGE—average number of data elements per record
STD. DEV.—Standard Deviation

Statistics given for the record lengths (TOT_LEN) are as follows:

COUNT—the number of non-zero length records
ZERO_CT—the number of zero-length records
MIN—minimum length of any one record
MAX—maximum length of any one record
AVERAGE—the average length of all records
STD. DEV.—Standard Deviation

The record lengths reported above are 8-bit bytes where the length is four bytes less than the OS record length value —i.e., the length of the OS variable length data record. The table of data elements gives the name, either the CAS or UGA hexadecimal identification number as appropriate for the file, the number of occurrences in the file (COUNT), and the length characteristics. The data element length information is taken directly from the length attribute given in the data element directory, and is, therefore, the number of 8-bit bytes required to store the data only. No control information is counted as part of the data element length, and the length does not include any padding to either double or full word boundaries.

The Volume 71 Chemical Substance File contains 428,763 records, or 428,760 index entries plus three SDF control records. Record lengths (IBM 360 OS record length) vary from 108 to 668 8-bit bytes, with an average of 238.37. The character set is the 95-character ASCII-8 representation. Table I gives the frequency and size characteristics of the data elements in the SDF version of the Chemical Substance segment as distributed by CAS. There are 23 different data elements in the Chemical Substance File, seven of which occur only in the control records. Registry Numbers are present in the Chemical Substance file for approximately 91% of the compound index entries, the remainder of the entries being primarily incompletely defined substances which were unregistered as of this time—e.g., MYSTOX LSL. The Chemical Substance segment contains only index entries for specific substances; index entries for classes of chemical substances—e.g., Phenols, Tin Compounds, Alkaloids—are contained in the General Subject file. Use of the CAS statistic of 178,135 unique Registry Numbers in the Chemical Substance segment gives an average of 2.4 citations per Registry Number for the CS file as a whole. Text modifications are present for 66% of the entries, which implies that the remainder (data element null) are entries for synthesis or that only general information was given in the original document. Individual data elements vary in length from 1 to 443 bytes, but it should be noted that as many as five data elements (Heading Parent, Substituent, Name Modification, Stereochemistry, and Line Formula) must be used to represent a compound completely.

Similar characteristics of the ISF General Subject File are given in Table II. There are 264,090 index entries repre-

Table I. Statistics for CAS ISF Chemical Substance File

Statistics for Data Element Count (DE_COUNT)

Count	Zero Ct.	Min	Max	Average	Std Dev
428763	0	4	14	9.16	1.03

Statistics for Total Length (TOT_LEN)

Count	Zero Ct.	Min	Max	Average	Std Dev
428763	0	108	668	238.37	43.75

Frequency and Size Characteristics

Data Elements	CAS ID	Count	Min	Max	Avg	Std Dev
Copyright Marking	0003/01	2	13	13	13.00	0.00
Sequence No. 1	0012/01	428763	8	8	8.00	0.00
Record Creation Date	0145/01	2	5	5	5.00	0.00
File Key Description	0149/01	2	8	8	8.00	0.00
CAS Issues Identifn.	014A/01	3	18	18	18.00	0.00
Tape Format Escape Code	014B/01	2	4	4	4.00	0.00
CA Publication Citation	0066/01	428757	12	17	14.00	0.02
CA Section/Subsection	0067/01	428760	8	8	8.00	0.00
Heading Parent	0161/01	428760	2	217	14.21	10.11
Heading Parent Sortkey	0161/02	428762	6	221	19.13	10.34
Text Modification	017D/01	282819	2	172	40.58	19.97
Substituent	0165/01	220542	3	443	29.00	21.22
Preferred Order Molform	0192/01	405642	1	71	8.89	5.38
Registry Number	0011/01	388541	9	9	9.00	0.00
Primary Publicn. Type	0295/01	108872	1	1	1.00	0.00
Name Modification	016E/01	169083	1	349	26.64	22.51
Stereochemistry	0170/01	32345	2	35	5.77	3.72
Qualifier	0171/01	107208	6	31	13.65	5.30
Functional Category	015F/01	51472	5	32	8.03	2.33
Line Formula	0169/01	18696	2	38	5.07	2.19
Homograph Definition	0160/01	310	2	45	9.83	8.00
DE ID Frequency Counts	0148/01	1	124	124	124.00	0.00
Number of File Units	0140/01	1	12	12	12.00	0.00

sented by 8339 unique Concept Headings. This gives an average of 31.7 citations per Concept Heading, but the standard deviation is large, as the number of index entries per Heading ranges from 1 to over 2780. Most of the entries (99.5%) contain text modifications.

UGA CONVERTED DATA BASES

Three physically distinct data bases were created from the ISF segments as delivered by Chemical Abstracts Service for retrieval studies using the UGA Text Search System.[4] Two of these files retain the general format, content, and organization of the original inverted ISF files, the conversion being required primarily to substitute the UGA data element identification numbers and to edit information in a few of the fields to a format compatible with other UGA data bases. The character set has also been translated from upper-and-lower case ASCII to upper case EBCDIC. The third data base created was a merged file containing the index entries from the two ISF files and the bibliographic data and keyword index entries from the corresponding *CA-Condensates* data base. This merged data base is also recorded in the UGA version of Standard File Format and is sequenced in document order on the CA abstract number.

The data bases which were constructed for search with the UGA Text Search System were designed to facilitate comparison of alternative search strategies, file organizations, and data contents. The two files which are essentially translations of the Chemical Substance and General Subject segments of the ISF contain individual records for each index entry. Thus, a document must be determined to be an answer (a hit) on the basis of the information in a single, highly pre-coordinated index entry. The document-ordered data base, on the other hand, allows post-coordination of all pertinent index entries for a given document since these are collected in one logical record. The fact that these index entries have been merged with the bibliographic information and keyword index entries from *CA-Condensates* provides an opportunity for comparing retrieval results from CA-Condensates and the ISF both singly and in various combinations. As mentioned previously, some constraints were imposed upon the creation of these search files in order to use available search software.

Characteristics of the UGA-converted ISF segments are given in Tables III and IV. Control records and their associated data elements were dropped during the conversion, decreasing the over-all size of the files. Preliminary analysis of the ISF indicated no apparent way to make use of either the control or sortkey data elements during retrieval. The Publication Type Code, which occurs in the CA version of the ISF files only for patents, books, and reviews, has been added for the remaining records (assumed journal since technical reports and conference proceedings cannot be distinguished from the available information). The search system requires explicit information in the record to effect a match—i.e., a "null" data element cannot be specified as a search requirement. A citation to the secondary publication has also been added to each record in a form suitable for printing on retrieval results—e.g., CHEM. ABSTS. 71(24) 116998B. One data element has been constructed for the chemical nomenclature, concatenating the five CAS data elements as necessary with intervening "comma blank" delimiters—e.g., GOLD(II), DIIODOBIS (O-PHENYLENEBIS(DIMETHYLARSINE))-, IODINE, TRANS-,. The decision to create a single data element for the nomenclature was based on previous studies of the *CA Formula Index* (volume 69) where retention of the separate data elements caused excessive redundancy in the search profile terms and contributed to both precision and recall failures.[5] The length of chemical names in this data element ranges from 2 characters to a maximum of 454 characters, considerably less than the sum of the maximums of the five separate segments in the CAS file. The remaining data elements in the Chemical Substance segment are unchanged from the original. The only significant change in the converted General Subject segment

Table II. Statistics for CAS ISF General Subject File

Statistics for Data Element Count (DE_COUNT)

Count	Zero Ct.	Min	Max	Average	Std Dev
264093	0	5	9	6.49	0.62

Statistics for Total Length (TOT_LEN)

Count	Zero Ct.	Min	Max	Average	Std Dev
264093	0	92	332	173.88	25.31

		Frequency and Size Characteristics				
Data Elements	CAS ID	Count	Min	Max	Avg	Std Dev
Copyright Marking	0003/01	2	13	13	13.00	0.00
Sequence No. 1	0012/01	264093	8	8	8.00	0.00
Record Creation Date	0145/01	2	5	5	5.00	0.00
File Key Description	0149/01	2	8	8	8.00	0.00
CAS Issues Identifn.	014A/01	3	18	18	18.00	0.00
Tape Format Escape Code	014B/01	2	4	4	4.00	0.00
CA Publication Citation	0066/01	264091	12	16	14.00	0.01
CA Section/Subsection	0067/01	264090	8	8	8.00	0.00
Concept Heading	0151/01	264090	3	60	11.62	5.94
Concept Heading Sortkey	0151/02	264092	8	58	16.17	5.41
Text Modification	017D/01	262825	1	179	43.24	19.29
Primary Publicn. Type	0295/01	48938	1	1	1.00	0.00
Homograph Definition	0160/01	366	4	21	6.72	3.77
Qualifier	0171/01	78815	1	102	12.43	7.07
Functional Category	015F/01	1391	6	21	8.54	1.07
DE ID Frequency Counts	0148/01	1	76	76	76.00	0.00
Number of File Units	014C/01	1	12	12	12.00	0.00

is the addition of a six-digit numeric Concept Heading Code, assigned on the basis of the concatenation of the Concept Heading and all associated Homograph Definition, Qualifier, and Functional Category data elements. Record size considerations for the merged *CA-Condensates*-ISF data base precluded carrying the full index entries into this file. Consequently, an identification number corresponding to the major components of each index entry was assigned for subsequent use in the merged file. It was felt that the Homograph, Qualifier, and Functional category data provided important modifiers to the very generic Concept Headings, so pre-coordinated index terms were constructed on this basis prior to code assignment. A total of 10,271 unique codes were assigned to the 264,052 General Subject index entries based on this criterion, giving an average of 25.7 entries per concept heading code for the General Subject segment.

The *CA-Condensates*-ISF data base was constructed by merging the bibliographic information from *CA-Condensates* with the encoded representations of the index entries from the two ISF segments. The *CA-Condensates* data base used for this purpose was the UGA Standard File Format versions corresponding to the odd-numbered and even-numbered sections of *Chemical Abstracts,* which were sorted into CA abstract number order and merged prior to addition of the ISF entries. Because of constraints within the logic capabilities of the UGA Text Search System for which this particular file was generated, it was not possible

Table III. Statistics for UGA ISF Chemical Substance File

Statistics for Data Element Count (DE_COUNT)

Count	Zero Ct.	Min	Max	Average	Std Dev
428760	0	4	12	9.98	0.87

Statistics for Total Length (TOT_LEN)

Count	Zero Ct.	Min	Max	Average	Std Dev
428760	0	100	620	238.81	37.37

		Frequency and Size Characteristics				
Data Elements	UGA ID	Count	Min	Max	Avg	Std. Dev
Publication Type Code	0009/01	428760	1	1	1.00	0.00
Secondary Citation	0020/01	428669	27	27	27.00	0.00
Secondary Journal Volume	0023/01	428669	2	2	2.00	0.00
Secondary Journal Issue	0024/01	428669	2	2	2.00	0.00
Sec. Journal Abstract No.	0026/01	428669	7	7	7.00	0.00
CAS Section	003A/01	428760	6	6	6.00	0.00
Registry Numbers	0038/01	428760	9	9	9.00	0.00
Compound Name	0067/01	428760	2	454	42.34	31.47
Text Modification	007F/01	282810	2	172	40.56	19.98
Molform Print	0041/02	405642	1	71	8.89	5.38
Qualifier	007E/01	107208	4	31	13.65	5.30
Functional Category	0070/01	51472	2	32	8.03	2.33
Homograph Definition	007D/01	310	2	45	9.83	8.00

Table IV. Statistics for UGA ISF General Subject File

Statistics for Data Element Count (DE_COUNT)

Count	Zero Ct.	Min	Max	Average	Std Dev
264090	0	5	11	9.30	0.47

Statistics for Total Length (TOT_LEN)

Count	Zero Ct.	Min	Max	Average	Std Dev
264090	0	68	348	200.38	23.66

		Frequency and Size Characteristics				
Data Elements	UGA ID	Count	Min	Max	Avg	Std Dev
Publication Type Code	0009/01	264090	1	1	1.00	0.00
Secondary Citation	0020/01	264052	27	27	27.00	0.00
Secondary Journal Volume	0023/01	264052	2	2	2.00	0.00
Secondary Journal Issue	0024/01	264052	2	2	2.00	0.00
Sec. Journal Abstract No.	0026/01	264052	7	7	7.00	0.00
CAS Section	003A/01	264090	6	6	6.00	0.00
Concept Heading	007B/01	264090	3	60	11.62	5.94
Text Modification	007F/01	262816	1	179	43.24	19.29
Concept Heading Code	0087/01	264090	6	6	6.00	0.00
Homograph Definition	007D/01	366	4	21	6.72	3.77
Qualifier	007E/01	78815	1	102	12.43	7.07
Functional Category	007C/01	1391	6	21	8.54	1.07

Table V. Statistics for UGA Merged *CA-Condensates*–ISF Vol. 71 File

Statistics for Total Length (TOT_LEN)

Count	Zero Ct.	Min	Max	Average	Std Dev
130105	0	260	3628	531.77	116.51

		Frequency and Size Characteristics				
Data Elements	UGA ID	Count	Min	Max	Avg	Std Dev
Primary Citation	0001/01	130105	22	190	57.99	15.16
Publication Type Code	0009/01	130105	1	1	1.00	0.00
Secondary Citation	0020/01	130105	38	54	38.67	3.20
Title	0030/01	130105	3	297	74.32	35.20
Authors, Inverted	0032/01	126743	2	175	24.21	14.21
Location of Work	0037/01	125788	2	235	33.69	16.10
Free Index Terms	0030/01	130105	5	299	49.38	22.02
Primary Document Language	0007/01	4622	2	11	3.25	0.72
Document Serial Number (TAN)	0011/01	5444	9	9	9.00	0.00
CAS Coverage	001D/01	5443	16	16	16.00	0.00
CAS Section Group	001F/01	5444	1	1	1.00	0.00
CAS Section	003A/01	5444	6	6	6.00	0.00
Document Availability	0010/01	612	14	97	47.86	14.20
Title, Foreign Language	0031/01	9	28	98	55.78	19.01
Author Name, Corporation	0034/01	715	9	137	28.40	15.81
Concept Heading Code	0087/01	114860	7	255[a]	16.00	10.65
Registry Numbers Qualified	0088/01	90407	10	3000[a]	69.80	101.63

[a] Artificially imposed maximums. See text for discussion.

to retain the hierarchical association of the various indexing levels—e.g., Concept Heading, Qualifier, Functional Categories, Text Modifications, etc.—for each individual index entry when these entries were merged with *CA-Condensates*. The choice was made to adapt the data base content to the available retrieval system for the evaluation studies, rather than to develop an entirely new search system for which the characteristics and long range need are as yet unknown. The Concept Heading Codes, which were assigned to the concept index entries on the basis of the Concept Heading plus associated qualifiers, functional categories, and homograph definitions, were concatenated into a single data element for each document record, separated by semicolons—e.g., (CHC010110;008502;000-017;004616;). Registry Numbers were treated similarly, except that qualifiers, functional categories, and homograph definitions were appended to appropriate Registry Numbers prior to creation of the single data element per document record—e.g., (RGN)007782414, PROPERTIES;023968362;-023868384;. The modifiers were appended to the Registry Numbers for potential use as a precision device in the post-coordinate searches. A total of 26,779 pseudo "Registry Numbers" were assigned to the 40,219 ISF compound file entries for which no CAS Registry Numbers were available to carry these index entries into the merged file. Characteristics for 17 of the 32 different data elements in the merged data base are given in Table V; the remaining data elements are primarily individual items of the primary and secondary document citations—e.g., Secondary Journal Volume "71"—which are carried in the file individually as well as combined in the print forms of the citation records—e.g., Secondary Citation = CA-CONDENSATES (CHABA8), 71(18) 084425U. The merged file created for the evaluation studies contains 130,105 document records. During the merge operation, 324 ISF index entries had to be dropped because there was no corresponding *CA-Condensates* record, and 91 ISF Chemical Substance entries were dropped because the CA abstract numbers were missing. There were 242 *CA-Condensates* records for which there were no ISF index entries. It also appears that a block of records is missing from one issue of the *CA-Condensates* tapes used as a source for the merged file, the reason for which is unknown at this time. The absence does not significantly affect the comparison study results, but it would be of importance in an operational environment. Some 58.0% (75,404) of the merged file documents contain index entries from both ISF segments, 30.3% (39,456) contain only General Subject entries, and 11.5% (15,003) contain only Chemical Substance index entries. The number of Registry Numbers per document (including the pseudo-Registry Numbers) varies from 1 to 541, a maximum which was not anticipated when the maximum length of the variable length data element in the UGA merged file was set at 3000 characters. The maximum length would have to be at least 5410 characters to accommodate the maximum number of unqualified Registry Numbers per document as stored in the UGA format. Registry Number counts in excess of 200 are contained in 15 documents. The average number of Registry Numbers per document is 4.63. The range in the number of Concept Heading Codes per document as assigned is 1 to 99, with a mean of 2.29 and standard deviation of 1.56. The assumed maximum data element length of 255 characters is also too low since a length of 693 would be required for the record with 99 Concept Heading Codes. Twenty-five documents have Concept Heading Code counts in excess of 20, three of which have over 50. The over-all mean on the number of index entries per document in the file is 5.25 with a standard deviation of 7.58.

Additional information for these data bases which has been tabulated or graphed includes the frequency distribution of documents with respect to CA section numbers; the frequency distribution of Concept Headings, Concept Heading Codes, Qualifiers, Functional Categories, and Registry Numbers with respect to CA sections; the frequency of index entries with respect to Concept Headings and Concept Heading Codes; and the correlation of individual Concept Headings with respect to individual CA sections. These data are presently being analyzed in conjunction with other vocabulary-related information for potential application to file partitioning, profile construction, and similar vocabulary-related tasks.

Table VI. Computer Processing Times for Principal Runs

	CPU (Min:sec)[a]	Elapsed (mins.)[a]
ISF-Chemical Substance Conversion	342:20	397.29
ISF-General Subject Conversion	217:40	246.69
CA-Condensates Merge	7:45	42.09
CA-ISF Merge	177:44	199.21

[a] IBM 360/65 MVT with HASP.

COMPUTER RUNS

All of the computer runs made as part of this study were done on the University's IBM 360/65, operating under OS MVT with HASP, in high speed core. All programs were written in PL/1 Level F, using numerous macro and subroutine facilities which are equally suitable for either SDF or SFF. Run times in terms of both CPU and elapsed time for the major computer jobs are given in Table VI.

ACKNOWLEDGMENT

The authors acknowledge the cooperation of Chemical Abstracts Service in making available some of the supplemental resources used in this portion of the ISF study. Sincere thanks also go to Howard Petrie, of Sheffield, England, who undertook much of the preliminary analysis of the General Subject segment of the ISF while a Visiting Foreign Scientist at the University of Georgia.

LITERATURE CITED

(1) "Evaluation of Alternative Retrieval Techniques for the Integrated Subject File," Computer Center, University of Georgia, Athens, Ga., 1971.
(2) "Standard Distribution Format Technical Specifications Revised," Chemical Abstracts Service, Columbus, Ohio, 1971.
(3) "Data Content Specifications for the CA Integrated Subject File in Standard Distribution Format," Chemical Abstracts Service, Columbus, Ohio, 1971.
(4) "UGA Text Search System," 4 Volumes, Computer Center, University of Georgia, Athens, Georgia, January 1971.
(5) Park, M. K., Carmon, J. L., and Stearns, R. E., Jr., "Chemical Compound Retrieval Based on CA Formula Index Nomenclature," Computer Center, University of Georgia, Athens, Georgia, July 1971.

Production of Printed Indexes of Chemical Reactions. I. Analysis of Functional Group Interconversions

R. CLINGING and M. F. LYNCH*
Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, England

Received November 27, 1972

A set of programs is being developed for the purpose of producing printed indexes of chemical reactions from a simple reactant/product data base. A program is described which identifies functional group interconversion reactions, hydrogenations, and dehydrogenations in a data base containing structures encoded as Wiswesser Line Notations. These reactions account for about 20% of a sample of 5104 reactions. Production of the data base is briefly described.

Because of the great variety of organic reactions, indexing them is a complex process and although several methods have been evolved, none fully solves the problems. The least logical approach is to name the type of reaction after the author who first reported it—e.g., Claisen Condensation and Diels-Alder Reaction—a method which has been much used in the literature.[1-3] As this method presupposes a knowledge of the name of the reactions, its usefulness, especially for general searching, is severely limited, and so it was not considered suitable for computer indexing. A second, more systematic, approach is to base the name of the reaction on the functional groups which are different in the reactant and product molecules. This approach comes close to that used by most chemists when searching outside their own specialized fields and so deserves special consideration. Patterson and Bunnett[4] have taken this a stage further by combining the names of the functional groups involved to form reaction names, but this becomes unwieldy in all except the simplest cases. A third approach[5,6] looks at the bonds which change during the reaction; although this is amenable to computerization,[6,7] it produces indexes which are still difficult to search.

A number of systems[7-10] have been, or are being developed, for the automatic or semiautomatic searching of various types of chemical reaction data bases. In addition, Corey *et al.*[11-15] are developing systems to predict the best ways of tackling complex syntheses. Still lacking is an automatic method for indexing reactions, especially with a view to producing easily used printed indexes. The first stage of an attempt to solve this deficiency follows.

Chemical reactions are analyzed by comparing the records of the reactant and the product molecules, and seeing which entities are changed. A more thorough approach would need to take into account neighboring groups which may affect the course of the reaction, but this is outside the scope of the present work. Initially it was intended to approach the problem using structures encoded as connec-

* To whom correspondence should be addressed.