The Automatic Generation of Keywords from Chemical Compound Names: Preparation of a Permuted Name Index with KWIC Layout

F. H. ALLEN* and W. G. TOWN†

Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

Received September 16, 1976

A computer procedure has been developed for the preparation of a permuted keyword index of chemical compound names. Lists of common chemical prefixes and suffixes have been derived and are used in the analysis of chemical syntax to establish keywords within a name. Output is in the form of a permuted index with KWIC layout. The keywords are arranged alphabetically down the page center with the rest of the name printed to the left and right. A wraparound facility is used to preserve maximum context in long names. The procedure has been developed using the Bibliographic File of the Cambridge Crystallographic Data Centre as data base, but has general applicability to files of compound names which obey normal chemical syntax.

INTRODUCTION

The use of permuted indexes with KWIC or KWOC layouts is well known.[1] Such schemes have, for example, been applied to chemical titles,[2] descriptor sentences,[3] chemical line notations,[4] and NMR chemical shifts.[5] In these cases analysis of the text string for keywords to act as indexing points follows some relatively well-defined syntax rules.

The location of keywords (defined in this context as information-rich variable-length character strings) in chemical names constructed using standard rules of nomenclature is not straightforward. Syntax analysis based on chemical punctuation, e.g., space, hyphen, bracket, is totally inadequate. Relatively simple names such as:

Threo-α,β-Dimethylacetyl*choline* Iodide or
Acetyl*selenocholine* Iodide

are indexed only at the points indicated upper case, while the most important keywords (italicized) are masked out. An early solution to this problem[6] required highly structured input where each indexable syllable or keyword was separated from the next by a hyphen. Another approach[7] is to input a list of "compound-name root words" or "word stems". The first approach requires consistent application of numerous syntax rules during file assembly to yield a name which is easily indexed but does not conform to accepted rules of nomenclature. The second approach requires continuous monitoring of new material to identify new keywords or keyword roots for addition to the input reference list. In both cases syntax analysis is essentially external to the process of index generation.

The procedure described here is designed to operate on compound names constructed in the normal manner. Syntax analysis is performed by reference to input lists of common chemical prefixes, suffixes, derivative names, and a list of element-name roots. The lists are used in conjunction with the chemical punctuation symbols to break down the name into its constituent syllables and also to decide which of these syllables represent useful keywords for indexing purposes. The reference lists are, to some extent, file specific but once established for a given database they require infrequent monitoring.

An index of this type was presented in the Cambridge Crystallographic Data Centre publication "Interatomic Distances 1960-65",[8] and the basic programming philosophy has been briefly described.[9] This philosophy has been greatly extended in the present program, and the resolution of the index has been much improved.

Table I. Summary of Bibliographic File Card Types and Information Content[a]

COMPND*	*Chemical compound name*, with normal chemical syntax, usually that assigned by the authors; a qualifying phrase[10] may follow the name in parentheses
SYNONYM	*Synonym for compound name*, normal syntax, included only where appropriate.
FORMUL*	*Molecular formula* expressed in terms of residues, i.e., discrete covalently bonded groupings
AUTHOR*	*Authors names*
JRNL*	*Literature reference*, journal name, journal code number, volume, page, year
REFER	*Cross-references* to crystallographic reference books including the Bibliographic volumes in the Molecular Structures and Dimensions series[11]
CLASS*	*Chemical classification flags*

[a] Card types marked (with an asterisk) are mandatory; the six-letter/two-digit *reference code* is present on all cards pertaining to a given file entry.

DATA BASE STRUCTURE AND CONTENT

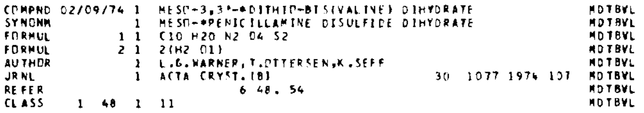
The procedure has been developed using the Bibliographic File[10] of the Centre as the data base. The file currently contains bibliographic information for some 15 000 organic and organometallic compounds whose structures have been determined by x-ray or neutron diffraction. The file structure and information content are fully described elsewhere,[10] but a brief summary is given here for clarity.

Bibliographic information for a given compound is stored on magnetic tape as a series of 80-byte card images. Each entry is made up of a number of card types shown in Table I together with their information content. Multiple use of the first five card types is allowed to accomodate all required information in any category. An example of a complete bibliographic entry is shown in Figure 1. Each entry is identified by an eight-character reference code[10] which appears in columns 73–80 of each card of the entry.

The present procedure is mainly concerned with the contents of the COMPND and SYNONM records. Compound names in the file are usually those given by the authors of the publication, but if this is faulty or ambiguous then a correct systematic name is derived. If the authors' name is "correct" no attempt is made to standardize it with any set of nomenclature rules. Synonym names are included only where

* Author to whom all correspondence should be addressed.

† European Communities Joint Research Centre, I-21020 Ispra (Varese), Italy.



```
COMPND 02/09/74 1   MESO-3,3'-*DITHIO-BIS(VALINE) DIHYDRATE                MDTBVL
SYNONM          1   MESO-*PENICILLAMINE DISULFIDE DIHYDRATE                MDTBVL
FORMUL        1 1   C10 H20 N2 O4 S2                                       MDTBVL
FORMUL        2 1   2(H2 O1)                                               MDTBVL
AUTHOR          1   L.G.WARNER,T.OTTERSEN,K.SEFF                           MDTBVL
JRNL            1   ACTA CRYST.(B)                  30  1077 1974 107      MDTBVL
REFER                          6 48. 54                                    MDTBVL
CLASS    1  48  1   11                                                     MDTBVL
```

Figure 1. Sample entry from the Bibliographic File of the Cambridge Crystallographic Data Centre.

appropriate, e.g., phenobarbitone for 5-ethyl-5-phenylbarbituric acid.

At present all names are held in upper case, together with a system of typesetting signals. All letters are assumed lower case unless appropriately signalled, subscript and superscript situations are catered for, and Greek characters are spelled in full. Interpretation of this scheme for all cards in an entry yields a steering tape for a photocompositor to produce the bibliographic volumes in the "Molecular Structures and Dimensions" series.[11]

SYNTAX ANALYSIS OF NAMES

The inadequacy of chemical punctuation signs as keyword indicators was mentioned above. If we re-express the examples given there in our extended format and introduce / as an additional "punctuation" symbol to separate concatenated syllables, we get:

threo-alpha,beta-di/methyl/acetyl/*choline*/iodide and acetyl/*seleno*/*choline* iodide

It is clear that, in addition to natural punctuation, we need a list of prefix syllables to use in a character matching process in order to break down long alphabetic strings. In these examples *di*, *methyl*, *acetyl*, *seleno* simply mask the major keyword *choline*. This gives rise to the concept of a START list, i.e., a list of syllables that may be regarded as keywords in their own right and which also cause a restart of the character-matching process. Thus the location of *acetyl* in the second example causes a pointer to move to the "*s*"; the remainder of the name string is then recompared with the START list. In cases where a match is not found, e.g., *choline*, the name string up to the next punctuation point, space, is regarded as a keyword.

Application of this philosophy to the examples yields eight keywords for the first and four for the second. Many of these keywords convey minimal information when included in an index, i.e., *threo*, *alpha*, *beta*, *di*, *acetyl*, *methyl*. Such syllables may be transferred from the START list to a STOP list; these syllables are *not* regarded as indexable keywords but do cause a restart of the character-matching process described above. For the two examples the transfer of the cited syllables to the STOP list reduces the indexable keywords to two and three, respectively. For our file, which includes many halide derivatives where the halogen was introduced to help solve the crystallographic phase problem, the syllable *iodide* is also in STOP. This reduces the indexable keywords (italicized in the examples) to one and two, respectively.

The development of the STOP and START lists was an iterative process, and initially each list entry consisted of a simple character string as described above. The complete lists currently in use are given in Tables II and III. Any entry may be a simple string or a complex string with syntax of its own.

The STOP List: Syntax and Content. Simple strings in the list are terminated by a space or by an asterisk. This difference distinguishes complete words, which are known to end in a space, from syllables in a name string. This is necessary because our file contains "descriptive" names; these arise from very brief conference reports where insufficient information is given for the derivation of a systematic name. Names of the form

Condensation product of A *and* B

do occur, and each italicized word is included in the STOP list with its trailing blank as terminator. This leaves the chemical names of A and B to be indexed normally. The word *of* is not included in the list, and this is explained below. The necessity for regarding the trailing blank as part of the STOP string is exemplified by comparing this name with

5-alpha-androstan-3,17-dione

The important keyword *androstan* would be lost if the syllable *and* was treated as a three-character string; the meaningless string *rostan* would be indexed. If the user wanted this he could input AND* to exclude the word "*and*", together with any name string beginning with *and*.

In the example of a "descriptive" name cited above, it was noted that the word *of* need not be included in the STOP list. This arises from implementation of the rule that there must be at least three consecutive alphabetic characters in a keyword. All one- or two-letter strings immediately followed by a chemical punctuation symbol are ignored. This implies that all one- or two-letter strings which are known *always* to terminate in a punctuation symbol may be omitted from STOP. This is important for our file not only to exclude the strings *of*, *in*, *at*, etc., which might occur in descriptive names, but also to automatically exclude the spelled form of many common Greek letters: *mu*, *pi*, *nu*, etc., elemental substituent prefixes, e.g., *NN*-, stereochemical descriptors, e.g., DL-, and a range of two-letter ring fusion descriptors as in:

4,8-dihydro-dibenzo(*cd*,*gh*)pentalene

This simply applied rule reduces the length of the STOP list and considerably reduces the computer time for character matching. It should be noted that the commonly occurring numerical prefix *di* does not fall into this category since, as shown in the above example, it is seldom followed by a punctuation point.

The extended or complex STOP list entries arise when a syllable acts as a masking syllable in one context and as the start of a valid keyword in a different context. For example, the syllable *penta* in the example above is not a mask, but acts as such in the name

pentachlorobromobenzene

Therefore the extended STOP entries take the form

S*A/B/C/D/E/...

where S is the STOP syllable which is not indexed unless it is followed by one of the context strings A, or B, or C, or etc., the context strings being separated by / symbols and the implied space terminating the final context string has no special meaning. Many of the context strings in Table II are chemically obvious and apply to any file of compound names; others are present to accommodate natural product names, e.g., *git* to ensure that *digitogenin* is treated as a single word, and more specifically for our file, to accommodate synonym names. Names such as D*ianin's* compound and Tri*quat* yield the somewhat unrecognizable context strings indicated in italics.

The entries in the present STOP list fall into a number of distinct categories:

(i) nomenclature prefixes, e.g., *trans*-, *endo*-, *syn*-, etc.
(ii) chemical prefixes, e.g., *methyl*-, *ethyl*-, *cyano*-, *chloro*-, etc.
(iii) numerical prefixes, e.g., *bis*-, *tri*-, *hepta*-, *octa*-, etc.
(iv) terminators, e.g., *lyl*-, *ene*-, *oate*-, etc.
(v) frequently occurring derivative names, e.g., *chloride*, *iodide*, *hydrate*, etc.

Table II. The STOP List

```
AALPHA*
ABEO*
ABETA*
ACETOXY*L
ACETYL*ACET/ENE/SAL
ACID*IUM
ACI-*
ACONTA*
ADDITION
ADDUCT
ALKOXY*
ALL-*
ALPHA*
AMINO*
ANALOGUE
AND
ANGULO*
ANHYDRO*
ANION
ANOMER
ANTI*BIOTIC/MON
AQUO*
AROMATIC
ARYL*
ATE*
ATO*
BALPHA*
BASE
BASIC
BBETA*
BETA*INE
BETWEEN
BIS*MUTH/UCC/ULF
BI*CARB/CYCL/NOX/OTIN/
    PHEN/PYR/URET
BLUE
BROMIDE*
BROMO*FORM
CAGED
CARBA*LD/MATE/MATO/MIC/
    MID/MOYL/MYL/ZOL
CARBONYL*
CDELTA*
CHELATE
CHLORIDE*
CHLORO*FORM/HEMIN/PHONE/
    PHYLL/QUINE/TINE
CIS*
CLATHRATE
CLOSO*
CLOVO*
COMMO*
COMPLEX
CONDENSATION
CYANO*GEN/XIME
DECA*LIN/LONE/LYL/NAM/
    NE/NOATE/NOIC
DEG.*
DEGRADATION
DEF*
DELTA*
DEOXY*
DERIVATIVE
DISORDERED
DI*ALUR/ANIN/AZAN/AZEN/
    CHOT/ELD/GIT/MED/
    OSG/OXAN/OXIN/OXOLAN/
    OXON/PT/QUA/THIZ
DODECA*NE/NO
DRUG
ENNEA*
ENDO*
ENE*
ENO*
EN*ANTHOLAC/DRIN/DUR/
    HYDRIN/MEIN/NIAT
EPI*MINO
EPOXY*
EPSILON*
ERYTHRO*ID/MYC
ESTER*
ETA*
ETHYL*ENE/IDEN/IDYN
EXO*
FACIAL*
FAC-*
FLUORIDE
FLUORO*ANTHENE
FORM
FORM)
FREE
FROM
FULLY
FUSED
GAMMA*
HAPTO*
HEAD*
HEDR*
HEMI*N
HEPTA*CENE/CHLOR/DEC/
    FULV/LEN/NE/NO
HEXA*CENE/COSAN/CRYL/
    DEC/MINE/MMINE/NE/
    NO/SON/TRIACONT/ZA
HOMO*
HYDRATE*
HYDRIN
HYDRIOD*
HYDROLYSAT*
HYDROLYSED
HYDROLYSIS
HYDRO*GEN/NIUM/XAM/XANATO/
    XIDE/XIM/XO/XYL
IDENE*
IDE*
IDO*
III*
INACTIVE
INCLUSION
INNER
IODIDE*
IODO*FORM/NIUM
IONS*
ITOL
IUM*
KETO*NATO/NE/XIME
KIS*
LACTONE
LAMBDA*
LAYERED
LYL*
MER*CAP/CUR
MESO*
METHOXY*
METHYL*ENE/ENOMYCIN/
    IDE/IDYN/IUM
MIXTURE
MONO*XIDE/XIM
NIDO*
NITRILE*
NITRO*GEN/MIN/NAT/NE/NYL/
    SAM/SITE/SO/SYL/XIDE
NN'*
NNN*
NONA*CTIN/DEC/NE/NOIC/NO/NYL
NOPP*
NUCLEAR
NYL*
OATE*
OCTA*COSAN/DEC/DIEN/
    LEN/LIN/NE/NO
OIC*
OL*EAN/EFIN/EIC/IGO
OMEGA*
ONAT*
ONO*PORD
ONE*
ORANGE
OXO*NY/NIUM
OXY*GEN
PARTIALLY
PENTA*CENE/DEC/LEN/
    MINE/MMINE/NE/NO
PER*CHLORAT/CHLORYL/I/
    KIN/LOL/OPYR/YL
PHENYL*ENE
PPP*
PRODUCT
PSR*
RACEMIC*
RAC*
RADICAL
REACTION
REARRANGEMENT
RED*UCTONE
RPS*
RSED
SALT
SECO*
SEDO*
SEPTA*MYC/NOS
SESQUI*
SIGMA*
SOLID
SOLUTION
SOLVATE
SRS*
SSS*
SUBST*
SYM*
SYNCLINAL*
SYN*ALAR
SYSTEM
TAIL
TERT*
TETRAZINC*
TETRA*CENE/CETATE/COSAN/
    CYCL/DEC/LIN/LONE/
    MINE/MMINE/NACT/Z
THREO*NIN/NYL
TRANS*
TRI*AZIN/AZOL/CHOD/COSAN/
    CYCLO/DEC/GON/LL/
    MESIC/OSE/OXAN/OXIN/
    PO/PT/QUAT/UR
UNNATURAL
UNDECA*N
WATER
WITH
XY*L
YL*
YNE*
YNOIC*
YNYL*
```

(vi) miscellaneous individual words, e.g., *form*, *salt*, *and*, *product*, etc.

(vii) Greek letters where the spelled version exceeds two letters, e.g., *alpha*, *beta*, etc.

(viii) strings of element symbols indicative of substitution, again only where these strings exceed two letters, e.g., *NNN-*, *NOPP'-*, *PPP-*, etc.

The presence of categories i–v is mandatory for any file and the content is not file specific except for the choice of entries in categories ii and v. It is in category ii that the main transfers between START and STOP occur, until a balance is struck between the number of keywords retained for indexing and the usefulness of the information conveyed by them. In practice this usually depends on the frequency of occurrence of a given keyword. Entries in categories vi–viii are largely file specific as explained above.

The START List: Syntax and Content. The START list also consists of simple and complex strings. Because the START list consists entirely of prefixes, i.e., all entries would fall into category ii of the STOP list if transferred, there is no need for special terminators in simple strings. The implied blank at the end of, e.g., *acetato* has no special meaning.

Complex strings in the START list arise in a similar manner to those in the STOP list, a simple string acting as a masking syllable in one context but as part of a major keyword in another. For example, the syllable *seleno* acts as a mask in

acetylselenocholine

but not in names such as

trimethylselenonium iodide or
dibenzoselenophene

Complex or extended START strings take an analogous form to those in STOP, i.e.,

S*A/B/C/D/... etc.

where S is the START string and A, B, C, D,... are context strings where the restart philosophy is *not* applied. The adoption of this definition and syntax means that complex START strings may be transferred directly to STOP without amendment.

Organization of STOP and START Lists: Ordering of Entries. It should be noted that in both lists the context strings are kept as short as possible to minimize character-comparison time. The context strings only need to be long enough to define the context with respect to the file content. For example, the syllable *penta* is included in STOP, but is the root of a valid keyword in the contexts *pentane*, *pentano*; the two letter context strings /*ne*/*no*/ are sufficient to exclude *pentanitrate*, etc. In suitable cases the strings can revert to a single letter to cover a number of possibilities. For instance, the syllable *undeca* in our file is never indexed unless it is followed by *n*, as in *undecane*, *undecano*, *undecandioic*, etc. Here we need the STOP entry

UNDECA*N

to cover all current situations in the file.

Both lists are ordered into 26 alphabetic groups based on the first letter of each entry (see Tables II and III). The program stores the starting address for each alphabetic group in both lists and uses a simple look-up to establish where character-comparison is to start. Comparison of the target name with either list ceases as soon as a hit is found or when a new initial letter is encountered in the lists.

Within the 26 alphabetic groupings in each list the entries are not stacked in strict alphabetic sequence. The reason for this is best exemplified by studying the START entries *azo* and *azonia* and the names

trans-N,N'-azo/morpholine and

Table III. The START List

```
ACENAPHTHO
ACETAMIDO*X
ACETATO
ACETO*A/I/N/PH/X
ADENOSYL
ALANYL
ALLO*XA
ALLYL
ALTRO
AMIDO*X
AMMONIO
AMO
AMYL*OSE
ANTHRAQUINONE
APO*RP
APRO
ARABINO
ARSA*N/Z
ARSENO
ARSINO
AZA
AZIDO
AZIRIDINO
AZONIA
AZO*CINE/L/N/PHT/TATE
BENZOXY
BENZO*CAIN/PHENO/Q/X
BENZ*AL/AM/ENE/IDIN/IL/
     IMID/IODOX/IT/OX
BORA*L/NATA/NYL/T/Z
BROSYL
BUTYL
CARBAMOYL
CARBOXY*L
CARBO*L/N/R/X
COBALTA*TE
COBALTO*C
CYCLO*BUT/HEPT/HEX/
     PENT/PHAD/PHAN/
     PROP/TET/TRI/XO
CYSTEINYL
CYSTYL
DEACETOXY
DEACETYL
DEHYDRO
DEMETHYL
DESACETOXY
DESOXY
DETHIO
DEUTERIO
DEUTERO*NIUM
ETHANE*DIO
ETHANO*L
ETHENO
ETHOXY
ETHYLENE*D/TR
ETHYNYL
ETIO
FERRA*TE
FORMAMIDO*Y
FORMYL
FRUCTO*SIDE
FURO*PH/X
FURYL
GALACTO*S
GERMA*CR/N/T
GLUCO*N/SA/SY
GLYCERO*L
GLYCERYL
GLYCYL*GLYCYL
GUANIDINO
GUANIDO
GULO*N
HISTAMINO
HISTIDYL
HYDRAZINO
HYDRIDO
HYDROGEN
HYDROXY*LA
IMIDO
IMINO*XYL
ISOPROPYL
ISO*CYAN/PROP/THIO/XAZ
LACTO*B/N/S
LEUCYL
LEVO
LYXO*SE
MANGANA*TE
MANNO*S
MERCAPTO*LE
MERCURI
MESITYL*ENE
META*BOL/LO/NIL
METHANO*L
METHANE
METHYLENE*DIA/TE
METHYL*ENE/ENOMYC/
     IDE/IDYN/IUM
NAPHTHALENE
NAPHTHYL
NEO*DYM
NITROSO*L
NONYL
NOR
ORTHO
OXA*LA/LIC/LIN/LUR/M/RS/Z
OXIDO
PARA*BAN/LD/QUAT/THION
PERCHLORATO
PERCHLORYL
PERI
PHENACYL*I/D
PHENOXY
PHENOXA*R/ZINE
PHEN*AC/AL/AN/AZ/
     ETO/IR/O/YL
PHOSPHA*MIDE/T/Z
PHOSPHINO*L/Y
PHOSPHONIA
PHOSPHORYL
PHOSPHO*CINE/L/N/R
PHOTO
PHTHALOYL
PICRYL
PIPERIDINYL
PIVALOYL
PLATINA
PROLYL
PROPYL*ENE/ID/PH/SAL
PROTO*P
PSEUDO
PYRIDYL
PYRO*GA/NE/P
RHODA*N/T
RHODIA
RIBO*S
SEMI*C/D
SELENA*N/TE/Z
SELENO*L/N/PH
SILA*N/T/Z
SILYL
SPIRO*D/NOL/SOL
STIBA
SUCCINYL
SULFA*M/N/TE/TO
SULFONYL
SULFO*L/N/X
TALO
TELLURA
TER*B/E/PE
TETRAZA
THIA*DI/MIN/NT/Z
THIONO
THIOXO
THIO*CAN/CARB/CIN/
     CTIC/CYAN/LAN/LAT/
     LE/LID/LIUM/LL/LO/
     LP/LS/N/PH/UREA/X
TOLYL
VINYL
XYLO*B/S
XYLYL
```

azonia/indane

In order to split these correctly at / the longer string, *azonia*, must precede the shorter string *azo*. If the strings are reversed, then the second name would generate a meaningless keyword *niaindane*, comparison would cease, and the name would not be compared with *azonia*. A complex example of this rule occurs with START prefixes which begin with *benz*; this example shows not only the stacking of long strings before shorter strings, but also shows that the START and STOP lists can interact to advantage in some circumstances. There are five prefixes, viz., *benzoyl*, *benzoxy*, *benzyl*, *benzo*, *benz*, which should be included *in that order* in our START list so that names such as

N-benzoyl/glycine
(+)-2-benzyl/glutamic acid
dibenzo/equinine and
1,2-benz/anthracene

are split in the correct places (/). In Table III only *benzoxy*, *benzo*, and *benz* are included in START. Since *benzoyl* and *benzyl* are always restart syllables (they have no context strings for our file), we can take advantage of the fact that the linking syllable *yl* is included, for other reasons, in STOP. Thus for the first name above the restart will occur at *ylglycine*, the *yl* will be ignored since it is in STOP, and the string *glycine* will be established as a keyword. The result is exactly that obtained by including the full string *benzoyl* in START. A similar trick could be used to remove *benzoxy* from START since *oxy* is also a STOP entry, but many genuine keywords start with *benzox*, e.g., benzoxazole, benzoxazine, etc., which we wish to index only at *benzo*. Hence we have the extended string *benzo*x* in the START list, which makes *benzoxy* a special case of *benzox*. This can only be rationalized by including the full string *benzoxy* in the START list ahead of *benzo* and *benz*.

Groups of context strings following a main STOP or START entry are stacked alphabetically, ordered on the first letter of the context string, viz.

HEPTA*CENE/CHLOR/DEC/FULV/LEN/NE/NO

Table IV. The Element ROOT List

AC	ACTIN/I/A	AU	GOLD*AUR/A/I/O	PR	PRASEODYM
AL	ALUMIN/A/I/Y	HF	HAFNIUM	PM	PROMETH
AM	AMERIC	HE	HELIUM	PA	PROTACTIN
SB	ANTIMON*STIB	HO	HOLMIUM	RA	RADIUM
A	ARGON	H	HYDROGEN	RN	RADON
AS	ARS	IN	IND/ATE/IUM	RE	RHEN
AT	ASTATIN	I*	IODINE	RH	RHOD
BA	BARIUM	IR	IRID/AT/IU	RB	RUBID
BK	BERKEL	FE	IRON*FERR	RU	RUTHEN
BE	BERYLL	KR	KRYPTON	SM	SAMAR
BI	BISMUTH	LA	LANTHAN	SC	SCAND
B	BOR/A/I/O/Y	LW	LAWRENC	SE	SELEN
BR*	BROM	PB	LEAD*PLUMB	SI	SIL/A/IC/OX/SE/Y
CD	CADM	LI	LITHI	AG	SILVER*ARGENT
CA	CALCIUM	LU	LUTETI	NA	SOD/ATE/IUM
CF	CALIFORN	MG	MAGNES	SR	STRONT
C*	CARBON	MN	MANGAN	S*	SULFUR
CE	CER/ATE/IUM	MD	MENDELEVIUM	TA	TANTAL
CS	CESIUM	HG	MERCUR	TC	TECHNET
CL*	CHLOR	MO	MOLYBD	TE	TELLUR
CR	CHRO	ND	NEODYM	TB	TERBIUM
CO	COBALT	NE	NEON	TL	THALL
CU	CUPR*COPPER	NP	NEPTUN	TH	THOR
CM	CURIUM	NI	NICKEL	TM	THUL
D	DEUTER	NB	NIOB	SN	TIN*STANN
DY	DYSPROS	N*	NITROGEN	TI	TITAN
EI	EINSTEIN	NO	NOBEL	W	TUNGST*WOLFRAM
ER	ERBIUM	OS	OSM	U	URAN/ATE/IUM/YL
EU	EUROP	O*	OXYGEN	V	VANAD
FM	FERMIUM	PD	PALLAD	XE	XENON
F	FLUORO	P*	PHOSPHORUS	YB	YTTERB
FR	FRANCIUM	PT	PLATIN	Y	YTTR
GD	GADOLIN	PU	PLUTON	ZN	ZINC
GA	GALL	PO	POLON	ZR	ZIRCON
GE	GERMA	K	POTASSIUM		

Element symbols marked (*), i.e. C,H,N,O,S,P,Br,Cl,I, are not included in the ROOT search (see text). No attempt has been made to derive systematic ROOT's for these elements.

This allows the program to cease the context search process as early as possible, thus saving computing time.

Location of Embedded Element Names: the Element ROOT List. The STOP and START lists of Tables II and III have proved surprisingly efficient in the breakdown of compound names into their constituent syllables, and some relevant statistics are presented below. At the end of the process there remained, however, a few cases where important keywords were still masked. About 100 such syllables were identified by visual inspection, and it was found that some 70% of these involved element names or name roots, e.g.,

acetyl ben*chro*trene
benzene*tellur*enyl bromide thiourea complex
10-phenox*ars*ine chloride
ox*ars*enanium bromide

All of these cases could be accommodated by additions to the START list, but frequently in areas (e.g., *ben*, *benz*, etc.) where this list is already complex. For instance, inclusion of the string *ox* in START, to accommodate the final example, would require it to be followed by a very extensive context list and would increase string-comparison time out of all proportion to the number of valid hits obtained.

In order to locate these embedded items, a list of element name ROOT's was developed. The full ROOT list is in Table IV. Syntax symbols * and / are used to distinguish alternative roots and alternative root endings, respectively; e.g.,

GOLD*AUR/A/I/O

indicate four roots: *gold*, *aura*, *auri*, *auro*. The string *aure* is excluded to avoid confusion with, e.g., *aure*omycin.

The provision of alternative ROOT endings minimized confusion between element names and organic roots, but some cases of identity still remained; e.g., the root *germa* must be included for germanium, but clashes with the germacranolide family of sesquiterpenes. To avoid this the individual chemical formula records (Table I) were scanned. The search is then restricted only to those roots which correspond to elements which are actually present, and certain roots, viz., those for B, Br, C, Cl, F, H, I, N, O, P, S, are excluded altogether on the basis of their frequency of occurrence. This means that more purely organic compounds are not searched at all, and that the element root/organic root confusion noted above is avoided. A corollary of this is that the ROOT list itself can be stripped of alternative endings in most cases. However, this gives a marginal speed improvement overall and the ROOT

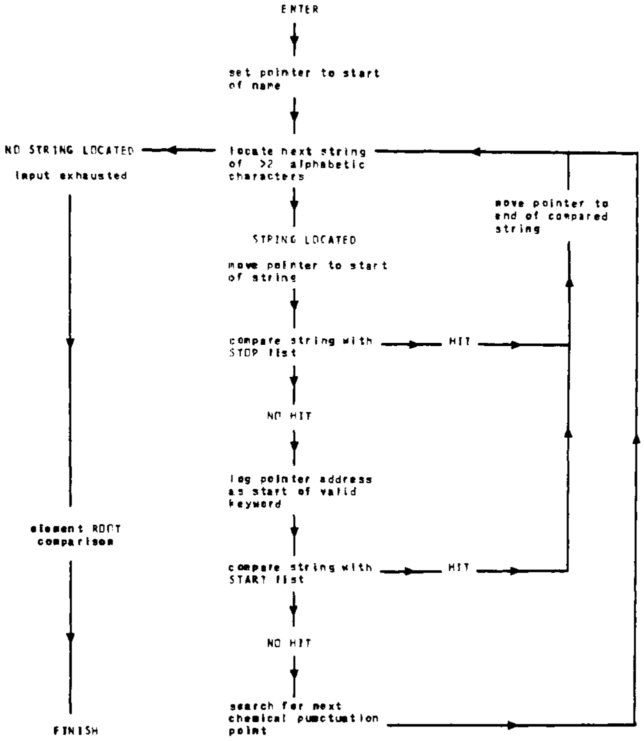

ENTER
set pointer to start of name
NO STRING LOCATED
input exhausted
locate next string of >2 alphabetic characters
move pointer to end of compared string
STRING LOCATED
move pointer to start of string
compare string with STOP list
HIT
NO HIT
log pointer address as start of valid keyword
element ROOT comparison
compare string with START list
HIT
NO HIT
search for next chemical punctuation point
FINISH

Figure 2. Flow diagram of keyword location procedure.

list has been left unaltered since it has some applicability, in its full form, in other areas of our work.

PROGRAM STRUCTURE

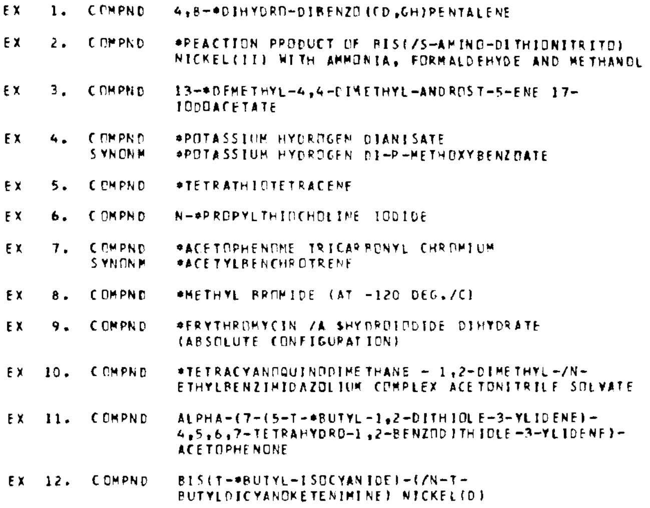
For each entry in the Bibliographic File the full compound (and synonym) name is reconstructed from the card image records. The chemical formula records are analyzed, and switches are set to indicate the presence of element types, for use in conjunction with the ROOT list. Information from other card types is extracted, according to the users requirements, to act as index reference points.

Syntax Analysis. The systematic application of the procedures described above is summarized in the flow chart of Figure 2. There are four basic steps:

(i) Location of strings of three or more alphabetic characters, these are potential keywords
(ii) Comparison of potential keywords with the STOP list
(iii) Comparison of valid keywords with the START list
(iv) Search for embedded element names or name roots.

Steps i–iii are reiterated according to the rules described above until the end of the name is reached. The start of each valid keyword is stored in terms of its character position in the name. The final step is performed as a single pass through the full name; new keywords are only added at this stage if not already validated by steps i–iii.

Generation and Formatting of the Keyword Index. The name is reconstructed with each of the keywords, established above, shifted to the center of the output name field. Some examples of names are given in Figure 3, and the result of syntax analysis and reformatting are in Figure 4. The character preceding the index point is preserved as a space, while the index point character itself is always converted to upper case. With the keyword positioned, the name is examined, and wraparounds and/or truncations are performed as required; as much of the full name as possible is preserved to clarify the keyword context. Some examples of wraparounds and truncations are shown in Figure 4, examples 2, 10, 11, 12. If we define the portion of the output name field preceding the keyword as the prefix, and the portion following the index point as the suffix, then four special symbols are used to define context in the following five cases:



```
EX   1.  COMPND   4,8-*DIHYDRO-DIBENZO(CD,GH)PENTALENE

EX   2.  COMPND   *REACTION PRODUCT OF BIS(/S-AMINO-DITHIONITRITO)
                  NICKEL(II) WITH AMMONIA, FORMALDEHYDE AND METHANOL

EX   3.  COMPND   13-*DEMETHYL-4,4-DIMETHYL-ANDROST-5-ENE 17-
                  IODOACETATE

EX   4.  COMPND   *POTASSIUM HYDROGEN DIANISATE
         SYNONM   *POTASSIUM HYDROGEN DI-P-METHOXYBENZOATE

EX   5.  COMPND   *TETRATHIOTETRACENE

EX   6.  COMPND   N-*PROPYLTHIOCHOLINE IODIDE

EX   7.  COMPND   *ACETOPHENONE TRICARBONYL CHROMIUM
         SYNONM   *ACETYLBENCHROTRENE

EX   8.  COMPND   *METHYL BROMIDE (AT -120 DEG./C)

EX   9.  COMPND   *ERYTHROMYCIN /A $HYDROIODIDE DIHYDRATE
                  (ABSOLUTE CONFIGURATION)

EX  10.  COMPND   *TETRACYANOQUINODIMETHANE - 1,2-DIMETHYL-/N-
                  ETHYLBENZIMIDAZOLIUM COMPLEX ACETONITRILE SOLVATE

EX  11.  COMPND   ALPHA-(7-(5-T-*BUTYL-1,2-DITHIOLE-3-YLIDENE)-
                  4,5,6,7-TETRAHYDRO-1,2-BENZODITHIOLE-3-YLIDENE)-
                  ACETOPHENONE

EX  12.  COMPND   BIS(T-*BUTYL-ISOCYANIDE)-(/N-T-
                  BUTYLDICYANOKETENIMINE) NICKEL(O)
```

Figure 3. Examples of compound and synonym names.

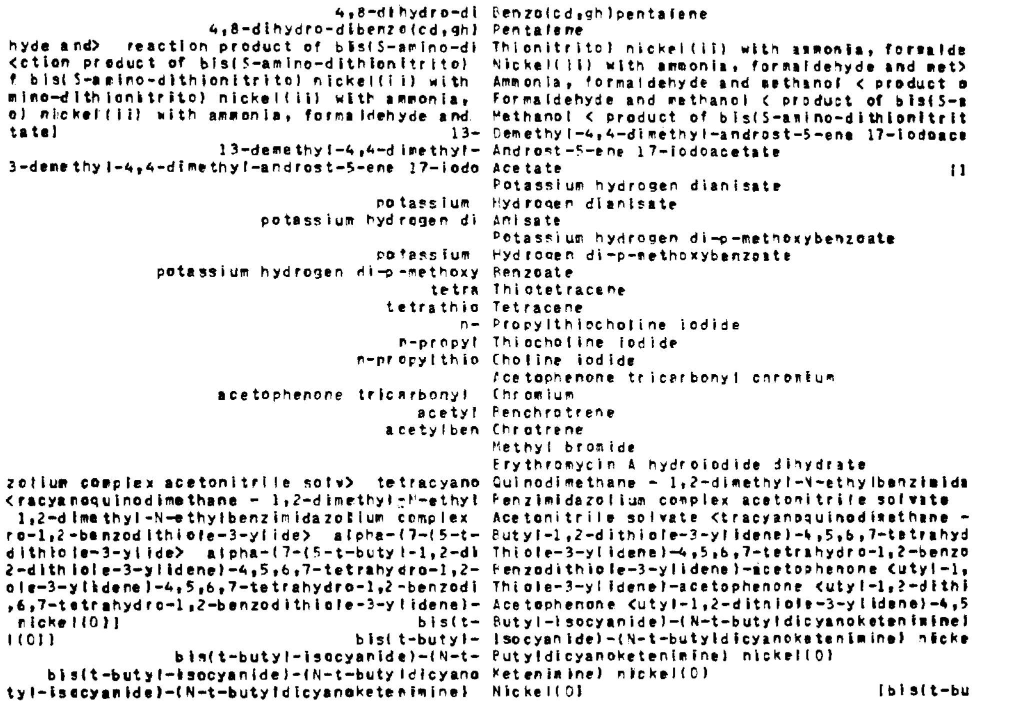
(i) Wraparound left, no truncation. The final part of the name is carried over to the prefix and ends with].
(ii) Wraparound left, name truncated. Name ends with > in prefix.
(iii) Wraparound right, no truncation. The start of the name occurs in the suffix indicated by [.
(iv) Wraparound right, name truncated. The actual start of the name is lost; name starts as early as possible with < in the suffix.
(v) Name truncated both left and right. Occurs in long names; prefix begins with < and suffix ends with >.

The reformatted names together with index reference points (example numbers in Figures 4 and 5) are written to disc structured as Figure 4. The final index, sorted alphabetically on keywords and shown in Figure 5, is produced using IBM sort/merge software as implemented on the University of Cambridge IBM 370/165 computer.

User Options. The user may specify the output print style, the total character width of the final index and the parameter(s) required as index reference points. There are three print options: (1) upper case with no interpretation of typesetting signals; (2) upper case with subscript and superscript indicators interpreted as parentheses, and other signals removed; (3) upper/lower case with subscript and superscript as above, and other signals fully interpreted. The printed index, consisting of the name field and index reference points, has a maximum total print width of 130 characters. This may be altered by the user to any number between 60 and 130 characters. The example of Figure 5 has a print width of 100 characters. In practice indexes with a total width less than 80 characters suffer significant context loss. A variety of index reference points are available, many being specific to the needs of the Centre. Commonly used options are the eight-character reference code, the entry number from the "Molecular Structures and Dimension" series,[11] or both.

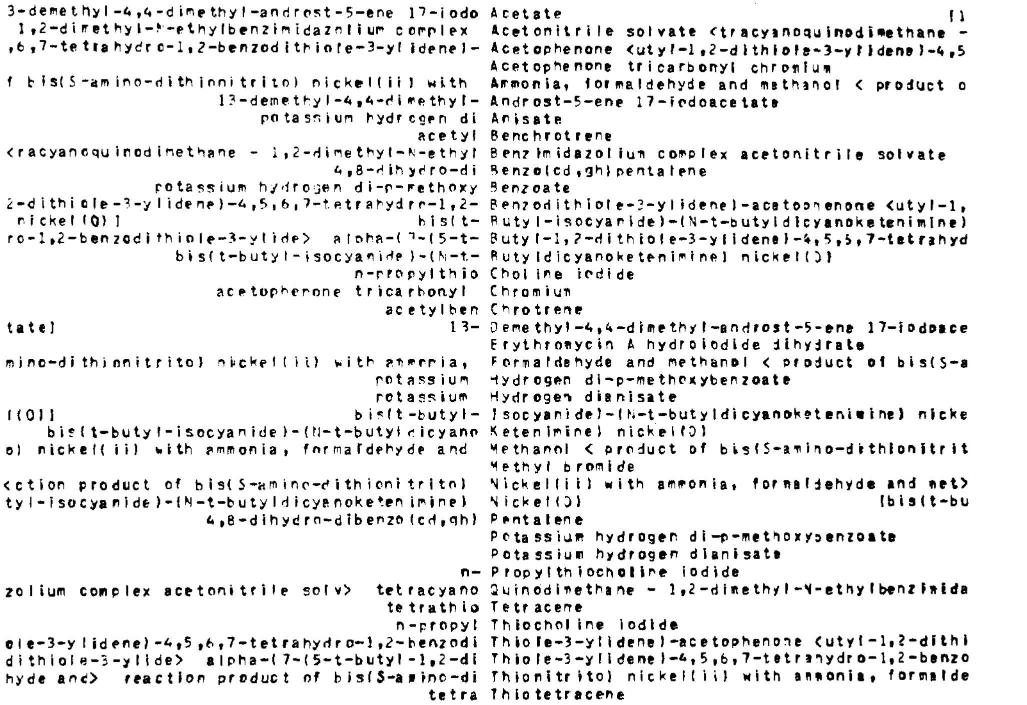
The choice of these items allows the user rapid access into the files[10,12,13] via the reference code, or simply access to the literature reference in printed form via the reference books.[11]

The provision of extensive user options means that the program may be readily transferred to other computers which



```
                               4,8-dihydro-di Benzo(cd,gh)pentalene                           1
                   4,8-dihydro-dibenzo(cd,gh) Pentalene                                       1
hyde and>  reaction product of bis(S-amino-di Thionitrito) nickel(ii) with ammonia, formalde   2
<ction product of bis(S-amino-dithionitrito)  Nickel(ii) with ammonia, formaldehyde and met>   2
f bis(S-amino-dithionitrito) nickel(ii) with  Ammonia, formaldehyde and methanol < product o   2
mino-dithionitrito) nickel(ii) with ammonia,  Formaldehyde and methanol < product of bis(S-a   2
o) nickel(ii) with ammonia, formaldehyde and  Methanol < product of bis(S-amino-dithionitrit   2
tate]                                     13- Demethyl-4,4-dimethyl-androst-5-ene 17-iodoace   3
                    13-demethyl-4,4-dimethyl- Androst-5-ene 17-iodoacetate                    3
3-demethyl-4,4-dimethyl-androst-5-ene 17-iodo Acetate                                     [1   3
                                              Potassium hydrogen dianisate                    4
                                    potassium Hydrogen dianisate                              4
                        potassium hydrogen di Anisate                                         4
                                              Potassium hydrogen di-p-methoxybenzoate         4
                                    potassium Hydrogen di-p-methoxybenzoate                   4
              potassium hydrogen di-p-methoxy Benzoate                                        4
                                        tetra Thiotetracene                                   5
                                    tetrathio Tetracene                                       5
                                           n- Propylthiocholine iodide                        6
                                     n-propyl Thiocholine iodide                              6
                                 n-propylthio Choline iodide                                  6
                                              Acetophenone tricarbonyl cnromium               7
                     acetophenone tricarbonyl Chromium                                        7
                                       acetyl Benchrotrene                                    7
                                    acetylben Chrotrene                                       7
                                              Methyl bromide                                  8
                                              Erythromycin A hydroiodide dihydrate            9
zolium complex acetonitrile solv>  tetracyano Quinodimethane - 1,2-dimethyl-N-ethylbenzimida  10
<racyanoquinodimethane - 1,2-dimethyl-N-ethyl Benzimidazolium complex acetonitrile solvate    10
 1,2-dimethyl-N-ethylbenzimidazolium complex  Acetonitrile solvate <tracyanoquinodimethane -  10
ro-1,2-benzodithiole-3-ylide>  alpha-(7-(5-t- Butyl-1,2-dithiole-3-ylidene)-4,5,6,7-tetrahyd  11
dithiole-3-ylide>  alpha-(7-(5-t-butyl-1,2-di Thiole-3-ylidene)-4,5,6,7-tetrahydro-1,2-benzo  11
2-dithiole-3-ylidene)-4,5,6,7-tetrahydro-1,2- Benzodithiole-3-ylidene)-acetophenone <utyl-1,  11
ole-3-ylidene)-4,5,6,7-tetrahydro-1,2-benzodi Thiole-3-ylidene)-acetophenone <utyl-1,2-dithi  11
,6,7-tetrahydro-1,2-benzodithiole-3-ylidene)- Acetophenone <utyl-1,2-ditniole-3-ylidene)-4,5  11
 nickel(0)]                            bis(t- Butyl-isocyanide)-(N-t-butyldicyanoketenimine)  12
l(0)]                            bis(t-butyl- Isocyanide)-(N-t-butyldicyanoketenimine) nicke  12
                bis(t-butyl-isocyanide)-(N-t- Butyldicyanoketenimine) nickel(0)               12
    bis(t-butyl-isocyanide)-(N-t-butyldicyano Ketenimine) nickel(0)                           12
tyl-isocyanide)-(N-t-butyldicyanoketenimine)  Nickel(0)                            [bis(t-bu  12
```

Figure 4. Results of syntax analysis of names in Figure 3; the example numbers are used as index reference points.



```
3-demethyl-4,4-dimethyl-androst-5-ene 17-iodo Acetate                                     [1   3
 1,2-dimethyl-N-ethylbenzimidazolium complex  Acetonitrile solvate <tracyanoquinodimethane -  10
,6,7-tetrahydro-1,2-benzodithiole-3-ylidene)- Acetophenone <utyl-1,2-dithiole-3-ylidene)-4,5  11
                                              Acetophenone tricarbonyl chromium               7
f bis(S-amino-dithionitrito) nickel(ii) with  Ammonia, formaldehyde and methanol < product o   2
                    13-demethyl-4,4-dimethyl- Androst-5-ene 17-iodoacetate                    3
                        potassium hydrogen di Anisate                                         4
                                       acetyl Benchrotrene                                    7
<racyanoquinodimethane - 1,2-dimethyl-N-ethyl Benzimidazolium complex acetonitrile solvate    10
                               4,8-dihydro-di Benzo(cd,gh)pentalene                           1
              potassium hydrogen di-p-methoxy Benzoate                                        4
2-dithiole-3-ylidene)-4,5,6,7-tetrahydro-1,2- Benzodithiole-3-ylidene)-acetophenone <utyl-1,  11
 nickel(0)]                            bis(t- Butyl-isocyanide)-(N-t-butyldicyanoketenimine)  12
ro-1,2-benzodithiole-3-ylide>  alpha-(7-(5-t- Butyl-1,2-dithiole-3-ylidene)-4,5,6,7-tetrahyd  11
                bis(t-butyl-isocyanide)-(N-t- Butyldicyanoketenimine) nickel(0)               12
                                 n-propylthio Choline iodide                                  6
                     acetophenone tricarbonyl Chromium                                        7
                                    acetylben Chrotrene                                       7
tate]                                     13- Demethyl-4,4-dimethyl-androst-5-ene 17-iodoace   3
                                              Erythromycin A hydroiodide dihydrate            9
mino-dithionitrito) nickel(ii) with ammonia,  Formaldehyde and methanol < product of bis(S-a   2
                                    potassium Hydrogen di-p-methoxybenzoate                   4
                                    potassium Hydrogen dianisate                              4
l(0)]                            bis(t-butyl- Isocyanide)-(N-t-butyldicyanoketenimine) nicke  12
    bis(t-butyl-isocyanide)-(N-t-butyldicyano Ketenimine) nickel(0)                           12
o) nickel(ii) with ammonia, formaldehyde and  Methanol < product of bis(S-amino-dithionitrit   2
                                              Methyl bromide                                  8
<ction product of bis(S-amino-dithionitrito)  Nickel(ii) with ammonia, formaldehyde and met>   2
tyl-isocyanide)-(N-t-butyldicyanoketenimine)  Nickel(0)                            [bis(t-bu  12
                   4,8-dihydro-dibenzo(cd,gh) Pentalene                                       1
                                              Potassium hydrogen di-p-methoxybenzoate         4
                                              Potassium hydrogen dianisate                    4
                                           n- Propylthiocholine iodide                        6
zolium complex acetonitrile solv>  tetracyano Quinodimethane - 1,2-dimethyl-N-ethylbenzimida  10
                                    tetrathio Tetracene                                       5
                                     n-propyl Thiocholine iodide                              6
ole-3-ylidene)-4,5,6,7-tetrahydro-1,2-benzodi Thiole-3-ylidene)-acetophenone <utyl-1,2-dithi  11
dithiole-3-ylide>  alpha-(7-(5-t-butyl-1,2-di Thiole-3-ylidene)-4,5,6,7-tetrahydro-1,2-benzo  11
hyde and>  reaction product of bis(S-amino-di Thionitrito) nickel(ii) with ammonia, formalde   2
                                        tetra Thiotetracene                                   5
```

Figure 5. Final sorted keyword index for names in Figure 3; the example numbers are used as index reference points.

may not have extended print-chain facilities or which operate on a print-line width of less than 130 characters.

STATISTICS

Some statistics relating to the syntax analysis procedure are in Table V. They are derived from an index based on the 14813 entries (15452 names) in the data base in May 1976. The figures show that an "average" name yields 5.8 potential keywords, and 2.4 are retained for indexing. It should be noted that, although the STOP and START lists contain 347 entries, only 8 comparisons are required to accept or reject each potential keyword.

For our file 92% of all keywords are generated using the STOP list in conjunction with the chemical punctuation symbols; 8% of keywords are generated using the START list while the element name analysis yields less than 0.2%. While this order is to be expected, since STOP contains the most common syllables, the actual figures obtained are atypical. This is due to the generous use of hyphens and spaces in names input to the data base; this partial breakdown of long syntax strings is performed to improve the readability of the bibliographic reference books.[11] For a more syntactically "correct" file, the contributions from the START and element-name ROOT sections would increase significantly.

The resolution of the index may be measured by its ability to recognize important general roots as keywords. This was assessed by comparing the index entries for selected keyword

Table V. Statistics Relating to the Syntax Analysis Procedure

Overall Statistics	
Compound names	14 813
Synonym names	639
Total names	15 452
Potential indexing points located	89 486
Comparisons with STOP and START lists	709 305
Valid keywords from STOP analysis	34 168
Valid keywords from START analysis	2 869
Additional keywords from ROOT analysis	76
Total keywords	37 113

Distribution of Keywords No. of keywords	No. of names	%
1	3703	24.0
2	5429	35.1
3	3836	24.8
4	1695	11.0
5	551	3.6
6	177	1.1
7	48	0.3
8	9	0.1
9	2	
10	1	
11	1	

roots (e.g., *adenos*, *penicill*, *urea*, *cholin*, *barbit*, etc.) with the results from the program BIBSER (written by Dr. W. D. S. Motherwell). This is a generalized character-string search procedure for this bibliographic data base; the selected strings were input to the chemical name search segment. The index was found to be 99.1% efficient over a sample of 1237 names. Analysis of the two sets of results showed the effectiveness of the START list in revealing embedded roots in certain cases; e.g., recognition of the string *adenos* increased from 89 to 100%, *penicill* from 82 to 100%, and *choline* from 74 to 97%.

It is interesting to note that only 25 names in the file (0.16%) fail to yield any valid index points; i.e., all syllables are present in STOP. In these cases, e.g., *methyl bromide*, example 8 of Figure 4, the first potential index point, *methyl*, is validated and the compound occurs once in the index. This result indicates that the choice of entries for the STOP list is well balanced for the present data base.

USE OF THE INDEX

The index provides rapid entry into the files of the Crystallographic Data Centre on the basis of chemical compound names. The procedure has four major uses for the Centre.

- As a file-search tool in its own right, the index provides instant answers to the very common question, "has the crystal structure of "compound X" been studied?" Even when the answer to this question is "No" the questioner is immediately presented with a full list of related compounds containing the same keyword.
- As an aid to computerized bibliographic searches, e.g., using the program BIBSER mentioned above. The user wishing to access references via compound names can obtain a good idea of the number of "hits" to expect. This knowledge may suggest a refinement of the question to extend or decrease this list, or may supply the necessary answers directly, without recourse to the computer search, if the number of hits is small enough for manual reference retrieval.
- To re-present current awareness listings in a brief form, easily scanned by the chemist. Each subscriber to our Current Awareness service[13] receives a full listing of each file update (about 250 compounds every 6 weeks) together with a copy of the index run against the update file.
- Within the Centre the index is also used as an aid to the registration of new material and the standardization of the file with respect to chemical syntax and classification.

DISCUSSION

The printed index, to which the statistics of Table V refer, required 108K bytes of storage on a 512K IBM 370/165 computer. Using the highly structured card image file as input, the syntax analysis, interpretation of typesetting signals, and reformatting required 6.6 min of central processor time to produce the 37 113 upper/lower case index records. These were sorted in 12.3 s using IBM Sort/Merge software. The programming language used is Fortran IV.

The index is now being maintained routinely, a new version being produced after each six-weekly update of the data base; the cpu and input/output overheads are nontrivial. Two changes to our bibliographic file will greatly reduce cpu time. Firstly we are devising a new file structure, internal to the Centre, in which the compound name will be held in its explicit upper/lower case form as a variable length record. Secondly the syntax analysis routine will be built into our file-updating programs so that the pointers to the first character of each indexable keyword may be stored on file. This was not possible in the development stage because the content of the STOP and START lists has been constantly changing. After a trial period of 9 months it was found that these lists needed minimal editing to cope with new syntax situations. We would need to run the syntax analysis process on the master file only when the lists are changed, and this should not be more than once per year.

The I/O overheads have been greatly reduced by the recent use of microfiche output. The index is written to magnetic tape in Cambridge in the format required to drive an FR80 film plotter at the SRC Rutherford Computer Laboratory. The present index of 619 pages occupies 3.2 192-page fiche, and multiple copies may be made photographically at low cost.

This index forms part of a fully integrated printed index system for the bibliographic file which will be described in a later publication.

ACKNOWLEDGMENT

We thank the Science Research Council for financial support and the staff of the University of Cambridge Computer Centre for their advice and cooperation. We also thank the staff of the Cambridge Crystallographic Data Centre for much helpful discussion and constructive criticism, especially Olga Kennard, David Watson, Sam Motherwell, and John Rodgers.

LITERATURE CITED

(1) M. F. Lynch, "Computer-Based Information Services in Science and Technology-Principles and Techniques", Peter Peregrinus, Stevenage, England, 1974, pp 49–50.

(2) A. K. Haas, *J. Chem. Doc.*, **5**, 160 (1965).

(3) B. Altmann, *J. Chem. Doc.*, **6**, 154 (1966).

(4) A. E. Petrarca, S. V. Laitinen, and W. M. Lay, *J. Chem. Doc.*, **11**, 148 (1971).

(5) H. Skolnik, *J. Chem. Doc.*, **10**, 216 (1970).

(6) S. Kirschner, S. H. Kravitz, and J. Mack, *J. Chem. Doc.*, **6**, 213 (1966).

(7) J. Villareal, "A General Model for Storage and Retrieval of Chemical Structures", Thesis, Texas A&M University, 1974, pp 113–115.

(8) O. Kennard, D. G. Watson, F. H. Allen, N. W. Isaacs, W. D. S. Motherwell, R. C. Pettersen, and W. G. Town, "Molecular Strucrtures and Dimensions", Vol. A1, Oosthoek, Utrecht, 1973.

(9) F. H. Allen, N. W. Isaacs, O. Kennard, W. D. S. Motherwell, R. C. Pettersen, W. G. Town, and D. G. Watson, *J. Chem. Doc.*, **13**, 211 (1973).

(10) O. Kennard, D. G. Watson, and W. G. Town, *J. Chem. Doc.*, **12**, 14 (1972).

(11) O. Kennard, and D. G. Watson, "Molecular Structures and Dimensions", Vol. 1–3, with W. G. Town, Vol. 4 and 5, Oosthoek, Utrecht, 1970, 1972, 1973, 1974; with F. H. Allen, and S. M. Weeds, Vol. 6 and 7, Bohn, Scheltema and Holkema, Utrecht, 1975, 1976.

(12) F. H. Allen, O. Kennard, W. D. S. Motherwell, W. G. Town, and D. G. Watson, *J. Chem. Doc.*, **13**, 119 (1973).

(13) O. Kennard, D. G. Watson, F. H. Allen, W. D. S. Motherwell, W. G. Town, and J. R. Rodgers, *Chem. Br.*, **11**, 213 (1975).