ON-LINE INFORMATION RETRIEVAL: A METHOD OF QUERY FORMULATION USING A VIDEO TERMINAL

J. L. Hall
A. E. Negus
D. J. Dancy

Culham Laboratory
Abingdon Berkshire
1972

ON-LINE INFORMATION RETRIEVAL: A METHOD OF QUERY FORMULATION USING A VIDEO TERMINAL

by

J L Hall A E Negus and D J Dancy
UKAEA Culham Laboratory, Abingdon, Berks. April 1972
(Submitted for publication in "Program")

SUMMARY

This paper discusses some particularly desirable features of the user-computer dialogue involved in query formulation in an on-line information retrieval system. The methods used in the Culham RIOT II system, designed to give on-line access to a data base of 25,000-50,000 references, are described. To facilitate query formulation a matrix technique is employed and the most recent titles may be displayed while the query statement is under construction and before a full search is requested.

1. BACKGROUND

The RIOT (Retrieval of Information by On-line Terminal) project has been under experimental development at Culham since 1968 using a KDF9 computer and various small data bases. This followed earlier mechanization of Culham Library procedures (1) under which routine SDI and the Culham Library Bulletin have been produced via computer on a regular production-line basis since 1966. These routine operations have resulted in archive magnetic tapes containing a very high proportion of the total literature of plasma physics produced since 1965.

One particular aim of the on-line experiments (which from 1971 have been part-supported by an OSTI grant) has been to examine the user-computer interface, especially the "dialogue", one of the prime factors which will influence users in making, or not making, long-term use of an on-line retrieval system.

In general the progress of the RIOT project (1968-1971) has been from teletypewriter interaction in 1968 to somewhat more sophisticated video terminal interaction from 1970 onwards. Some examples of the "dialogues" experimented with during the development have been published (13,14). It has been clear from these experiments that considerable care is needed in

the construction of dialogues and a prototype user-test carried out in December 1971 just prior to the shut-down of the KDF9 provided reactions which have helped to determine the "image" of the user-computer dialogue being developed in early 1972 for the ICL System 4-70 computer which replaced the KDF9 at Culham.

This change to another computer conveniently provided the opportunity to move to a new phase; effectively the design and implementation of RIOT II aimed to be operational on the Culham data base of 25,000-30,000 references by the summer of 1972. It is hoped that RIOT II will:

(a) provide access to the many items in the "recent" literature known to be in the Library, and indeed included in the Culham Library Bulletin, but not easily findable since staff effort is not available to maintain full card indexes to recent journal and report literature. Abstracting and indexing journals are not sufficiently current, particularly in the case of report literature where extensive exchange agreements ensure very prompt receipt at Culham of non-UK reports and preprints.

(b) enable useful "instant" retrospective retrieval to be carried out on the Culham data base which for the users of the Culham Library is a particularly comprehensive and sizable data base.

A full description of RIOT II will be published in due course, as will the results of operational use of the system. This paper, however, discusses only one aspect of RIOT II - the "user-image" with particular emphasis on the query formulation stage.

It is useful to mention briefly some of the factors which have helped to shape the library mechanization programme at Culham. The principal factors have been limited staff availability, the desire to offer improved library and information services, the availability of suitable computing facilities, and a situation in which no single tape service covers the journal, report and book literature sources of interest to Culham staff.

2. USER-COMPUTER DIALOGUE

Many examples of man-machine information retrieval dialogues are available; the following references (2-18) in no way represent a full bibliography but do illustrate that there is no common approach to the problem of constructing a retrieval dialogue. This is hardly surprising bearing in mind the widely differing constraints applying to different systems, e.g. different search logic applied, different data base contents,

and different computer facilities and peripherals available. A point of particular importance is that some on-line systems have been designed on the assumption that the only users will be information specialists (e.g. in a specialised information centre) while other systems aim to be usable by the non-specialist. In some systems it appears that the principal initial aim has been to demonstrate technical feasibility and scant attention has been paid to user dialogue and query construction.

Most on-line systems use some form of Boolean search logic. Often quite complex "nesting" is allowed; this is a perfectly adequate approach for centres in which only information scientists use the system.

However, where self-service use is possible by non-specialists it is perhaps too readily assumed by system designers that the average user will easily master the details of query formulation. This assumption can be questioned on the following grounds:

(a) Many library users put little or no effort into finding out how to use conventional library tools to best advantage so it is not reasonable to assume that they will put the necessary effort into finding out how to make efficient use of an on-line information retrieval system. Most systems in fact offer relatively easy access; the danger is that having learned a few routines the user sticks to these and does not use the full power of all the routines available to him. (This behaviour-pattern is certainly typical of many users' approach to conventional subject indexes, abstracting journals etc. and indeed to general-purpose on-line computer systems.)

(b) The users who presently use on-line systems may in many cases be the more forward-looking, information-conscious users, and perhaps not really representative of the "average" user who may quickly reject "complicated" on-line systems. It is of course commendable to build a system helping the forward-looking users (often comparatively more useful to the organization than the average user) but libraries have a clear duty to serve all users and system design must take account of this.

3. USER IMAGE: DESIRABLE FEATURES OF ON-LINE TERMINAL

Normally a teletypewriter or video terminal is used. A teletypewriter has the advantage of retaining a permanent record of all stages of query formulation. On the other hand there is the noise factor to be considered. More importantly, the formulation of a complex enquiry can prove difficult for the occasional user. A video terminal has the advantage of being

silent; as will be seen later it can be used in a much more flexible fashion in constructing a query while a permanent record of the search formulation can be provided if a slave printer is used. If a video terminal is employed it should have a legible, steady screen image, preferably with upper and lower case.

Culham experience with a number of teletypewriters and video terminals has led to the conclusion that the best approach in the design of RIOT II would be to concentrate on a video terminal with a linked printer. The video terminal used in the RIOT II simulation test reported in this paper is a Beehive Model III (Fig.1) with 20 lines of 80 characters per line, upper and lower case, and with useful additional facilities such as "tab" and "reverse video". The "tab" facility allows the construction and display of a special "protected" matrix and the ability to skip rapidly from "box" to "box" within this matrix. The "reverse video" facility permits display of black-on-white as an alternative to the normal white-on-black, thus allowing additional emphasis where desired; indeed the reverse video block can be made to "blink", a useful method of drawing attention to a user error e.g. an invalid command.

4. USER IMAGE: DESIRABLE FEATURES OF DIALOGUE

In general, a number of particularly desirable features can be identified:

(a) Easy entry ("logging in") to the system.

(b) Speed of response should be as rapid as possible.

(c) The search logic should be appropriate to the user audience. If the users are information scientists skilled in the construction of complex queries, e.g. nested queries, then the logic can allow this. If the users are largely unskilled and unlikely to make frequent use of the system then the designer should preferably try to avoid the need for complex nesting or weighted-term searching, and instead concentrate on providing a comparatively simple, well-explained and unambiguous form of search logic, such as a quite simple Boolean search or an easy-to-understand "quorum search", e.g. "find all references with *X* terms out of the following *Y* terms". If the system is to serve both skilled and unskilled users a compromise search procedure may have to be chosen, or two separate search options provided.

(d) The actual mechanics of constructing a search question should be as simple as possible especially if the system is likely to be used by

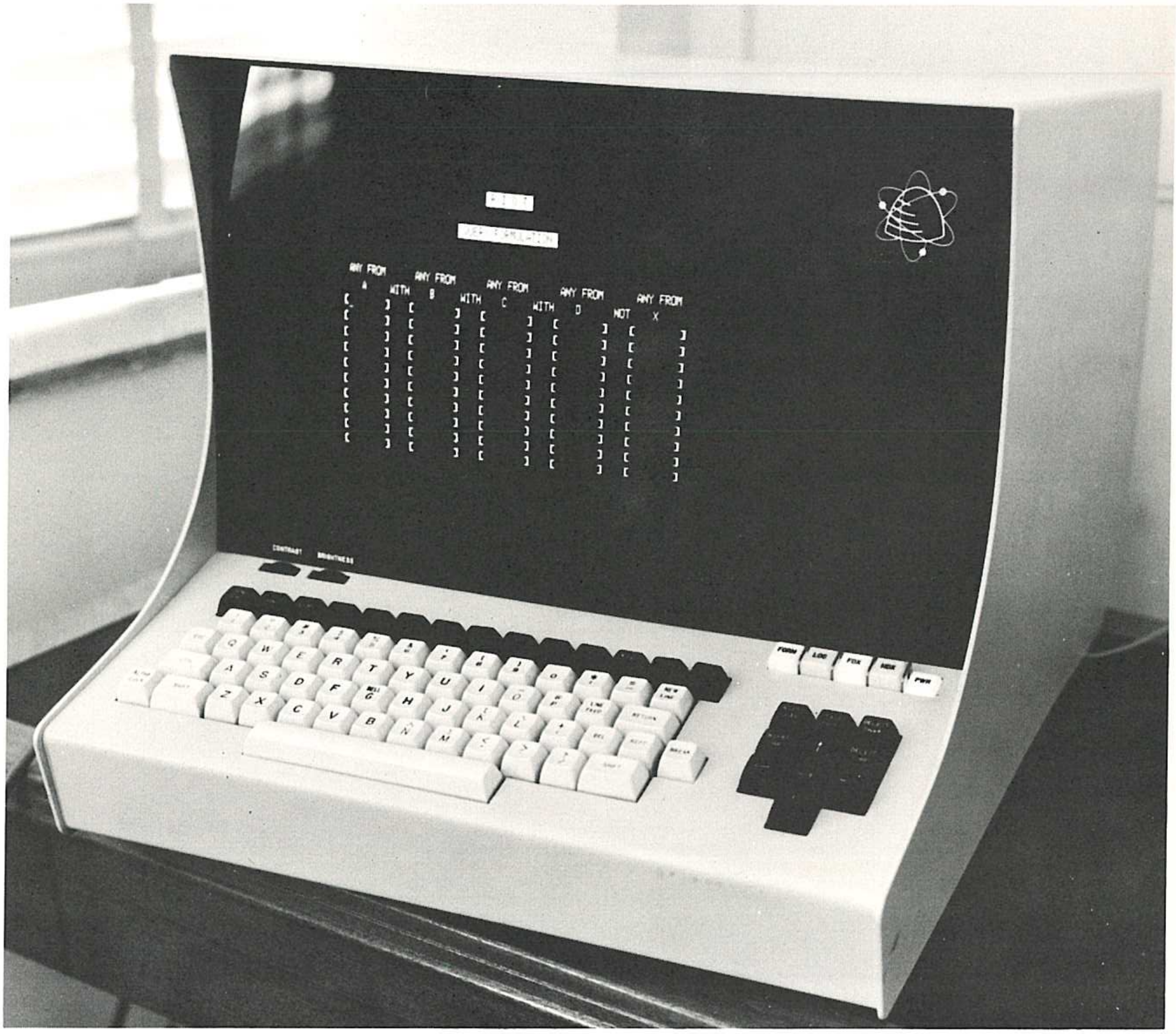


Fig.1 Beehive Model III terminal. (The query formulation shown was an early form of the matrix display technique later developed into the matrix shown in Figs. 3 and 4.)

CR 72 63

self-service users. This applies particularly to amending a search formulation.

(e) It would be helpful to most searchers to be told, virtually instantaneously, how many "captures" can arise from the use of a particular search word, while formulating the full search statement.

(f) In addition, the searcher should be helped by being given an indication of other related words that might be used in constructing a full question profile.

(g) An indication of the total capture expected from any particular combination of search terms would be useful.

(h) As query formulation proceeds, some users would find it useful if entry of a term into the search formulation caused automatic display of the most recent titles in the data base containing that term.

(i) A file of "found" references can be scanned quite quickly on a video terminal and it is of particular advantage if the system can provide a print-out of those references, and only those references, of interest preferably immediately on a linked printer but on a batch basis if immediate output is not possible. (This relieves the searcher of the need to take notes - one of the most tiresome aspects of the conventional literature search.)

(j) The ability to "browse" would be useful.

(k) Where a user regularly puts a complex search to the system, search "macros" could usefully be made available to avoid tedious keying-in (in most systems, of course, SDI will be a more satisfactory way of meeting such a user's needs).

(l) Ease of exit from the system.

It is doubtful if any of the many existing on-line information retrieval systems have all these desirable features. A point to note is the argument (14) that "as systems become more sophisticated they often become more difficult to use. The implication, which is often overlooked, is that inexpert use of sophisticated search capabilities can easily nullify any savings arising from the sophisticated logic and programming techniques employed. But whether the logic is simple or complex the skill of the user in formulating his query will still be vital, as will be the provision of early feedback alerting him to possible errors or shortcomings in the query formulation". Thus the system image at query formulation stage can be seen as a particularly vital part of the total on-line information retrieval process.

It is often found that large systems provide a number of the above

features by basing the system design on the straightforward approach of employing a thesaurus (with up-dating as necessary) together with detailed indexing of input material on a continuing basis. In such systems the thesaurus, sometimes held on-line, provides the necessary basic framework. In some cases the search routine is arranged so that a search always stops to report progress after a proportion of the data base has been searched, thus allowing the searcher to decide whether his query formulation requires revision before proceeding to make a full and possibly (in the case of a large data base) expensive search.

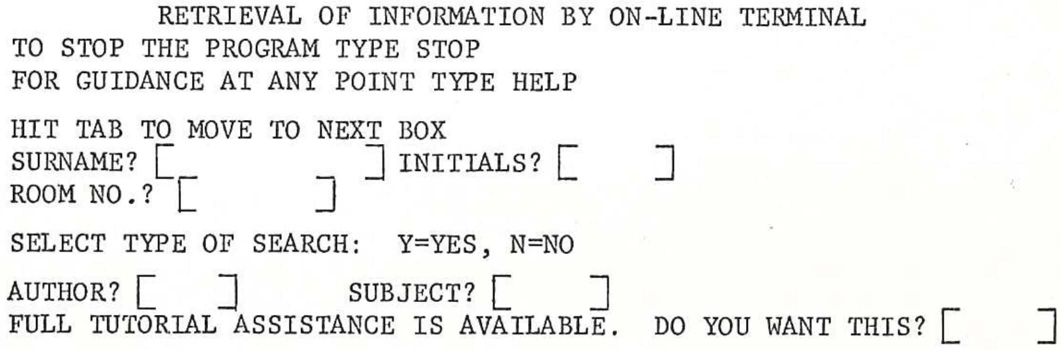
The thesaurus with input-indexing and inverted file storage approach is, of course, a valid line of action, particularly for large units concerned with a very wide spectrum of knowledge or with very large document input quantities. But this approach is often expensive (particularly in staff time), and is possibly over-powerful for smaller units not seeking a general answer to a large scale information retrieval problem but only a simple and robust method of creating and searching a data base of relatively modest size. The basic literature "core" for many units is no more than 50,000-100,000 selected references. In such smaller units the non-core literature will embrace many peripheral disciplines and computer retrieval in these disciplines on an on-site basis will usually be seen as having a low priority; indeed non-core questions may never be handled on an in-house on-line basis but preferably put to an external unit specializing in that particular knowledge area if they cannot be answered using conventional abstracting/indexing tools available in-house.

5. THE CULHAM RIOT APPROACH TO ON-LINE DIALOGUE, PARTICULARLY QUERY FORMULATION

In the case of Culham Library with relatively limited resources, particularly of staff, it is desirable to construct the RIOT system in such a way that (like the Culham SDI system) it can utilize input surrogates based on natural language with optional enrichment, together with computer programs and files requiring minimum maintenance. The system is thus being designed to operate over reasonably long periods of time with a minimum of library and information staff involvement. While RIOT II is not expected to meet fully all the desirable aims outlined in Section 4 above, an attempt has been made to provide optimum facilities accepting the inevitable constraints falling upon a relatively small library and information unit.

The Culham approach in designing RIOT II is to provide a system with moderately powerful search facilities, aimed primarily at use by information and library staff but also capable of being used directly by library users on a self-service basis. Optimum retrieval will probably arise in cases where there is close team-work between a RIOT search officer, operating the system and making suggestions on strategy, and the enquirer, making critical decisions during search formulation (and later, of course, as the actual search proceeds).

The initial screen display contains the details shown in Fig.2.



```
              RETRIEVAL OF INFORMATION BY ON-LINE TERMINAL
TO STOP THE PROGRAM TYPE STOP
FOR GUIDANCE AT ANY POINT TYPE HELP

HIT TAB TO MOVE TO NEXT BOX
SURNAME? [              ] INITIALS? [     ]
ROOM NO.? [        ]

SELECT TYPE OF SEARCH:  Y=YES, N=NO

AUTHOR? [    ]         SUBJECT? [    ]
FULL TUTORIAL ASSISTANCE IS AVAILABLE.  DO YOU WANT THIS? [     ]
```

Fig.2 Initial screen display. The information output to the screen is "protected", the user only being able to type in the areas enclosed by square brackets.

In the case of a subject search, RIOT uses a conventional Boolean combination of search terms linked by AND/OR/NOT. The searcher is allowed initially to use up to 3 AND groups together with 1 NOT group. In each group, OR terms (synonyms and alternatives) are allowable up to a total of 8 terms in each group. Thus the general logic employed (where WITH=AND) is

```
  A    \
WITH    |
  B     |
WITH     >    where up to 7 OR terms are allowed for each of
  C     |     A, B, C, X if desired.
 NOT    |
  X    /
```

The limitation of 3 AND groups, 1 NOT group and 8 terms in each group is purely arbitrary and the searcher can expand his search formulation if necessary by adding additional groups and synonyms.

In practice, two significant user difficulties have been encountered in previous work on the RIOT system:

(a) the difficulty some users experience in appreciating the correct use and consequential effect of the usual logical operators, and

(b) difficulty in amending a query.

Many users, of course, experience no difficulty particularly those who are frequent users. Occasional users can, however, experience considerable difficulty.

For these reasons the version of RIOT II now being designed makes use of a convenient matrix display when a subject search is called for. The standard matrix display is shown in Fig.3. The searcher can "build-up" his query formulation by "tabbing" from box to box or "rubbing out" terms subsequently found to be unwanted; only when he is satisfied with the query formulation does he ask the computer to carry out the full search.

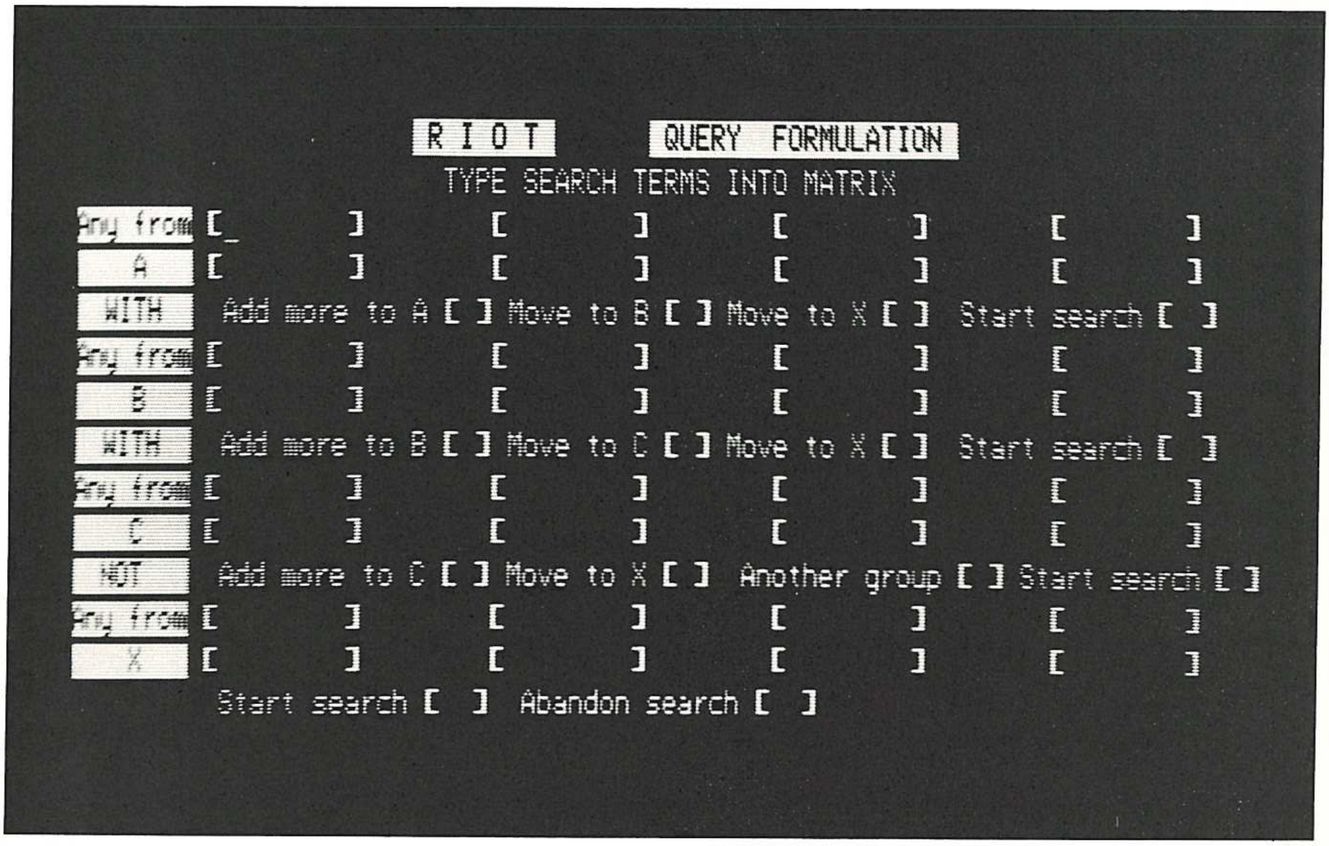
The matrix parameters can be altered dynamically where the user specifies more alternatives or groups than the initial matrix allows, existing search terms being "closed up".

Fig.4 shows a simulation of a search. It can be seen that the number of captures to be expected in respect of each search word is given in an eye-catching reversed video display with the same number also being entered at an appropriate part in the matrix. This does not imply that there is a full dictionary of all possible words - instead, the design allows an on-line dictionary of "major" distinctive words* and word stems which it is believed will cover a considerable proportion of the words entered in real-life searches using this particular data base; where a search word is not in the "major" word dictionary an appropriate message will be displayed in the reversed video block, e.g. NOT IN MAJOR WORD FILE, TRY FULL SEARCH.

An additional feature to note in Fig.4 is the display of the most recent input items, achieved for "major" dictionary words by keeping an associated on-line record of the last five references to use each major word. Entry of a major word into the search matrix not only causes the number of occurrences to be displayed in the reversed image block but also presents the user with brief title details in the lines shown numbered 1-5 (shown blank in this simulation). It is expected, in the particular situation prevailing at Culham, that sometimes one or more of these will, without further search, be valuable to the user and the user will be provided with a facility for instant examination and print-out of any of these items, after which the search can be resumed if desired.

It is also reasonable to hope that the display of recent titles will help the user to choose "related" words. In larger systems this is, of course, often achieved by thesaurus techniques, necessary because the amount of literature being processed is so very great. The RIOT system, with

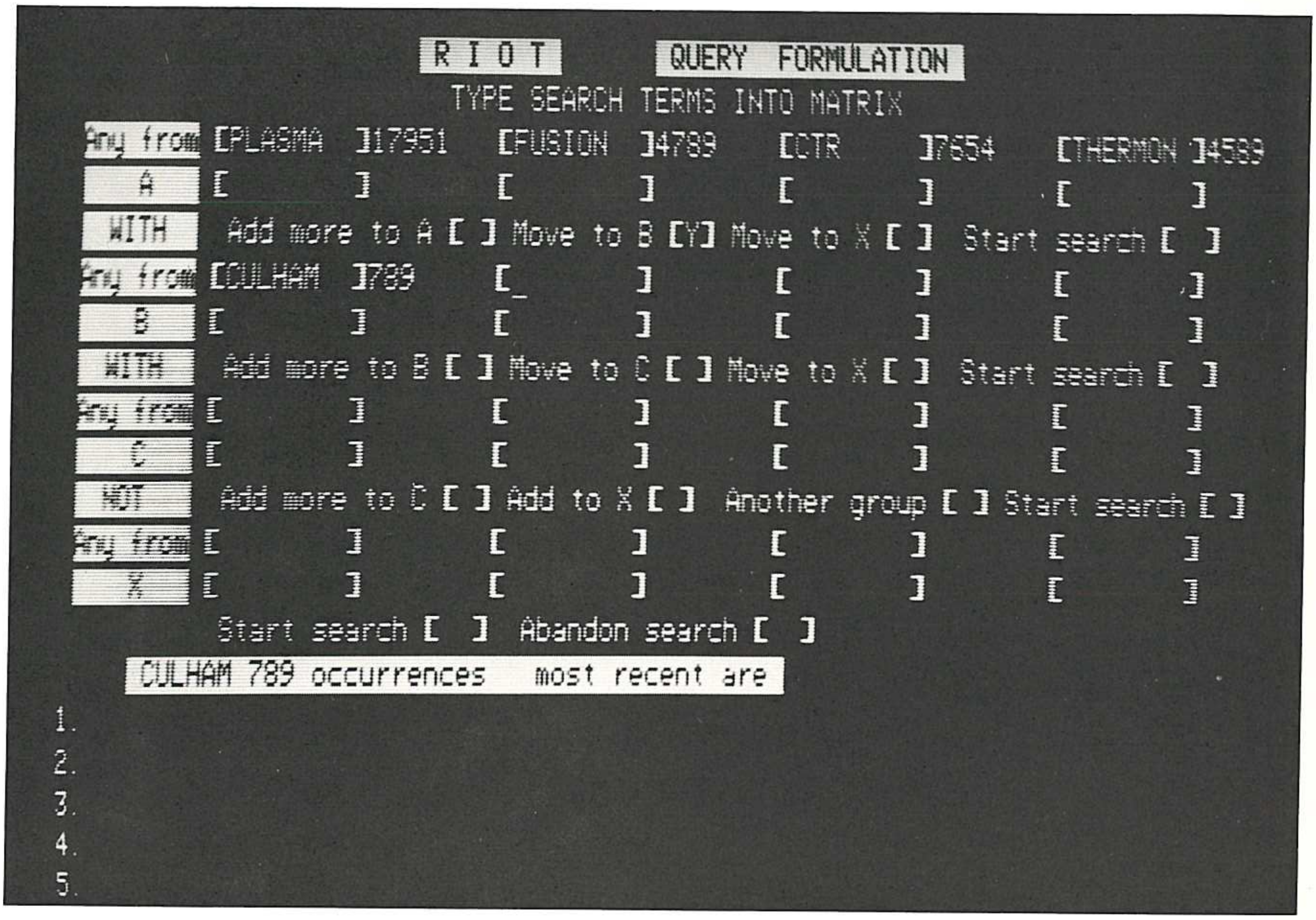
*i.e. major words which are distinctive; some major words like "Plasma" and "Laser" occur so frequently as to be non-distinctive. Statistical study of the present data base indicates that the major word dictionary need only contain about 800-1,000 words.



```
                    R I O T      QUERY  FORMULATION
                    TYPE SEARCH TERMS INTO MATRIX
Any from [_       ]     [       ]      [       ]      [       ]
    A    [        ]     [       ]      [       ]      [       ]
  WITH    Add more to A [ ] Move to B [ ] Move to X [ ]  Start search [ ]
Any from [        ]     [       ]      [       ]      [       ]
    B    [        ]     [       ]      [       ]      [       ]
  WITH    Add more to B [ ] Move to C [ ] Move to X [ ]  Start search [ ]
Any from [        ]     [       ]      [       ]      [       ]
    C    [        ]     [       ]      [       ]      [       ]
  NOT     Add more to C [ ] Move to X [ ]  Another group [ ] Start search [ ]
Any from [        ]     [       ]      [       ]      [       ]
    X    [        ]     [       ]      [       ]      [       ]
          Start search [  ]  Abandon search [  ]
```

Fig.3 Basic "query formulation" matrix displayed to the searcher on asking for a subject search. The searcher can move easily from "box" [ ] to "box" simply by using the "tab" key, the matrix itself being "protected data".

CR 72 63



```
                         R I O T          QUERY  FORMULATION
                              TYPE SEARCH TERMS INTO MATRIX
  Any from [PLASMA   ]17951   [FUSION  ]4789   [CTR      ]7654   [THERMON ]4589
     A     [         ]        [        ]       [         ]       [        ]
   WITH     Add more to A [ ] Move to B [Y] Move to X [ ]  Start search [ ]
  Any from [CULHAM   ]789     [_       ]       [         ]       [        ]
     B     [         ]        [        ]       [         ]       [        ]
   WITH     Add more to B [ ] Move to C [ ] Move to X [ ]  Start search [ ]
  Any from [         ]        [        ]       [         ]       [        ]
     C     [         ]        [        ]       [         ]       [        ]
   NOT      Add more to C [ ] Add to X [ ]  Another group [ ] Start search [ ]
  Any from [         ]        [        ]       [         ]       [        ]
     X     [         ]        [        ]       [         ]       [        ]
            Start search [ ]  Abandon search [ ]
      CULHAM 789 occurrences   most recent are
1.
2.
3.
4.
5.
```

Fig.4 Simulation of search. Note that when the searcher entered the search word CULHAM, the number of items in the data base containing this word was "posted" within the matrix and also brought to his attention in the reversed video display (recent titles are not shown in this particular simulation but will be displayed automatically in the finally developed version of RIOT II).

CR 72 63

limited staff effort, relies on (a) the native wit of the searcher, (b) the fact that the data base is comparatively modest in size, and (c) the hope that the display of "recent titles" as provided for in Fig.4 will help to trigger off a certain number of related words in the user's mind.

While it would be useful to provide an indication, before search, of the total capture likely to arise from combination of terms, this is not available in the RIOT II system (indeed most systems apparently providing this facility do so after having in fact carried out a full search). Whether "expected capture" is a matter of real significance in searching small data bases is open to question; it may be economically unjustifiable to provide this facility in systems such as RIOT where the cost of computer time used in actual search is expected to be less than £1.

6. CONCLUSIONS

After discussing the desirable features involved in on-line query formulation, this paper outlines a method of Boolean query formulation using a matrix display technique. A noteworthy feature is the display during query formulation of a "feedback" facility which in the case of a "major" word advises the user how many references are likely to be found and also the most recent titles containing the last word added to the search profile. It is expected that the essential simplicity of this approach will be of particular value in the case of data bases of modest size where the expense of constructing and maintaining a thesaurus and inverted files may not be economically justifiable.

REFERENCES

1. ANTHONY, L.J., CHENEY, A.G. and WHELAN, E.K. Some experiments in the selective dissemination of information in the field of plasma physics. Information Storage and Retrieval, Vol.4, No.2, p.187-200, June 1968.

2. BURNAUGH, H.P. The BOLD (Bibliographic On-Line Display) system. In: "Information Retrieval"; Ed. SCHECTER, G., Academic Press, p.53-66, 1967.

3. CARVILLE, M., HIGGINS, L.D. and SMITH, F.J. Interactive reference retrieval in large files. Information Storage and Retrieval, Vol.7, No.5, p.205-210, December 1971.

4. COOK, K. An experimental on-line retrieval system for Psychological Abstracts. In: Proceedings American Society Information Science, Vol.7, Washington, ASIS, p.111-114, 1970.

5. FREEMAN, R.R. AUDACIOUS - An experiment with an on-line interactive reference retrieval system using the Universal Decimal Classification as the index language in the field of nuclear science. American Institute of Physics, April 1968. PB 178 374.

6. HARLEY, A.J. Dialogue with a computer. NLL Review, Vol.1, No.4, p.123-136, October 1971.

7. HERBERT, E. Information transfer. International Science and Technology, No.51, p.26-37, March 1966.

8. HIGGINS, L.D. and SMITH, F.J. On-line subject indexing and retrieval. Program, Vol.3, Nos.3,4, p.147-156, November 1969.

9. ISOTTA, N.E.C. Europe's first information retrieval network. ESRO/ELDO Bulletin, No.9, p.9-17, April 1970.

10. Massachusetts Institute of Technology. Project INTREX. Semi-Annual Activity Report, PR-6. p.14 et seq. September 1968. (Some additional useful information is contained in PR-11. p.40-43. March 1971.)

11. MATHEWS, W.D. The TIP retrieval system at MIT. In: "Information Retrieval"; Ed. SCHECTER, G., Academic Press, p.95-108, 1967.

12. McCARN, D.B. Networks with emphasis on planning an on-line bibliographic access system. Information Storage and Retrieval, Vol.7, No.6, p.271-279, December 1971.

13. NEGUS, A.E. A real time interactive reference retrieval system. The Information Scientist, Vol.5, p.29-44, March 1971.

14. NEGUS, A.E. and HALL, J.L. Towards an effective on-line reference retrieval system. Information Storage and Retrieval, Vol.7, No.6, p.249-270, December 1971.

15. RUBINOFF, M., BERGMAN, S., FRANKS, W. and RUBINOFF, E.R. Experimental evaluation of information retrieval through a teletypewriter. Communications of the ACM, Vol.11, No.9, p.598-604, September 1968.

16. Stanford University. Institute for Communication Research. SPIRES 1969-70 Annual Report. p.124-129. June 1970.

17. THOMPSON, D.A. Interface design for an interactive information retrieval system: a literature survey and a research system description. Journal of the American Society for Information Science, Vol.22, No.6, p.361-373, November-December 1971.

18. THOMPSON, G.K. Some cost estimates for bibliographic searching in a large-scale social sciences information system. Information Storage and Retrieval, Vol.6, No.2, p.179-186, June 1970.