Table VI. Languages Used in Sections 14 and 15

(% frequency is given in parentheses)

Rank	Previous study		Section 14		Section 15 (papers)	
1	English	(44)	Russian	(37)	Russian	(45)
2	Russian	(20)	English	(37)	English	(21)
3	German	(17)	German	(12)	Japanese	(12)
4	French	(6)	French	(6)	Polish	(5)
5	Japanese	(3)	Japanese	(3)	Chinese	(5)
6	Italian	(3)	Italian	(3)	Romanian, German	(4)
	11 Others	(7)	13 Others	(3)	9 Others	(8)

The relative importance of various languages is shown in Table VI. The order for Section 14 has not changed since 1958 except for the reversal of English and Russian. This change is quite drastic, however, the ratio of English to Russian going from approximately 2:1 to 1:1. This change results almost solely from the increased importance of the *Zh. Neorgan. Khim.*

Actually Russian is now slightly more common than English, the difference having been lost in rounding off. Essentially the same order holds for patents (Table III) except that Russian runs a very poor third after German. The over-all importance of Japanese is worth noting. In view of the fact that French makes such a poor showing compared to Russian it seems unreasonable to accept it even as an alternative to Russian for the Ph.D. language requirements. This is especially so since there are no indispensible compendia in French such as there are in German.

REFERENCES

(1) R. F. Trimble, *J. Chem. Educ.*, **37**, 419 (1960).
(2) F. S. Wagner, Jr., "A Study of Patent Documentation," presented at the 133rd National Meeting of the ACS, April, 1958.

Chemical-Biological Activities A Computer-produced Express Digest*

By G. MALCOLM DYSON and MICHAEL F. LYNCH

Chemical Abstracts Service, Research and Development Division, The Ohio State University, Columbus 10, Ohio

Received September 17, 1962

It has been recognized for some time that a more rapid survey of the literature dealing with the biological activity of chemical compounds is needed. The normal procedures following publication of an original paper, until a printed abstract is available, occupy about five months, and cannot be compressed into a much shorter time; additional time is necessary before the index is available, so that an abstract system is much more a tool for retrospective searching than for current awareness. Some progress has been made toward a more rapid current awareness service by the introduction of *Chemical Titles*,[1] but as it is presently constituted, this publication is limited in its scope by the very obvious fact that a title is neither an abstract nor a paper. It is probable that each of the many subsections of the chemical discipline could benefit by a prompt current awareness service; we selected the interrelation of structure and biological activity as one such subsection and have conducted our experiments exclusively with it, in the hope that the results obtained might encourage the production of analogous journals for other subsections of chemistry.

DESIDERATA

The desiderata of a publication promoting current awareness in the field of chemical/biological activities are:

1. Promptness
2. A complete account of the primary source, including the name(s) and address of the author(s), from which the data have been obtained
3. A complete description of the substances employed, with, where possible, their structural formulas
4. A description of the biological conditions under which the observations were made and an account of the results obtained
5. Reasonable legibility of the completed publication
6. A good set of indexes
7. A method of storage of data such that the cumulation of indexes is rapid and comparatively inexpensive
8. A wide coverage

Although not included in the desiderata listed above, it has been kept in mind that in gathering and processing data for this publication, the data should be accumulated in a form which can be transferred without change to our tape storage system. Each of the above heads will be considered in turn.

1. Promptness.—This will be ensured by having copies of the originals delivered direct to the journal office, and worked on by a group of data analysts who have the necessary technical and linguistic ability to prepare the digests. The analysts encode those parts of the information which are required as instructions for the keypunchers.

Our aim is to publish the material within 14 days of the receipt of the journal in our office.

It is proposed to publish bi-weekly, and it is expected that cumulative indexes will be prepared.

* Presented before the Division of Chemical Literature, ACS National Meeting, Atlantic City, N. J., September 11, 1962.

2. Bibliographic Details.—The journal title is given in full at the head of each group of abstracts in the digest section of the publication, but in individual references within the section the four-letter coden is used. This coden system, originated by Bishop[2] and developed conjointly by the ASTM and the Research and Development Division of CAS, has been used for some time, successfully, both in *Chemical Titles* and in an ASTM publication, *Numerical List of Abstracted Ultraviolet and Visible Spectra Indexed on Wyandotte-ASTM Punched Cards.*[3] The ASTM is to publish shortly a reference book of coden, which covers some 20,000 technical publications.

The authors' names and address are given at the head of each digest, with (when known) the date of acceptance by the primary journal.

3. Chemical Descriptors.—The nature of the chemical substances examined is given as fully as possible. In most cases, the name used is that used by the author; for many complex structures, the time available between receipt of material and publication is too short for weighty problems of nomenclature to be resolved. Since, therefore, the authors' nomenclature may be unfamiliar, the molecular formula and official IUPAC notation[4] are added for each substance described. The notation affords a brief, unique description of each organic compound, inclusive of stereochemical detail, from which the structure can be regenerated with ease. Indexes based on the notation are simple to use, and group similar structures in a manner not achieved by conventional name indexes. Furthermore, substructure analyses at various levels of sophistication are feasible, thus enabling far-reaching structure–activity relationships to be uncovered once an appreciable amount of data has been entered into the sytem. In addition, the *CA* Registry Number (See Appendix 1) is added. A typical entry may therefore read:

$C_6H_5 \cdot CO \cdot S \cdot CH_2CH_2NH_2$ 0001172 $C_9H_{11}NOS$

B6:CEQ;S/2C$_2$N S-benzoyl-2-mercaptoethylamine

which is adequate to specify the structure. The structural formula will be placed on the opposite page to the appropriate digest, and will have the Registry Number as its distinguishing mark.

4. Biological Descriptors.—One of the main objectives of the research leading to the publication of *Chemical-Biological Activities* has been to develop a system of coding biological activity concepts in such a way that the computer could handle them. It is not proposed, in this general introduction, to deal in detail with the method used; this is deferred to a later part of our paper. Here it suffices to say that the system used is that of a partly facetted machine language, with computer dictionary memory which can translate the condensed machine code to fairly respectable English. It was decided at the outset of these investigations that no product would be considered satisfactory unless it was printed out by the computer devices in direct language; thus, although the machine memory holds the biological concepts in condensed code (machine language) this code never appears in the product.

5. Legibility.—Experience with many printouts, especially that of *Chemical Titles*, has taught us that the use of capital letters and figures with a few punctuations marks not only restricted the language used, but was also difficult to read. With this in view, we collaborated with the IBM Company in the production of a machine (Modified 1403 Printer) which would not only give the normal capital letters and figures but in addition, lower case letters, numerical subscripts and superscripts, underlining, some Greek and italic letters and a substantial range of punctuation and allied marks. This is now available as the "120 Character Chain (CAS Character Set Model 1)." The character set is shown in Fig. 1.

We have experimented with this printer, which operates at 240 lines per minute, so that a full page of the new publication is printed, apart from editing, in approximately 20 seconds. The output of the printer is reduced to 72.5% during photography in preparation for litho-offset reproduction. Since the output print of the 1403 is "broad," attempts have been made to condense this type with anamorphic lenses.

The formulas are set and photographed by the method described by one of us in 1961.[5]

In all, there is a substantial gain in readability and we consider our introduction of this computer-printer to be an important step forward.

6. Indexes.—There are five ways in which a searcher may enter the system: by directly scanning the structure pages, or by the use of author, molecular formula, notation, chemical name and general concept indexes. The chemical name and concept indexes are combined in one general alphabetical list. Thus, a searcher can approach

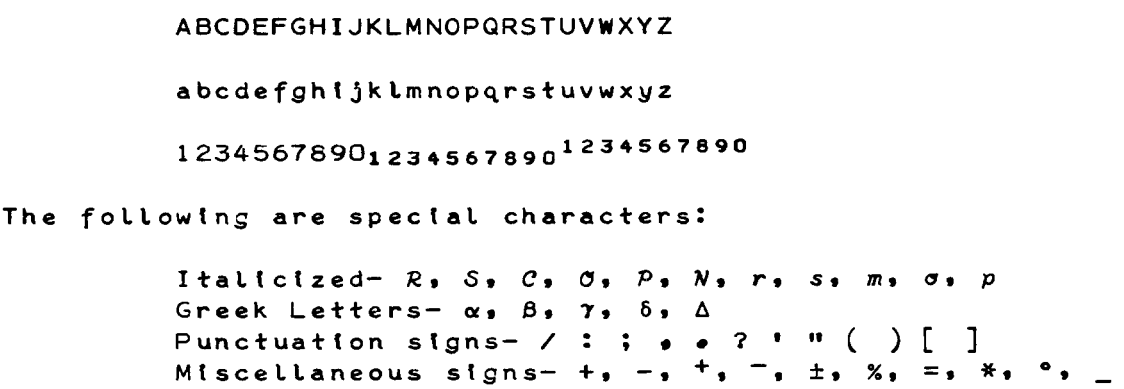

Fig. 1.

his field either from the chemical, biological or author aspect; he may look for compounds, animals, infective agents, biological structures, or writers. The indexes are compiled by computer and the concept index is of the keyword type so that a certain amount of context is given with the indexed word; this serves to help in distinguishing those index references which are particularly pertinent to any given search.

7. Data Storage.—Since the indexes are produced by IBM 1401 computer and IBM 1403 printer, the cumulation of indexes is a relatively simple process. It is anticipated that cumulated indexes will be produced at least at semi-annual intervals. Although it does not add anything to the publication as such, it may be noted that the system of input which is used for CBAC is entirely compatible with our storage and retrieval system so that when sufficient time has elapsed, correlative searches on structural features and biological concepts will be possible without additional records.

8. Wide Coverage.—Initially, data will be obtained for CBAC from a group of 300 journals, some of which are not currently abstracted by *CA*. The list of journals used will be expanded when CBAC has passed the initial stages. It is hoped that the provision of this new service will render it unnecessary otherwise to abstract papers which contain merely lists of substances and the results of screening tests. It may even prove that authors of such papers, which are, in fact, data sheets, may be content to have them published in CBAC. This applies particularly to screening tests which (as with anti-mitotic agents) are mostly entirely concerned with absence of biological activity. It seems desirable that authors should be encouraged to publish this type of material only in express publications of the nature we now describe. Since all such information will be automatically transferred and accumulated in a general scheme for retrieval of chemical knowledge, it will in no sense be lost.

THE BIOLOGICAL CONCEPT CODE

Whereas the nature of chemical structure is such that it can, and should, be placed into the system precisely, the concepts of the interaction of a chemical substance and its environment need organization before they can be made available for computer handling. In general principle, our method of depositing such data is to arrange them in a predetermined order, so that the following attributes are treated in turn (see Table I), after the statement of the exact nature of the substance or substances concerned.

Characteristic codes are associated with each of these operators; some of these are reproduced in Appendix 2 to this paper.

The codes for organs (operators E and F), conditions, diseases and side-effects (operators G and R), and operations (operator K) are adopted from the *Standard Nomenclature of Diseases and Operations*,[6] while those for microörganisms are derived mainly from *Bergey's Manual of Determinative Bacteriology*.[7]

Greatest interest centers perhaps on the codes for the action and modification of action of the test-compound (operators D and U), which were devised for maximum economy in coding and dictionary-match time. Five types of codes are distinguished here :

Table I

A	Comparison of activity
B	Type of drug
C	Degree of activity
D	Action
E	Organ
F	Isolated organ
G	Disease or condition
H	Pathogenic organism
J	Pathogenic compound
K	Operation
L	State
M	Sex
N	Subject
P	Host
Q	Comment
R	Side-effects
S	Mode of administration
T	Metabolite
U	Modification of action of test compound

(I and O are not used to avoid confusion with "one" and "zero.")

(a) simple verbs, expressed by independent one- or two-digit codes, such as
D1 "increases," and D5 "inhibits"

(b) compound verbs, expressed by independent two-digit codes, such as
D47 "has diuretic action" and
D50 "has carcinogenic action"

(c) complex verbs, expressed by a combination of one- and two-digit codes, such as
D112 "increases growth of" and
D212 "decreases growth of"

(d) verbs denoting modification of the biological activity of another compound, expressed by combination of the one-digit codes given in (a) and the compound verbs given in (b), such as
D5*42 "inhibits anticholinergic action of" and
D1*37 "increases analgesic action of"
Furthermore, modification of the action of the test-compound by another can be expressed by means of

(e) a past participle at a secondary position in the sentence, denoted by operator U. Thus,
U03 "action increased by" and
U12 "action not blocked by"

Certain operators cause a standard phrase to be prefixed to the decoded content at printout; thus, "isolated" is prefixed to "organ" by use of F instead of E, and "caused by" inserted by H or J, accompained by" by R, "given" by S, and "metabolized to" by T.

It should be noted that terms for which entries have not yet been made in the dictionary can be accommodated in any operator by using parentheses. Thus,

D30(hyperglycemic) denotes "has hyperglycemic action" and N(giraffe) denotes "giraffe"

Dictionary updating ensues in retrospect on the basis of the frequency of occurrence of additional terms.

Conjunctions are encoded by means of special characters, prepositions by means of lower-case letters as shown in Appendix 2.

It is the present duty of the encoder to complete for each entry into the journal a form which constitutes the instructions to the keypunch operator. This machine language appears thus*:

(Compound A)A5(Compound B)D18[140]Nc113S15ΔD212N113-ΔD111E680Na113....

The interpretation of these signs can be followed from the codes in Appendix 2, but to simplify matters for the reader, the interpretation is given in Table II. An adtional example is given in Table III.

Table II

Code entry	Significance
(Compound A)	Compound A
A5	, less effective than
(Compound B)	Compound B
D18	has toxic action
[140]	[LD_{50}, 140 mg./kg.]
Nc113	on rat
S15	given intravenously
Δ	, and
D212	decreases growth of
N113	rat
Δ	, and
D111	increases size of
E680	liver
Na113	of rat

Table III

Code entry	Significance
(Compound A)	Compound A
B72	, diabetogenic,
D115	increases content of
(Compound B)	Compound B
Eb510	in plasma
Na113;	of rat;
D30(hyperglycemic)	has hyperglycemic action
Mc1	on male
N113	rat
U11	, action blocked by
(Compound C)	Compound C

While the machine language is sufficiently concise for storage in magnetic tape, it is, of course, unintelligible to anyone but an expert document analyst. With this in mind, we have devised a program that acts as a dictionary between the tape reader and the printout. This dictionary interprets the symbolic language and instructs the printer to print in open language. In this way, a readable product has been obtained. We do not claim that the sentences are completely grammatical, but we have programmed the machine to distinguish between singular and plural subjects and to adjust the verbs accordingly and have also programmed sufficient prepositions and conjunctions to enable the use of phrases and secondary clauses.

The use of the symbolic machine language is essential in that it uses only about one-fifth of the symbols of the corresponding open language, with consequent saving of 80% of tape and machine time. In addition, the discipline of such coding ensures that the concepts are encoded uniformly in such a way that retrospective retrieval will be certain. For example, even if open language was not precluded on other grounds, confusion would arise when the same concept is expressed in different ways, as with "renal" and "kidney," "liver" and "hepatic," "lung" and "pleural."

Some simplification of complex information (thesauric organization of input) is unavoidable with these procedures, but the display of data is intended to be substantially indicative and is specific in a few areas only. Provision for incorporation into a scheme such as this of all possible details would decrease its timeliness and destroy the "browsability" of the product.

The following example of a typical entry in machine language and as printed out in open language is given in Fig. 2.

* The specification of the two compounds has been omitted from this example to avoid confusion.

```
JACS-0084-2601-10 STRUCTURES OF PARTHENIN AND AMBROSIN.
Herz W, Watanabe H, Miyazaki M, and Kishuda Y;
Florida State Univ., Tallahassee, Fla. Recd.
Feb. 2, 1962.
"1723405=C15H18O4=A752 13ZQ10EC8C35EQ911Q4=*
parthenin"D74ΔD40%R(high toxicity)ΔC1D30(antibacterial)
ΔC0D47ΔC0D30(saluretic)c(saline-loaded)M2N113/
"1723405===*parthenin"C0D512(leukemia P-1534)ΔC1D512Gk8091
Q[but not sufficiently for further investigation]/


STRUCTURES OF PARTHENIN AND AMBROSIN. JACS-0084-2601-10
Herz W, Watanabe H, Miyazaki M, and Kishuda Y; Florida State Univ., Tallahassee, Fla.
Recd. Feb. 2, 1962.

1723405=C15H18O4=A752 13ZQ10EC8C35EQ911Q4= parthenin has CNS depressant action and has
antiadrenergic action but accompanied by high toxicity and weakly has antibacterial
action and does not have diuretic action and does not have saluretic action on
saline-loaded female rat.

1723405===parthenin does not inhibit growth of Leukemia P-1534 and weakly inhibits growth
of adenocarcinoma [but not sufficiently for further investigation].
```

Fig. 2.

The authors have received a grant (RG-9772) from the National Institutes of Health of the Department of Health, Education, and Welfare in support of this research. In addition, many of the members of the Research and Development Division of this Service have given valuable assistance during the course of the work and we would like to refer particularly to the especially hard work done by the computer programming section under the direction of Mr. W .E. Cossum.

APPENDIX 1

The *CA* Registry Number is a seven digit serial number randomly applied to all structures, so that each structure has its unique Registration Number. Such a number serves as an address for the compound throughout our storage and retrieval system, of which CBAC is an integral part. It is hoped to publish a Handbook of *CA* Registry Numbers in due course.

APPENDIX 2

Extracts from Dictionary

Operator A:
1. much more effective than
2. more effective than
3. equal in action to
4. similar in action to
5. less effective than
6. much less effective than
7. unlike

Operator B:
31. general anesthetic
32. local anesthetic
33. sedative
34. hypnotic, etc.

Operator C:
0. do(es) not
1. weakly
2. moderately
3. strongly
4. very strongly

Operator D:

Simple one- and two-digit codes, used independently
0. alters
1. increases
2. decreases
3. stimulates
4. initiates
5. inhibits
6. blocks
7. maintains
8. potentiates
9. prevents
10. causes
11. damages, etc.

Two-digit codes, used independently
31. has general anesthetic action
32. has local anesthetic action
33. has sedative action, etc.

Two-digit codes, used in combination only
-10. ..... symptoms of
-11. ..... size of
-12. ..... growth of
-13. ..... action of (drug)
-14. ..... function of (organ)
-15. ..... content of, etc.

Operator L:
1. ovum of
2. fetal
3. newborn
4. immature
5. young
6. mature
7. senile

Operator M:
1. male
2. female

Operator N:
102. human
103. monkey
104. rhesus monkey
105. dog
106. cat, etc.

Operator S:
1. by aural installation
2. by cardiac implantation
3. by inhalation
4. by injection
5. intra-arterially
6. intra-articularly, etc.

Operator U:
01. action potentiated by
02. action not potentiated by
03. action increased by, etc.

Conjunctions:

Δ and
$\complement$ but
+ and
$\pm$ or
= followed by

} within the content of an operator

Prepositions:
a of
b in
c on
d by
e with
f measured by
g against
h to

BIBLIOGRAPHY

(1) R. R. Freeman and G. M. Dyson, Development and Production of *Chemical Titles*, a Current Awareness Index Publication Prepared with the Aid of a Computer, *J. Chem. Doc.*, **3**, 16 (1963).
(2) C. Bishop, *Am. Doc.*, **4**, 54 (1953).
(3) *Fourth Supplement to Numerical List of Abstracted Ultraviolet and Visible Spectra Indexed on Wyandotte-ASTM Punched Cards*, ASTM, Philadelphia, Pa., March, 1961.
(4) *Rules for I.U.P.A.C. Notation for Organic Compounds*, issued by the Commission for Codification, Ciphering and Punched Card Techniques, Longmans, London, and Wiley, New York, N. Y., 1961.
(5) *Chem. Eng. News*, **39**, 60 (March 27, 1961).