ACKNOWLEDGMENT

I would like to express my thanks to Mrs. Diane Farmer of Polaroid for her help and suggestions in connection with setting up this system. Thanks are also due to my manager, Mr. Charles Zerwekh, Jr., for the support and encouragement given and to the Polaroid Corporation for allowing me to publish this work.

LITERATURE CITED

(1) Meeting of the American Society for Information Science, Oct 1-4, 1969, San Francisco, Calif.

(2) Yerke, T. B., *et al.*, FAMULUS: A Personal Documentation System . . . Users' Manual, Pacific Southwest Forest and Range Experiment Station, Berkeley, Calif., 1969, PB-202 534.

(3) Rase, W. A., *Angew. Informatik*, **14**, 459-465 (1972).

(4) Falor, K., "Modern Data," March 1970, 62-72; Aug 1970, 48-59.

(5) DRS is a proprietary information retrieval software, developed by the Aeronautical Research Associates of Princeton, Inc. (ARAP).

(6) DRS Users' Manual, distributed solely by ARAP.

(7) The DRS LINK Command, distributed solely by ARAP.

(8) DRS System Generator, distributed solely by ARAP.

(9) DRS Error Messages, distributed solely by ARAP.

(10) DRS System Manual, distributed solely by ARAP.

(11) CWIK Note, Aug 1967, and the corresponding card deck.

(12) "Merck Index," 7th ed., Merck & Co., Inc., Rahway, N. J., 1960.

(13) Barnard, A. J., Kleppinger, C. T., and Wiswesser, W. J., *J. Chem. Doc.*, **6**, 41-48 (1966).

(14) IBM System Reference Library, IBM 1130 Disk Monitor System, Version 2, Programmer's and Operator's Guide, 10th ed, May 1972.

Performance of an SDI System with Interactive Features

VIERA ŠAŠKOVÁ

Institute of Inorganic Chemistry, Slovak Academy of Sciences, Bratislava, Czechoslovakia

JIŘÍ KOSÍK

Economic Research Institute of Chemical Industry, Bratislava, Czechoslovakia

Received March 7, 1974

A user-oriented interactive system was developed and tested. For two years, 280 profiles were searched in the *CA-Condensates* data base. The performance and effectiveness of the system were evaluated in relation to the data base used, to the hardware configuration and software package, to the user population, and, finally, in relation to the aid offered by the information center. Various ways and means that lead to a better satisfying of the user's needs, such as the iterative way of searching, quantification of user's needs, searching at various specifity levels, etc., are discussed.

It has been generally accepted that the aim of all information systems is to bring the right information to the right man at the right time. In other words, the primary aim of an information system is the satisfaction of the user's needs. Emphasis on the needs of the user in the information process, in contrast to the previous stressing of the document-handling side, is far more than a pure theoretical or terminological question. This shifting of attention brought about great changes in information handling; new methods were developed with the aim of increasing the adaptability and flexibility of information systems. Interactivity of information systems, man-machine dialogues, machine-aided formulation of the user's profiles, all these were introduced in order to respect the user's wishes and to facilitate his search for information. On the other hand, it was found how little we know about the user, his personality, background, and his literature habits.

Summing up the above, a modern information system should fulfill the following requirements:

1. As a response to the user's request, it should give a relevant and complete set of information.

2. In relation to the user, the system should be active—it should, *e.g.*, be able to make suggestions and point out the errors committed in the profile formulation. It is further desirable that the formulation of profiles be machine-aided and the alterations of profiles be easily performed.

We tried to comply with the above requirements and introduced at least some of the above specified features into the system developed in cooperation with the Institute of Inorganic Chemistry of the Slovak Academy of Sciences and of the Economical Research Institute of Chemical Industry in Bratislava. The system ran under the working name CACS. For some 15 months about 280 profiles were matched against the *CA-Condensates* data base, and a current awareness service was supplied for the users.

We evaluated the performance of the system and tried to find the best ways of satisfying the user's needs. Emphasis was put on the interface between the user and the system. From the analysis of the performance, we concluded that the factors influencing the effectiveness may belong to four major categories:

(1) The data base that is searched
(2) The information system, the software package, and hardware configuration
(3) The personality of the user and his needs
(4) The information center, their assistance given to the user

DATA BASE

CA-Condensates in SDF were searched. The advantages of an external ready-made reference service are obvious: results of the work of many highly qualified abstractors, indexers, and editors are at our disposal; we need not analyze the primary literature. The disadvantage of such an external reference service is that, even if it does not suit us, we have to accept the indexing mode of the outside service organization. There is no doubt that a data base prepared by

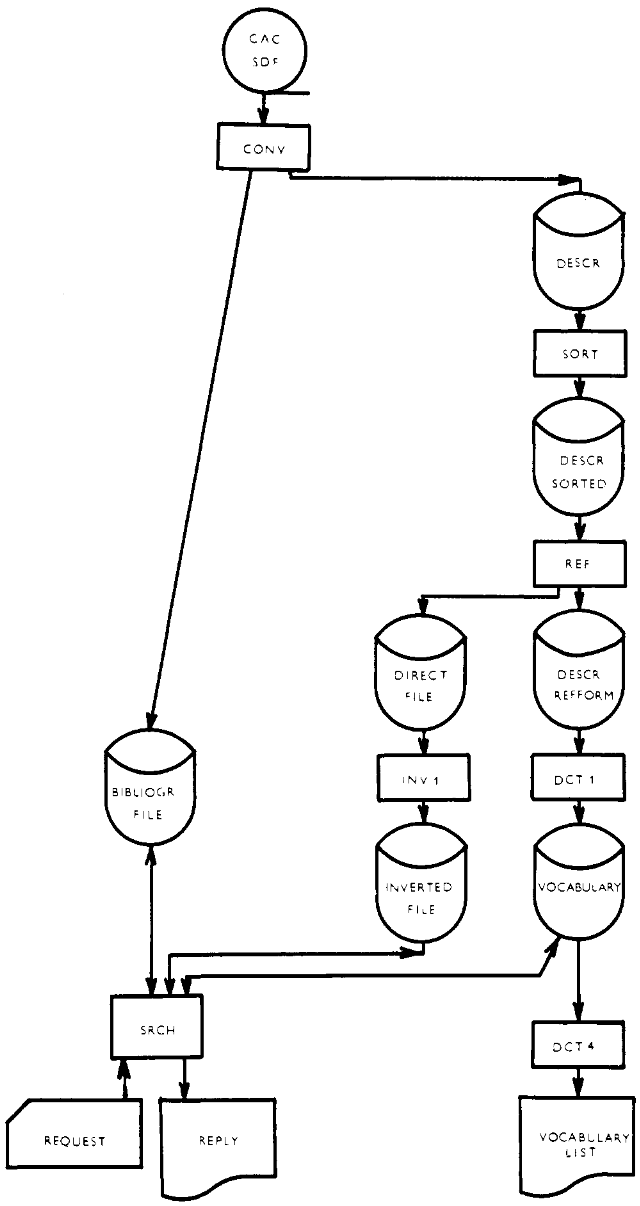

Figure 1. General flowchart of the system.

an internal information service will more closely meet the user's needs. The present volume of chemical information and the costs of scientific staff personnel for analyzing and indexing the documents make an internal information service in most cases ineffective or even impossible.

The *CA-Condensates* use for indexing an uncontrolled vocabulary. Consequently, one subject may be expressed by a variety of synonyms or near-synonyms. In the "Search Guide" published in 1967 to be used with *Chemical Titles* and *CA-Condensates,* the main term (primary synonym) was marked by an asterisk. We expected that the indexers would preferably use these main terms. This was not the case and synonyms, *e.g.,* device—apparatus, dewatering—dehydration, etc., were both used by the indexers. Needless to say, this practice may cause loss of pertinent information. Failure in retrieving a pertinent piece of information also may be caused by clerical mistakes and errors which may happen during the keyboarding operations. Abstracts indexed, *e.g.,* by "boundry" instead of boundary, by "ethylne" instead of ethylene, would not be retrieved by the correct forms of these words.

In the discussed system, each issue of the *CA-Condensates* was reorganized and an inverted file was formed. This rearrangement allowed for printing an alphabetic dictionary of keywords used for indexing abstracts in the CAC. We very soon discovered that these vocabularies might be very helpful in the course of the profile formulation. By browsing through these dictionaries of indexing terms, not only many errors and misprints were discovered, but, what is much more important, we were able to see the wide range of terms used for indexing abstracts. Thus we were able to check, *e.g.,* how often the singular and plural forms of one term were used. After we introduced the right-hand truncation into our system, we used the dictionary even more frequently than before. Looking up the dictionaries from several issues of the CAC we were able to find by which form of truncation to obtain all the words we wished to get without false drops.

THE SOFTWARE PACKAGE AND HARDWARE CONFIGURATION

We used an IBM 360/30 working under DOS. For reasons of economy, we tried to find an existing software which could be used. It had to be a user-oriented and interactive information system. Since for a variety of reasons we were not able to find a convenient system, we decided to take parts of IRMS (Information Retrieval and Management System), modify them, and, with the aid of additionally written programs, fit them into a system which would be operational with the available hardware configuration.

From the original seven programs contained in the IRMS package, the following were used: DCT1, DCT4, INV1, and SRCH. Two new programs, CONV and REF, were written. The IBM SORT program was also involved. The IRMS search program SRCH was heavily modified. By the CONV program the original magnetic tape of *CA-Condensates* is decoded and the data base is reorganized in that a bibliographic file is created and inputs for a descriptor file and for an inverted file are prepared. An additional reorganization necessary for the descriptor file is done by SORT and REF programs.

The descriptor file is managed by the DCT1 and DCT4 programs. The INV1 program is used for the preparation of the inverted search file on a disk. Request data are processed by a SRCH search program. All searching, including that from the bibliography file, is done by direct access.

Figure 1 shows the general flowchart of the presented system.

Request and printout structures are similar to those used in the IRMS with small modifications. They will be described in the following paragraphs.

MACHINE AND SYSTEM CONFIGURATION

The system is designed to operate on an IBM 360 running under DOS. Programs are written in the Assembler Language. The minimum configuration is the System 360 Model 22, 22 Kbytes of main storage.

The minimum secondary storage requirements are three 2311 disk drives and a tape drive.

THE INTERACTIVITY OF AN INFORMATION SYSTEM

There exists quite a number of interactive information systems. Though no exact definition of these systems has been given yet, they all seem to have certain features in common, as pointed out by King and Bryant.[3] They are especially the speed with which they respond to requests, the ability to give on request certain system parameters, *e.g.,* the number of documents indexed by a given set of indexing terms, the iterative way of searching, and conversational querying.

Especially when using an external reference service, it is of great importance to have an interactive information system. The user may or may not be familiar with the organization, indexing policy, and practice of the file he is going

```
A L P H A B E T I C   D I C T I O N A R Y                    PAGE      3

301 ANALGESIC                      351 ANTHRACENE
302 ANALGIN                        352 ANTHRANILIC
303 ANALOG                         353 ANTHRAPYRROLE
304 ANALYSIS                       354 ANTHRAQUINONE
305 ANALYTICAL                     355 ANTHROSOL
306 ANALYTICITY                    356 ANTIAGING
307 ANALYZER                       357 ANTIANEMIC
308 ANALYZING                      358 ANTIARRHYTHMIC
309 ANATASE                        359 ANTIBACTERIAL
310 ANATION                        360 ANTIBARYON
311 ANAYZER                        361 ANTIBIOTIC
312 ANCHOR                         362 ANTICARIES
313 ANDERSON                       363 ANTICARIOGENIC
314 ANDESITE                       364 ANTICATHODE
315 ANDROSTENE                     365 ANTICORROSION
316 ANEMIA                         366 ANTICORROSIVE
317 ANEROBIC                       367 ANTICROSSLINKING
318 ANESTHETIC                     368 ANTICRUSTING
319 ANGLE                          369 ANTIDEPRESSANT
320 ANGULAR                        370 ANTIFERROELEC
321 ANHARMONIC                     371 ANTIFERROMAGNET
322 ANHARMONICITY                  372 ANTIFERROMAGNETIC
323 ANHYDRIDE                      373 ANTIFERROMAGNETISM
324 ANHYDRITE                      374 ANTIFOGGANT
325 ANILINE                        375 ANTIFOULANT
326 ANILINO                        376 ANTIFRICTION
327 ANIMAL                         377 ANTIFUNGAL
328 ANION                          378 ANTIGEN
329 ANIONIC                        379 ANTIGLARE
330 ANISIDINE                      380 ANTIHISTAMINE
331 ANISOLE                        381 ANTIHYPERTENSIVE
332 ANISOTROPIC                    382 ANTIINFECTIVE
333 ANISOTROPY                     383 ANTIINFLAMMATORY
334 ANNEALED                       384 ANTIKAON
335 ANNEALING                      385 ANTIMICROBIAL
```

Figure 2. Alphabetic dictionary of the indexing terms.

to search. In his effort to elaborate the optimum request the user should be actively aided by the computer.[8] In most interactive systems the searches are iterative. The user makes modifications of his query and alters the directions of the search on the basis of information retrieved by the previous request. This stepwise tracing down of the desired information has proved to be very useful.

PROCESSING OF THE REQUEST IN THE CACS SYSTEM

We mentioned already the difficulties encountered in the request formulation caused by the fact that the CAC uses a free language for indexing. The usefulness of an alphabetic dictionary of indexing terms compiled by the computer has already been mentioned, while the CAC was reformatted and an inverted file formed. In systems where a thesaurus or a controlled vocabulary is used, the requestor may be guided by the computer to find the most suitable terms for his problem in that parts of the thesaurus, *e.g.*, broader, narrower, or related terms, are displayed to him. Since we had no such possibility, we had to make the most of the existence of the alphabetic dictionary of indexing terms. Part of such a dictionary is shown in Figure 2.

Each keyword used in the request was checked against the alphabetic dictionary. In case that no such word occurred in the list (*i.e.*, in the dictionary of indexing from one issue of the CAC which was being processed), the computer printed: "... descriptor wrong or unknown." Very often we found that the term used in the request was misspelled and this is an error that escapes only too easily our attention. When an otherwise correct term had been repeatedly announced as "wrong or unknown," it was advisable to examine whether the user's wishes were adequately expressed.

A request constructed according to a set of rules was then processed by a set of search programs.

A request profile consists, apart from an identification line, of up to 15 parameters in which up to 10 descriptors are joined by either of the two logical connectors, AND and OR. The last line of a request profile is a "query line" (Qu-line), in which the parameters are joined by the logical connectors AND, OR, and NOT. Parentheses may be also used in the query line. Two request profiles are shown in Figure 3.

Since the user's needs are, among others, determined by the quantity of information he wishes to get, a quantitative determination, *i.e.*, an output definition parameter of the request profile, is possible. When the user constructs a new request profile and does not know how much literature he

```
00=SEARCH NUMBER 07 SAV      DEPT A/3                                         0500007
01,OR=THERMOD*,THERMODNS,THERMODYNAMIC,THERMODYNAMICS.                        0500007
01,OR=HEAT,ENTHALPY,ENTROPY.                                                  0500007
02,OR=EQUIL,ACTIVITY,COMPRESSIBILITY,MP,FUGACITY,THERMAL.                     0500007
03,AD=MOLAR,VOL.                                                              0500007
04,AD=POTENTIAL,CHEM.                                                         0500007
05,AD=VAPOR,PRESSURE,OXYGEN.                                                  0500007
06,OR=DEHYDRATION,EVAPN,FORMATION,FUSION,HYDRATION,HYDROXYLATED.              0500007
06,OR=MELTING,SINTERING,SUBLIMATION,VAPORIZATION.                             0500007
07,OR=DECOMPN,DISSOCN,MIXING,SURFACE,INTERFACIAL,GRAIN.                       0500007
08,AD=GRAIN,GROWTH.                                                           0500007
09,OR=CARBONATE*,HYDRATE*,HYDROXIDE*,KAOLINITE*.                              0500007
09,OR=MONTMORILLONITE.                                                        0500007
10,OR=OXIDE*,SPINEL*,ALLOY*,MELT,MELTS.                                       0500007
99,50=(01OR02OR03OR04OR05)AD(06OR07OR08)AD(09OR10)                            0500007
00=SEARCH NUMBER 08  SAV      DEPT A/3                                        0500008
01,OR=DENSITY.                                                                0500008
02,OR=MEASUREMENT,DETN,APP,CALCN.                                             0500008
02,OR=COMPUTER*,INSTRUMENT.                                                   0500008
99,20=01AD02                                                                  0500008
```

Figure 3. Two request profiles.

may expect to occur in one issue of CA, then it is advisable to process the request profile without specifying the output definition parameter. The request is then processed in the usual way, and as a response to his request the user will obtain the total number of abstracts retrieved by the profile. The abstracts themselves are not printed out. According to the principles of interactivity of the information systems, at this stage the user makes his further decisions depending on the information obtained in the previous step. Thus, *e.g.*, when the user learns—as a response to a highly specific request, such as "phase equilibria of calcium, magnesium, silicon oxide systems"—that there are some 85 items in the file, he knows that there must be some error in his request. Most probably a wrong logical connector was used—in the above case "OR" instead of "AND." Or else the requester expressed his wishes incorrectly and a modification of the profile is indicated. In the next step the output parameter is specified; *i.e.*, the request is quantified. Limiting, *e.g.*, the number of documents that are to be retrieved to 25, the requestor determines that he wishes to obtain up to 25 documents. If more than 25 documents are retrieved, the computer prints out only the total number of the retrieved items and awaits the further decision of the user.

This determination of the acceptable quantity of information, quantification of the request, became, after we had learned to use correctly all the possibilities it offered, an extremely useful tool for the stepwise improvement of some difficult request profiles. For a better understanding of its working we must remember that the query line is processed level by level in an ascending order and from left to right for a given level. Owing to this, the user may obtain answers to his request at different steps of the search process. The computer prints out answers at that level at which the number of documents retrieved is lower or equal to the output specification parameter. Let us consider a practical example. If the requestor sought information on the wear resistance of refractory materials, the corresponding profile would be

01, OR = refractory, refractories
02, AD = wear, resistance
QU = 01 AD 02

In the first approach the total number of retrieved documents is printed out, *e.g.*

"Last answer N = 3"

The next step would be the specification of the output definition parameter. The requestor is either (a) satisfied with the three documents and accordingly he specifies the output definition parameter

QU, 5 = 01 AD 02

(b) or he thinks the three documents unsatisfactory and he may use a higher output definition parameter, *e.g.*

QU, 60 = 01 AD 02

Table I

	Research teams	%
Academic research institutes	34	39
Industrial research institutes	45	50
Technical university	9	11
	88	100

In the latter case, after the completion of the first step, *i.e.*, after the computer has located all documents indexed by the terms "refractory" or "refractories" and the number of these documents is lower or equal to 60, the answer obtained at the first level of the search is printed out as the so-called "temporary answer." The level of the search is always indicated in this temporary answer. The request is then processed to the end in the usual way and in the "last answer" the documents which satisfy all conditions expressed in the query line are indicated.

When performing a search in a data base with a systematic ordering or a data base for which a thesaurus has been given, the requestor may determine beforehand the hierarchy level at which he wishes the search to be performed. Searching a data base indexed by an uncontrolled vocabulary and without a systematic ordering, such as the CAC, we have to accept for the search the specifity degree given by the indexers; *e.g.*, we cannot ask for "transition metals" if the indexers chose to use the individual names of transition metals, and *vice versa*. On the other hand, as a consequence of a search question, a certain ordering of the data base takes place. At least two classes of documents are always formed: the retrieved and unretrieved documents. Most retrieval systems, especially those with coordinate indexing, will produce multilevel ranking of the documents in the data base.[2] If we search for documents of some properties (epr, nmr, etc.) of coordination compounds of copper, our search for the profile:

```
01, AD = COORDINATION, COMPOUNDS.
02, 0R = COPPER
03, 0R = EPR, NMR.
    QU = 01 AD 02 AD 03
```

will produce a segregation of the document collection into four ranked sets: a large set of documents on coordination compounds, a smaller set on coordination compounds of copper, and finally a set of documents on the epr and the nmr of the coordination compounds of copper. This segregation is based on the overlap and on the degree of this overlap between the indexing terms of the documents in the data base and the indexing terms (keywords, descriptors) and their prescribed combination in the request profile. Searching for the above profile will produce one set of documents with no overlap with the profile and three sets of documents with an increasing degree of overlap. For the given profile these three sets represent three levels of hierarchy. The widest is the first set on coordination compounds in general. This is subsequently narrowed down by performing the prescribed logical operation (the combination of two sets of descriptors with the aid of the logical connector "AND").

Thus, with the aid of the "temporary answer," the requestor has the possibility of obtaining answers to his request with a varying degree of specificity. On the basis of these, he may optimize his search strategy.

In the course of our test runs, we found the display of answers at various levels of the search to be most helpful. Though CAS supplies its users with excellent information material and data base description, a requestor is never quite sure whether the way he has expressed his needs would coincide with the system's and indexer's language and practice.

Table II. Distribution of Research Teams According to Their Special Field of Chemistry

	%
Inorganic chemistry including coordination chemistry	25
Organic chemistry	23
Biochemistry	11
Macromolecular chemistry	12
Petrochemistry	14
Chemical engineering	15
	100

TRUNCATION

Most retrieval systems use the option of searching for fragments of words. In the CACS system, we use right truncation. Left truncation was not feasible in our system, since we reformatted the original tapes of the *CA-Condensates* and created an inverted file.

We found that truncation needs careful handling. We used to check the truncated terms against the alphabetic dictionary of keywords from several issues of the CAC in order to prevent the retrieval of unwanted terms. In general our experience agrees with those reported in the literature.[6] We made several control searches with truncated and full terms. These runs showed that the results were nearly identical when profiles with both truncated and full terms were used. Naturally, it is much more convenient for the searcher to use the truncated form of term rather than to think of all possible variations. Truncation is actually indispensable only where, owing to the system of the organic nomenclature, some suffixes and prefixes designate various functional groups or the structure of substances.

PERSONALITY OF THE USER AND HIS NEEDS

It was beyond our possibilities to take part in some widereaching user research, though we were well aware that all problems connected with the role of the user in the process of scientific communication deserve the greatest attention.[1,9]

We shall attempt a description of the user population and a characterization of their needs. The user population is described with regard to their affiliation, professional age, branch of chemistry, and the degree of their professional involvement. The information needs of the users are characterized by various features, *e.g.*, by the kind of research for which information is sought (basic *vs.* applied), by the object of the request (substance, method, theory). The search requests—user's profiles—are then examined with regard to these characteristics.

The user population belonged to an academic research institution (the Slovak Academy of Sciences), and to those research institutions which, in a form of cooperation, participated in the research performed by the Slovak Academy of Sciences. On the whole, 88 research teams expressed their information needs in 280 search profiles. The average number of researchers within a research team was 4–5, so that some 450 persons were involved in the information usage and evaluation.

The distribution of these teams according to their affiliation is shown in Table I. Distribution according to their special field of chemistry is shown in Table II. Though there is no distinct borderline between basic and applied research, our users were classified into these two classes according to the prevailing character of their research. The division is as follows: basic research, 48%; applied research, 52%. Similarly, the classification of researchers according to their special branches of chemistry is only approximate (Table II), since out of the total of 280 search profiles 120

Table III. Characteristics of Users

Research group	Members within the res group	Educational level	Professional age	Professional involvement (M = medium H = high, VH = very high)
I	6	3 Ph.D. 3 graduates	25, 20, 12 8, 5, 5	H
II	6	5 Ph.D. 1 graduate	25, 20, 12 16, 15, 5	H
III	3	1 Ph.D. 2 graduates	 25, 20, 5	M
IV	4	3 Ph.D. 1 graduate	17, 15, 14 10	VH
V	5	3 Ph.D. 2 graduate	23, 10, 7 5, 3	M
VI	7	7 Ph.D. 2 graduates	20, 18, 15 10, 10, 9, 8	VH
VII	3	1 Ph.D. 2 graduates	25, 10, 5	VH
VIII	3	2 Ph.D. 1 graduate	20, 18, 5	VH
IX	7	3 Ph.D. 4 graduates	20, 18, 15 15, 12, 12, 10	H
X	1	Ph.D.	9	VH

Table IV. Characteristics of Users

Research group	Affiliation	Branch of chemistry	Type of research
I	Academic research institute	Chemistry of silicates	Basic + applied
II	Academic research institute	Physical chemistry + silicates	Basic
III	Academic research institute	Synthetic hydrosilicates	Basic + applied
IV	Academic research institute	Natural hydrosilicates	Basic + applied
V	Academic research institute	Spectroscopy, spectrochemistry	Basic + applied
VI	Academic research institute	Electrochemistry	Basic + applied
VII	Academic research institute	Inorganic chemistry	Basic + applied
VIII	Academic research institute	Crystallography, structure res	Basic
IX	Industrial research institute	Pulp and paper chemistry	Applied
X	Technical university	Coordination chemistry	Basic

may be classified as belonging to physical or theoretical chemistry as well.

Since a thorough description and classification of users requires a fairly great amount of cooperation from the users and this could not be expected from all of them, we chose ten research teams who were willing to cooperate and answer our questions.

We interviewed representatives of these ten research teams in order to find out how many people were working within a team, their educational level, their professional age, the branch of chemistry in which they were working, their type of research, and finally the degree of their professional involvement. The latter was estimated with regard to the membership in professional organizations, to active participation in scientific meetings, to the publication of papers, and to other professional activities, such as lecturing at the technical university or the editorship of a professional journal, etc. The aforementioned characteristics are listed in Tables III and IV.

As it may be seen from Table III, a research team consisted on the average of four persons. The group usually worked under the leadership of an experienced researcher (professional age between 20 and 25 years). Out of 45 researchers 23 possessed degrees corresponding to those of Ph.D. or Professor; 22 were graduates in chemistry. As to the character of the research, three teams characterized their research as basic, one as applied, and six as basic combined with applied.

We have tried to learn as much as possible about the information needs of our users ever since we started this work. The reason is obvious: it is difficult to express information needs in a search profile if one does not know them. We used for this purpose the technique of unobtrusive interviews since any kind of questionnaire is most unpopular and is resented very much. We wanted to see clearly especially into the following points.

1. Which are the specified functions of information processing—does the user wish to obtain current awareness of literature in his special field, or else does he wish to get new ideas for his work in progress or for some new work.

2. How much information does he want—just the information closely connected with his research or does he wish some background information or information on a broader field.

3. What is the object of the request—a substance, a method, or a theoretical problem.

As regards the first question, 95% users answered that they expected both, current awareness and new ideas and stimulation. In the smaller group of ten teams (our experimental group), all users wanted the system to fulfill both functions.

The answers to the second question when examined and correlated with the division of researchers into basic and applied research groups showed that scientists in basic research wished to obtain a complete set of information They are willing to accept even up to 60% nonpertinent items in the search output, providing the system guarantees that no pertinent items are left unretrieved in the data base.

Contrary to the foregoing group, the researchers in applied research wish to obtain current awareness service tailored to meet exactly their needs. The occurrence of nonpertinent items in the search output is resented much more than a possible loss of pertinent information.

The user may seek information on a substance, a method, or a theoretical problem. We divided the 22 profiles of the experimental group into these three categories and in Table V we listed the number of keywords per profile, the average number of responses per profile, and the degree of specificity of a profile, by which we mean whether the search algorithm is simple or not. By "S" (for simple) we designated search profiles in which only the connector "OR" is used. By "R" (for restrictive) we designated profiles in which the connectors "AND" or "NOT" are used once. If the connectors "AND" or "NOT" were used twice or more than twice, the profile is regarded as "highly restrictive" ("HR").

According to our experience the substance oriented profiles are easy to formulate, while requests for information on some theoretical problem or phenomenon are more difficult. A scientist interested, *e.g.*, in some problem of thermodynamics or kinetics wants to get basic theoretical papers; he is not interested in the application of routine calculations. The question is how to formulate such a profile. Most probably the user will have to check all items indexed by the respective terms ("thermod*," "kinetic*") or by some terms designating more special parts of these fields.

Table V. Users' Needs Expressed by Search Profiles

Research group	No. of search profiles	Type of search profile	Keywords profile	Responses profile	Specificity of the profile
I	3	1 theoretical	10	2	HR
		1 methodological	4	1	HR
		1 substance oriented	2	33	S
II	5	2 theoretical	48	11	HR
			34	2	HR
		2 methodological	7	3	R
			9	8	R
		1 substance oriented	5	1	R
III	2	2 substance oriented	22	2	HR
			9	6	R
IV	1	1 substance oriented	41	50	R
V	2	1 methodological	8	29	R
		1 theoretical	3	3	HR
VI	2	2 substance oriented	5	47	R
			18	45	HR
VII	2	1 theoretical	9	6	R
		1 substance oriented	35	77	R
VIII	2	2 theoretical	11	1	R
			12	2	R
IX	2	1 methodological	18	11	HR
		1 substance oriented	6	8	R
X	1	1 theoretical	55	6	HR

Table VI

	Average keywords/profile	Average responses/profile
Group of 8 theoretical profiles	23	4
Group of 5 methodological profile	9	9
Group of 9 substance-oriented profiles	16	29

He will have to select the few items he actually wants and to reject the rest. We know from experience that it is futile to add terms such as "fundamental, basic, theoretical, new development, new trends," etc.

Computer searches for information on methods are much easier than one would assume with regard to the experience we have had with manual searching of, *e.g.*, the printed issues of *Chemical Abstracts*. The methods themselves are, in most cases, unambiguously defined by their names, so that a proper wording of a SDI profile is relatively easy. The abstracts related to these methods are scattered over many sections of CA, which, for a manual search, is a heavy handicap. For a computer it is a matter of mere seconds, especially where an inverted file is used.

Data assembled in Tables V and VI seem to confirm the above. When we take the group of theoretical profiles with the aid of an average of 23 keywords per profile, four responses per profile were obtained. In the methodological group the ratio is nine keywords:nine reponses; in the substance-oriented group an average of 16 keywords per profile is needed to obtain 29 responses per profile.

THE INFORMATION CENTERS AND ASSISTANCE TO THE USER

The user can hardly be supposed to know in detail the information system he is going to use. If his searches are to be effective, the assistance of information specialists familiar both with the system and the user's needs is indispensable. We take the specialist's knowledge of the system for granted. The question is how to acquire a thorough knowledge of the needs of the individual user of users' groups.

```
                CACS SYSTEM
00=SEARCH NUMBER 181 SAV      DEPT W                         0500181
01,OR=CRYSTALLOG*,X,STRUCTURE.                               0500181
02,AD=SINGLE,CRYSTALS.                                       0500181
03,OR=POWDER.                                                0500181
04,OR=DIFFRACTION,DIAGRAM.                                   0500181
05,OR=SILICATE*.                                             0500181
99
QU,50=(01OR02OR(03AD04))AD05



LAST ANSWER          N=     3

   4308  FELSCHE, J.
         RARE EARTH SILICATES WITH THE APATITE STRUCTURE
         CAS PUB.CIT.:CA07718119197Z   PUB.CL.CD.:J   PUB.DATE:000072
         J. SOLID STATE CHEM.
         SER./VOL.NO.:5                      PATENT NO.:
         ISS./REP./PART.NO.:2
         PAGES:266-75
         APATITE            STRUCTURE           RARE
         EARTH              SILICATE


   4492  SMOLIN, YU. I.
         SHEPELEV, YU. F.
         TITOV, A. P.
         REFINEMENT OF THE CRYSTAL STRUCTURE OF THORTVEITITE SC2SI2O7
         CAS PUB.CIT.:CA07718119382F   PUB.CL.CD.:J   PUB.DATE:000072
         KRISTALLOGRAFIYA
         SER./VOL.NO.:17                     PATENT NO.:
         ISS./REP./PART.NO.:4
         PAGES:857-8
         THORTVEITITE       STRUCTURE           SCANDIUM
         SILICATE


   4520  MERLINO, STEFANO
         NEW TETRAHEDRAL SHEETS IN REYERITE
         CAS PUB.CIT.:CA07718119410P   PUB.CL.CD.:J   PUB.DATE:000072
         NATURE (LONDON), PHYS. SCI.
         SER./VOL.NO.:238                    PATENT NO.:
         ISS./REP./PART.NO.:86
         PAGES:124-5
         REYERITE           STRUCTURE           POTASSIUM
         CALCIUM            SILICATE            HYDROXIDE


END OF REQUEST
```

Figure 4. Search output with "Last answer."

The obvious method, the filling in of questionnaires, is not quite satisfactory and not very popular. In our experience, and ours is not unique,[4] a continuous dialogue between the user and the information specialist is needed. The more accurately and completely the information specialist knows what the researcher is doing, the better he is able to help him in formulating his requests.

We made a clear distinction between the user's request and its machine-processable form, *i.e.*, a profile that would comply with the above cited rules (the number of parameters, the restrictions imposed on the number of the keywords used, the query line in the proper form, etc.). We instructed our users as regards the working of the system and the construction of machine-processable profiles.. If they were not able to construct their requests according to the rules of the system, we did not insist on their doing so. We insisted, on the other hand, that the users describe as accurately and completely as possible topics on which they would like to get current information. We strongly recommended that each user should define unambiguously: the substance or phenomenon that is the object of research, the method or methods used in the research (or even the apparatus, if this is of importance for the research), and finally the amount of information required. We had questionnaires for this purpose, but we accepted the specification of the user's request in any form he chose to use.

On the basis of the request specifications the profile was constructed by the information specialist. At this stage the search strategy was decided upon with regard to the type of research the requestor was doing (fundamental or applied) with regard to the object of the request (substance, method, theory). Of special importance was the output parameter specification on which it depended whether a "Temporary answer" or solely the "Last answer" was obtained. In Figures 4 and 5 we show outputs with the "Last answer" and those with a "Temporary answer."

Though computer handling of chemical information has become a routine matter in many cases, in general the users

```
                 CACS SYSTEM
00=SEARCH NUMBER 180 SAV      DEPT W                              0500180
01,OR=ZEOLITE,FAUJASITE*,MORDENITE*,LINDE.                        0500180
02,OR=MOL.                                                        0500180
03,OR=SIEVE*.                                                     0500180
04,OR=PREPARATOIN,SYNTHESIS,STRUCTURE.                            0500180
99
OU,50=(01OR(02AD03))AD04
FAUJASITE              DESCRIPTORS WRONG OR UNKNOWN
MORDENITE              DESCRIPTORS WRONG OR UNKNOWN
PREPARATOIN            DESCRIPTORS WRONG OR UNKNOWN



TEMPORARY ANSWER    N=    24    QU=01,02,03

      42  EGINA, S. P.
          ALIEV, A. M.
          NESTEROVA, L. A.
          STRUCTURAL STUDY OF POLYMERIZATION PRODUCTS ON DIFFERENT CAT
          ALYSTS
          CAS PUB.CIT.:CA07718114931Z   PUB.CL.CD.:J   PUB.DATE:000072
          NEFTEPERERAB. NEFTEKHIM. (MOSCOW)
          ISS./REP./PART.NO.13
          PAGES:55
          BUTANE              BUTENE              POLYMN
          CATALYST            MOL                 SIEVE
          ALUMINUM            CHROMIUM            NICKEL
```

Figure 5. Search output with "Temporary answer."

have to get used to this method and educational activity is still necessary. User education has to be administered carefully and in small doses since the scientists think (and we agree with them) that an information system has to help them and it should not make their literature search more tedious or time-consuming.

EVALUATION OF THE EFFECTIVENESS OF THE SYSTEM

Recently, much attention has been devoted to the evaluation of information systems. Many attempts have been made for finding appropriate measures by which values could be set on the performance of an information system.[2,3,6] Evaluation is possible only with regard to the aims and purposes that are to be attained by the system. There is no such thing as an absolute value of an information system. A comparison with some other system is necessary.

In the aforegoing paragraphs the aims of the system were discussed with special regard to various groups of users. The description of the system was also given and various factors that, in our opinion, affect the effectiveness of an information system were discussed. We had to choose very carefully the measures for the evaluation, since such procedures are not only costly, but time-consuming as well.

The evaluation of the effectiveness should not be an end in itself, it should rather contribute to a better understanding of the retrieval process and to improvements of this process.

The frequently used measure applied for some profiles had been precision *vs.* recall ratio. We found this measure not quite satisfactory for various reasons: it disregards completely the quantification of the users' needs, the judgment whether an item is relevant or not depends on the personal qualities of the user, and, as to the recall, it is nearly impossible to find in a data base like the *CA-Condensates* for control purposes by manual searching all items pertinent for one profile.

Introducing a computerized information system is justified only when it produces a better information service and when it saves effort, time, and money.[5]

From among the wide range of methods for the measurement of the system performance reported in literature,[7] we chose, for reasons of economy, a simplified measurement of an expected search length proposed by Cooper.[2] The basic idea is that there is a certain amount of wasted search effort when searching a collection of documents at random until the needed relevant documents are found. Using, in our case, a computerized retrieval system, we expected to reduce this amount of wasted effort. We took three profiles for which we compared the total of relevant retrieved items with the total number of abstracts in those CA sections in which the relevant answers occurred. If we had wanted to

Table VII[a]

Request profile	No. of relevant R	S	S/R
Atomic absorption	172	14,006	81
Structure of viruses	140	8,798	62
Refractories	265	15,497	58

[a] R = number of retrieved relevant documents, S = total number of documents (abstracts) in the sections of CA in which relevant documents occurred, S/R = search reduction factor.

Table VIII. Scattering of Relevant Documents over Sections of CA

No. of abstracts retrieved in one section of CA for profiles on		
Refractories	Atomic absorption	Structure of viruses
158 (1×)[a]		
	68 (1×)	
		52 (1×)
		43 (1×)
	25 (1×)	
23 (1×)		
18 (1×)		
13 (1×)		
	11 (1×)	11 (1×)
9 (1×)		
		8 (2×)
7 (1×)	7 (1×)	
	6 (1×)	6 (1×)
5 (1×)	5 (2×)	5 (1×)
4 (2×)	4 (3×)	
3 (3×)	3 (3×)	3 (1×)
2 (4×)	2 (4×)	2 (1×)
1 (7×)	1 (16×)	1 (2×)

[a] Numbers in parentheses indicate in how many sections of CA the respective quantity of relevant documents occurred.

find all these items by manual search we would have had to search all the above sections of CA (see Table VII).

Evaluating the expected search length we were surprised by finding that the relevant answers to our request profiles were scattered over so many sections of CA. At first sight it was clear that the scientist would never be able to make a manual search in all those sections of CA in which some pertinent abstracts had been found by the computer. We expected that to occur only in the case of the profile on atomic absorption, since we asked for the application of a method without any limitations as to substances. In Table VIII the number of items retrieved per section may be seen. As we see, 54% of the relevant abstracts on the atomic absorption are to be found in two sections of CA; in the search for information on both the refractories and the structure of viruses, 68% of the relevant information is concentrated in two sections. If the researcher were satisfied with receiving 54 and 68% of relevant items from the data base, then most probably a manual search of those sections where the pertinent information is clustered would be adequate and computerization would not be indicated. On the other hand, if completeness of information is required—and that is most often the case in fundamental and applied research where an unwanted duplication of the research effort is to be avoided—then the application of a computerized retrieval system is adequate. As it was emphasized in the previous paragraphs, the user's needs are to be respected and complied with. For this purpose an interactive computerized information system seems to be most convenient.

ACKNOWLEDGMENT

We thank all those whose cooperation and support made this study possible. We wish to express our sincere gratitude to Dr. M. Zikmund for his constant encouragement and advice.

LITERATURE CITED

(1) Balke, S., "Benutzerprobleme der Dokumentation und Information," *Nachr. Dok.*, **24**, 2 (1973).
(2) Cooper, W. S., "Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems," *Amer. Doc.*, **19**, 30 (1968).
(3) King, D. W., and Bryant, E. C., "The Evaluation of Information Services and Products," Information Resources Press, Washington, D. C., 1971.
(4) Meyer, R. L., Meskin, A. J., Mracek, J. J., Schwartz, J. H., and Wheelihan, E. C., "A Systematic Approach to Current Awareness and SDI," *J. Chem. Doc.*, **11**, 19 (1971).
(5) Skolnik, H., "The What and How of Computers for Chemical Information Systems," *J. Chem. Doc.*, **11**, 185 (1971).
(6) Stumpf, W., "Entwicklung und Erprobung von Methoden zur Auswertung in- und ausländischer Datenbänder für Retrievalzwecke im Bereich der anorganischen Chemie und ihrer Grenzgebiete," *Chem.-Zt.*, **96**, 301 (1972).
(7) Swets, J. A., "Information Retrieval Systems," *Science*, **141**, 245 (1963).
(8) Thompson, D. A., "Interface Design for an Interactive Information Retrieval System: A Literature Survey and a Research System Description," *J. Amer. Soc. Inform. Sci.*, **22**, 361 (1971).
(9) Wersig, G., "Zur Systematik der Benutzerforschung," *Nachr. Dok.*, **24**, 10 (1973).

Handling Commercial Product Names at Chemical Abstracts Service†

RUSSELL J. ROWLETT, JR.,* and DAVID W. WEISGERBER

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received January 11, 1974

Because Chemical Abstracts Service (CAS) abstracts and indexes a wide variety of technological and scientific literature, it is important that CAS be able to quickly and accurately equate the many commercial product names encountered in the literature with the actual complete chemical structures and the corresponding CA Index Names. This is accomplished readily within CAS processing by means of the CAS Chemical Registry System, a computer-based system that uniquely identifies chemical substances on the basis of their molecular structures. To the user of CAS products, the CA Index Guide provides a similar, although manual, link between the many commercial product names and the CA Index Names and Registry information.

Chemical Abstracts Service (CAS) publishes abstracts of the world's primary scientific literature which contains chemical and chemical engineering information and provides a variety of indexes to the original documents. All chemical substances for which new information is presented in the literature are indexed in *Chemical Abstracts* (CA) by name, molecular formula, and other indicators of structure. In order for a chemical substance name index to be useful, it must have all entries for a single substance appear reliably and consistently at one place in the index. Scattering of information in the index at synonymous substance names simply destroys the utility of the index because the user would never know whether he had located all references to a specific substance. This is particularly true of a large index such as CA, which now has over 630,000 individual Chemical Substance Index entries per six-month volume.

Equally important as having all index entries for a single substance appear at only one place in the index is the requirement that entries for related substances appear in proximity to facilitate generic searching. This ability to group related substances in an index is best accomplished by using fully systematic chemical substance names rather than their usual commercial or trivial names. Table I compares some common commercial and trivial names with the corresponding fully systematic names for the same substances. To obtain this reliability for its indexes, CAS has developed a comprehensive set of naming rules based on the nomenclature principles established by the International Union of Pure and Applied Chemistry.[1]

† Presented at the 166th National Meeting of the American Chemical Society, Symposium on Importance of Nomenclature of "Commercial" Chemicals in Chemical Safety, Chemical Disasters, and Chemical Literature, Chicago, Ill., Aug. 29, 1973.

* To whom correspondence should be addressed.

The problems arise when chemical substances are identified in the scientific and technical literature only by trivial or commercial product names such as Cinnamene or Dowanol EM, or perhaps only generic descriptions such as "the insecticide Gammexane" or "Polygard antioxidant." Placing such substances at their commercial product names in the CA indexes would simply scatter the chemical information. Such scattering of entries would make it almost impossible for a chemist searching CA to find all data for which he is looking. Just as a CA indexer must know where to place such substance index entries, so must the user of CA know where to look for such entries. Both require some way to rapidly and reliably equate the many commercial and trivial names from the original literature with the correct structure information and the CA Index Names.

Table II illustrates the type of problem an indexer, as well as a chemist preparing to search CA, might typically encounter. The five commercial product names are for the same common solvent. Some would be recognized immediately; others would probably not be recognized because they are less frequently used. But, for the purposes of indexing and searching CA, each of these names must be reliably converted to the CA Index Name "Ethanol, 2-methoxy-" and to the molecular formula $C_3H_8O_2$. The chemical substance name is inverted in the CA index; in normal text, this substance name will appear uninverted as "2-methoxyethanol."

While systematic nomenclature is essential to the production of an effective index, many chemists do not wish to become nomenclature experts. To facilitate their search of the CA indexes, the CA Index Guide identifies many commercial products and their CA Index Names. The CA Index Guide, which accompanies the CA Volume Indexes, is a collection of cross-references from the chemical substance