Of course, those situations where there were too many documents retrieved, additional descriptors might be added to reduce the number of documents per search.

Abrasives	No. of doc.	No. of searches
In 128 searches	0–5	64
	6–10	14
	11–20	22
	20–40	16
	40+	12
		128 searches

Synthetic Fibers	No. of doc.	No. of searches
Total 231 searches	0–5	102
	6–10	31
	11–20	43
	20–40	35
	40+	20
		231 searches

In many instances one mechanized search per patent application was sufficient. However, in other instances more than one search per patent application was used.

CONCLUSION

The above described technique has resulted in amortizing the indexing costs in one year.

LITERATURE CITED

(1) Frome, Julius, "A Punched Card System for Searching Steroids," *U. S. Patent Office Research and Development,* Rept. No. 7, 1957.

(2) Frome, Julius, "A Punch Card System for Phosphorus Compounds," *J. Chem. Doc.* 1, 84-7 (1961).

(3) Frome, Julius, "Random Access Mechanization of Phosphorus," *J. Chem. Doc.* 1, 76 (1961).

(4) Frome, Julius, and O'Day, Paul T., "A General Chemical Compound Code Sheet Format," *J. Chem. Doc.* 4, 33-42 (1964).

(5) Frome, Julius, "Semi-Automatic Indexing and Encoding," *U. S. Patent Office Research and Development,* Rept. No. 17, 1959.

(6) Frome, Julius, "Mechanized Searching of Phosphorus Compounds III Punched Cards *vs.* Random Access Computer," *J. Chem. Doc.* 1, 88-90 (1961).

(7) U. S. Dept. of Commerce, Patent Office *Manual of Classification,* Washington, D. C., July 1971.

Evaluation of the Database *CA Condensates* Compared with *Chemical Titles*

INGE BERG HANSEN

The Documentation Department, The National Technological Library of Denmark (DTB), Lyngby, Denmark

Received October 20, 1971

The performance of *CA Condensates* and *Chemical Titles* based on analysis of precision and "relative recall CT/CC" for a collection of 46 search profiles was studied over a period of one year. Special emphasis was laid on the function of the keyword phrases of CC and the users' attitude towards literature categories not represented in CT. The results are discussed in terms of the value of the systems for Danish users seen from the users' and the documentalist's point of view.

When Chemical Abstracts Service in September 1968 made their *CA Condensates* tapes available for use by national documentation centers and private companies all over the world, The National Technological Library of Denmark (DTB) decided to make a study of the new database to see if the user community in Denmark would respond favorably to the idea of having current access by computer to the world's chemical literature.

DTB had in 1968 been running the *Chemical Titles* tapes (CT) in cooperation with I/S DATACENTRALEN on commercial terms for the Danish scientific and industrial community for two years and had, therefore, developed a considerable amount of know-how in connection with computer-based information retrieval.[5-8,10]

The *Chemical Titles* service had from the start been run on a cost-recovery basis in the sense that the subscribers paid the actual computer costs, whereas the library paid the subscription charges to Chemical Abstracts Service. When the library decided to take up the more comprehensive and more expensive *CA Condensates* service, we were aware that the experiment might show that it might not be feasible to run the service on an ecomonically sound basis in a small industrial and scientific community like the Danish.

PURPOSE

Our primary intention in the present study was to compare *CA Condensates* and *Chemical Titles* and, further, to

examine the users' reactions towards the two services. The present report includes the results of the following aspects of the study:

1. Precision/recall analysis of *CA Condensates* compared with CT
2. Importance of the keyword phrases included in *CA Condensates*
3. User interest in reference to
   a. Condensates specific literature like patents, government reports, etc.
   b. Literature in "nonfamiliar" languages
4. User reactions towards literature retrieval systems like Condensates and CT in general

DESCRIPTION OF THE DATABASE

CA Condensates originally emerged as a by-product from the production of the printed version of *Chemical Abstracts,* and the first version which appeared in 1968 was, therefore, a rather crude product which by no means was intended as a commercial SDI service.

The database contained in 1968 approximately 250,000 references to the chemical literature in approximately 12,000 journals, dissertations, books, conference reports, government reports, and, in addition, 40,000 chemical patents. The information about the individual article consisted of CA abstract number, author, author's affiliation, title, journal coden, full bibliographic reference, and finally 4 to 7 keyword phrases corresponding to the entrances in the keyword index in the printed *Chemical Abstracts* (Figure 1a). The keyword phrases are constructed by the indexer from the original article and the CA abstract; it should, however, be noticed that the keywords are not selected from a controlled vocabulary. The average time lag between the primary publication and the appearance of the reference on the tape is reported by CAS to be 105 days (for comparison with *Chemical Titles,* see Table I).

As I/S Datacentralen was in the possession of a standard software which could be used for *CA Condensates* with minor conversion of the tape, it was decided not to use the software supplied by CAS. Because some sort of conversion was necessary, we decided to modify the database content to remove some of the information, which was considered as redundant by us, in the same operation. Apart from giving the output a more pleasant appearance, the major changes were removal of the name of the authors' institution and concentration of the keyword section to the words not already present in the title (Figure 1b). The reason for the removal of the first item was that DTB in connection with the literature retrieval systems supply all literature wanted by the users of the systems and neither we nor the users therefore need this kind of information.

The concentration of the keyword section to those words which were important for the retrieval of the references was based on the philosophy that the function of the keyword lines in connection with a tape service primarily is found in addition of words which may result in a hit which otherwise might have been lost. We are aware that the concentration almost entirely spoils the other function of the keyword phrases which is to supply the reader with additional information about content of the cited literature, but it was our impression after a careful scrutiny of corresponding titles and keyword phrases that the concentration process did not remove meaningful information not already present in the title to any large extent.

This version of *CA Condensates* was used by DTB throughout the experimental period and up to April 1971. By this time the Danish version of the CAS Standard Distribution Format was ready for use. An illustration of the output from a tape in the new format is shown in Figure 2, but apart from this no details regarding the SDF format will be given here, as this would be outside the scope of the present report.

a. original version

```
074286Q  7017 1  BIJOAK011100036903710004
074286Q  7017 4  BIOCHEM. J.
074286Q  7017 2  INTERACTIONS BETWEEN THE LY-
  SINE-RICH HISTONE F1 AND DEOXYRIBONUCLEIC
  ACID. = =
074286Q  7017 3  JOHNS EW, FORRESTER S.
074286Q  7017 4  (ROY. CANCER HOSP., LONDON,
  ENGLAND)
074286Q  7017 5  PROTEINS DNA COMPLEXES
074286Q  7017 5  RNA SYNTHESIS TEMPLATES
074286Q  7017 5  DNA HISTONE COMPLEXES
074286Q  7017 5  HISTONE DNA COMPLEXES
```

b. Danish version

```
BIJOAK/074286Q
  JOHNS EW              FORRESTER S.
    INTERACTIONS BETWEEN THE LYSINE-RICH
      HISTONE F1 AND DEOXYRIBONUCLEIC ACID. = =
    *PROTEINS *DNA *COMPLEXES *RNA *SYNTHESIS
      *TEMPLATES
  BIOCHEM. J., VOL 0111,0003,1969,PAGE 0371-0004
```

Figure 1 *CA Condensates*

METHODS

Selection of Profiles. To obtain a reasonable impression of the function of *CA Condensates* as an SDI service, we found that a comparison with the existing *Chemical Titles* service would offer a good basis for an evaluation of the system. In addition, we found that the user population for our purpose, where we were specifically interested in investigating whether it would be economically feasible to run the condensates system in a Danish environment, would be a group of people who were familiar with the existing possibilities for mechanized literature retrieval and who already

Table I. Comparison of CA Condensates and Chemical Titles (1968)

	CA Condensates	CT
No. of journals covered	1200 + Patents (25 countries) Books Dissertations Government reports Conference reports	700
No. of titles per year	250,000	125,000
Time interval from publication date until appearance on the tape	105 days	35 days
Information about the individual document	Title Author Full bibliographic Keywords CA-abstr. No.	Title Author Journal (CODEN)
Issues/year	52	26

CA-section No. ——————
CA-abstr. No. ——————
CODEN ——→ ANCHA/067748/64
HADDON, WILLIAM F. WILEY, MABRY WAISS, ANTHONY C.
WET. REG. RES. LAB., AGRIC. RES. SERV., ALBANY, CALIF.
AFLATOXIN DETECTION BY THIN LAYER CHROMATOGRAPHY MASS SPECTROMETRY. (ENG)
* AFLATOXINS QUAL ANAL
COTTONSEED MEAL THIN LAYER CHROMATOG MASS SPECTROMETRY
ANAL. CHEM. VOL 43, NO 2, 268-70, (71).

Figure 2. Illustration of the Danish version of the New CAS Standard Distribution Format (SDF)

Term cards (col. 1 - 4: duplicate query no.)

A B C . . . N	T S A C . . .	1 2 3 4	Term weight	Search term (char. string to be matched with text on searched tape, max. 60 char.)
5	6	7	10	11
A	T	2		FATTY ACID
A	T	4		LIPID
A	T	4		STEROL
B	T	2		ANAL
B	T	2		ESTIMAT
B	T	2		DETERMINAT
B	T	2		CHROMATOGR
N	T	2		PLANT

The subscriber to this profile will retrieve all references containing one word from group A + one word from group B, but not the word plant—e.g., "Gas liquid *chromatography* of long-chained *fatty acids* in the brain of growing rats" but not, e.g., "*Analysis* of *lipids* in plant material."

Figure 3. Illustration of a search profile

had expressed interest in and need for computer-based, SDI services.

We, therefore, offered the Danish research institutes and private companies, who in April 1969 were using the CT-service at DTB, to have one profile on *CA Condensates* free of charge for one year on the condition that during that year they would participate in the evaluation of the system. The role of the users in the experiment should be: to return information about the number of relevant items in each search on a form attached to the output, and to return a copy of the output indicating the references found to be of value after the user had seen the full article or the CA-abstract.

The interest for participation in the experiment on these conditions, which in our opinion would involve a great deal of work for the user, turned out to be surprisingly high. The number of institutions in Denmark which in 1969 were using our SDI service amounted to 24 representing approximately 130 individual profiles on CT. (A profile is in the present context defined as a group of up to 20 search terms (Figure 3)). The 24 user institutions were equally divided between public research institutes and private industrial companies. Of the 24 user institutions, 22 were interested in participating in the experiment. We did not make any attempt to urge the remaining two (one industry, one research representative) to participate; neither did we make any attempt to get an explanation for their lack of interest as we found that this action might push somebody into the experiment somewhat against his interests.

As some of the user institutions with a large number of profiles on CT were granted a few additional profiles in the experiment and as other institutions chose to let two different profiles participate for each half a year, we ended up with a group of profiles of the following composition (Table II).

Four of these profiles had to be left out in the final analysis as it turned out to be impossible to obtain the necessary feedback from the participants in spite of repeated attempts. The remaining 34 profiles as seen from Table II do not form any ideal population if the primary intention was to get an impression of the over-all performance of the database, and we found that it would be against our intentions with the study if we were to add a rather large group of profiles representing the CA sections which are

Table II. Distribution of the Experimental Profiles among Various Subject Fields

Public Research Institutes		Private Industrial Companies	
Biochemistry/Microbiology (pure and clinical)	6	Biochemistry/Microbiology (pure and technical)	8
Organic chemistry	1	Metallurgy	5
Physical and theoretical chemistry	9	Glass and Ceramics	1
Building materials	2	Reprography materials	1
Radiochemistry	1	Fertilizers	1
Semiconductors	1		
Chemical processing (theoretical)	1		
Pollution	1		
Total	22		16

underrepresented here (the organic chemistry sections and the macromolecular sections), as firstly, the users of these profiles most likely would respond very differently from the rest of the user group and, secondly, would not be true representatives of a potential user community.

On the other hand, there was obviously a missing link in a study of users' reaction to Condensates if only existing subscribers to CT were included in the experimental group as we, in that case, would miss those for whom CT was inadequate, and we, therefore, tried to compensate for this by asking those users who took commercial subscription to Condensates during the experimental period to participate in the experiment by supplying feedback based on output alone for a period of 3 to 6 months.

The participating "commercial group" had the following composition:

Research		Industry	
Biochemistry/microbiology (pure and clinical)	5	Biochemistry/microbiology (technical)	3
Physical chemistry	1	Metallurgy	2
Analytical chemistry	1		
Total	7		5

The final group of users thus consisted of 26 research profiles and 20 industrial profiles and gave a fair representation of the performance of *CA Condensates* in three of the five main section groups. The organic chemistry sections and the macromolecular sections are hardly represented. We made several attempts to attract organic chemists to use the database, but failed to create any lasting interest in this field. Judging from the test searches we made in the organic chemistry field in connection with teach-ins and study groups, we tend to confirm the general impression[9] that *CA Condensates* cannot be considered as a suitable database for organic compounds until it will be possible to search by the compound registry numbers.

The lack of interest for the macromolecular sections in Denmark can be explained by the fact that the industries in this field are rather small and do not require current awareness at the advanced level represented in *CA Condensates*.

Construction of the Profiles. Profiles that had been run on *Chemical Titles*—and reached an acceptable form there—were transferred directly for search on *CA Condensates* after addition of the abbreviations necessary for search in the keyword phrases and replacement of the segmented

ICHAA3-0003-0129 69 129-35
KARAYANNIS NM MINKIEWICZ JV PYTLEWSKI LL
LABES MM
4-ETH OXY PYRIDINE N-OXIDE COMPLEXES OF METAL HALIDES AND PER CHLORATES.=

Figure 4. Illustration of the word segmentation used in *Chemical Titles*

word forms used in CT (Figure 4) by the normal CAS standard spelling. A few profiles had to be modified further when the word combination required for CT resulted in too much noise in Condensates; but in the majority of the cases, it had to be realized that a larger amount of nonrelevant hits had to be accepted if relevant material should not be lost. In certain areas the profile performance could be improved by limiting the search to either odd or even issues of CA.

For new profiles our standard procedure in the construction of the profile was to ask for a list of relevant articles, a written description of the subject in normal language and a draft of a profile written by the user with the aid of a user's guide prepared by DTB. The final draft of the profile was then constructed after personal contact between the user and the information specialist at DTB. In most cases, the user's first draft of the profile showed a very good understanding of the search techniques and the whole philosophy behind the information retrieval procedure. We did, however, make the same observation as reported by Barker *et al.*[4] in the Nottingham study, that users generally tend to make profiles much too specific and that the majority of the profiles had to be broadened considerably to get sufficient material for the construction of the final profile.

The actual search strategy has been reported in detail elsewhere and will be discussed only briefly here. The search logic is a straightforward and/or/not logic which can be combined with the application of weights to the individual search terms to a nested Boolean logic. The search program had, during the experimental period, facilities for search in text (title + keyword), author, or journal coden. It was possible to search for word or word stem allowing free suffix and/or prefix.

The average number of search terms per final profile was 16.6. A variety of search logics was used as seen from Table III. Of the 15 weighted profiles, only one came from the "old" CT profiles. This does not imply that weight is more useful in connection with use of Condensates compared with CT (although sometimes more necessary), but rather that the author, who took part in the construction of the majority of the new profiles used in the experiment, prefers using weighted profiles in general.

Table III.

Search Logic	No. of Profiles
A	7
A+B	15
A+B+C	3
A/NOT	1
A+B/NOT	5
A/W	9
A+B/W	4
A+B/−W	1
A/W/NOT	1

W = weight used

All profiles were initially run on odd as well as even issues of Condensates. After approximately 5 months, we examined the output carefully and those profiles which never retrieved relevant answers in one of the series were removed from the redundant file. For certain profiles where some search terms were either redundant or resulted mainly in "noise" in one of the issues, the profiles were run in two different versions in the two series. In such cases the combination of the two versions are, however, still considered as one profile for analytic purposes.

User Feedback. Our original intention regarding user feedback was that users should return statistics over the number of relevant items per search based on the output alone, as soon as the output was received. In this context, we requested the user to use the term relevant in the sense that all references, which judging from the output appeared to be of interest to the user, should be regarded as relevant. This means that items which were relevant according to the profile and the user's description of the subject, but not considered as relevant by the user where not classified as relevant, although they in our opinion ought to be. The same applies to references in languages not familiar to the individual users. If the user did not mark a reference which according to the output should be of interest as relevant on the assumption that literature appearing in certain languages were not worth reading, we accepted the user's judgment (although we specifically had requested people to have a look at this sort of literature, before it was dismissed). This first evaluation of the output was used for the calculation of precision I (see later).

The second step in the procedure was that users were supplied with the full article corresponding to the relevant reference, when it could be obtained within the Scandinavian countries; in other cases, the user was supplied with a copy of the CA-abstract. This means that we did not make any attempt to investigate whether some of the references classified as uninteresting on the basis of the output really would have been of interest to the user. According to the Nottingham Study,[4] it probably would have raised the figures for precision II by approximately 10%.

When all relevant articles (or abstracts) had been scanned by the user, it was our idea that a copy of the output would be returned with indications of (a) references classified as relevant based on output (marked +) and (b) references classified as relevant after scanning of the articles (marked R). This latter step did not work successfully. In many cases (particularly in the industrial group), we never managed to get any feedback from this part of the project, and in many cases where we finally got it, it came so late that we were unable to update the profiles in time to get some results based on the revised and improved profiles. Generally speaking, we must admit that we were too ambitious in setting up this part of the project, and we finally had to accept that the calculation of precision II based on final relevance could be based upon 13 profiles only (11 research and 2 industry). For ten other profiles, we received a classification of more or less relevant, but for the remaining profiles we were only able to use the first relevance estimates.

Comparison of Condensates and CT. The comparison of Condensates and CT was based upon precision and recall, whereas no figures for comparative currency were calculated, as in this respect we accepted the figures of 105 and 35 days, respectively, given by CAS. The comparative figures for precision and recall were calculated by the documentalist in the library who examined the evaluated output and worked out a list of how many of the relevant items retrieved through Condensates would have been retrieved by the corresponding profile in CT; in addition, it was established to which extent the retrieval of items found in Condensates, but not in CT, were due to the presence of the keyword phrases in Condensates and to which extent they represented citations not present in CT at all.

To obtain a certain uniformity in the material, the average figures for each profile were calculated on the basis of material for either one volume of Condensates or at least 100 relevant items. For a few very restrictive profiles we continued, however, the collection of data for up to one year.

The calculation of recall was based on the assumption that in the author's opinion it is unrealistic to aim at establishing any exact figures for an absolute recall and that it, therefore, would be more useful to calculate a comparative "recall CT/Condensates" estimate as follows:

$$\frac{\text{recall (CT)}}{\text{recall (Condensates)}}\% =$$

$$\frac{\dfrac{\text{relevant ref. (CT)}}{\text{total no. of rel. ref.}}}{\dfrac{\text{relevant ref. (Condensates)}}{\text{total no. of rel. ref.}}}\% =$$

$$\frac{\text{relevant ref. (CT)}}{\text{relevant ref. (Condensates)}}\%$$

Mean and standard deviations were calculated for the experimental group of profiles, and in addition the mean values for the industry and research groups, respectively, were calculated and compared by means of a t-test.[11]

User Reactions. It was our original intention to include an examination of the users' opinions of the systems by means of a questionnaire. The answers to a preliminary questionnaire which was sent out at the end of the experimental period gave, however, such an obscure picture of the users' wishes and opinions that it would have been meaningless to try to obtain any conclusions which would have any statistical importance and we, therefore, decided not to send out a final questionnaire. Instead, we decided to limit ourselves to draw some conclusions on the basis of personal interviews combined with the comments received through the preliminary questionnaire.

RESULTS AND DISCUSSION

Precision and Comparative Recall. As seen from Table IV the investigation made it possible to draw some important conclusions. First of all, the precision obtained in Condensates was almost identical to the precision obtained in CT with the corresponding profile. This was slightly surprising to us, as we had expected that the precision in Condensates would be significantly lower in Condensates than in CT due to the larger amount of secondary and "local" literature cited in Condensates. Further, it appears that the number of relevant answers to a profile is approximately twice as high in Condensates as in CT. This is in accordance with the fact that Condensates contains twice as many references as CT, but, on the other hand, as the keywords in Condensates also contribute to the retrieval of the additional relevant items in Condensates, we must conclude that the precision in the CT-covered part of the references is higher than in the "non-CT-covered" references of Condensates.

When the results for the industry and the research group are compared, two things should be noticed. First, the figures indicate that the precision in the industry group is lower than in the research group. This result is not statistically significant for the present size of population,

Table IV. Comparison of Precision and Recall in *Chemical Titles* and *CA Condensates* (CC)

	Research				Industry	
	Total, %	Basic Chem., %	NBC, %	Commercial, %	Total, %	Commercial, %
	$\bar{X}$ S.D.	$\bar{X}$ S.D.	$\bar{X}$ S.D.	$\bar{X}$ S.D.	$\bar{X}$ S.D.	$\bar{X}$ S.D.
Precision I CC	31.6 ± 25.1	25.6 ± 23.5	36.1 ± 25.2	28.0 ± 11.3	24.6 ± 12.2	30.9 ± 14.0
Precision I CT	35.0 ± 26.1	29.8 ± 26.9	38.8 ± 24.7	30.2 ± 13.0	25.7 ± 14.4	34.7 ± 15.5
Recall I	50.0 ± 15.5	61.3 ± 9.3	41.7 ± 13.8	43.6 ± 10.2	26.9 ± 11.4	34.8 ± 9.7
No. of profiles	26	11	15	12	20	8
Av. No. of relevant hits examined per profile	129				92	
Precision II CC	12.6 ± 12.2	15.0 ± 16.0	9.8 ± 2.4		11.9 ± 10.6	
Precision II CT	14.8 ± 12.3	17.2 ± 15.7	11.9 ± 4.9		11.2 ± 11.5	
Recall II	54.0 ± 18.8	66.2 ± 9.2	39.4 ± 16.8		24.3 ± 14.7	
No. of profiles	11	6	5		8	

$\bar{X}$ = Mean. S.D. = Standard deviation. I = Based upon output relevance. II = Research, based upon final relevance; Industry, based upon major outlet relevance.

but it supports the impression we got during the experiment regarding the literature habits of industrial vs. research chemists, namely that the research chemist employed in public institutions have much more time and/or interest in receiving background information for his research, whereas the industrially employed chemist generally only marks "straight-to-the core" references as relevant.

Second, the average relative recall CT/CC is significantly higher in the research group than in the industry group. This clearly demonstrates that as far as literature coverage is concerned industry is much better served by Condensates than by CT which for a large number of the industrial profiles in our study proved to be quite inadequate.

When comparing the figures for final precision and recall based on actual knowledge for the cited articles, we find that the precision figures are significantly reduced for CT as well as Condensates precision. In the industry group, we had too few estimates of final relevance to calculate meaningful figures, but here the average figures for major relevance corresponded closely to the figures for final precision in the research group. In the research group, recall II was slightly higher than recall I, but the difference was not significant.

As we found some very high standard deviations for almost all of the results in Table IV, we thought that we might obtain some information about the utility of Condensates in various fields if the two main groups could be divided in smaller more uniform groups, and it turned out that the basic chemistry group within the research group was significantly different from the rest of the research group in having a much higher CT/Condensates recall. We further compared the remaining part of the research group (non-basic chemistry = NBC) with the industry group to see if there still could be noticed a difference between industry and research and found that as far as recall was concerned, the two groups were significantly different. We further observed that the precision values for the NBC group and the industry group possibly were significantly different ($t_{0.975} > t > t_{0.95}$); analysis of a larger population may therefore show that the two groups actually are different also as far as precision is concerned.

As we further supposed that the precision and recall values might reflect a possible difference between the commercial subscribers and those who did not decide to subscribe to Condensates after the end of the experiment, we compared commercial and noncommercial profiles within the industrial and research groups. This comparison showed some interesting trends although the differences were not statistically significant. In the research group, we found precision and recall slightly lower than in the total R group, whereas the corresponding values in the industry group were higher in the commercial group than in the total I group. For the research group the lower recall can be ascribed to the fact that the commercial group mainly consists of non-basic chemistry profiles, and the recall should, therefore, correspond closely to the recall value for the NBC group. The explanation for the lower precision value may be that the users who subscribe to Condensates primarily are interested in a high recall and, therefore, not particularly interested in constructing profiles with high precision values. For the industrial group, we interpret the results as an indication that the commercial industry users are more information-minded than the average industrial chemist, but we are aware that this may be considered as a rather subjective interpretation, particularly since the economic factor may have influenced the choice of the database considerably.

We shall not go into details regarding the level of the precision values, as precision values in connection with SDI services have been subject to much discussion elsewhere.[12] Our values correspond to the values reported by others, although the industrial values may be slightly lower than the values generally reported for SDI services.[9,13] We do, however, find that the majority of the values in our study are thoroughly acceptable for an SDI service, and as long as the average scientific author has not yet learned to be more consistent in his choice of titles we find that this precision level is not only acceptable, but also preferable. Further, we find that although precision is an excellent measure for profile performance from the documentalist point-of-view, the users opinion about the optimal levels for precision values to a large extent will depend upon the absolute number of answers to a profile; for instance, a precision of 10% will probably be acceptable if the profile retrieves 10 answers, but not if the number of hits exceeds 100.

Importance of the Keyword Phrases. Table V shows that the keyword phrases for almost all of the user categories contribute to the relevant hits with approximately 20%; this finding is consistent for relevance based upon output as well as relevance based upon knowledge of article or abstract. We do, however, also see that the keywords are responsible for approximately 10 times as many irrelevant as relevant answers. There is an apparent difference between the basic chemistry group and the other users which is not statistically significant, but the results

Table V. Importance of the Keyword Phrases of *CA Condensates*

		Research					Industry		
		Total		Basic Chem.		NBC	Commercial	Total	Commercial
		$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$ S.D.	$\bar{X}$ S.D.	$\bar{X}$ S.D.	$\bar{X}$ S.D.
% of relevant hits retrieved by Keywords (k)	I	19.2 ±	8.8	16.0 ±	7.0	21.5 ± 9.3	20.3 ± 9.3	26.7 ± 11.5	26.7 ± 7.7
	II	19.3 ±	5.9	20.5 ±	5.7	17.8 ± 5.7		22.4 ± 9.6	
Irrelevant hits retr. by k / Relevant hits retr. by k		11.0 ±	13.8	16.6 ±	18.1	6.6 ± 6.2	6.2 ± 5.6	7.5 ± 5.6	4.9 ± 3.9
Maximal recall CT/CC, %	I	61.0 ±	16.9	73.6 ±	10.9	51.8 ± 14.4	54.7 ± 10.7	35.8 ± 15.9	46.1 ± 15.0
	II	67.7 ±	22.9	83.9 ±	10.0	48.4 ± 18.5		30.2 ± 17.8	

No. of profiles and explanation of symbols are given in Table IV. The maximal recall CT/CC is calculated by addition of the percentage of relevant CC-hits covered by CT, but not retrieved by the profile, to the recall values given in Table IV.

clearly indicate that the keyword phrases function less satisfactorily in the basic chemistry fields than in the other profiles. The ratio between irrelevant and relevant hits may be reduced in the new SDF version of Condensates where it will be possible to limit the search to certain sections, but this remains to be seen, when more material is available regarding the use of the SDF format.

As far as the first version of Condensates is concerned, we must conclude that the keywords have been a very valuable tool in improving the recall of a profile; it is an expensive tool owing to the large number of irrelevant hits retrieved by the keywords, but on the other hand, this solution will probably be cheaper than the alternative solution which would be to increase the number of search terms of a profile to cover the otherwise "lost" items, as this alternative not only would raise the expenses of search time, but also might retrieve as many irrelevant hits through the added terms as through the keywords.

The keyword analysis further shows the per cent of the relevant hits in Condensates that could have been retrieved by CT by expansion of the CT profile. Table V shows that the recall could have been improved by approximately 10% for industry as well as research profiles by expanding the CT profile. The maximal recall II further shows a probably significant difference between the BC and the NBC groups ($t_{0.99} > t > t_{0.975}$).

Distribution of Relevant Hits between Various Languages and Various Literature Categories. In connection with the examination of the users' interest in literature only covered by Condensates, we made a separate study of the distribution of the relevant items among the following groups:

a. Journal literature in the main Western European languages (English, German and French)
b. Journal literature in other Western European languages
c. Eastern European journal literature*
d. Japanese journal literature*
e. Other Asian literature*
f. African literature (incl. South Africa)
g. Latin-American literature
h. Patents**
i. Government reports**
j. Dissertations
k. Congress reports**
l. Books

*Includes journals published in the main Western European languages—e.g., *Journal of Biochemistry (Tokyo)*.
**Covers all languages.

The results in Table VI came as a surprise to us. First of all, the difference in the distribution between the industrial and research group was much more pronounced than we had expected—the two significant features being the dominance of Western journal literature in the research group on one side and, on the other, the almost equal distribution between Western, Eastern, and patent literature in the industry group. The Eastern literature contributed to the relevant hits with approximately 20% in both groups, which corresponds closely to the Eastern European representation in *Chemical Abstracts* as a whole.[3] Also the Japanese literature contributed to the relevant hits in the same proportion as represented in *Chemical Abstracts*.

Regarding the Condensates specific publications, we found that the conference reports were highly appreciated by both industry and research representatives, patents by the industry, and dissertation abstracts to a certain extent by the research representatives. The government reports were, contrary to our expectations, not classified as relevant by very many participants; as these reports are normally regarded as extremely useful and subject to an extensive demand in our library, we assume that the lack of interest in this type of literature is explained by the fact that the reports often are cited by Condensates so late that a corresponding journal article often has been cited several months before.

Users Attitude towards CA Condensates. The opinion of the Condensates service among the user population in

Table VI. Distribution of Relevant Answers Between Various Types of Publications (%)

	Research				Industry	
	I*		II**		I	
	$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.
Western European J. (Eng., Ger., Fr.)	60.1 ±	13.7	69.0 ±	12.4	33.5 ±	19.8
Western European J. (other languages)	2.2 ±	1.8	1.9 ±	1.8	2.1 ±	3.0
Eastern European J.	17.1 ±	10.5	15.2 ±	13.6	24.4 ±	16.0
Japanese J.	4.3 ±	2.7	4.5 ±	3.5	5.0 ±	6.0
Other Asian J.	1.3 ±	1.4	1.6 ±	2.6		
Patents	2.7 ±	5.1			26.3 ±	21.4
Government reports	2.5 ±	4.8				
Dissertations	2.0 ±	2.0				
Congress reports	5.9 ±	4.3	6.1 ±	8.2	6.6 ±	7.4
No. of profiles	25		10		20	

*I = The figures are based upon output-relevance. **II = The figures are based upon final relevance. $\bar{X}$ = Mean. S.D. = Standard deviation. Only publication types representing more than 1.0% are included.

our experiment was judged by means of a preliminary questionnaire and personal interviews. Generally, we found the reactions from the original CT users extremely unsystematic, and the criteria used for making the choice between the two systems appeared to be so personal that many user evaluations were of limited help in our analysis of the value of the service.

Table VII shows the CT-user group's choice between the two systems by the end of the experimental period. Here we see, for instance, the very characteristic finding that although the relative CT/CC recall in none of the groups exceeds 62%, a considerable proportion of the participants state that they do not find more relevant references in CA Condensates than in CT. This discrepancy between calculated results and user statements has several explanations. First of all, many users made current use of other abstracting services, such as the printed versions of *Current Contents, Nuclear Science Abstracts,* etc., and they would, therefore, often not know exactly where they had seen the references previously. Secondly, the majority of the CT users ran about five profiles in CT, and they would, therefore, have retrieved many of the references covered by CT, but not by the experimental profile through other profiles in CT. As seen in Table V, the basic chemistry group could have retrieved approximately 73% of the relevant hits found in Condensates by expansion of the profile in CT, and this may explain the rather negative attitude towards Condensates in this particular group.

Another typical reaction from the users was that they were annoyed with the higher number of irrelevant hits, and the irritation did not seem to be compensated for by the many extra relevant hits retrieved by Condensates.

The two factors which generally are considered as disadvantages in Condensates are, of course, the longer response time and the higher price for the annual subscription. It is, however, our opinion that this reaction primarily is a psychological problem. Considering that it often will be sufficient to run a 10-term profile in Condensates instead of a 20-term profile in CT and that this profile normally will retrieve twice as many relevant references, we do not think that the price difference should be of importance. The difficulty is rather that the Condensates price of 1250 Danish Kroner for a 10-term profile compared with 800 Danish Kroner for a 20-term CT profile has reached a prohibitively high level.

As far as the second factor is concerned, it is our opinion that the disadvantage of the longer response time of Condensates often has been overestimated. Most research workers (including ourselves) will, of course, in principle state that a difference of two months is a severe disadvantage for the slower service, but it is our impression that as soon as our users have subscribed to our services for a couple of months, they are so engaged in their own fight against the information explosion that it does not matter very much whether an article is cited a month sooner or later. This does not mean that we deny that certain types of information are needed urgently, particularly by the industry, but since the experiment here clearly has shown that the coverage offered to the industry by CT is rather inadequate one would in any case from a documentalist point of view tend to give priority to coverage compared with urgency. We, therefore, very much appreciated that the questionnaires returned from the participants confirmed our opinion in this respect as only 20% declared that Condensates was too slow.

Apart from these more obvious objections to the use of Condensates we met many odd reactions. As examples could be mentioned that a user who had a precision of 100% stated that the profile retrieved too many redundant relevant answers. Another did not want to use Condensates, because it was too difficult to read the output (see Figure 1b). Other users who from the experiment could see that the value of CT was of limited value to them, whereas Condensates worked quite well, continued happily with CT after the end of the experiment. As a whole, we found so many different explanations for the choice of one system compared with the other, that we tend to conclude that it probably will be the best solution if the documentation center decides which systems should be run in the environment concerned, possibly after discussion with users, who express sufficient interest to approach the center and tell what they want. Unfortunately, our experiences do not really favor the ideas of user democracy in connection with library and information services.

The remaining opinions expressed in the questionnaire largely correspond to the calculated results. Only a few were dissatisfied with the amount of literature to nonfamiliar languages and with the patent coverage; both things are contrary to our expectations in this respect, but correspond to the results in Table VI.

Among those users who found that Condensates was a valuable tool in their information supply, five switched entirely over from CT to Condensates by the end of the experiment. Regarding those who wanted to continue the experimental profile in Condensates after the end of the experiment, we found, however, surprisingly little impact of the experiment among the colleagues of the participants. Of course Table VII shows that 14 new subscriptions were made by the participating institutions after the experiment, but as these 14 profiles represent only four subscribers, the figure gives a slightly misleading picture of the interest in Condensates among the participants.

Table VII. Summary of Questionnaires

	Research			
	Basic Chemistry	Others	Industry	Total
Same profile ctd. after exp.	0	5	8	13
New subscriptions among the participants	3	—	11	14
Corresponding profiles disctd. in CT after expt.	1	6	9	16
Comments				
CC too slow	3	2	2	7
CC too expensive	4	2	5	11
Too much "noise" in CC	4		1	5
All relevant answers found in CT	7	4	—	11
Too many ref. in nonfamiliar languages			4	4
Patent coverage not satisfactory			2	2
No. of participants	10	11	15	36
Av. recall CT/CC	62.9	40.6	24.2	

USER NEEDS

What then can one assume that users want from an SDI service (more specifically Danish users) if one should use the experiences gained in the present study? First of all one can state that the service must work smoothly without involving the user in too many problems to add to his own research problems. In our study, we tried not to bother the participants with more questionnaires, circulars,

and changes in search technique than absolutely necessary, but judging from the response time of messages sent to the users, it does not seem as if messages concerning the services are popular reading among the users, not even when very valuable improvements are introduced.

What the users want in terms of precision and recall is difficult to state, as we had users only concerned with optimal recall as well as users mainly interested in high precision and a minimal output. We, therefore, find, that Condensates after the introduction of the section numbers on the tapes can offer both groups the service they want. On the other hand, a user belonging to the second category working with basic chemical research should subscribe to CT which will supply him with faster information than Condensates. For the industrial user in the second category, the problem will be more complex, as CT does not offer an efficient coverage of industrial topics, and although his needs in terms of precision and recall can be met with the SDF format, he may still find that the references appear too late.

When thus watching the users' claims regarding precision and recall from a purely academic point of view, we must conclude that precision and recall should not be used as a measure for the value of the system to the ultimate user. The two parameters should, however, be considered in the documentalists evaluation of the system, as he will have to estimate the literature coverage required for the entire user population he wants to serve; the precision is important when he shall decide how much money his institution and his potential users can afford to spend on a literature service. In this latter context, the precision also will be of importance in the users opinion of the service, where the users have to pay the service on the basis of the number of hits; at present the Danish price for information retrieval services is independent of this and this may explain the lack of interest in obtaining a high precision among some of our users.

What in our experience is considered as having a high priority among our user population is the smooth supply of the literature, which in our library forms an integrated part of our services. Literature order forms are mailed with every search output, and the interest in ordering the literature is extremely high. In 1970, we supplied 240 copy-pages per profile. As a matter of fact, we consider the literature service offered by our library in connection with the information retrieval service as one of the main reasons for our existence, as our users if they were less interested in this part of the service would have great economic advantage of using the state-subsidized information retrieval services offered by neighboring countries.

In a discussion of user needs it is, of course, not only necessary to study the reaction among the users of computer-based services but perhaps even more to study the reactions from chemists who do not find that the mechanized services will satisfy their requirements. A larger sociological study of information needs of chemists and chemical technologist is outside the scope of the present study, as this has been done elsewhere,[1,2] but a few comments heard during meetings and teach-ins should be noted. First of all there is the almost traditional claim for security from a large part of the industry. In our opinion, this point has been strongly overemphasized in connection with CT and Condensates, and we, therefore, found it interesting that a group of representatives of some of the largest industrial documentation departments recently in a working group under EUSIDIC (The European Association of Scientific Information Dissemination Centres) stated that considering the expenses involved in running the databases in closed environments the confidentiality claim had to be reduced in favor of collecting larger groups of users for databases run by larger centers.

Another comment heard generally is that the organic chemistry sector is not covered sufficiently in Condensates and CT. We agree that CT in this field is unacceptable, and that even Condensates is not working too well; it is our intention to initiate cooperation with centers running databases on substructure searching to serve the requirement of Danish organic chemists.

The third criticism often raised against Condensates from industry is the late and insufficient coverage of the patents. In this connection, we would like to remark that the aims of *Chemical Abstracts* is to cover the chemical patent literature as well as other types of chemical literature, whereas it by no means pretends to function as an actual patent service which enables people to make objections to other people's patents in time. As almost 25% of the relevant hits in the industry group is found among the patents, we must conclude that at least the participants in this experiment appreciate the references to the patents found in Condensates. This does, of course, not say anything about whether the coverage is satisfactory or not, and although only two of our users in the questionnaire described the patent coverage as unsatisfactory, a careful study of the coverage of the British patents made by one of the users showed at least in this specific area only a 50% coverage of the patents which in the opinion of the user ought to be covered. The same examination gave an impression of the currency of the patent references in Condensates compared with the three patent services IPRO (International Patent Research Organization), Univentio, and Pullitzer normally used by the company concerned (Table VIII), and it was found that the currency of the British and American patent references was quite acceptably regarded as literature information, whereas the references to German and French patents in at least this particular field must be described as too slow.

Table VIII. Difference between the Currence of Patent Citations in *CA Condensates* and the Patents Services IPRO, Univentio, and Pullitzer

	CC − P.S.,* Months		
	$\bar{X}$	S.D.	No. of Patents
British patents (BR)	1.0	± 0.8	10
US patents (US)	2.0	± 3.6	20
French patents (FR)	6.4	± 4.1	33
German patents (GW)	5.6	± 3.1	8

*Date of CC citation − date of patent service citation; i.e., the difference is positive for CC citation appearing later than patent service citation. $\bar{X}$ = Mean value. S.D. = Standard deviation.

CONCLUSION

Summarizing, the results indicate that *CA Condensates* is a very good database covering a wide spectrum of interests in chemistry and the bordering areas. It cites the references about two months later than CT, but in the opinion of the majority of our users this difference should not be regarded as a severe disadvantage.

For the industrial user, the results show that CT does not offer a sufficient coverage; in spite of this, a large proportion of the former CT users continued using CT which can be explained by economy as well as habit.

CA Condensates covers a number of publication types not included in the CT service; some of these were regarded as very useful, particularly the dissertation abstracts and the patents.

ACKNOWLEDGMENT

The author is indebted to the Danish Council for Scientific and Industrial Research for the financial support of the work. We further wish to express our thanks to the staff of the Documentation Dept., DTB, especially Birgit Pedersen, for assistance during the experiment and to the users participating in the experiment for their evaluation of the output. We are extremely grateful to the staff of I/S DATACENTRALEN for their active interest in the development of improved techniques in the utilization of the CAS tapes.

LITERATURE CITED

(1) Amick, D., "Multivariate Statistical Analysis of the Use of a Scientific Computer-Based Current Awareness Information Retrieval System," *J. Amer. Soc. Inform. Sci.* **21,** 171–8 (1970).
(2) Arnett, E. M., "Computer-Based Chemical Information Services," *Science* **170,** 1370–6 (1970).
(3) Baker, D. B., Tate, F. A., and Rowlett, R. J., Jr., "Changing Patterns in the International Communication of Chemical Research and Technology," *J. Chem. Doc.* **11,** 90–8 (1971).
(4) Barker, F. H., Kent, A. K., and Veal, D. C., "Report on the Evaluation of an experimental Computer-Based Current Awareness Service for Chemists," United Kingdom Chemical Information Service, University of Nottingham, 1970.
(5) Berg Hansen, I., "Chemical Literature on Tape," *Kem. Teollisuus* **26,** 997–1005 (1969).
(6) Berg Hansen, I., "Computer-Based Chemical Information Systems at the Danish Technological University Library," *IATUL Proc.* **5,** 14–20 (1970).
(7) Berg Hansen, I., "A Comparative Study of Some Information Retrieval Systems in Chemistry and Biomedicine," *Ind. Chim. Belge* **37,** 25–30 (1972).
(8) Boman, M., "Computer-Based Chemical Information Retrieval," *Sv. Kem. Tidskr.* **80,** 379–83 (1968).
(9) Johansson, A., Kallner, A., and Markusson, K., "Literature Documentation Service through Chemical Abstracts Condensates—an Evaluation," *Kem. Tidskr.* **82,** 24–6 (1970).
(10) Skov, H. J., "An Electronic SDI Service for the Danish Chemical Industry and Research," *Libri* **18,** 204–15 (1968).
(11) Spiegel, M. R., "Theory and Problems of Statistics," Schaum Publishing Co., New York, 1961.
(12) Swets, J. A., "Effectiveness of Information Retrieval Methods," *Amer. Doc.* **20,** 72–89 (1969).
(13) Veal, D. C., "United Kingdom Experiences in the Operation of a Retrieval and Dissemination Service Based on CAS Search Tapes," *Neue Tech. A.* **11,** 281–95 (1969).

Subject Compatibility between *Chemical Abstracts* Subject Sections and Search Profiles Used for Computerized Information Retrieval

INGE BERG HANSEN
The Documentation Department,
The National Technological Library of Denmark (DTB), Lyngby, Denmark

Received October 20, 1971

The need for introducing subject section numbers on the *CA Condensates* tapes was studied by analysis of the distribution of relevant answers to 41 search profiles among the 80 subject sections of *Chemical Abstracts*. The average profile requires 10 CA-subject sections for adequate coverage. The average printing expense per profile could be reduced 25% by searching the individual profiles in the appropriate subject sections.

In connection with our study of the CAS database *CA Condensates*,[2] a detailed investigation was made of the compatibility of the 80 subject sections used in *Chemical Abstracts* with the subject interests expressed in each individual search profile.

The background for this investigation was that the first tape version of *CA Condensates* did not contain the subject sections to which the references belonged, and as the occurrence of redundant answers from subject sections far from the user's field of interest aroused considerable irritation from the users of *CA Condensates*, CAS was repeatedly requested by the documentation centers to introduce the section numbers on the Condensates tapes.

To see how useful the section numbers actually would be, DTB decided to examine how the relevant answers to the individual profiles were distributed among the 80 subject sections of CA. As this investigation was accomplished by the time DTB was ready for use of the new CAS Standard Distribution Format, which includes the section numbers, the investigation turned out to be a valuable means to obtain the best possible use of this new facility in the new CAS Standard Distribution Format.

METHODS

For every profile, a base corresponding to approximately 100 relevant answers (based upon output relevance) was selected, and the percentage distribution among the 80 subject sections was calculated for

Total answers
Relevant answers (output relevance)
Answers marked relevant after reading of the article or, if insufficient data regarding final relevance were available, answers considered as highly relevant judged from the output

For all of the profiles, it was illustrated graphically (ex-