A user interface for simultaneous multi database searches

Werner Robert Svoboda

SUMMARY: 1. *Introduction.* – 2. *Development and use of internal databases by intermediary services.* – 3. *Retrieval from multiple external databases.* – 4. *Other functions.*

1. INTRODUCTION

In the years 1981 and 1982 the German Society for Information and Documentation (Gesellschaft für Information und Dokumentation) conducted a project with the goal of developing a software package for the support of intermediary user services.

Although this project did not deal specifically with the legal field, its results are also of interest for this area, for intermediary services are gaining in importance in the legal field as well (in many countries they are the largest clients of available systems). The role of such services as intermediary among various national and/or international systems is also steadily growing.

In the empirical investigations carried out during the project two distinct focal points among possible support functions crystallized:

— the function cluster "development and use of services' own databases" and
— the function cluster "retrieval and processing of search results from multiple external databases" with the functions:
• referral as internal orientation aid for searching
• automated LOGON procedures
• standardized documentation and retrieval language
• merging and sorting of retrieved references from various databases, including the elimination of redundancies.

These two clusters of functions do not necessarily represent alternatives; they overlap to various degrees at the different intermediary services. Both clusters are supplemented by a number of other functions, e.g. internal administrative functions, support during document delivery, training, etc. Numerous software packages have been developed for the functional cluster "developmente and use of internal databases" – also for mini and micro-computers – including very comfortable input, search and output facilities as well as new approaches such as "relational database structure" or "hardware solutions to search processes".

As of yet there exist no operational software packages for the entire complex "retrieval and processing of search results from multiple external databases", although past funding and the currente infrastructure (e.g. DIANE) concentrate on the use of external databases by intermediary services. User aids for multi database search are therefore of particular importance.

The specifications developed in this project for multiple database retrieval could form the basis for the implementation of a virtual retrieval system based on the Common Command Language for search in networked heterogeneous databanks. The virtual retrieval system would be developed by means of a translator interface in the search computer, which would allow standard and user friendly access to various databanks using various retrieval systems.

2. Development and use of internal databases by intermediary services

Intermediary services are not always satisfied with the information supplied by large databank services. The available central databanks are often criticized:

— information supplied by the databanks is too general for the needs of users;
— the quality of the presentation of references could be improved;
— the retrieval of references alone is of little use if it is difficult or even impossible to obtain the originals of the retrieved documents.

The use of small-scale computers offers intermediary services the possibility of storing and retrieving from their own online data collections. The adavantages are:

— optimal adjustment of scope and depth of evaluation to needs during inclusion of information in the databank
— systematic storage and availability of infomation arising from the intermediaries' own wealth of experience, gathered in the course of his work
— cessation of difficulties arising from links to external mainframe computers (dependency, changes in system, low priority, implementation of programs for internal use, etc.)
— relatively low investments for hardware and software
— no operating costs for telecommunication and computer time
— constant availability.

3. Retrieval from multiple external databases

The investigation indicated a distinct need for all functions wich would facilitate multi database search, especially in light of the current costs incurred by complicated databank switching. These functions include above all:

— referral as internal orientation aid during search

— automatic LOGON (due to the large number of line disturbances, high telecommunications costs and as support during multiple searches)
— standardized documentation and command language (because of ideally simultaneous multiple retrieval)
— merging and sorting as well as elimination of redundancies (due to overlapping databanks during multiple searches).

3.1 *Referral as internal orientation aid and as service*

Access to instruments which lead to information is of twofold importance for intermediary services:

— as internal support for conducting searches
— as an additional service which increases the total value of the intermediary service.

This means that such referral aids must, on the one hand, provide the intermediary with tips as to which sources he should use for a particular search and how these sources are to be used. On the other hand, they should allow the intermediary to refer users to alternative or supplemental sources of information, in cases in which he himself cannot satisfy a client's information needs or in order to offer the client extended information opportunities. However, a complete referral instrument should not only contain information on online retrievable databases; it should also include a broad spectrum of information sources, e.g. offline information services, experts and informal sources. This means that such referral aids will have to be based on the experiences and contacts of each individual intermediary service and therefore must be developed and implemented there as well.

Different forms of referral instruments are conceivable:

— printed referral aids (e.g. database guides or availability lists for ordering literature),
— referral databanks with retrieval possibility,
— referral tools which are tailored to the needs of the individual intermediary service and which are only available there.

Database guides which are manufactured by the producer on magnetic tape could also be used on the mini- or microcomputer of an intermediary service. The referral process takes place in roughly two steps, i.e. answering the questions

— in which databases (of those available) there is theoretically a chance of finding relevant information, and
— in which databases search should actually be carried out, because chances of finding relevant information are good (because of the expense, serach cannot be conducted in all possibly appropriate databases, even if the necessary links existed).

Database guides are generally a good means of answering the first question; however, in general a number of different database guides must be consulted, since they often cover a geographical area or a network (e.g. the ODIN-Datanbankführer, EURONET DIANE Guide), and not a particular subject field. Moreover, there is always the basic choice between thorough database guides, which are out of date, and up-to-date ones, which offer only superficial information.

For the answering of substantial questions, database guides are often considered inadequate, since the descriptions of their content is either too general, inexact or not clearly presented; in addition, they are often not up-to-date enough.

3.2 *Automatic LOGON procedures*

Automated Logon procedures save time and money. Especially during the use of international information networks, it is advantageous when user identification, password, address of computer, etc. can be entered by pressing only a few keys, since very long strings are characteristic of precisely such systems.

Automatic Logon is also important when a thinking break (or other short pause) is necessary during a session. The ability to reenter the system quickly by means of automatic logon encourages more frequent logoffs and thus saves money.

A further argument for automated logon procedures is that they provide a more reliable storage place for addresses, identifications and passwords than do paper for this information.

The economization of connect time achieved by automatic logon procedures and interactive preformulation of search queries before dialing the host are estimated in the literature at 35-70% of total connect time depending on the conditions (especially on average connect time during search, whether the user is experienced or not, etc.).

3.3 *Standardized documentation and command language*

The creation of national and internatioinal network affiliations in the last few years was an important requirement for access to domestic and foreign databases, i.e. these databases are now reachable at a reasonable cost (cost being mostly dependent on amount of data and not on distance). This created three problem areas with regard to language, caused by

— various natural languages
— various documentation languages
— various command languages.

Only the difficulties involving differing documentation and command languages shall be discussed here; the possibility of automatic natural language translation will not be treated.

3.3.1 *Various documentation languages*

There are two possible approaches to achieving a standardized (from the user's point of view) documentation language for various differently structured databases:

— a more formal approach, in which the attempt is made to make the various (also at different locations) databases appear to the user as one virtual database, so that all of them can be searched through with a single query (and one command);
— a more content-oriented approach, in which the attempt is made to generate automatically the optimal search terms for each database after a query is entered.

In practice the attempt is being made to implement one of the two approaches and to take the other as much as possible into consideration. The first approach's (virtual databank) major problem is to simulate a databank (also considering different networks), whereas the advantages are only apparent when a meta command language exists (or is being developed).

The main problem to be solved in the second approach is the automatic conversion of input search terms for a particular database (subject switching). Although freetext search offers certain - though limited and based on coincidence - possibilities, (one must be confident that the chosen search term is allowed/exists in all databases) other attempts have been made to find a better solution to the problem:

— development of word concordancies (i.e. a neutral dictionary which performs translations from the vocabulary of one database into that of another),
— clustering
— word mapping.

To date there is probably no really operational system which offers a convincing solution to the problem of subject switching, so that for the project the first approach - the formal approach - was chosen.

3.3.2 *Various command languages*

The different host computers operating under the auspices of Euronet-DIANE currently use more than 10 command languages (and this will continue to be the case in the near future); in addition there are the two large US online services with their languages DIALOG and ORBIT as well as numerous other national database services with their own systems. Although all these command languages are, in principle, similar, they vary greatly in detail and the designations of individual commands are very confusing (e.g. DISPLAY in STAIRS means something very different in DIRS, and in GOLEI there are three commands for this function).

These differences are not objectively justifiable and lead in practice to limited access possibilities, since a searcher cannot master and keep up practice in more than two languages (especially considering their similarity in conjunction with their illogical dissimilarities).

There are four approaches to overcoming this problem:

a) a standard interface to a virtual retrieval system which would guarantee the smallest common denominator among the retrieval systems in question,
b) the translation of one command language into another (analogous to the situation with programming languages),
c) an attempt to form a standard by developing and implementing a minimum or maximum catalogue (existing systems would then either represent extensions to or a subsystem of the standard),
d) the monopoly of one system, which would replace all currently existing systems.

Attempts to realize the third and fourth approaches were outside the scope and objectives of the project. In the project the first approach was implemented.

3.4 *Merging and sorting retrieved references from different databases, including the elimination of redundancies*

In principle, merging and sorting references retrieved from various databases is a simple sort function, which requires intermediate storage of output with a subsequent comparison of document references. The sorting and elimination of redundancies would have to be performed by the comparison of various fields or by the development of an internal duplication check code. However, in detail this sorting function does pose some problems when the compared document references are not according to exactly the same formal rules, which is, as a rule, the case with references from various databases.

The document references found during a search arrive at the searching computer in an unstructured (i.e. structured in such a way that the destination computer cannot necessarily recognize the structure) form, so that the comparison of certain categories of a reference from one database with the corresponding categories of references from another database is not easily accomplished. There are two conceivable solutions:

— either the categories selected for comparison and sorting (e.g. author: 10 characters, title: 50 characters, date of pubication) are retrieved again individually (inernally, i.e. the user is unaware of it) from the central databanks when a merge redundancy elimination command is issued, or they are routinely recalled and intermediately stored upon issuance of a display command (e.g. by setting a switch at the outset of the dialogue), so that the content of the categories is available in an intermediate store and can be identified by the searching computer,

— or the needed category contents are subsequently identified by the searching computer for sorting and comparison (provided the documentation languages and formal indexing rules allow this), e.g. by seeking out the looked for categories by means of category identifications or other identifying characteristics communicated to the searching computer.

Both approaches require that the structure of the document references are known, i.e. above all that such a function is only usable for databases already decided upon prior to implementation (therefore it must be easy to include new databases). Moreover it would have to be decided which database had priority, i.e. that is from which one the redundant references should not be removed.

Since double references are not necessarily redundant, because different indexing practices may mean that different references to one and the same document may contain different information (e.g. an abstract, many more or different descriptors, etc.), this aspect would also have to be considered when choosing the databases for which this function is to be implemented.

Since such a function would probably require a large amount of development work (above all due to the necessary generality of the software and the intellectual effort needed to implement the function for concrete databases), the intellectual elimination of double references (and the elimination of irrelevant search results) at the terminal for printouts should definitely be particularly comfortable.

The implementation of the function could be accomplished by means of an extended print command, within which the references to be printed (or deleted) could be specified. It would be useful to allow the user to store the desired information in a user file by means of selection commands in order to then print it. Following the selection process he would then have a further opportunity to check the selected information for printing and could modify (or resort) it before the actual print command is given.

4. Other functions

4.1 *User training*

One of the many methods, materials and techniques for the education and further training of potential system users are computer-supported programmed instructions and training databanks with varying didactic detail.

The advantages of such programs include:

— help in overcoming the bottleneck caused by training new users
— practicing with training programs is considerably cheaper than using large databanks for training
— they are not as dependent on time and place as are courses
— they spare the user exposing his lack of knowledge and awkwardness in front of a large group.

The disadvantages, on the other hand, include the primitive character of most training programs and the reluctance of users with no computer experience to use computers for learning to use computer aids.

There exists very little empirical information on the use of training programs; the use of practice databanks is generally advocated, above all out of cost considerations.

4.2 *User guiding during search*

Most of the currently used retrieval and documentation languages are complicated to use for unpracticed or infrequent users. They all have their own often very formal or mnemonically difficult syntax (the necessity to place things in a certain, for the user non-intuitive, order; the significance of special characters such as blanks, commas, etc., which have no meaning for the user, etc.), and react to false or incomplete inputs invariably with error messages (which generally only give a rough idea of what was done wrong, but not what the correct input should look like).

The correct usage of a retrieval language and the utilization of all search aids require good training, long experience and constant practice and further training. These prerequisites are nearly prohibitive for new users or those who only occasionally search: it is hardly possible for such users to achieve relatively satisfactory results with their rudimentary knowledge. In a series of investigations this situation has been empirically substantiated by showing that the majority of users only use the simplest retrieval possibilities of a given system and, in addition, that they always proceed in the same rigid pattern. This occasionally goes so far that users refer again and again to a search example and use the sample search queries as a parameter for their own search (e.g. by simply exchanging search terms, but leaving everything else, especially the syntax, unchanged).

Therefore, aside from the necessary familiarity with the content and philosophy of the different systems, for the effective use of retrieval and documentation languages, instruments are needed which would also allow the occasional user to search in available information systems to his own and to the satisfaction of his goals.

There are above all three (mutually complementary) means of accomplishing this, which are usually regarded as belonging to the area of artificial intelligence:

— allowing input to be as close as possible to natural language, i.e. unformatted
— "intelligent" system reactions to syntactically false input
— in the case of poorly selected search strategies, substantial proposals from the system to optimize search queries.

The problems, involved in natural language input and the state of the art in research and development in this field cannot be intimated here; a thorough theoretical solution to the problems is not to be expected in the near future.

4.3 *Obtaining original literature*

The online ordering of literature following information retrieval is still a generally unsolved problem. As can be seen from the results of various investigations, most intermediary users consider automated support for acquiring documents in addition to the retrieval of sources very important.

Online ordering, or the automation of certain steps during ordering, requires cooperation between intermediary services, host organizations and document suppliers. This cooperation can take different forms. For the large part, existing online ordering systems are run by databank services. During the search process, the searcher can call the ordering program and store desired titles with additional information about the recipient, etc., online in a "message file" (or "parking file" or "electronic mailbox (maildrop)"). The orders are then called up by the addressee whenever he wants and executed (or printed out, e.g. as ordering forms with subsequent distribution to addressees by mail, e.g. Dial Order, SDC Electronic Mail Drop).

One of the disadvantages of such systems is regarded to be the coupling of the search and ordering processes. The end user has no way of determining the relevance of the retrieved references or to choose among them. It also has not been conclusively shown that online literature ordering significantly reduces delivery time.

In cases in which the online databank service offers no literature, it should be possible for the intermediary service to determine the owner (or supplier) and the location of a document. To accomplish this, one of the following is necessary:

— either the literature database itself provides the information (e.g. by the inclusion of corresponding fields in the records, which, however, is only possible if the database producer performs the service or the host organization does so during databank implementation), or
— a separately operating catalogue database is made available, which can be

searched in online and which provides information on location, suppliers, etc. (in addition, information concerning lending conditions, fees, prices, etc. could be offered).

In conclusion it may be said that the real value of online literature ordering lies in the provision of the name of the supplier of the desired document; a reduction of delivery time for orders is comparatively less important; delivery time can only be optimized by the suppliers themselves. The communication of fulltexts online, which could provide a considerable improvement, does not appear possible on a large scale either now or in the near future, due to legal, organizational and technical (communication costs, rather slow transimission rates) problems.