searchers, the types of search topics, and so on.

REFERENCES AND NOTES

(1) Rollins, G. "Some Economies of Online Searching", *Public Libr. Q.* **1983**, *4*, 13-18.
(2) Bellardo, T. "Scientific Research in Online Retrieval. A Critical Review", *Libr. Res.* **1981**, *3*, 187-214.
(3) Hawkins, D. "Online Information Retrieval Bibliography, Fourth Update", *Online Rev.* **1981**, *5*, 139-182.
(4) Kruse, K. W. "Online Searching of Pharmaceutical Literature", *Am. J. Hosp. Pharm.* **1983**, *40*, 240-253.
(5) Flynn, T.; Holohan, P. A.; Magson, M. S.; Munro, J. D. "Cost Effectiveness Comparison of Online and Manual Bibliographic Information Retrieval", *J. Inf. Sci. Principles Practice* **1979**, *1*, 77-84.
(6) Almond, J. R.; Nelson, C. H. "Improvements in Cost Effectiveness in On-Line Searching. I. Predictive Model Based on Search Cost Analysis", *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 13-15.
(7) Almond, J. R.; Nelson, C. H. "Improvements in Cost-Effectiveness in On-Line Searching. II. File Structure, Searchable Fields, and Software Contributions to Cost-Effectiveness in Searching Commercial Data Bases for U.S. Patents", *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 222-227.
(8) Buntrock, R. E. "Cost-Effectiveness of On-Line Searching of Chemical Information: An Industrial Viewpoint", *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 54-57.
(9) Michaels, C. J. "Searching CA Condensates On-Line vs. the CA Keyword Indexes", *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 172-173.
(10) Buckley, J. S., Jr. "Planning for Effective Use of Online Systems", *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 161-164.

IDC Inorganic Chemicals Data Base. 2. Utilization of Chemical Abstracts Service Data Bases for the IDC Inorganic Chemistry Documentation System

FRITZ EHRHARDT* and HORST ROSCHKOWSKI

Fachinformationszentrum Chemie G.m.b.H., 1000 Berlin 12, Federal Republic of Germany

Received December 3, 1984

IDC (Internationale Dokumentationsgesellschaft fuer Chemie m.b.H., Sulzbach, Federal Republic of Germany) makes use of machine-readable data provided by Chemical Abstracts Service (CAS) for supplementing its inorganic chemistry documentation system. Data from CAS (formerly CASIA and CACON and now offered as CA Search) are brought into the format of the IDC documentation system by machine procedures and can be corrected, supplemented, or deleted intellectually if necessary. The procedures involved in this process are described. In addition to the IDC inorganic data base containing bibliographic data, inorganic compounds, their reactions, and nonstructural index terms (concepts), supplementation of the IDC GREMAS file, which correlates CA Registry Numbers with encodings used for inorganic or organic compounds, is carried out by IDC. This file enables IDC to encode a compound once only at its first appearance in the system. Furthermore, CA Concept Headings are transferred intellectually only at their first appearance into Alpha Numbers of the IDC inorganic chemistry documentation system and on further appearances are generated by computer procedures.

Since 1973, the IDC (Internationale Dokumentationsgesellschaft für Chemie m.b.H.) has made use of a special documentation system for inorganic chemistry, the construction of which was described in Part 1 of this series.[1] At the beginning, only patent information was searchable in the data base. This deficiency was criticized by many users. After the economical alternatives for extending the IDC Inorganic Chemicals System were examined, it was decided to supplement the inorganic data base by existing machine-readable versions of data bases of Chemical Abstracts Service (CAS). CAS data bases are made available for evaluation to IDC by an agreement concluded with CAS in 1975. It was the aim to employ the CAS data bases in a manner that would allow, in the final result, the searching in and construction of an inorganic chemicals data base build up analogously to that of the patent data base for inorganic chemicals.

The intellectual effort in supplementing the inorganic chemicals data base as well as the cost for additional work (keyboarding, data processing) was to be kept at a minimum. The unequivocal characterization of chemical substances in the CAS system by CA Registry Numbers made it possible that each chemical substance had to be encoded intellectually only once at the time of its first appearance and could subsequently be found in the data file. This single encoding of inorganic substances was to be supplied in the desired format to the GREMAS Register of IDC (see BMFT Research report of Schwier[2] and Jungblut[3] on the usage of "Data of the Chemical Abstracts Service (CAS). Structural Data Base." 1979).

In order to examine the extent to which the CAS data base could be employed in creating a data base for the IDC search system, a study was conducted with the CAS tape service CA Subject Index Alert (CASIA), containing all entries necessary for constructing a semiannual CA index. In this study, only those entries were examined that had been made within a certain period for CA Section 49 (Industrial Inorganic Chemicals). The study showed that about one-third of the substance and concept information considered to be important for storage in the inorganic chemicals data base was to be found in the CA data base. Intellectual evaluation of the texts belonging to Registry Numbers or Concept Headings furnished a further third of information. The remaining information was taken from the corresponding CA abstracts.

It was concluded that CASIA was suited for supplementing the inorganic data base but not without additional intellectual evaluation of CA abstracts. Some major reasons for this are as follows:

(1) In general, there are no CA index entries for reactants in inorganic synthesis, even when mentioned in CA abstracts. Since this information is asked for by the users of the IDC inorganic data base, it has to be added to the file. In many cases, reaction products with fractional indexes have no index entries (no own Registry Numbers) in the CA system and thus have to be added to the IDC inorganic file intellectually.

(2) The encodings of solid solutions and mercury alloys are just two further examples for differences between the CA and IDC systems, making supplementations necessary in constructing the IDC inorganic data base. Solid solutions of

0095-2338/86/1626-0063$01.50/0

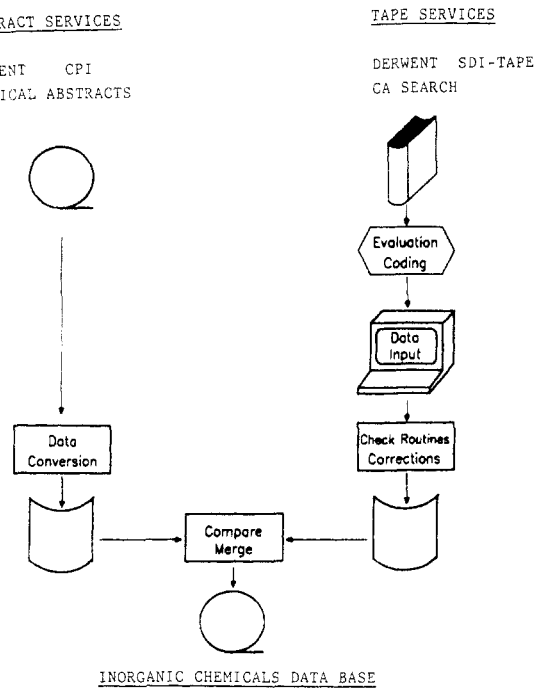

Figure 1. Construction of the IDC Inorganic Chemicals Data Base.

Zn–Cd sulfides, for example, are encoded in the CA system by the Registry Number of the constituents (ZnS, CdS), whereas the IDC inorganic data input and retrieval system allows encoding and retrieval of solid solutions by a more or less general formula, e.g., Zn 0.3 Cd 0.7 S for a defined solid solutions and Zn0Cd0S0 for an undefined one. In CA indexes, information on amalgam is generally found at the Registry Numbers of the Hg-free parent alloys. In the IDC inorganic data base, amalgams are encoded by a molecular formula, containing the element symbol Hg.

(3) CA makes use of General Concept Headings such as Electrolysis. Only general information on electrolysis can be found in CA indexes at this place, while information on special electrolytic processes (e.g., electrolytic NaOH production) is not offered. This information has to be added to the data file by intellectual evaluation of the CA abstract.

(4) In some cases, indexing depth in CA is not extensive enough for IDC purposes even though corresponding Concept Headings are available. For example, for articles dealing with "flotation" or "extraction", there may be no index entries at the corresponding Concept Headings.

A second CAS data base available for the construction of an inorganic data base was the tape service of Chemical Abstracts Condensates (CACON). This data base offered bibliographic data. Combination of the data of CACON and CASIA was achieved by employing the CA abstract number. Merging the tape services CASIA and CACON by CA to yield the new tape service CA Search made an adjustment of file-building software necessary.

DATA BASE

Input and output are part of the IDC inorganic documentation system. The connection link between these two parts is the IDC inorganic data base. Construction of this data base can be achieved without supplementing existing machine-readable data bases (Figure 1). The data base of Chemischer Informationsdienst[2] was constructed in this manner, supplementing the centrally erected data base with information of special interest to the users of the system. A complete data input by keyboarding is avoidable if the literature to be evaluated is accompanied by machine-readable data and if part of the information is available by machine-type transfer.

In order to evaluate abstracts of the *Central Patents Index* of Derwent Ltd., bibliographic data were selected in machine-readable form, and data corresponding to compounds and concepts were stored by data input by keyboarding after intellectual encoding.

By using existing CA files, the data corresponding to compounds and concepts can, after intellectual examination, be transferred by computer procedures into the IDC inorganic data base. These data can be supplemented by intellectual encoding.

The format of the inorganic chemicals data base is a version of the standard format used in searching with the retrieval system SIR (Serial Information Retrieval),[4] also employed in other magnetic tape services—CA Search, Chemical Industry Notes, etc. This format has been in use without fundamental changes since 1967, thus having the advantage that auxiliary and service programs, e.g., those for corrections, selections, keeping archives, and statistical evaluations, are generally employable. The data base has a normal file organization. All data from a literature evaluation are stored in a serial-type manner. The abstract numbers in the printed abstract services are used for ordering within the data base. The individual volumes are stored in separate files.

The data obtained from a literature evaluation (documentation unit) are stored in the form of "records", which are character strings represented by a read command in the work program. Each record consists of an identifier of four characters, which characterizes categories (e.g., document title) and possible record seqences of the same category (e.g., author names). The records of a literature evaluation are ordered by the identifier.

As far as possible, machine-readable CAS data are employed directly, such as bibliographic data, section/subsection numbers, and contexts of abstract-related concept information in the form of Concept Headings (concepts in *Chemical Abstract Index Guides*). Since CA Concept Headings correspond to Alpha Numbers (Thesaurus concept codes) of the IDC inorganic data base, a data file was planned that established a concordance between Concept Heading and Alpha Number.

Usage of such controlled vocabularies as Concept Headings and definite CA Registry Numbers for defined substances makes it possible that concepts and substances have to be encoded intellectually only once at the time of their first appearance. When they occur again, the stored encoding can be obtained from the data files.

At the same time, these data files receive the data input via keyboarding conducted on floppy disks. Printouts serving as manuscripts for encoding and keyboarding are produced. Merging the machine-readable data with that of the intellectual encoding is done with an entry number found in the printout.

Programming work was eased by application of already existing programs, e.g., those formerly used in constructing the IDC inorganic data base from Derwent CPI for complicated check routines and data conversions in generating standardized molecular formulas and linear structural formula with a range of figures in substance encoding. Standard software for work files, employed for years and offering auxiliary programs for selecting, printing, correcting, and keeping archives, was also helpful.

As already stated, a supplementation of the CAS data based on the printed CAS abstracts is necessary. Since no entry number and no corresponding machine-readable data are present, these supplements cannot be made on the manuscript printouts. Therefore, a completely separate indexing of data of the supplementation was planned, whereby the new data were fed into the data base with the help of the CAS Abstract Number. The supplements introduced are System Numbers,

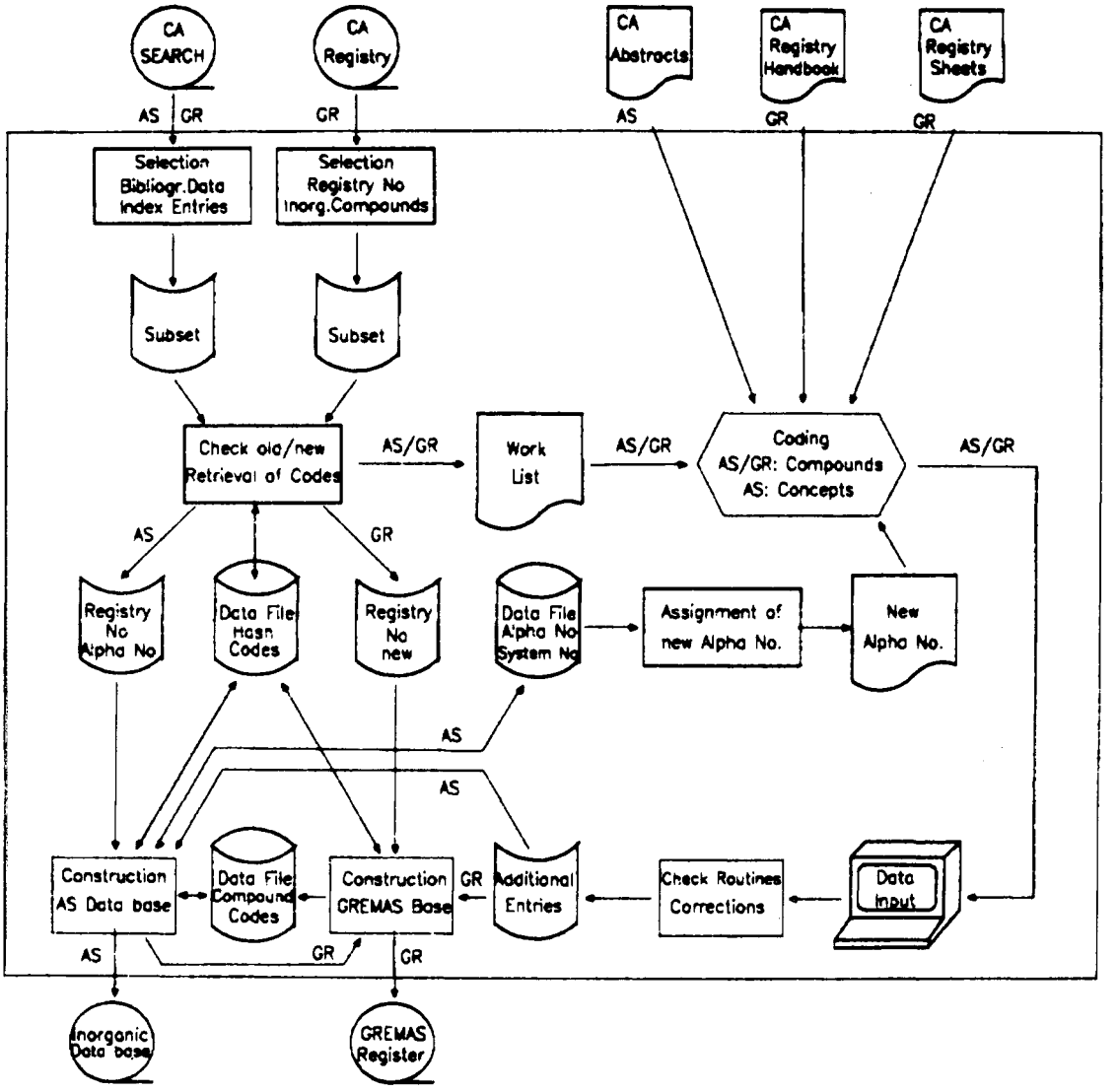

Figure 2. Flow chart illustrating the inorganic chemicals data base and the GREMAS Register.

abstract-relevant concepts, and complete substance encodings, which do not contain a CA Registry Number in contrast to data files derived from CAS material.

The delivery of data to the substance and structure data base (GREMAS Register) of the IDC is controlled by the data file. When appearing for the first time, the substance encodings are converted into the desired data base format and then delivered as update. A selection of Registry Numbers from the CA registry data base serves as a further initial file in this process and is produced as follows: for organic substances, a GREMAS encoding is produced automatically from the topological structure encoding of the CA registry data base. At the same time, inorganic compounds are selected and their Registry Numbers withdrawn for the purpose of transfer to inorganic compound encoding. This data base does not contain further information, i.e., compound name and linear formula, necessary for encoding. This information is made available to the encoder by the CAS *Registry Handbook* and the microform of the Registry Sheet.

DATA FLOW

The scheme shown in Figure 2 illustrates the steps necessary in the construction of the inorganic chemicals data base as well as the GREMAS Register labeled with AS and GR, respectively. The magnetic tape service CA Search and the selection of Registry Numbers from the CA registry data base serve as input. *CA Abstracts*, *Registry Handbook*, and copies of certain Registry Sheets are available in printed form for intellectual evaluation.

Table I. Data Elements from CA Search

identifier no.	data element name	type of processing[a]
0055	coden	M
0056	publication classification code	M
0059	personal author name	M
005B	document title	M
005D	abbreviated title of publication	M
005F	series/volume number	M
0060	issue/report/part number	M
0061	pages	M
0063	original language code	M
0065	price	M
0066	CA publication citation	M
0067	CA publication section/subsection	M
0508	date of meeting or volume	M
0074	name of publisher	M

[a] M = automatic processing.

DATA FLOW IN THE INORGANIC CHEMICALS DATA BASE

Work Lists. Data elements of the data base CA Search that are relevant for processing are listed in Tables I and II. Additionally, the processing type of the data elements is noted. Therein, M stands for automatic transfer, I for intellectual evaluation at every appearance, and I/A for a single intel-

Table II. Data Elements from CA Search

identifier no.	data element name	code in work list	type of processing[a]
0011	registry number	REG	M
0066	publication citation	CAN	M
0067	publication section/subsection	SEC	M
0151	concept heading	CTH	I/A
015F	functional category	CAT	I/A
0160	homograph definition	HOM	I/A
0161	heading parent	PAR	I/A
0165	substituent	SUB	I/A
0169	line formula	LIN	I/A
016E	name modification	MOD	I/A
0171	qualifier	QLF	I/A
017C	stereochemistry	STE	I/A
017D	text modification	TMD	M/I
0192	preferred order molform	FOR	I/A

[a] M = automatic processing. I = continuous intellectual evaluation. I/A = intellectual evaluation for the data file.

lectual evaluation only at its first appearance in the data file.

For the production of profile services, the bibliographic data of CA Search are transferred into the standard format of the retrieval system SIR. From this data base, bibliographic data of all literature units are selected, meeting the following conditions: section number 19 (Fertilizers, Soils, and Plant Nutrition), or section number 49 (Industrial Inorganic Chemicals), or section number 67 (Catalysis and Reaction Kinetics), or section number 78 (Inorganic Chemicals and Reactions) and type of publications unlike P (no patent literature). The conditions for selection are flexible. By input of the CA Abstract Number, nonautomatically selectable entries from sections can be assigned for processing, and already selected ones can be ignored.

The data are stored in a pending file, from which information is, prior to erasing, continuously retrieved for the construction of an inorganic chemicals data base. The format of this file corresponds to that of the inorganic chemicals data base. Intellectual manipulations are not necessary.

Entries without Concept Headings or Registry Numbers are selected by the same criteria. In the case of alloys, an additional entry containing the so-called Functional Category "Base" has to be made, since CAS indexes alloys by permutating the components, whereby the main component is labeled as "Base", thus avoiding multiple encodings.

Relevant entries are checked against a data file, yielding for every information unit (Concept Heading or Registry Number) the following information: present in the system or not; if present, the abstract it appeared in first, the processing state, and the number of appearances. In the case of Concept Headings, the processing of which has been concluded by assignment of an Alpha Number, the data file itself makes the Alpha Number available.

The entries of the data file are of a very condensed nature and consist of only 15 characters. Therefore, it is impossible to store and retrieve encodings from this data file. This is made possible at a later stage by means of a second data file.

Due to a lack in space, the text of Concept Headings can also not be stored. Instead of a text, a Hash Code is stored. This is a text-specific, condensed representation gained by application of an algorithm. In this fingerprint-type identification, the text cannot be regenerated from the code; this transformation can only be achieved by a text/code sequence.

For further intellectual evaluation, lists are printed during work routine. On these lists, it is noted for substances whether an encoding has to be made or if an encoding is already available. In special cases it is stated that a substance has already appeared but that for special reasons, e.g., inaccurate information, encoding was not possible and had to be conducted now. For Concept Headings, an already present Alpha-Number will appear in the printout or, otherwise, the reference "NEW" or "IN PROCESS". The lists consist of two parts: the left-hand side contains the information printout, while on the right-hand side handwritten encodings can be made.

For substances marked as NEW in the work list, the encoder has to list, next to the general concept code and a possible mark for special substance classes, the linear structural formula at the right-hand side of the list. Writing conventions are kept at a minimum. A program standardizes the input form with respect to void spaces between single elements and expressions within brackets and prints the index number 1 when element symbols were written without an index. Only general concept codes are assigned for substances labeled as "OLD".

In cases, where Alpha Numbers cannot be assigned automatically to a Concept Heading, this is achieved by intellectual effort. The assignment of Alpha Numbers is conducted centrally. They are regularly fed into data files for Alpha Numbers and System Numbers. Update versions of the concept lists are made available in printed form to all editors.

After examining the CA abstracts, the editor decides if the storage of additional information is necessary, especially with respect to the assignment of System Numbers and the encoding of substances serving as starting material in reactions. With the help of standardized commands, the editor can delete certain entries or even whole abstracts not of interest to the Inorganic Chemicals Data Base.

CHECKING AND CORRECTING

Reading in of data stored on floppy disks, checking, and correcting are conducted with programs used for processing patent literature. These programs have to be adjusted to the additional procedure. Deletions, replacements, and insertions are possible.

The data are checked for completeness and plausibility. The number of routine-based checks amounts to 66, consisting mainly of routines for generating standardized molecular formulas and ranges of figures for linear structural formulas.

Typical error messages are, for instance, element symbol missing, concept heading code invalid, invalid mark, brackets unpaired, index missing, and error in the index for the range of figures.

CONSTRUCTION OF THE INORGANIC CHEMICALS DATA BASE

This step is conducted by several programs operating in series. Initially, both work files, i.e., those for encoding of work lists and additonal index entries being free of formal errors after checking and correcting, are separately converted into the desired file format. The line format of the keyboarding consisting of records of fixed length is lost in this process, whereby two work files with records of variable length and record identifiers are formed.

Processing of the encoded work list is more complicated, whereby the entry number is compared with the work file, which was listed in the printout of the work list. From this file, data such as Registry Numbers and retrieved Alpha Numbers are used.

At the same time, the data file is updated by storage of current information. New Alpha Numbers are read into the file, when necessary in the production of concept heading lists and in the printout of relevant hits during the search for substitution of Alpha and System Numbers by clear text.

As already stated, the encoding for substances (standardized molecular formula and linear structural formula, possibly with range of figures) cannot be stored in the condensed data file due to lack in space. For this purpose, a second text data file is employed, which is also accessible by CA Registry Number.

At this processing state, information is retrieved from as well as stored in the file: for substances labeled as OLD in the work list and to be encoded only by record identifier, the substance encoding is retrieved via Registry Number and stored. Substances labeled as NEW in the work list are encoded by the editor on the basis of the linear molecular formula, whereby the checking program generates the standardized molecular formula and the range of figures. These data, together with the Registry Number, are read into the text data file and thus made available for recall. The encoding of a chemical substance already known to the system replaces the old encoding. Both states of encoding are recorded, making easy corrections possible.

At present, three files are available having the format of the Inorganic Chemicals Data Base and being arranged sequentially according to abstract number: file with bibliographic data; file consisting of the encoding of the work lists and the corresponding data of the work file with supplementation from the text data file; file with additional index entries. These three files are merged.

During this merging step, the record identifiers for Concept Headings and chemical substances are numbered consecutively. The date is added to the final record generated.

The data are now complete and processed batchwise for SDI. Subsequently, merging with the overall data base is conducted, making retrospective searches possible.

DATA FLOW IN THE PRODUCTION OF THE GREMAS REGISTER

As already stated, inorganic chemicals are also stored in the IDC-GREMAS Register, which correlates Registry Numbers with GREMAS Codes (for organic compounds) or STRUFO and SUFO (as defined in Part 1 of this series for inorganic compounds). Processing here is conducted generally parallel to the construction of the Inorganic Chemicals Data Base. The same data files are employed. The programs are identical, and the same search routines are applied.

In order to complete the IDC GREMAS Register more rapidly, an additional selection of CA Search data serves as input under the following conditions: records with CA Registry Number that have either the functional category "BASE" or a molecular formula without the element carbon.

Only substances appearing for the first time are treated here. The linear structural formula is encoded and, if necessary, supplemented by clear text and an identifier. Keyboarding, checking, and correcting are conducted as in the Inorganic Chemical Data Base.

PROGRAMMING TECHNIQUES

Assembler language is used for all of the 28 programs employed.

STRUCTURE OF THE DATA FILES

The master files used in the system are disk files of the indexed sequential access method (ISAM) and of a block size of 2048 bytes. File-update programs (for correcting, statistics, reorganization) exist for the master files.

TEXT-FREE DATA FILE

The purpose of this file is to check whether or not a CA Registry Number or a Concept Heading is to be processed for the first time. Since the file has to be addressed at various locations within the system, attempts were made to keep record length as short as possible in order to save storage capacity and run time.

The records have a length of 15 bytes and the following structure:

1 5	6 9	10 12	13	14 15
Key	abstract number Alpha Number	counter	status	run number

Remaining unchanged for substances and Concept Headings are the following fields:

6–9, Abstract Number. The volume number (two digits) and the abstract number (six digits) are packed decimals without sign and of a length of 4 bytes: CA 086-10-057499 as

6			9
86	05	74	99

10–12, Counter. The number of recalls of a record from the master file is stored as packed decimal: 158 as

10		12
00	15	8C

13, State of Processing (Status Byte). Each bit position has a significant digit that is marked in the case of relevancy:

position	meaning
0	record written again (unprecise information)
1	concept heading record
2	substance not encoded as inorganic compound (organic compound)
3	substance encoded as inorganic compound (stored in text data file)
4	substance additionally encoded according to GREMAS
5	functional category "base" (=alloy)
6	functional category "compound" (=normal compound)
7	functional category "others" (=special compound classes)

Positions 3 and 4 are set and the byte is stored as

$$\frac{13}{18}$$

for substances encoded as inorganic compound and additionally as organic compound by GREMAS and the processing of which is concluded.

14–15, Run Number. For every program run and for every program in which the master file is opened, the run number entry is increased by 1. This procedure is necessary in order to repeat a run in case of technical failure. Otherwise, it would not be possible to decide whether a record had been stored in an earlier run or if the entry had been made in the interrupted run. For example, run number 71 as

14	15
00	47

The Registry Number—consisting of nine digits, the last of which acts as check digit in error checking—serves as a key to access to the file and is stored as a packed decimal without sign. Here, the first half-byte is always zero. The Registry Number 630-08-0 is stored as

1				5
00	00	63	00	80

The example for the substance carbon monoxide is

1 5					6 9				10 12			13	14 15	
00	00	63	00	80	86	05	74	99	00	15	8C	18	00	47

Content of Information. The substance of Registry Number 630-08-0 was processed in run number 71 in CA abstract 86-057499 as an inorganic compound with additional GREMAS encodement and recalled 158 times.

For Concept Headings, the Hash Code is used as a key. The method was developed for the retrieval system SIR as a screen in the rationalization of text retrieval.

In the first 3 bytes of the Hash Code, the type of letter is stored, necessary for the character string to be registered. For a large number of technical terms in English language, the letter frequencies were analyzed and the letters assigned to bit positions according to their frequencies:

seldom-		-often	
J Q W K Z X V G	B F U Y P M H D	C L R S T N A O	(I, E)
Byte 1	Byte 2	Byte 3	

The most often occurring letters are I and E. Due to a lack in significance, no bits are assigned to those letters. For every character string, the bit positions are set into state 1, corresponding to letters present therein. For special symbols, no bits are set. For letters occurring repeatedly, the position can only be set once.

In this retrieval system, search conditions (terms) and the words in data, which are searched for, are provided with these 3-byte codes. Prior to the expensive text comparison, a comparison of the codes is conducted. Only in the case of a match is further searching necessary. Since access proceeds via tables, letter frequency of the least often occurring letter is employed to yield a better distribution of terms. This procedure is of no advantage but also of no disadvantage when used in the data file.

The efficiency of the preselection is documented by the following figures, as derived by the text retrieval: 100 000 comparisons between terms and words in data led to 30 relevant hits. All but 48 cases, for which continued checking was necessary, were rejected by the front-running comparison of the 3-byte codes.

As an example for nonrelevant information, RETRIEVAL and RELATIVE have the same codes, yet differ in character string lentgh and letter sequence. Byte 5 of that code contains the length of the character strings in binary form and is used for screening. An example for nonrelevant information possible when this comparative method is used is RETRIEVAL and RAT-LIVER. Both character strings have a length of nine digits and identical masks, since no bit is set for the hyphen in RAT-LIVER. All but 32 of the 48 cases identical in the letter comparison were eliminated after an additional length comparison. By comparison of the initial letter, encoded in the fourth byte, additional cases are solved. The first and second letters are stored hexadecimally in the first and second half bytes, respectively, by the following procedure:

A = 1	G = 5	M = 8	S = C
B = 2	H = 6	N = 9	T = D
C = 3	I = 6	O = 9	U = E
D = 4	J = 7	P = A	V = E
E = 4	K = 7	Q = B	W = F
F = 5	L = 8	R = B	XYZ = F

An example is the character string that starts with KI, byte 4:

76

An example of nonrelevant information is RESERVE and REVERSE: the character strings contain the same letters, they are of same length, and their first two letters are identical.

Cases of the described nature are not to be expected for the fixed vocabulary of CAS Concept Headings. Nevertheless, should they occur, they would be processed separately via tables in the program. As an example, the Concept Heading "KINETICS OF OXIDATION" contains the code

1				5
14	41	9F	76	15

For the processing duration of an Alpha Number, field 6–9 contains, as for substances, the Abstract Number.

Assignment of the Alpha Number, consisting of a letter, four digits, and a check character, is stored as follows:

byte	6	letter of the Alpha Number (EBCDIC)
bytes	7–8	four digits, packed without sign
byte	9	check character (EBCDIC)

An example is R0700Y as

6			9
D9	07	00	E8

An example of a complete Concept Heading encodement is

1				5	6			9	10		12	13	14	15
14	41	9F	76	15	D9	07	00	E8	00	06	5C	40	00	21

A Concept Heading of a length of 21 characters ($=15_{16}$) and the initial letters KI contains the letters K, X, F, Y, D, C, S, T, N, A, O and (possibly I and E). The process was conducted in run 33 ($=22_{16}$). The Alpha Number R0700Y (=REACTION RATE) was assigned and recalled 65 times.

At this time, the master file contains about 230 000 Registry Numbers and 3400 Alpha Numbers. The useful effect of avoiding multiple encodements is encouraging: depending on section, 98–100% of the Concept Headings and 80–100% of the Registry Numbers are recallable. The file covers 8600 blocks of 2048 bytes.

DATE FILE FOR SUBSTANCE ENCODEMENTS

Input proceeds via records of variable length. Behind the field for record length on bytes 1–4, the Registry Number is stored on byte 5–13 as EBCDIC, representing the access key. Bytes 14–16 are either empty or contain concept codes for substances. Starting with byte 17, the substance encodement is listed in the form of the records 3xxx of the data file.

For the purpose of field separation, the characters * and $ are used. The standardized molecular formula, in the form of artificial element symbols, follows either the range of figures (opened with $) or the linear structural formula (opened with *). A possible clear text entry is separated by a further * character. An example for $KBrO_3$ is

	1 4	5 13	17 38
$KBrO_3$		007758012	ADO1PRO3QTO1*K1BR103

At this time, this file consists of 185 000 records and covers about 2900 blocks of 2048 bytes.

DATA FILE FOR ALPHA NUMBERS AND SYSTEM NUMBERS

The file consists of records of 80-byte length. The key is located in bytes 1–6 of the records and contains Alpha Number or System Number. The text starts from column 8 on. An example is

1 6	8 31	32 80
R0700Y	REACTION RATE	

The file covers about 100 blocks with 2048 bytes and contains 1700 records.

```
o1oo / oo1                                    IDC - ANORGANIKA-BAND CAS VOL 9o

CA o9o - o4 - o3332o     SECTION o78.oo9    TYPE JOURNAL    CODEN NKAKB8
--------------------
CATALYTIC OXIDATION OF CARBON MONOXIDE OVER LANTHANUM MANGANESE
STRONTIUM OXIDE(LA1-XMNSRX03)

SHIMIZU, TAKASHI / MORIWAKA, YUKIHIRO

NIPPON KAGAKU KAISHI 1978 (11), 1462-6 (JAPAN)

H2ooo HETEROGENE KATALYSE, KATALYSATOREN

A125o ADSORPTION
      OF OXYGEN, ON STRONTIUM-CONTG. LANTHANUM MANGANESE OXIDE
      CATALYSTS

Ro7oo REAKTIONSGESCHWINDIGKEIT
      KINETICS OF OXIDATION, OF CARBON MONOXIDE WITH STRONTIUM-CONTG.
      LANTHANUM MANGANESE OXIDE CATALYSTS

03231 OXIDATION
      OF CARBON MONOXIDE WITH STRONTIUM-CONTG. LANTHANUM MANGANESE OXIDE
      CATALYSTS

Ko28o KATALYSATOREN
      CATALYST FOR THE OXIDATION OF CO

(R4) C1 01 63o-o8-o / (22,R5) LA1 MN1 03 12o31-12-8 / (R4) LA1 (N1 03)3
                                                    ----------------
/ (R4) SR1 (N1 03)2 / (R4) MN1 (N1 03)2 / (R4) H2 02 / (22,R5) LAo-1 MN1
                                          ----------
SRo-1 03
```

Figure 3. Retrieval printout in the SIR system.

```
CA o9o - o4 - o3332o     SECTION o78.oo9    TYPE JOURNAL    CODEN NKAKB8
--------------------
CATALYTIC OXIDATION OF CARBON MONOXIDE OVER LANTHANUM MANGANESE
STRONTIUM OXIDE(LA1-XMNSRX03)

SHIMIZU, TAKASHI / MORIWAKA, YUKIHIRO

NIPPON KAGAKU KAISHI 1978 (11), 1462-6 (JAPAN)
```

Figure 4. Data taken from CA file without intellectual manipulation: section subsection (as 1. System Number), publication classification code (translated in clear text), CODEN, document title, personal author names, abbreviated title of publication, publication date, issue number, pages, and original language.

H2ooo HETEROGENE KATALYSE, KATALYSATOREN

	CODE	KG	TEXT
	H2000V	N	
1	H2000j	+n+	

Figure 5. Excerpt from the data processing form. A System Number was assigned by additional encodement. The upper line contains the System Number (with check digit) as assigned by the editor; below that, the keyboarding protocol is listed. N represents the category description for System Number. The text "Heterogeneous Catalyses, Catalysts" was obtained from the data file for Alpha Numbers/System Numbers.

INORGANIC COMPOUND CLASSES

An advantage of encoding inorganic compounds by employing the standardized molecular formula is that it makes it possible to handle compound classes in the same manner as specific compounds during storage and retrieval. For example, alkali metal halides, alkali metal chlorides, and sodium chloride can be searched for without loss of information either separately or together via a search condition.

Since compound classes are listed under Concept Headings in the CAS system, a procedure had to be developed to generate records for inorganic substances from these headings. This is accomplished as follows: at its first appearance, an Alpha Number is assigned to the Concept Heading. This Alpha Number, with an initial letter X and a number ≥0100, is stored in the master file and assigned when the substance appears again.

Furthermore, a substance encoding is carried out and read into the text file by program after generating a standardized molecular formula, with X Alpha Number as the key instead of the customary CA Registry Number. During the construction of the inorganic chemicals data base, the Concept Heading record is replaced by the corresponding record for substances. In this case, the editor assigns the concept codes.

COLLECTIVE ALPHA NUMBERS

A series of CAS Concept Headings, e.g., taxonomic headings, are too specific and not of interest to the IDC inorganic chemicals data base. Thus, taxonomic headings in the IDC data base are listed under the general heading "Plants". Loss of information is avoided by storing the text of the Concept Heading as context, next to Qualifier and Text Modification.

The Alpha Number R0700Y (=reaction rate), present in the processing example, is also a collective number: kinetic investigations of special reactions (example: Kinetics of Oxidation) have their own Concept Heading in the CAS system in contrast to the IDC inorganic data base. Therefore, the

```
A125o ADSORPTION
      OF OXYGEN, ON STRONTIUM-CONTG. LANTHANUM MANGANESE OXIDE
      CATALYSTS

 1166..................................................................... 1166

ANR  IN BEARBEITUNG  113 / 9o-oo5o66   CA9o-o3332o.
CTH  ADSORPTION                                    .
TMD  OF OXYGEN, ON STRONTIUM-CONTG. LANTHANUM MANG.
     ANESE OXIDE CATALYSTS                         .
                                                   .
 1167..................................................................... 1167
```

Figure 6. Excerpt from the CA file work file. Sheet number 1166 of the work list for Concept Heading entry "Adsorption" contains the information that the Alpha Number is not yet available but in process, according to CA abstract 90-005066 (run number 113). Since the runs are processed chronologically, Alpha Number A 1250 N was available for recall during construction of the data base; the text modification was transferred as context.

```
Ro7oo REAKTIONSGESCHWINDIGKEIT
      KINETICS OF OXIDATION, OF CARBON MONOXIDE WITH STRONTIUM-CONTG.
      LANTHANUM MANGANESE OXIDE CATALYSTS

03231 OXIDATION
      OF CARBON MONOXIDE WITH STRONTIUM-CONTG. LANTHANUM MANGANESE OXIDE
      CATALYSTS


 1167.....................................................................

ANR  Ro7ooY                            CA9o-o3332o.
CTH  KINETICS OF OXIDATION                         .
TMD  OF CARBON MONOXIDE WITH STRONTIUM CONTG. LANT.
     HANUM MANGANESE OXIDE CATALYSTS               .
                                                   .
 1168.....................................................................

ANR  03231Z                            CA9o-o3332o.
CTH  OXIDATION                                     .
TMD  OF CARBON MONOXIDE WITH STRONTIUM-CONTG. LANT.
     HANUM MANGANESE OXIDE CATALYSTS               .
                                                   .
 1169.....................................................................
```

Figure 7. Example for the processing of Concept Headings, the assigned Alpha Number of which was available for recall. R0700Y is an Alpha Number, whose Concept Heading, next to the Text Modification, is stored in context.

```
Ko28o KATALYSATOREN
      CATALYST FOR THE OXIDATION OF CO

2
3 k0280q  +s+  catalyst for the oxidation of co
```

Figure 8. Example of an additional encoding of a concept registration with Alpha Number and context. S is the category description. The text "catalyst" was obtained from the data file for Alpha Numbers/System Numbers.

```
(R4) C1 01 63o-o8-o / (22,R5) LA1 MN1 03 12o31-12-8


 1164..................................................................... 1164

REG  63o-o8-o     ALT                  CA9o-o3332o.
PAR  CARBON MONOXIDE                               .
QLF  REACTIONS                                     .
TMD  OXIDN. OF, WITH STRONTIUM-CONTG. LANTHANUM MA.
     NGANESE OXIDE CATALYSTS                       .
FOR  CO                                            .
                                                   .
 1165..................................................................... 1165

REG  12o31-12-8   NEU                  CA9o-o3332o.
PAR  LANTHANUM MANGANESE OXIDE                     .
LIN  LAMNO3                                        .
TMD  PREPN. AND USE AS OXIDN. CATALYSTS FOR CARBON.
      MONOXIDE                                     .
FOR  LAMNO3                                        .
                                                   .
 1166..................................................................... 1166
```

Figure 9. Example for the processing of two substances. Carbon monoxide is already known, and thus, only the concept code R 4 (=reaction substance) is noted. The second substance is new. In addition to the concept codes 22 (=preparation) and R 5 (=auxiliary agent for reactions, catalyst), a linear molecular formula is recorded.

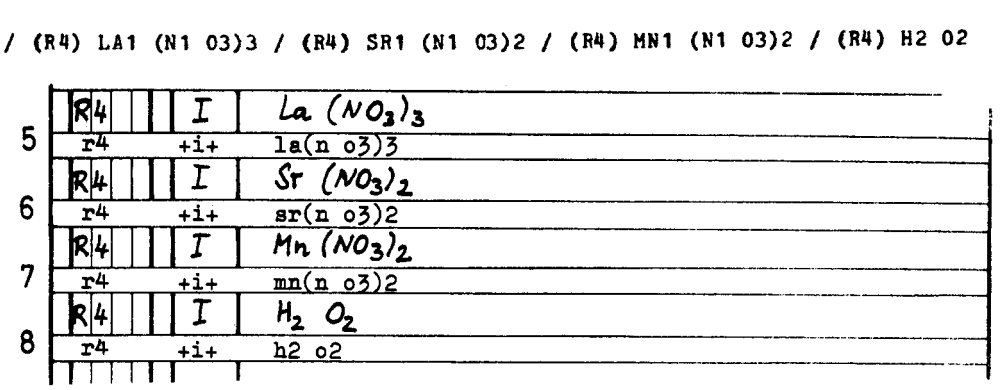

Figure 10. Example of an additional encoding of four substances with concept code and linear molecular formula via the keyboarding form. The category description is I. Addition of index number 1 and treatment of void spaces in the molecular formula are conducted during the check routine. No Registry Number is stored for additionally encoded substances.

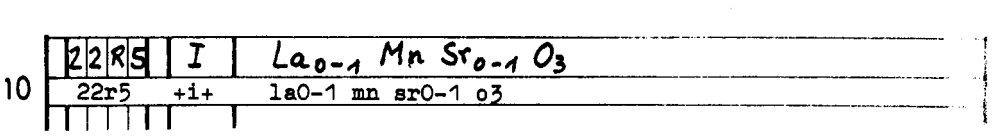

Figure 11. Example of a formula with a range of figures. For the additional encoding listed above, a range of figures is stored additionally to the standardized molecular formula and the linear structural formula: standardized molecular formula, BG00 (Sr), DJ00 (La), LD00 (Mn), and PR00 (O); range of figures, 00000 01000 00000 01000 01000 01000 03000 03000; linear structural formula, LA0-1 MN1 SR0-1 O3. By convention, all element indexes are converted to zero in the standardized molecular formula, when a range of figures is present.

character string "Kinetics of Oxidation" is transferred additionally to the context in the registration of "Reaction Rate" in order to avoid loss of information.

ORGANIC COMPOUNDS

Naturally, organic compounds also appear as reactants or reagents in the processed abstracts. Since organic compounds cannot be sufficiently evaluated by means available to the inorganic system, abstracts they appear in are also processed in the IDC system according to GREMAS. Thus, they can additionally be searched for in that system.

All CA Registry Numbers of organic compounds, labeled as "organic compound" by the editor, are combined into one substance record, in order to prevent complications when searches in both data bases followed by correlation of relevant abstract numbers are necessary. This record contains as context a listing of all CA Registry Numbers of organic compounds present in the abstract in increasing order. In data retrieval, the presence or absence of organic compounds can be demanded.

PROCESSING EXAMPLE

The CA abstract CA 90-04-033320 serves as example to demonstrate the means by which encoding and keyboarding work combines the data of a literature evaluation. The evaluation contains bibliographic data, a System Number (encoded), four Alpha Numbers (three from Concept Headings, one from additional encoding), and seven substance encodements (two from the CA file, five from additional encoding). Figure 3 illustrates the printout of a relevant hit in the form of a retrieval answer when the system SIR is used. Data elements contributing to relevancy are underlined. In Figures 4–11, the processing of the individual data in the system is illustrated, on the basis of the order of the printouts.

The described system makes it possible to construct an inorganic chemicals data base on the basis of the magnetic tape service CA Search, with the possibility of supplemental indexing. This data base corresponds, with respect to format, to that of the IDC Inorganic Chemicals Data Base for patent literature. Searching is identical for both data bases. Multiple encodements of identical concepts are avoided by employment of data files.

ACKNOWLEDGMENT

We thank the German Ministry of Research and Technology for supporting developmental work regarding the application of the system to journal literature.

REFERENCES AND NOTES

(1) For part 1, see Roschkowski, H.; Simmler, W. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 154–158.
(2) Schwier, Report BMFT-FB ID 79-006, Bundesministerium für Forschung und Technologie, Berlin.
(3) Jungblut, A. Report BMFT-FB ID 83-002, Bundesministerium für Forschung und Technologie, Berlin.
(4) *IDC-Systembeschreibung Teil BV2*; Internationale Dokumentationsgesellschaft für Chemie m.b.H.: Frankfurt/Main, 1977.