is as an access route to the ACS journals covered. A number of respondents suggested that coverage of some non-ACS journals would improve the service. A significantly higher percentage of subscribers than drops (33% *vs.* 14%) judged the subject area coverage of SAA to be adequate. With regard to display of the article titles, about 50% of both the subscribers and drops indicated a preference for display of titles in groups according to general subject areas, rather than in the original table of contents format. This approach will be considered, along with others, in any future modification of the service. About 20% of both subscribers and drops would prefer standardization of the amount of information provided in the various tables of contents, as well as in the formats of the tables.

SAA's frequency of publication (semimonthly) did not appear to be a factor influencing renewal, nor did the way in which the service was perceived (about 70% of both groups considered SAA as an information system). The last question dealt with pricing of the SAA service if suggested changes were made. This approach was used to see if there was a difference between subscribers and drops in their attitudes toward pricing of information services (the first-year subscription price was 20% less than the price of a renewal subscription). The results indicated that the price of the service did not influence renewal. This was borne out by the absence of any reference to cost in the "comments" section of the questionnaire.

As stated above, the 1972 renewal rates were disappointing: about 42% for member subscriptions and 32% for nonmember subscriptions. The 1973 renewal rates (as of August 31) are much better (about 59% for members and 73% for nonmembers), but still not satisfactory. However, enough new subscriptions have been received in 1973 so that total 1973 subscriptions exceed 1972 subscriptions.

Based upon the results of this study, a number of factors can be identified as having an influence on subscription renewals. First, the subject range of articles published in the ACS journals did not adequately cover the subject interests of a number of drops. Secondly, drops were less satisfied with the information content, arrangements, and format of the tables of contents. And, thirdly, the availability of other information sources influenced renewal.

While a number of factors influence subscription renewals, they also influence first-time subscribers to any publication. Design of a successful information system requires information about individuals' career goals, perceived needs in reaching these goals, and their reactions to the system. Studies such as the one described can help organizations in understanding the motivations of users and how information services can aid them.

One of the prime requirements of any system is its ability to adapt to a changing environment. Ways and means of disseminating information that were successful in the past are now questionable in light of what is known today. Adequate feedback via periodic evaluation is necessary so that modifications can be made in time to provide useful and efficient systems.

Since the information business is like any other business, it must have its own source of decision-making information if it is to serve its customers well.

LITERATURE CITED

(1) Kuney, J. H., and Weisgerber, W. H., "System Requirements for Primary Information Systems. Utilization of the *Journal of Organic Chemistry*," *J. Chem. Doc.* **10**, 150 (1970).

(2) Kuney, J. H., and Dougherty, V. E., "An Experiment in Selective Dissemination—The *ACS Single Article Service*," *J. Chem. Doc.* **11**, 9 (1971).

Computer Search Center Statistics on Users and Data Bases*

PETER B. SCHIPMA
IIT Research Institute, Computer Search Center
Chicago, Ill.

Received October 9, 1973

Statistics gathered over five years of operation by the IIT Research Institute's Computer Search Center are summarized for profile terms and lists, use of truncation modes, use of logic operators, some characteristics of *CA Condensates*, etc.

The Computer Search Center, at IIT Research Institute in Chicago, has been operating for some five years. Basically, the Center is a collection of people, hardware, and computer programs that provides SDI and retrospective search services to paying customers. That last phrase, "paying customers" is very important. Although the Center was established with funds provided by the National Science Foundation, it now operates on a cost-recovery basis from subscription income. The figures that I present, therefore, are not the result of a controlled experiment, but reflect the real world. You will note that many of them contain a price- or cost-dependent bias, but in this period, the dollar, or its lack, is a definite fact of life. This paper presents summary data collected on the users of the Computer Search Center and the data bases that are searched therein. Further information can be obtained from NTIS.[1]

THE USERS

The users of the Computer Search Center are scientists (Figure 1). These scientists are employed at various organizations, primarily industrial— and that probably be-

* Presented before the Division of Chemical Literature, 166th Meeting, ACS, Chicago, Ill., Aug. 28, 1973.

Industry	75%
Universities	15%
Government	10%

Figure 1. CSC user affiliations

Industry	5
Universities	2
Government	10

Figure 2. CSC user per profile

cause our major effort in marketing has been aimed at industrial organizations. It is interesting to note paths followed by these chemists in obtaining information services. We can very seldom sell directly to the chemist—the initial contact and final sale is usually with a director of research or similar individual. Similarly, our products usually reach the chemist through an intermediary—the company librarian or information specialist. Yet we maintain contact with the chemist himself when it comes to the technical aspect of his profile or question. I think that this reflects industry's well-planned division of labor and responsibility. Being pseudo-academics at IITRI, it took us a while to get used to this, but it works out quite well in practice.

We measure our service in units called profiles, upon which our services are based. A profile is a rather amorphous entity, since we cannot say how many questions it comprises or how many ideas it expresses. Generally a profile represents a concept. Figure 2 shows the difference in kinds of organizations as to how many people are served by a profile. In the academic world, the two users are almost invariably a professor and his research assistant. In the industrial community, a profile represents the interest of a project team. In government, the teams are apparently larger.

Two of the many things recorded for each profile are the terms per profile and the number of hits it generates each run. Figure 3 gives the data extracted for a few issues from Volume 76 of *CA Condensates.* Note that the average number of terms per profile hovers near 25 and that the hits generated per profile are fewer than 50. Part of the reason for these numbers is cost-dependent, as we shall see later. Also note the variation from odd-numbered to even-numbered issues in the hits per profile column. CA is issued in two parts, odd-numbered issues covering Biochemistry and Organic Chemistry Sections, and even-numbered issues covering Macromolecular Chemistry, Applied Chemistry, Chemical Engineering, Physical and Analytical Chemistry Sections. The number of terms needed to prepare a profile does not vary much from one group to another, but the number of hits returned certainly does. Even-numbered issues, however, generally contain more citations than do odd-numbered issues. Therefore, we have also shown the hit per profile data normalized by data base size. The difference decreases, but is still very present. People searching even-numbered issues ask more general questions and retrieve more output.

CA Vol. 76 Issue	Terms/Profile	Hits/Profile	Normalized Hits/Profile
14	23.7	46.3	40.1
15	24.0	26.9	27.2
16	23.3	47.9	41.1
17	23.4	32.3	29.8
18	23.1	40.9	37.8
19	24.0	31.0	28.1
20	24.2	49.7	45.4
21	24.6	34.3	28.4
22	25.3	47.7	42.5
23	25.4	34.0	30.4
24	25.9	47.2	45.7
25	26.8	21.8	34.2
26	26.9	51.2	51.9

Figure 3. Profile term and hit data

Up to 25 terms and 50 hits*	$290.00
Each additional 1-10 terms	120.00
Each additional 50 hits*	120.00

*Per issue searched—2600 per year

Figure 4. CSC charge basis

Chemical Abstracts Condensates Volume 76

Truncation mode	Coden, %	CA section, %	Term type text, %	Author, %	Corporate author, %
None	65.0	16.2	26.3	32.0	84.6
Left	0	16.2	2.6	0	0
Right	0.6	11.6	54.8	68.0	15.4
Both	34.4	56.0	16.3	0	0

Figure 5. Percentage of terms of various term types *vs.* truncation mode used

As to cost-dependency of these figures, our subscription prices (Figure 4) are based on input and output units. A basic profile has from 1–25 terms and can generate up to 50 output citations per weekly issue searched. A basic fee of $290.00 is charged for such a profile. Additional input units of 1–10 profile terms or 1–50 hits are charged at a rate of $120.00. Most of our users carefully prepare their profiles to have 25 terms—frequently by combining two or three questions. Although we do have provision for hit cutoff, very few people make use of it. Thus, the fact that number of hits is less than 50 is probably not artificial.

There are two other features of profiles, truncation used and logic expression construction. We allow full truncation, on either or both sides of a term. Figure 5 indicates the use of truncation made by our users, according to various data elements searched. CODEN are either searched in full or truncated on both sides. The latter is used in searching for the characters "XX" in the third and fourth positions of CODEN that denote patents. Authors, personal and corporate, are not truncated on the left, since it is fairly meaningless to do so, but frequently the user knows a last name but is unsure of first name or initial. Since the author names in CA are inverted (last name first) such users pick them up by truncating on the right. Truncation for textual terms is heavily weighted on right truncation to handle plurals, "ing" endings, etc., as you would expect, but a surprisingly high amount of use is made of "left" and "both" side truncation. Although left truncation is fairly expensive to implement in a search algorithm, it appears to be a worthwhile investment.

Our profiles are term sets tied together via a logic expression. The Boolean operators "and," "or," and "not" are used, and can be combined in any way, precedence denoted by use of parentheses. Any degree of nesting of parenthesized expressions is permitted, so that questions of high complexity can be phrased. The "or" operator is more used (Figure 6) with "and" a close second. "Not" is used infrequently, and when it is used it is nearly always applied to only one term or term subset. The distributions of the "or" and "and" operators are much more even.

Even with the capability of phrasing complex questions via parenthesized logic (Figure 7), the highest percentage of users require no parentheses at all (nearly a quarter of the total profiles). However, about 5% of our users require 10 or more sets of parentheses, so for these people, this

Chemical Abstracts Condensates Volume 76

No. times logic operator used in a profile	*AND* Percent of profiles	*OR* Percent of profiles	*NOT* Percent of profiles
0	13.1	22.4	67.5
1	25.0	22.0	27.6
2	21.3	17.2	3.4
3	14.6	5.6	1.1
4	9.7	8.6	0.4
5	5.2	7.8	0
6	5.2	5.2	0
7	2.6	3.4	0
8	1.1	1.1	0
9	0.7	0.4	0
10 (or more)	1.5	6.3	0
Percent use of logic operators using all operators in the run	43.7	49.7	6.6

Figure 6. Percent of profiles using and, or, and not logic *vs.* number of times each operator was used in a profile

Chemical Abstracts Condensates Volume 76

No. sets of parentheses	No. profiles	% Profiles
0	65	24.3
1	39	14.6
2	50	18.7
3	29	10.8
4	24	9.0
5	18	6.7
6	10	3.7
7	7	2.6
8	3	1.1
9	9	3.3
10 (or more)	14	5.2
Total	268	100.0

Figure 7. Use of parenthetic logic in profiles

Chemical Abstracts Condensates Volume 76

Highest degree of nesting of parentheses	No. profiles	% Profiles
0	65	24.3
1	89	33.2
2	70	26.1
3	29	10.8
4	13	4.9
5	2	0.7
Total	268	100.0

Figure 8. Use of nested logic in profiles

Volume	Precision
71	37.6
72	30.5
73	27.5
74	27.5
75	30.5

Figure 9. Precision *vs.* volume (*CA Condensates*)

Volume	Percent Return
71	92.0
72	84.1
73	81.3
74	71.5
75	60.7
76	49.2

Figure 10. Return of evaluation forms *vs.* volume (*CA Condensates*)

	Chemical-Allied Products			
Characteristic	Essential	Beneficial	Not important	No opinion
Regularity	32	42	5	21
Timeliness	37	21	21	21
Consistency	32	37	5	26
Thoroughness	64	10	5	21
Labor saving	69	5	5	21
Coverage	43	26	5	26
Cost reduction–labor	26	32	10	32
Cost reduction–publications	16	21	26	37

Figure 11. Evaluation of SD system characteristics by percent of respondents

feature is very useful. The distribution (Figure 8) of nested parenthesis use follows the same pattern—it is necessary at the 1–2 nesting levels for over 50% of the users, and a very small percentage have 5 levels of nested logic, a very complex profile indeed.

Before we switch over to discussion of the data base, there are a few additional bits of interesting information on users. Two are measured data and two come from a market survey we conducted. Precision (Figure 9) has remained fairly near 30% for several volumes. Although we have an occasional user who demands 100% precision, most users are sophisticated and use SDI services to stay abreast. They don't want to miss anything, and so write general questions, knowing that two-thirds of their retrieval will be discarded. In terms of effort taken to return evaluation forms (Figure 10) users are less inclined to do so as (1) they become long term users, and (2) they switch from free-trial to a paying basis. We started the service by providing free subscriptions in return for evaluations of the service. By Volume 74, nearly all profiles were on a subscription basis. Even though the purpose of the evaluation forms is to improve profiles, many users do not take the time to return them. When they got free service they regarded it as a duty. When they became paying customers, they did not. Although from our point of view, all users should return these forms to keep their profiles finely tuned, some of the lack is due certainly to longtime use of, and satisfaction with, given profiles.

From our market survey (Figure 11) we found thoroughness and labor-saving aspects of SDI service to be the most appreciated. These questions (Figure 12) were asked of people who had been receiving free service. In addition to pointing out areas in which we could improve our service, most respondents also subscribed, the best measure of the system!

Most of the things that we can say about the *CA Condensates* data base are quantitative. We do not, of course, modify the data base in other than operational ways. This is not to say that we have no appreciation for the content of Condensates; in fact, our profile coordinators probably have some of the most comprehensive views of CA in existence, since they work out profiles with users who have a very wide variety of interests. But most of our data base statistics are concerned with the efficient operation of the Computer Search Center. We are concerned with lengths

Totals exceed 100% because of multiple responses

	Question		Ind.
1	CA available	Yes	42
		No	—
2	Prior manual search	Yes	23
		No	18
3	Monitor searches	Yes	23
		No	18
4	Dispense with manual searches	Yes	22
		No	18
5a	Card format satisfactory	Yes	37
		No	3
5b	Index terms	Useful	36
		Not	6
5c	Terms causing hits	Useful	33
		Not	9
6	Maintain card file	Yes	32
		No	9
7	Card file useful	Yes	22
		No	10
8	Look up citations	Yes	40
		No	1
9	Hard-copy retrieval	Personal	26
		Librarian	22
12	Modifications could improve profile	Yes	28
		No	13
13	Distribution of cards prompt	Yes	41
		No	—
14	Profile liaison	Sat.	34
		Unsat.	4
15	Subscription desirable	Yes	32
		No	5

Figure 12. Summary of user evaluation of *CA Condensates* current awareness service

DataType		Average Length in	
Number	Name	CA 76:23,25	CA 76:24,26
1	Coden	38	30
2	Title	80	72
3	Author	37	36
4	Journal title	20	20
5	Keyword	47	50
8	Corporate author	43	36
13	Availability, etc.	24	27
14	Cross reference	8	8

Figure 13. Average length of data entries by data type in *CA Condensates*

	Odd-Numbered Issues (CA 76:23,25; 77:01)	Even-Numbered Issues (CA 76:24,26)
Citations	5600	7500
Length		
Mean	280	270
Standard deviation	60	58
Maximum	600	600
Data fields/ citation	7.1	7.1
Keywords/citation		
Mean	4.9	5.8
Standard deviation	2.3	3
Maximum	25	30

Figure 14. Statistics on length, data fields per citation, and keywords per citation in *CA Condensates*

Data Type		Percent Occurrence in	
		CA 76:23,25	CA 76:24,26
Number	Name		
1	Coden	100	100
2	Title	100	100
3	Author	99	98
4	Journal title	98	97
5	Keyword	100	100
8	Corporate author	98	96
13	Availability, etc.	90	82
14	Cross reference	9	16

Figure 15. Percent occurrence of data types in *CA Condensates*

No. Issues (in CA Vol.)	Citations	Tokens	Types	Type/Token Ratio
2 (Vol. 72)	9,067	91,760	16,753	1:5.48
6 (Vol. 72)	31,402	479,856	60,876	1:7.88
13 (Vol. 72)	67,456	877,734	92,216	1:9.52
13 (Vol. 73)	66,796	963,698	100,498	1:9.59
26 (Vol. 72 & 73)	134,252	1,841,432	153,268	1:12.01

Figure 16. CA type-token relationships

Term	Frequency
Forming	374
Intermediate	374
Oxo	374
Pregnancy	373
Chick	372
Bed	371
Critical	371
Hormones	370
Industry	370
Solids	370

Figure 17. Term sets with similar frequencies

of fields, since this affects search time (Figure 13). This figure shows some average lengths of various data elements. Over time, these lengths have remained fairly consistent, with the exception of keywords, which has slowly increased. CA is adding more indexing to Condensates, and that can only improve its utility. As we process each issue (Figure 14) of Condensates, we generate statistical tables such as this, to try to spot any trends in lengths of various fields and whole citations. We also (Figure 15) check for percentage occurrence of various data elements. While most remain fairly consistent, when a new data element is added, it appears infrequently at first and then increases in use as CA procedures become fixed.

We have also made several detailed analyses of word occurrences in CA, for use in inverting the file for retrospective search. In this table (Figure 16), the types (unique words) and tokens (total appearances of all words) are shown for various segments of the CA file. While it is true that as you view larger segments of the file, the same words are used more often (the type:token ratio increases), we have also found that chemists invent many thousands of words annually. Most of them, of course, are chemical names. The incidence of low-frequency new words in CA is very high, and any inverted file design has to allow for many additions each year.

Finally, just to prove that one can find respite from all these statistics, note the peculiar set of terms with similar frequencies (Figure 17) that appeared on the computer printout one day when we were running a frequency-ordered term frequency distribution list on *CA Condensates.*

LITERATURE CITED

(1) Williams, M. E. Schipma, P. B. Preece, S. E., Becker, D. S., Llewellen, P. A., and Stewart, A. K., "Educational and Commercial Utilization of a Chemical Information Center," IIT Research Institute, Computer Search Center, Chicago, Ill., June 25, 1968–June 25, 1972; NSF Contract No. **NSF-0554,** July 30, 1972; available NTIS or ERIC Clearinghouse, **ED 068 132,** hardcover price $16.45.

Profiling, the Key to Successful Information Retrieval*

C. H. O'DONOHUE
Box 26583, Philip Morris Research Center, Richmond, Va. 23261

Received October 9, 1973

A major tool employed to enter an information source is the search profile. The development of an adequate profile depends upon the aids supplied by the data bases. These aids vary in their content and depth and their proper use is essential for relevant information retrieval. The data bases examined are *CA Condensates, Index Medicus,* and BA data bases. Several searches are presented with a study of their comparative profiles.

The useful yield of any literature search will depend upon the requestor's ability to communicate effectively with the information system (data base). The usual communication tool used by the requestor is known as the profile or search question.

DEVELOPING THE PROFILE

Needs. The requestor has certain particular needs which must be met. The search strategy (Figure 1) begins when the requestor contacts a profiler or becomes a profiler himself. As part of this search strategy, an understanding of the structure of the indexing language used in the data base is required. The requestor/profiler has no control over the input into the system nor does he have any control over the method of input. He must operate within the constraints placed on him by the data base.

He needs to know what keywords (descriptors, indexing terms, codens) are available for the input phase and how these terms are defined and related. The profile structuring process is very similar to that of document indexing. Thus, before a requestor can begin, he is dependent on the data base for the items needed to help him structure his profile. In most cases he will have to subordinate his wishes to the idiosyncrasies of each data base he encounters.

What help do the various commercially available data bases give to a potential user? If one subscribes to magnetic tape services, then a word-frequency listing can be obtained. Also data bases supply certain aids to their magnetic tape subscribers that are not available to the hard copy (published edition) subscribers. Depending upon whether one uses outside commercial services or does manual searches in-house, the profile structuring problems are quite different.

Profiling Aids. The three major data bases that will be discussed are *Biological Abstracts* (BA), *Chemical Abstracts* (CA), and *Index Medicus.* The indexing/profiling aids available from these services are varied in depth of content.

Biological Abstracts provides a "Guide to the Indexes for Biological Abstracts and Bio-Research Index" and "A Guide to the Vocabulary of Biological Literature." The Guide to the Indexes only illustrates the "how-to-use" approach to the four indexes of *Biological Abstracts* and *Bio-Research Index.* The second, the Guide to the Vocabulary, is of major value in that it points one to the right road for choosing the proper keywords.

B.A.S.I.C., the subject index for BA, is a permuted, keyword-in-context index to published titles. These titles contain augmenting keywords added by BA. The Guide points out the pitfalls involved in the B.A.S.I.C. indexing concept, as for example, in a study involving dogs, such keywords as dog, dogs, canine, puppy, and beagle would have to be checked to obtain the relevant citations. The Guide contains word frequency listing and related term listing and was designed to aid those using either the published indexes or the magnetic tape data base.

As might be expected, *Chemical Abstracts* has many aids available for the user. Subscribers to the complete CA also receive the "Index Guide," "Index to Ring Systems," "Formula Index," and the "Registry Number Index." Also available are the "CAS Search Guide," "The Desktop Analysis Tool (DAT)" and a "CAS Chemical Substance Name Selection Manual for the Ninth Collective Index." Generally, each of these aids is needed at one time or another. Figure 2 lists two dyes chosen from DAT. Fewer than 6% of the possible names for each of these substances appeared in the "Index Guides." Thus, a compound with a common name having neither a registry

*Presented before the Division of Chemical Literature, 166th Meeting, ACS, Chicago, Ill., Aug. 28, 1973.