"Substructure Searching of Computer-Readable Chemical Abstracts Service Ninth Collective Index Chemical Nomenclature Files", *J. Chem. Inf. Comput. Sci.*, **15**, 73-84 (1975).
(14) "Chemical Abstracts Service Search Aids for the Ninth Collective Period (1972-1976)", Chemical Abstracts Service, Columbus, Ohio, 1974.
(15) D. C. Owsley, J. J. Bloomfield, A. J. Solodar, and E. Block, "CA Nomenclature", *Chem. Eng. News*, **53** (18), 3 (1975).
(16) M. Dub, "CA Nomenclature", *Chem. Eng. News*, **53** (18), 3 (1975).
(17) A. Robson, and J. S. Longman, "Automatic Aids to Profile Construction", Office for Scientific and Technical Information Report No. 5230, 1975.
(18) F. Wilcoxon, "Individual Comparisons by Ranking Methods", *Biometrics Bull.*, **1** (6), 80-83 (1945).

Comparison between CACon and CASIA Files for Development of New SDI Service in 1977

TILLY BAYARD and JEANNIE PERSOZ*

Association Francaise de Documentation Automatique en Chimie (AFDAC), 75116 Paris, France

Received December 21, 1976

A comparison between CACon and CASIA files is described, based on processing 40 profiles on both files. A new SDI service dealing with chemical structures and using CASIA tapes will be developed in 1977: chemical nomenclature, registry numbers, and molecular formulas will be searchable, thus allowing specific compounds and substructure searching at reasonable costs.

AFDAC is a part of CNIC, the French center licensed by Chemical Abstracts Service (CAS) since 1972. Our mission is to provide the French chemical community with computerized chemical information.[1]

At present we process 470 profiles on *Chemical Abstracts Condensates* (CACon) tapes, 70 profiles on *Polymer Science and Technology* (POST) tapes, and 20 profiles on *Chemical Industry Notes* (CIN) tapes.

Since 1974 we have processed on-line retrospective searches on CACon, and on other scientific, technical, and technico-economic files: BIOSIS, CAIN, COMPENDEX, INSPEC, METADEX, NTIS, PREDICASTS, etc.

AIM OF THE EXPERIMENT

In the field of chemistry, the problem of textual information searching has been given satisfactory solutions, both in Selective Dissemination of Information (SDI) and in retrospective search, with the CACon tapes.[2]

The situation is quite different for specific compounds or substructure searching. One of the most satisfactory solutions would be to process the CAS Registry File, with use of appropriate screens and possibility of atom-by-atom search. Several teams are working on this problem.[1,3-7] Two European industrial centers[5,6] provide their users with a retrospective substructure search service using the CAS Registry File. But, until now, such publicly available services did not exist.

Another way of searching the CAS Registry File is via chemical nomenclature.[8] Several American federal agencies use it for searching subfiles of the Registry File.[9-11] This solution seems especially interesting for SDI: even if services using the CAS Registry File with chemical codes become publicly available, updates of the file probably will not be very frequent because of their cost. Searching chemical nomenclature, available in the Chemical Abstracts Subject Index Alert file (CASIA), could be a good substitute for the update time interval.

Therefore, within a working group of the Union des Industries Chimiques (the French chemical manufacturers association which sponsored the foundation of AFDAC in 1970), it was decided, at the end of 1975 to compare the CASIA and CACon files for SDI, in terms of recall, relevancy, currency, and processing costs. This investigation also was expected to develop a thorough knowledge of CASIA, which starts being searchable on-line through the LOCKHEED/DIALOG system.

The development of a new operational service by AFDAC, using CASIA tapes, scheduled for January 1977, will result from this study.

DESCRIPTION OF THE FILES

File Coverage. The CACon and CASIA files include all the documents abstracted in the printed issues. They are organized sequentially, in the *Chemical Abstracts* (CA) number order. The CACon tapes are published weekly, while the CASIA tapes are published bimonthly and correspond roughly to an odd CACon tape plus an even CACon tape. However, index entries corresponding to some documents are published later in CASIA than in CACon. Thus, the corresponding CA index data in a CASIA tape does not include all the index entries for documents covered in the two corresponding CACon tapes, but it may include some for previous CACon issues. This study allowed us to compare the publication delay times of CACon and CASIA.

File Content.[12] The CACon file includes, for each citation, the CA number, title, author(s), bibliographic reference, organization, CA Section and subsection numbers, language, document type, issue keywords, and, for patents, the application country, number, classes, and priority dates. The title and issue keywords are written using uncontrolled vocabulary.

The CASIA file includes, for each citation, the CA number, CA section number, and index entries that will appear in the CA Chemical Substance, General Subject, and Molecular Formula Volume Indexes. These entries use controlled vocabulary at the heading level, though a thesaurus is not used. The chemical compounds are described according to the CAS index nomenclature rules, while general subjects are chosen from a list of predefined headings. The user can acquire a knowledge of this controlled vocabulary with the Index Guide, Volume Indexes, and CASIA Search Aids made available by CAS.

The indexing of documents in the CASIA file may be roughly described in the following way:

(a) Chemical substances are indexed by the CA preferred index name assigned according to the CAS substance index nomenclature and consisting of a heading parent, possibly followed by a substituent, and/or name modification, and/or qualifier, and/or stereo data element. The nomenclature is

```
HEADING PARENT (HP)
     2,4,6(1H,3H,5H)-Pyrimidinetrione
SUBSTITUENT (SB)
     5-ethyl-5-(1-methylbutyl)-
NAME MODIFICATION (NM)
     monosodium salt
REGISTRY NUMBER (RN)
     000057330
MOLECULAR FORMULA (MF)
     C11H18N2O3.Na
TEXT MODIFICATION (TM)
     anterior pituitary and sympathetic nervous system
     response to anesthesia from
CONCEPT HEADING (CH)
     adrenal cortex
TEXT MODIFICATION (TM)
     anesthesia effect on anterior pituitary in relation to
```

Figure 1. CASIA content.

followed by the CA Registry Number and the Molecular Formula of the compound. These are often associated with a natural language text modification.

(b) General subjects are described by the concept headings from the controlled vocabulary and usually accompanied by natural language text modifications.

A sample of indexation is given in Figure 1.

EXPERIMENTAL METHOD

The most important users of our CACon SDI service were consulted for proposing questions, and the AFDAC staff also chose a number of them. The 38 questions (Figure 2) may be divided into five groups: (1) specific compounds; (2) substructures or classes of compounds; (3) general subjects (concepts); (4) specific compounds and concepts; (5) substructures or classes of compounds, and concepts.

The profiles were processed with the text-reading software PRETEXT II, created by the Institut Francais du Pêtrole, FRANLAB, used for our CACon SDI service. Software for translating the CASIA files into search program format was written by J. Delaunay, FRANLAB information specialist, according to AFDAC specifications.

For this investigation, all the data elements on the CACon and CASIA files were searchable, except for CA numbers, bibliographic references, patent numbers, and priority dates. Right and left truncation, Boolean operators, and Context operators (WITH and IGNORE) could be used (Figure 3). The Context operators were particularly useful for CASIA. The unit searched by the Context operators was the "sentence", which consisted of: (a) either a title or a keyword phrase in CACon; (b) heading parent, substituent, name modification, stereo, and text modification, or concept heading and text modification in CASIA. The WITH logic allows linkage of all the information related to specific index headings.

The profiles were coded for both CACon and CASIA, tested on some tapes, amended if necessary before the beginning of the experiment. Search profile examples from each of the five groups are given in Figure 4.

The profiles were run on 12 CACon tapes (Volume 84, issues 15–26) and on the 6 CASIA tapes for the same time period (Volume 76, issues 08–13). CACon and CASIA output samples are given in Figure 5.

RESULTS

(1) Recall and Relevancy. The relevancy of the retrieved items was assessed by the AFDAC staff, using the title and keywords printed in CACon outputs, the index entries printed

```
1) Specific compounds
   101  1,2-polybutadiene
   102  1,2-dimethoxybenzene
   103  phosphorus pentachloride and pentafluoride
   104  thallium (I) inorganic salts
   105  specific fluorinated compounds (sections 1-5)

2) Substructures or classes of compounds
   111  barbituric derivatives
   112  nitrofurfurylidene amino substructure
   113  halohydroxyphenothiazine
   114  two-rings heterocyclic systems : C3NS-C4N2,
        C3NS-C5N, C4S-C4N2, C4S-C5N
   115  thiophenecarboxylic acid derivatives
   121  fluorinated compounds (sections 1-5)
   122  organoboranes
   123  ellipticin derivatives
   124  clathrates
   125  nitroso compounds

3) General subjects (concepts)
   201  waste biogasification by bacteria
   202  breeders or fast nuclear reactors
   203  immobilized enzyme electrodes
   204  radiation physiological effects
   205  granulometric measurements

4) Specific compounds and concepts
   301  nitrogen oxides reduction
   302  hydrogen peroxide used as an oxidant in organic chemistry
   303  depollution of sewages from surface treatment of some
        metals (Cd, Cr, Zn)
   304  PVC use for food packaging
   305  air pollution by carbon monoxide
   306  carbon dioxide used in metallurgy

5) Substructures or classes of compounds, and concepts
   311  indolic hallucinogens
   312  anthraquinonic dyes
   313  thiadiazolic bactericides
   314  benzanthracene derivatives, carcinogenic activity
   315  patents on 2,2-dimethylcyclopropanecarboxylic acid and esters
   321  cycloolefins polymerization
   322  non-phosphorylated insecticides metabolism
   324  structure-activity relationships of crosslinked polymers
   325  phthalates toxicity
   326  sulfur compounds use in cosmetics
   327  carbamates and ureas with herbicidal properties
```

Figure 2. List of questions.

```
Term truncation   *

Search term types
     NOR   registry number term
     SEC   CA section number term
     TXT   text term

CASIA search terms
     HP =  heading parent
     SB =  substituent

Boolean operators
     /     OR
     +     AND
     +-    NOT

Context operators
     AAnn  asymmetric WITH
     ASnn  symmetric  WITH
     IAnn  asymmetric IGNORE
     ISnn  symmetric  IGNORE

     nn = number of spaces (or punctuation marks) between the
          terms linked by the context operators
     nn = 00  in the same word
     nn = 99  in the same "sentence"

Examples :

     001  AA03  002

     term 002 must follow term 001, with no more than 3 spaces
     (or 3 punctuation marks) between them.

     001  AS01  002

     term 002 must precede or follow term 001, with no more than
     1 space (or 1 punctuation mark) between them.

     001  IS00  002

     term 002 must not be present in the same word as term 001
```

Figure 3. Profile coding.

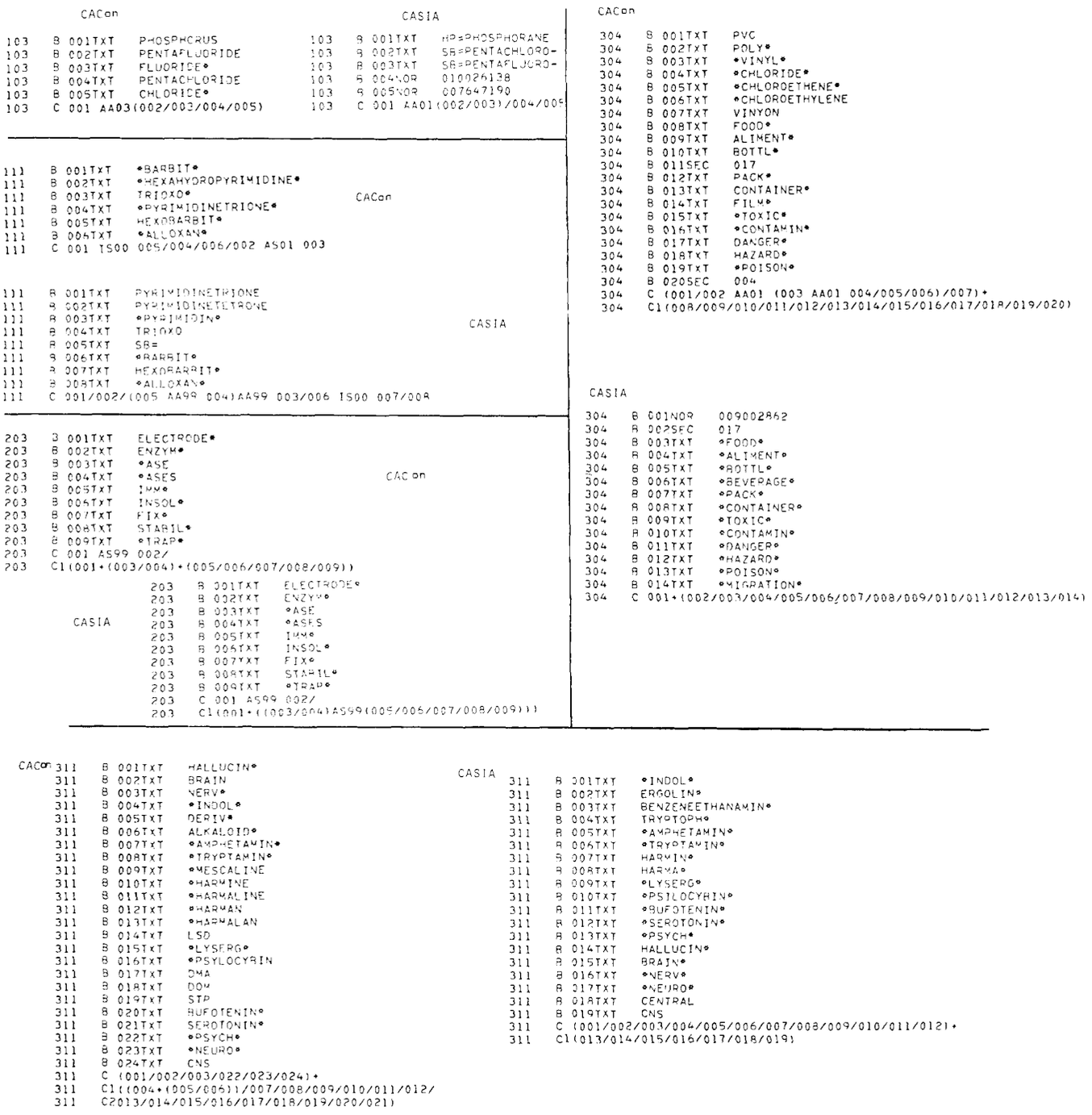

Figure 4. Search profile examples.

in CASIA outputs, and the printed CA abstracts when necessary.

The recall was evaluated with regard to all the relevant hits retrieved on both CACon and CASIA. We did not examine the printed issues for missed references.

The results are displayed in Figures 6 and 7. The questions are separated into three groups:

Questions dealing only with specific compounds or substructures: the CASIA results are much better than the CACon results, especially the recall percentages, 88 and 52%, respectively.

Questions dealing only with general subjects: the CACon results are much better than the CASIA results, especially the recall percentages, equal to 82 and 54%, respectively. This conclusion should be attenuated for three reasons: first, the number of profiles of this type was low in this experiment; secondly, we used more Context operators in CASIA profiles than in CACon profiles, which could lower the recall; thirdly, we were probably more clever for coding the CACon profiles because of our much longer experience with this file.

Questions dealing with both structures and concepts: as for the first group, the CASIA results are considerably better, with 84% recall compared with 64% for CACon.

For most of the profiles, the relevancy is equally good for CACon and CASIA. In some cases, it could still be improved by modifying the coding.

We paid special attention to hits missed on each data base.

Hits missed on CACon were due to the different indexing policy for compounds in CACon and in CASIA. All specific compounds described in primary documents are indexed in CASIA using CAS nomenclature; they are not systematically indexed in CACon, and they are often described by their trivial names. For instance, this type of information loss is important for questions 101–125.

CAC * 4(015)*099431K:001* ANSWERS **a**

EFFECT OF HALOTHANE AND DEHYDROBENZPERIDOL ONCIRCULATION AND CENTRAL SYMPATHETIC ACTIVITY.
SCHULTZ-EM ESCH, J.: TAUBERGER, G.: CLOSTERMANN, M.(ANESTHESIEABT. ME PROCHR. UNIVERSITAETSKLIN.),(BONN GER.).
ANAESTHESIOL. REANIMAT. (BUR. 000075) 93, 188-92 (GE)

HALOTHANE CIRCULATION NERVOUS SYSTEM *DEHYDROBENZPERIDL CIRCULATION NERVOUS SYSTEM *CIRCULATION ANESTHETIC *SYMPATHETIC SYSTEM ANESTHETIC *CENTRAL NERVOUS SYSTEM ANESTHETIC

CA **094(015)*099431K*S:001* CS7608* **b

*/HP=ETHANE/SB=2-BROMO-2-CHLORO-1,1,1-TRIFLUORO-/RN=000151677/MF=C2HBRCLF3/TM=CIRCULATION AND CENTRAL SYMPATHETIC NERVOUS SYSTEM RESPONSE TO, DEHYDROBENZPERIDOL IN RELATION TO,
***/CH=NERVOUS SYSTEM/TM=CENTRAL SYMPATHETIC, DEHYDROBENZPERIDOL AND HALOTHANE EFFECT ON ***/CH=CIRCULATION/TM=DEHYDROBENZPERIDOL AND HALOTHANE EFFECT ON ***/HP=2H-BENZIMIDAZOL-2-ONE/SB=1-(1-(4-(4-FLUOROPHENYL)-4-OXOBUTYL)-1,2,3,6-TETRAHYDRO-4-PYRIDINYL)-1,3-DIHYDRO-/RN=000548732/MF=C22H22FN3O2/TM=CIRCULATION AND CENTRAL SYMPATHETIC NERVOUS SYSTEM RESPONSE TO, HALOTHANE IN RELATION TO

PROFIL : CSA 105 CARTE : 078
AFDAC 88, avenue Kléber, 75116 PARIS Tél. : 553.65.19 ACS, AFDAC

Figure 5. (a) CACon output sample. (b) CASIA output sample.

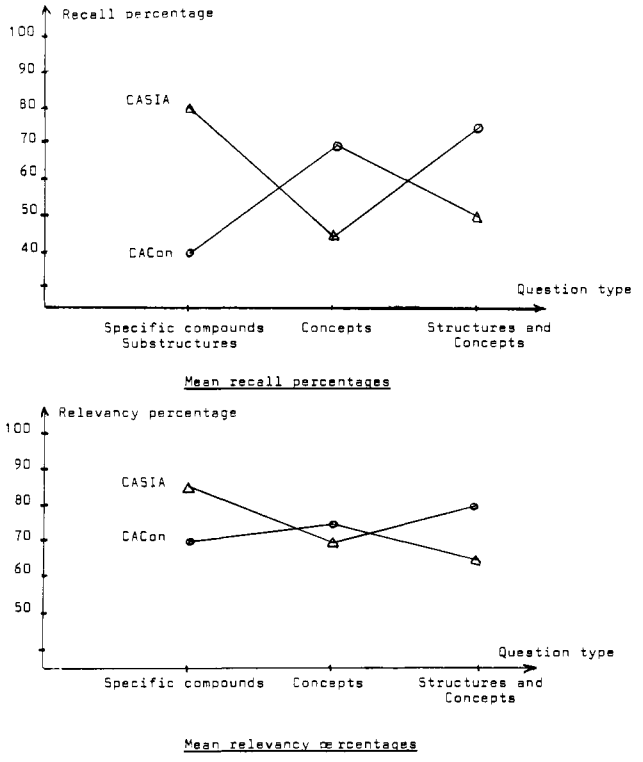

Figure 7. (a) Mean recall percentages, (b) Mean relevancy percentages.

Question Numbers	CACon			CASIA		
	Retrieved hits	Relevancy (%)	Recall (%)	Retrieved hits	Relevancy (%)	Recall (%)
101	29	84	70	52	15	80
102	8	25	29	7	86	86
103	13	23	30	9	100	90
104	124	18	71	21	95	64
105	125	100	58	192	100	88
111	146	96	61	205	95	84
112	17	94	40	39	100	97
113	6	17	50	1	100	50
114	11	100	55	16	100	80
115	0	–	–	18	100	100
121	157	96	40	337	97	86
122	48	77	48	89	73	84
123	4	100	100	5	80	100
124	31	97	83	28	89	69
125	194	100	58	319	100	96
101-125	913	81	52	1338	93	88
201	4	75	100	4	75	100
202	78	97	96	45	98	56
203	9	89	80	9	100	90
204	185	82	80	132	68	47
205	29	55	53	30	83	83
201-205	309	83	82	220	78	54
301	80	84	85	72	100	91
302	15	80	86	9	89	43
303	210	78	83	215	78	86
304	34	53	78	29	66	83
305	40	70	46	66	88	95
306	18	89	69	29	79	100
311	170	82	48	292	97	96
312	56	95	67	73	97	90
313	18	33	35	17	94	94
314	30	73	42	47	100	89
315	6	100	60	7	100	70
321	15	93	93	5	100	33
322	83	79	96	4	75	4
323	29	17	45	8	87	64
324	11	100	17	70	81	90
325	19	89	85	19	89	85
326	41	63	57	85	51	94
327	96	54	49	98	96	89
301-327	971	74	61	1145	87	84

Figure 6. Results: relevancy and recall.

Hits missed on CASIA were due to several factors:

• Because the publication delay time was longer for CASIA than for CACon, only 91.5% out of all the 2193 CACon hits were covered by CASIA during the same period.

• The use of trivial names in CACon permitted retrieval of compounds very similar to the specific compounds searched. For instance in the question concerning specific fluorinated compounds, using the term DEXAMETHASONE on CACon permitted retrieval of citations about dexamethasone and dexamethasone phosphate. The use of Registry Numbers on CASIA permitted retrieval only of citations about dexamethasone itself. For the CASIA profiles, recall could still be improved by searching both the CAS nomenclature and the trivial names that may appear in the text modifications.

• We have noticed that sometimes chemical compounds are described only by their trivial names in CASIA items, for example, alloxan (2,4,5,6(1*H*,3*H*)-pyrimidinetetrone) in profile 111 concerning barbituric derivatives. In this case, the recall was good because both the CAS nomenclature and the trivial name were searched. (CAS abstractors mentioned, after the experiment, a special entry for alloxan, indicated in the Index Guide: "2,4,5,6(1*H*,3*H*)-Pyrimidinetetrone diabetes from—see Diabetes mellitus").

• Compound classes are better indexed in CACon than in CASIA. For example, the terms FLUOROCARBONS (profile 121) or CYCLOALKENES (profile 321) are more frequently used in CACon than in CASIA.

• The most important factor for hits being missed on CASIA is the weakness of the general subject indexing, which results in the rather small recall for questions 201–205. We noticed this particularly for question 202 about breeders and fast reactors, and for question 322 about the metabolism of nonphosphorylated insecticides. In the latter case, the compounds are described by their trivial names in CACon and by their CAS preferred index names in CASIA, but the insecticide activity is indexed much more frequently in CACon than in CASIA. Perhaps, recall would have been better for the CASIA profile if we had used the term *ICID* instead of INSECTICIDE*. However, it seems that in CASIA emphasis is laid on specific compounds more than on activities.

(2) Delay Times for Citations in CASIA. All CA numbers present in each of the six CASIA issues processed were listed. These listings were useful for understanding missed hits as well as for studying the publication delay in CASIA.

On the average, 88% of the citations in one CACon issue appear in the corresponding CASIA issue. In a few cases (2

Table I

CACon issue	% of citations present in CASIA	CASIA issue
85-18	73	76-09
	16	76-10
85-19	86	76-10
85-20	87	76-10

Table II

File	No. of records (citations)	Record length (characters)	Translating cost ($)	Searching cost per profile per issue ($)
CACon odd	6 300	540	94	2.2
CACon even	8 700	540	112	2.4
CACon (odd + even)	15 000	540	206	4.6
CASIA	18 200	960	320	5.6

out of 12 CACon tapes), this ratio of citations is distributed over the CASIA issues noted in Table I.

Totalling all the tapes processed, there were 90 160 citations in the 12 CACon issues (85, 15–26) and 109 865 citations in the 6 CASIA issues (76, 08–13), leading to the average of 15 000 citations for 2 CACon issues (odd plus even) and 18 200 citations for one CASIA issue. The annual number of CA citations appearing in CASIA is still greater than the actual number of documents referred to in two CA volumes (i.e., 426 440 citations in CASIA Volume 75 vs. 392 234 in CA Volume 82 plus Volume 83). The discrepancy, due to the delay in index processing, will be reduced to zero in 1978.

On a typical CASIA tape, 73% of the items were cited in the same time period as they were in CACon, while 27% were delayed citations from previous issues. For a typical CACon issue, 88% of the items appeared in the same time period for CASIA, 7% were 1 to 2 months late, 2% were 2 to 3 months late, and 3% appeared in CASIA more than 3 months after they were cited in CACon.

The updating of CASIA file content does not seem to follow a precise pattern. For five CASIA issues, the oldest citations were dated January 76. In the CASIA tape 76-11 (May 76), we found citations issued in Volume 83 (from July 75). A thorough analysis of the most delayed citations does not explain why they were indexed much later. Out of 24 citations 17 referred to journal literature items, 5 to patents, 1 to a book, and 1 to a dissertation; 13 of the documents were in English, 5 in German or French, 4 in a slavonic language, and 2 in Japanese; two of these items were published in *Nature*, two in the *Journal of the American Chemical Society*. The difficulty involved in handling the primary document (type, language) does not seem to increase the delay for citation in CASIA. The major reason for delaying CASIA citations lies more probably in the difficulty of naming and indexing the chemical compounds; mishandling may also occur during the process of editing the references and the index entries.

(3) Comparison of Processing Costs for CACon and CASIA. Processing a CACon or CASIA issue with the PRETEXT batch software involves two stages:

- Translating SDF to PRETEXT format; CPU time depends on the number of records and on the size of these records.
- Searching the profiles; CPU time depends on the number of profiles and on the size of the tape.

Costs were calculated using the FRANLAB charge basis for Control Data 7600 during the second 1976 quarter. To evaluate the searching cost per profile on CACon, the 47 CACon profiles (25 odd, 22 even) corresponding to the 39 CASIA profiles were processed independently of the batch of AFDAC SDI profiles for issues 84-21, 22, 25, 26.

Table II gives the mean cost for translating CACon and CASIA files into the search format, and the average searching cost per profile per issue.

The searching costs are higher for CASIA than for CACon, especially for profiles run on either CACon even or CACon odd. The discrepancy is due to the greater number of records on CASIA tapes, their greater size, and complexity.

```
 1  Pharmacodynamics
 2  Hormone Pharmacology
 3  Biochemical Interactions
 4  Toxicology
 5  Agrochemicals
17  Foods
22  Physical Organic Chemistry
23  Aliphatic Compounds
24  Alicyclic Compounds
25  Noncondensed Aromatic Compounds
26  Condensed Aromatic Compounds
27  Heterocyclic Compounds (one Hetero Atom)
28  Heterocyclic Compounds (more than one Hetero Atom)
29  Organometallic and Organometalloidal Compounds
30  Terpenoids
31  Alkaloids
32  Steroids
33  Carbohydrates
34  Synthesis of Amino Acids, Peptides, and Proteins
59  Air Pollution and Industrial Hygiene
60  Sewage and Wastes
61  Water
62  Essential Oils and Cosmetics
63  Pharmaceuticals
64  Pharmaceutical Analysis
78  Inorganic Chemicals and Reactions
79  Inorganic Analytical Chemistry
80  Organic Analytical Chemistry
```

Figure 8. List of searchable CA sections on CAISA file, in 1977, at AFDAC.

(4) Conclusions. The results of our comparative study of CACon and CASIA files show that: delay times for citation in CASIA are the same as those in CACon for almost 90% of the documents; searching CASIA gives very good results for specific compounds or substructures, with nomenclature search terms, even though a general subject is also required; searching CACon gives better results for profiles dealing with general subjects or classes of compounds; processing costs are higher for CASIA than for CACon, because of the greater size and complexity of the records; on-line retrospective searching of CACon is essential to improve manual searching in volume indexes, or on-line searching of the CASIA file, for questions concerning general subjects or classes of compounds.

NEW SDI SERVICE FOR 1977

A regular SDI Service on CASIA alone cannot be considered: first, because bibliographic data, not included in CASIA, will not be printed on outputs; secondly, because general subject searches seem to give less satisfactory results than when searching CACon; thirdly, the processing cost is much higher for one CASIA issue than for the two corresponding CACon issues.

However, we think it is essential to use the specific contribution of the CASIA file to chemical information: nomenclature and registry numbers. Thus we agree with the conclusions of the study on the CAS Integrated Subject File (ISF) reported by the University of Georgia.[13]

Therefore, AFDAC will develop in 1977 a new SDI service on a part of the CASIA file, while going on with our CACon SDI service. For economic reasons, we will limit the content of the CASIA file by choosing the searchable sections, the list of which is given in Figure 8, and the searchable data elements (we will keep only the data concerning the compounds: heading parents, substituents, text modifications, registry numbers, molecular formulas).

The text-reading software PRETEXT II will be used, allowing left and right truncation, Boolean operators, and Context operators, particularly useful for searching the chemical nomenclature.

The mean annual price will be approximately $200.00 per profile, including coding cost and processing cost.

ACKNOWLEDGMENT

We are pleased to acknowledge the funding of this study by the Union des Industries Chimiques, and to express our gratitude to Mr. Delaunay of FRANLAB for adapting the PRETEXT II software to CASIA processing, to the AFDAC staff for help in the comparison, and to several persons of CAS for discussion of our work and help with translating this report from French into English.

LITERATURE CITED

(1) J. E. Dubois, "French National Policy for Chemical Information and the DARC System as a Potential Tool of this Policy", *J. Chem. Doc.*, **13**, 8–13 (1973).
(2) B. G. Prewitt, "Searching the Chemical Abstracts Condensates Data Base via Two On-Line Systems", *J. Chem. Doc.*, **15**, 177–183 (1975).
(3) M. Milne, D. Lefkovitz, H. Hill, and R. Powers, "Search of CA Registry (1.25 Million Compounds) with the TSS", *J. Chem. Doc.*, **12**, 183–189 (1972).
(4) G. W. Adamson, V. A. Clinch, S. E. Creasey, and M. F. Lynch, "Distributions of Fragment Representations in a Chemical Substructure Search Screening System", *J. Chem. Doc.*, **14**, 72–74 (1974).
(5) R. Fugmann, "The IDC System", J. E. Ash and E. Hyde, Ed., Chemical Information Systems, Ellis Horwood Ltd., London, 1975.
(6) H. R. Schenk and F. Wegmüller, "Substructure Search by Means of the Chemical Abstracts Service Chemical Registry II System", *J. Chem. Inf. Comput. Sci.*, **16**, 153–161 (1976).
(7) J. Powell, J. F. B. Rowland, and M. A. Veal, "Correlated Structure-Text Searching for Chemical Information", United Kingdom Chemical Information Service Report, Nottingham, July 1976.
(8) W. Fisanick, L. D. Mitchell, J. A. Scott, and G. G. Vander Stouw, "Substructure Searching of Computer-Readable Chemical Abstracts Service Ninth Collective Index Chemical Nomenclature Files", *J. Chem. Inf. Comput. Sci.*, **15**, 73–84 (1975).
(9) B. M. Vasta, D. F. Walker, and R. Schultheisz, "Conversion of TOXLINE Chemical Dictionary File into CHEMLINE", presented at the 168th National Meeting of the American Chemical Society, Atlantic City, N.J., Sept 11, 1974.
(10) S. A. Tancredi, R. H. Amacher, J. H. Schneider, and B. M. Vasta, "CANCERLINE: a new NLM/NCI Data Base", *J. Chem. Inf. Comput. Sci.*, **16**, 128–130 (1976).
(11) R. S. Heller, G. W. A. Milne, R. J. Feldmann, and S. R. Heller, "An International Mass Spectral Search System (MSSS). V. A Status Report", *J. Chem. Inf. Comput. Sci.*, **16**, 176–178 (1976).
(12) "Chemical Abstracts Service Specifications Manual for Computer-Readable Files in Standard Distribution Format", Chemical Abstracts Service, Columbus, Ohio, Feb 1976.
(13) W. C. Zipperer, M. K. Park, and J. L. Carmon, "The CA Integrated Subject File. II. Evaluation of Alternative Data Base Organizations", *J. Chem. Doc.*, **14**, 15–23 (1974).

Symposium on Information Handling and Processing by the Food and Drug Administration

Introductory Remarks†

SANDER W. BELLMAN

Food and Drug Administration, Rockville, Maryland 20857

Received January 14, 1977

This symposium on information handling and processing by the Food and Drug Administration was organized to help acquaint the scientific community with how FDA receives and utilizes scientific information in its enforcement of the Federal Food, Drug, and Cosmetic Act.

In addition to being the Bicentenial celebration of our Nation's independence and the 100th anniversary of the founding of the American Chemical Society, 1976 marks the 70th anniversary of the first Federal Food and Drug Act. On January 1, 1907, the Bureau of Chemistry of the Department of Agriculture, headed by Dr. Harvey Washington Wiley, began the enforcement of the 1906 law. Dr. Wiley was an early proponent of consumer protection and was the driving force behind the adoption of the Act of 1906. The FDA of today is a continuation of that first group organized by Dr. Wiley, although today it is organizationally located in the Public Health Service in the Department of HEW.

By 1938, it had become apparent that the Act of 1906 needed strengthening. Thus, the Federal Food, Drug, and Cosmetic Act of 1938 was adopted and, with various amendments added through the years, remains today the principal law governing the interstate movement of food, drugs, and cosmetics in the United States.

The Food and Drug Administration, charged with the enforcement of the Act, has therefore involved itself with the evaluation of scientific information in order to make judgments as to the safety and efficacy of the products it regulates. For example, FDA is involved with evaluation of chemical and toxicological data submitted with New Drug Applications (NDAs), New Animal Drug Applications (NADAs), and Food Additive Petitions. It must evaluate data in order to establish standards of purity and identity as well as safety and efficacy.

After these standards are established, they must be enforced through the collection of scientific evidence by use of investigative techniques and analyses by chemists, microbiologists, physicists, entomologists, and engineers.

As scientists we are aware of the fact that there are no absolute truths in science. Facts are collected and evaluated toward the goal of reaching as sound and as rational a decision as possible. Many of the FDA decisions are controversial and are criticized by proponents of one cause or another. FDA must, therefore, take pains to ensure that its decisions are based upon as sound a scientific basis as is possible with the evidence available.

In order, therefore, to present to you an overall view of the handling of scientific evidence in FDA, I have organized this symposium, loosely, into three parts: (1) setting of standards and tolerances; (2) development of sound investigative methodology; and (3) use of these standards, tolerances, and methods in the day-to-day enforcement of the act.

The first section, on the establishment of tolerances and standards, is covered by Dr. Banes, who, although not presently an FDA employee, is nonetheless intimately responsible for the setting of standards of identity, purity, and strength for the majority of drugs used today. These standards are developed partially with FDA data and are used by FDA in

† Presented before the Division of Chemical Information, 172nd National Meeting of the American Chemical Society, San Francisco, Calif., Aug 31, 1976.