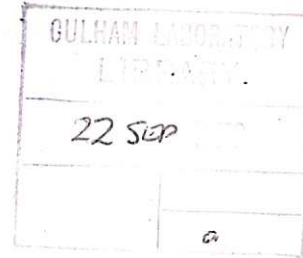


CULHAM LIBRARY  
REFERENCE ONLY



A REAL TIME INTERACTIVE REFERENCE RETRIEVAL SYSTEM

by

A. E. Negus\*

A B S T R A C T

A system for on-line interrogation of files of references stored on a KDF9 disc is being developed at the Culham Laboratory of the U.K.A.E.A. The system is currently available on an experimental basis only, due to limitation of disc space. Searches may be made from any terminal connected to the main computer, and the references printed, or the retrieved references can be shown one at a time on the screen of a DEC Type 338 Precision Display, the user choosing which to print.

Paper presented at the Fourth Conference of the Institute of Information Scientists, Reading, April 11, 1970.

Deputy Librarian, U.K.A.E.A., Culham Laboratory, Abingdon, Berks.  
April, 1970.



## A REAL TIME INTERACTIVE REFERENCE RETRIEVAL SYSTEM

### INTRODUCTION

Since April of 1966 the Library at Culham Laboratory has, on a routine basis, been offering a computerised SDI service to plasma physicists<sup>(1,2)</sup>. During the running of the service a stock of 20,000 references in machine readable form has been built up on magnetic tape and it has long been felt that use could be made of this store of material in an on-line retrieval system<sup>(3)</sup>.

OSTI were also interested in the possibility of on-line searching of this store of references. Unfortunately, due to lack of suitable on-line facilities at Culham at the time very little work was in fact carried out during the period of OSTI interest.

However, since the introduction of the COTAN 3 on-line system<sup>(4)</sup> in the autumn of 1968 it has been possible to experiment with various retrieval methods, and a stage has now been reached where a reasonably satisfactory system, albeit searching only a small number of references, has been developed. This test-bed system was developed with a number of objectives in mind. Since the Autumn of 1969 the main aim has been that it should give the Culham Library staff some real-life experience of the advantages and disadvantages of on-line computing, experience that would be useful in the design of a comprehensive library system for Culham. Subsidiary objectives were that it should give a response in a short time; that it should be simple to use and that it should require the minimum of professional effort in the preparation of the input to the



data bank. This paper concentrates on the practical development of a pilot system - at this early stage no tests have been carried out on "relevance", "recall" and other factors of interest.

#### EARLY ATTEMPTS

Towards the end of 1968 it was felt enough experience had been gained of the on-line system to attempt to retrieve from a sample of the routine input and so a small block of information, containing only 25 references, was copied from magnetic tape on to the disc of the KDF9. A short FORTRAN program was written which could be used in batch mode to retrieve references by matching on two search terms and a synonym for each of these. Although this was a batch job a facility is available at Culham for entering such jobs into a "remote queue" from a remote teletype console. In this queue jobs are given priority according to the length of time estimated for execution and, as this search of 25 references could be completed in about 50 seconds it was possible to obtain a relatively quick turn round under normal circumstances. In practice it was usual to be able to look at the captured references about 5 minutes after entering the query and putting the job into the queue. Although this early program did show the way towards future developments it was in many ways cumbersome in use. For example search terms had to be input to a rigid format and the appearance of the retrieved references was rather unsatisfactory. Nevertheless, it did provide a guide to the future and reactions to this system were generally favourable and encouraging.

#### THE PRESENT SYSTEM

The present system which has been developed from these early beginnings operates in real time and has been so designed that it





would in theory be possible to search all those references which have been accumulated during the running of the SDI service. However, due to limitations on computer storage available it has only been possible so far to run the program on a maximum of 750 references. These are in the field of computational physics and were specially prepared for this exercise.

#### HARDWARE

The program operates on the Culham Laboratory's KDF9 computer running under the EGIION 3 operating system with the COTAN 3 multi-access system. The machine has a core store of 32 K and a backing fixed disc of 4 million words capacity. However, due to the space required by the operating programme for interactive on-line jobs only 14 K of core is available to the user program. This has necessitated the use of many disc transfers in the search program and has consequently slowed it down rather more than one would like. In fact some 75% of the run time of the program is used in carrying out disc transfers. It is hoped that when a similar program is run on the System 4-70 computer now being commissioned at the Laboratory this ratio of run time to disc transfer time will be much more favourable.

Normal on-line access to the computer is via Type 33 teletypes installed at various points around the site, including the library.

The video terminal used is a DEC Type 338 Precision Display. This is intended primarily for the display of graphical information and for the interactive construction or alteration of graphs. It is not intended for text display and consequently has several disadvantages when used for information retrieval. The most serious





of these are that alpha numeric characters must be software generated; that it only has one buffer and consequently there is a pause as each retrieved reference is transferred separately from the KDF9 disc to the display; and that it is only possible to show about 200 characters on the screen without flicker making them almost unreadable. Modern video terminals have none of these major drawbacks and additionally are silent in operation unlike most of today's teletype terminals. Their use should make on-line techniques much more suitable for use for information retrieval, particularly in library environments.

#### SOFTWARE

The program has been written as far as possible in FORTRAN and could be relatively easily transferred to the Culham System 4-70. It relies on a large number of Usercode subroutines for input/output operations and for the use of the display, but on a third generation machine system routines are available for these types of operations. The reason for the choice of FORTRAN rather than any other high level language was simply that the KDF9 at Culham only has FORTRAN and ALGOL compilers available and the original SDI programs, to which this retrieval program is complementary, were written in FORTRAN.

#### DATA STORAGE

The references processed by a sub-set of the SDI programs are stored in several files on the disc. Each file at present holds about 100 references, but their size is only limited by the allocation of space to users. The program could equally well search files containing 1,000 or even 100,000 references.

Each reference is stored in a block of 150 KDF9 words of eight characters each, i.e. 1200 characters. Fig.1 shows an example in



a 10x15 matrix. Words 1-50 contain various code numbers and non-trivial title words truncated to seven characters. The first set of words (from the 8th-25th matrix word positions) is used in the SDI program to aid speed of searching, and its contents depend on the user profiles in use at the time of entering a reference into the system. Only the second full set (from the 26th-48th matrix word positions) is searched by the retrieval program.

Fig.1

Representation of the storage of a reference on disc. The format allows up to 150 KDF9 words of eight characters each.

Although this use of fixed fields for data storage is uneconomical in its utilization of disc space it was felt that the saving that could be made by using variable length fields would be so small in real terms that it was not worth the effort involved in designing a new program for writing the references to disc. Also the use of such a program would have meant that any search program based on this file would not have been applicable to searching those



references stored during the running of the SDI experiment. However, the advantages of variable fields are fully recognised and will certainly be taken into account in the design of the Culham Library programs to be run on the new System 4-70.

#### SEARCH TECHNIQUES EMPLOYED

Search is made on the words occurring in the 26th-48th matrix word positions. These consist of words in the title of the article truncated to 7 terms. Both singular and plural forms of words of less than 7 characters must be input as search terms. There is a short stop list in the SDI program consisting of trivial words such as "on", "and", "by", etc. In addition to this, words beginning with the stems collisi, electro, magneto and plasma are split rather than truncated. This is to allow for the search to differentiate between words such as magnetohydrodynamic, magnetosphere and the like. Hyphenated words are treated as two separate words. The search method employed uses simple Boolean logic; that is OR, AND and NOT links between search terms. Provision is made in the program for 4 groups of terms, each group being linked with AND, to be used and 1 group of exclusion terms. Words within each group of terms are linked by OR logic.

i.e. A1(OR A2 OR ... A9) AND B1(OR B2 OR ... B9) AND C1(OR C2 OR ... C9) AND D1(OR D2 OR ... D9) BUT NOT X1(OR X2 OR ... X9)

All terms other than A1 are optional.

Each group may contain up to 9 words. This may seem a large amount but of course there may be two or three synonyms for each term and each of these may also require a plural to be entered. Naturally the program only searches as many times as is necessary





for the number of search terms that have been input, and, as the title words themselves have been pre-arranged in alphabetical order, search is only made through the list as far as is necessary.

#### MAN-MACHINE INTERFACE

Although it is recognised that this type of system may well be used mainly by information specialists it was felt that it should be designed so that the casual user could operate it with the minimum of external help. The program was therefore designed so that each response required was as simple as possible. Most of the prompts put out by the computer obviously require the reply YES or NO, or they require a part of that prompt in reply. Where a prompt is to be used many times, and therefore brevity is important, the possible replies are given first. The reading of these replies is such that only the first character is recognised, thereby eliminating some errors that could be caused by mis-spelling. Where an incorrect reply is given the machine comes back with the response INVALID REPLY-RETYPE?

There is, of course, one part of the program in which it is not possible to check for spelling and this is in the keying in of the search terms themselves. Here there are some facilities for making corrections but these are slightly complicated and therefore have not been given in the description of the program which is available to users at run time. They are, however, standard COTAN amendment facilities and practised users of the on-line system would automatically use them. The problem has been partially overcome by giving a confirmation of the input query to those users who have asked for it and giving the opportunity to start from scratch again if the query is not what was intended. Of course if an inadvertant mis-spelling is only noticed after search has commenced the query can be





amended, so as to search for the correct term as well.

#### USING THE SYSTEM

(letters in brackets refer to the appropriate section of Fig.2).

The program is available to all users by typing in-SEARCH. LIBRARY, followed by pressing the return button (a). This in fact logs them out as themselves and logs them into the system as the user LIBDEM 2. The first prompt output is ARE YOU USING THE VISUAL DISPLAY? to which the answer is of course either YES or NO (b). This is followed by the prompt DO YOU REQUIRE AN EXPLANATION OF THIS SYSTEM? (c). If the answer is YES either a frame is displayed on the screen (Fig.3) or a few lines are typed out on the teletype (d) giving a very brief description of how the system works and how to use it.

#### Input of Query

The machine then comes back with the simple prompt TYPE WORD? and the user keys in his first search term (e). This is followed by <LINEFEED> or, as with most replies, <RETURN>, which will come back with the same prompt TYPE WORD? The user must then type in his second search term if he wants it to be linked to the first by OR logic (f). Otherwise he types in AND followed by <LINEFEED>, then his next term on the new line. Further search terms are then input in the same way, interspersed with links as necessary. Input of search terms is terminated by typing END on a new line followed by <RETURN>. Where the user has asked for an explanation of the system his query will be confirmed either on the display screen (Fig.4) or by listing on the teletype (g) and he will be given the opportunity of starting again (h). Then search is made of the first file of references, consisting at present of 100 references, and the result given in terms of number of references found and number of references searched (i).



The user is then given the option to look at these (j) and then the option to refine the query, continue with the search as it stands, or start again with a different query (k). This latter policy may be preferable to making a lot of amendments to the query, particularly for the inexperienced user. He does of course at this stage have the option of terminating the search altogether (l).

If he gives the command to continue at this stage a further 7 files of references are searched and again the final score in terms of numbers found and numbers searched will be given. When a teletype is being used all these retrieved references will be listed. It is possible to skip some or all of the output using standard COTAN facilities. If the display is in use the first retrieved reference will be shown on the screen (Fig.5) and brief instructions on how to get the next reference will be given on the teletype. The user may then move through these references one at a time printing them out as he so desires, or he can move through them fairly rapidly without printing any and then, having seen them all, go back to the first retrieved reference looking at them again more slowly and printing out as required (Fig.6).

#### Modification of a query (Fig.7)

In a retrieval system based on natural language for searching provision must be made for a large degree of refinement of the query. The program at present allows for the input of as many new terms within a group or new groups as the program structure itself will allow. However this may not always be enough and although no provision has been made for removing terms from a group it is possible to restart the query with a clean slate after searching the first file. It was felt easier to use this method rather than to



allow for removal of terms as this would have necessitated the use of a more complicated response (perhaps typing in minus and then the word to be removed, typed of course exactly as it had been at the original stage, mis-spellings included). A search term can be partially removed by repeating that term in the exclusion group although this is not the same, of course, as not having it there in the first place.

The procedure for amendment is possibly the most difficult part of the whole search. When the response REFINe is given to the prompt DO YOU WISH TO CONTINUE THE SEARCH, REFINe THE QUERY OR START AGAIN? the program allows for words to be added first to the last group of 'AND' terms to be input. The prompt is ADD TO GROUP BEGINNING (word)? Termination of addition to this group is caused by typing END, and the opportunity is then given of adding to the previous group of original terms, assuming there was one.

If the user wants to add more terms to be linked to the existing ones by AND logic he types AND followed of course by pressing <RETURN>. The machine comes back with the prompt NEW GROUP? and words can then be typed one per line. If amendment to or creation of an exclusion group is wanted the user types NOT followed by <RETURN>. After adding a new group the user gets another chance to amend the existing groups before the first file is searched again (fig.7)

The inexperienced user is of course given brief instructions (fig.8).

#### SPEED OF THE SYSTEM

The speed of machine response to individual components of the man-machine dialogue is almost instantaneous as far as the users are





concerned, the only time delays occurring when the actual search of a file is being carried out; and even here the time between pressing RETURN and the computer response giving the number of references found and searched in the first file is usually of the order of less than a minute. The largest time delay is when references are being printed out; due to the fact that the normal teletype only prints 10 characters per second, most references take of the order of 20 to 30 seconds to be printed. Also, when using the teletype the instructions in the use of the program take about 2 minutes to print, although of course these would not be used by the experienced user. When using the visual display these times are greatly reduced due to the fact that one whole frame is shown at once, usually after a time lag of 2 to 3 seconds after pressing the return button.

Although this system was designed to be used by information officers the system is deliberately constructed so that non-specialists can make use of it. Program components added to cater for the casual users do not detract from the facilities available to the experienced user nor do they make it tedious in use as it is possible to by-pass many of the prompts by supplying the answers first. For example, one may type

```
- SEARCH LIBRARY <LF>
Y                <LF>
N                <LF>
PLASMA          <LF>
PLASMAS         <LF>
END             <LF>
N               <LF>
C               <RET>
```

and the computer will search all the source files and display the



first retrieved reference on the display screen about one minute later. Useability by non-specialists will be investigated later.

#### FURTHER DEVELOPMENT

The present program, using second generation equipment, does show that on-line retrieval systems are technically feasible, and, of course, on a third generation computer with modern video terminals they would give a much speedier search. We have not as yet carried out quantitative tests on the retrieval efficiency of this system, but, using the display with its option to print, what is now thought of as a low level of "relevance" may become acceptable. It should be stressed that this present system is a pilot scheme - the direct access storage available on the KDF9 is too small to allow for a significantly larger data base to be used. However, considerable amounts of disc space will be available on the System 4-70 computer and this will enable further study of the technical feasibility and economics of using reasonably large on-line stores particularly for library use<sup>(5)</sup>.

#### CONCLUSION

To conclude, the pilot system described has met its initial objectives - it has a relatively fast response, it is relatively simple to use, input is simple, and last but not least the practical experience gained has been very valuable indeed.

#### ACKNOWLEDGEMENTS

I would like to thank J. L. Hall, Librarian at Culham for his advice and encouragement while developing this program and the many members of the Computing and Applied Mathematics Group at Culham who have helped with problems in the writing of the program, particularly T. J. Martin who introduced me to the power of visual displays.



## REFERENCES

1. ANTHONY, L.J., CARPENTER, D.H. and CHENEY, A.G. Aslib Proc., Vol.20, No.1, pp.40-64 (Jan.1969).
2. ANTHONY, L.J., CHENEY, A.G. and WHELAN, E.K. Information Storage and Retrieval, Vol.4, No.2, pp.187-200 (June 1968)
3. CHENEY, A.G. Culham Laboratory CLM-LM 1/68
4. POOLE, P.C. Information Processing 68, pp.531-535, Amsterdam, North Holland, 1969.
5. HALL, J.L. and NEGUS, A.E. in preparation



1	2	3	4	5	6	7	8	9	10
BJP	U	U	19-21 A%PRIL. '1967				PCOMPUTE	PLASMAS	THEORY
					A51	COMPUTE	EXPERIM	MICROSC	PLASMAS
THEORY									
								FEIX	WEINSTEI
A51									
---	C%OMPUTER EXPERIMENTS ON THE MICROSCOPIC THEORY OF PLASMAS'. BY								BY
M.R. FZEIX.	G.A. MZASSEL AND	'R.H. W%EINSTEIN'.					IN	NASA SP-153 S%	
YMPOSIUM ON	C%OMPUTER	S%IMULATION OF	P%LASMA	AND	M%ANY	-B%ODY	P%ROBLEMS'.		
A%PRIL. '1967.	WZILLIAMSBURG.	'V%A'.	%PP'.3-24.						

----- indicates positions available for enrichment terms.

' and % are case shift symbols.

Fig. 1

Representation of the storage of a reference on disc. The format allows up to 150 KDF9 words of eight characters each.





```

(a) *** -SEARCH.LIBRARY

(b) //ARE YOU USING THE VISUAL DISPLAY? NO

(c) //WOULD YOU LIKE AN EXPLANATION OF THE USE OF THIS PROGRAM? YES

(d) SEARCH IS MADE ON WORDS IN TITLES TRUNCATED TO SEVEN
CHARACTERS. WORDS CONTAINING THE FOLLOWING STEMS ARE
SPLIT RATHER THAN SHORTENED. E.G.ELECTRON CAN BE FOUND BY
SEARCHING FOR ELECTRO AND V.
COLLISI,ELECTRO,MAGNETO,PLASMAS.
*****

SEARCH TERMS ARE INPUT ONE PER LINE. AND - ON A NEW
LINE - GIVES A NEW GROUP, NOT AN EXCLUSION GROUP
AND END TERMINATES THE INPUT.
TERMS WITHIN A GROUP ARE LINKED BY OR LOGIC.

(e) //TYPE WORD? PLASMA
(f) ? PLASMAS
    ? AND
    ? WAVE
    ? WAVES
    ? OSCILLATIONS
    ? END

(g) TITLES WILL BE MATCHED IF THEY CONTAIN
WORDS IN COMBINATIONS AS FOLLOWS-
PLASMA
OR
PLASMAS
AND

WAVE
OR
WAVES
OR
OSCILLA

IF THIS IS NOT WHAT YOU WANT
TYPE RESTART AND PRESS RETURN.
OTHERWISE PRESS RETURN TO CONTINUE.
(h) //?

      4 REFS FOUND
(i) 100 REFS SEARCHED

(j) //DO YOU WANT TO SEE THESE NOW? NO

(k) /DO YOU WISH TO CONTINUE THE SEARCH, REFINE THE QUERY OR START AGAIN
(l) //? NO

/SEARCH ENDED
***

```

Fig.2  
Entering a query





ENTER: 1. TO ENTER A NEW RECORD IN THE FILE  
TO ENTER: 1. TO ENTER A NEW RECORD IN THE FILE

SEARCH: 1. TO SEARCH FOR A RECORD

DELETE: 1. TO DELETE A RECORD FROM THE FILE  
DELETE: 1. TO DELETE A RECORD FROM THE FILE

ALL RECORDS IN THE FILE ARE DELETED

END: 1. TO END THE PROGRAM

1. TO ENTER A NEW RECORD

END

1. TO SEARCH

1. TO DELETE

1. TO END

1. TO END

END

1. TO ENTER A NEW RECORD

1. TO SEARCH

1. TO DELETE

1. TO END

Fig. 3  
Instructions on the display





[illegible]

5 5

[illegible]

000000

**NAME** \_\_\_\_\_ **DATE** \_\_\_\_\_

IF THIS IS NOT AN INDIVIDUAL RESIDENT AND PRESS REPORT  
OF A FIRE PRESS REPORT TO BE MADE

Fig. 4

Confirmation of a query on the display





ON THE NUMERICAL SOLUTIONS OF THREE-  
DIMENSIONAL BOUNDARY-VALUE PROBLEMS IN ELECTRIC  
THEORY BY THE SUPERPOSITION OF REPRESENTATIONS  
RUSSIAN BY V. P. GATUNOV AND I. N. APPROXIMATE  
METHODS OF SOLVING DIFFERENTIAL EQUATIONS  
IZDAT. AKADEM. NAUK UKRAINI, KIEV, 1960  
PP. 18-25

REFERENCE

Fig. 5

A retrieved reference on the display



10 REFS FOUND  
100 REFS SEARCHED

/DO YOU WANT TO SEE THESE NOW? YES

TO PRINT THE CONTENTS OF THE SCREEN TYPE P BEFORE RETURN  
S BEFORE RETURN SHOWS FIRST REF AGAIN  
C BEFORE RETURN MOVES TO NEXT PART OF SEARCH  
/RETURN?

/RETURN? P

-----  
-----  
DIFFERENCE METHODS ON A DIGITAL COMPUTER  
OR LAPLACIAN BOUNDARY VALUE AND EIGENVALUE  
PROBLEMS. BY G.E. FORSYTHE. IN COMM.  
PURE APPL. MATH., VOL.9, 1956, PP.425-434.  
/RETURN? C

DO YOU WISH TO CONTINUE THE SEARCH, REFINE THE QUERY OR START AGAIN  
/?CONTINUE

Fig. 6

Selecting references from the display.





//CONTINUE, REFINE OR START? HELP

//TO REFINE THE QUERY TYPE R  
TO CONTINUE THE SEARCH TYPE C  
TO START AGAIN TYPE S  
TO TERMINATE PROGRAM TYPE N  
THEN PRESS RETURN  
? R

WHEN REFINING THE QUERY -  
END MOVES TO NEXT EXISTING GROUP  
AND GIVES A NEW GROUP  
NOT GIVES NEW OR MODIFIES EXISTING NOT GROUP

//ADD TO GROUP BEGINNING PLASMA ? FUSION  
? THERMONUCLEAR  
? AND

//NEW GROUP? CULHAM

//ADD TO GROUP BEGINNING CULHAM ? END  
? END

TITLES WILL BE MATCHED IF THEY CONTAIN  
WORDS IN COMBINATIONS AS FOLLOWS-

PLASMA  
OR  
PLASMAS  
OR  
FUSION  
OR  
THERMON  
AND

CULHAM

IF THIS IS NOT WHAT YOU WANT  
TYPE RESTART AND PRESS RETURN.  
OTHERWISE PRESS RETURN TO CONTINUE.

///?

1 REFS FOUND  
46 REFS SEARCHED

//DO YOU WANT TO SEE THESE NOW? YES

-----  
PQ WHETHER MAGNETO-HYDRODYNAMIC POWER WAS CONSIDERED IN  
ASSESSING FUTURE OF PLASMA PHYSICS WORK AT CULHAM LABORATORY.  
HC DEB. 774 C.456W. 3 DEC. 1968

//CONTINUE, REFINE OR START? S

Fig.7.  
Amending a query.





WHEN REFINING THE QUERY  
END OF QUESTION AND EXISTING GROUP  
AND GIVES A NEW GROUP  
NAME GIVES NEW OR OLD NAME USING NEW GROUP

Fig. 8

Instructions for alteration of a query





Report of the discussion on the paper of A.E. Negus

P.S. Davison (Scientific Documentation Centre, Dunfermline) said that searching a hundred references was equivalent to scanning four pages of foolscap, or two contents lists of J.Chem.Soc., which can be done quite quickly. How much does this require in computer time and cost?

Negus: The c.p.u. time taken in searching varies with the number of references found. Typical times are:

Search 100 and find 1 - 1 sec

Search 100 and find 20 - 3 secs.

These times are increased considerably when using the visual display but this is primarily due to the software character generation. The cost of this might be of the order of a shilling. The times involved do not appear to increase linearly, the time taken to search increasing rather more slowly than the size of the data bank.

R. Crompton (I.B.M.) asked if the search was on titles only.

Negus: No - titles can be enriched by adding free language descriptors. However the number of words that can be searched is limited to 22.

Mr Crompton asked the reason for the 7 character limit to word recognition.

Negus: The KDF9 is a word machine, each word containing 8 characters. In the SDI programs one character was required for diagnostic purposes when constructing search profiles, leaving 7 for title words. The potential loss of information is very small in a system based on word stems or truncations.

Mr Crompton asked what kind of queuing system was used.

Negus: The program has high priority but must of course wait behind any other jobs of equal priority. Delays are negligible as the use of this type of program is limited due to the high overheads involved.

F. Liebesney (Able Translations Ltd) asked what happened if the search found no references

Negus: The user is told that none have been found and is then given the usual option to amend the query etc.

V.P. Hardman (I.C.L.) asked whether there was any intention to study the best sort of terminal for use in a system like this

Negus: There are no firm plans to conduct a study, but we naturally have some ideas and intend to look at various machines; primarily videos but also "quiet" fast teleprinters.

