OCCASION

This publication has been made available to the public on the occasion of the 50th anniversary of the United Nations Industrial Development Organisation.

DISCLAIMER

FAIR USE POLICY

CONTACT

Please contact publications@unido.org for further information concerning UNIDO publications.

For more information about UNIDO, please visit us at www.unido.org

UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION
Vienna International Centre, P.O. Box 300, 1400 Vienna, Austria
Tel: (+43-1) 26026-0 · www.unido.org · unido@unido.org

FID
Fédération Internationale d'Information et de Documentation
Founded September 1895

FEDERATION INTERNATIONALE D'INFORMATION ET DE DOCUMENTATION (FID)
International Federation for Information and Documentation
P.O. Box 90402, 2509 LK The Hague, Netherlands
Tel. (070) 140671, Telex: 34402 KB GV NL Attn. FID

UNITED NATIONS
INDUSTRIAL DEVELOPMENT
ORGANIZATION

INTERNATIONAL
FEDERATION FOR
INFORMATION AND
DOCUMENTATION

16759

SELECTING THE MOST APPROPRIATE DATABASES TO ANSWER INDUSTRIAL INFORMATION REQUESTS

submitted by

International Federation for Information and Documentation (FID)
General Secretariat
P.O. Box 90402
2509 LK The Hague, Netherlands

Stella Keenan, Secretary General

THE HAGUE/VIENNA
1988

ACKNOWLEDGEMENTS

This report was prepared for the United Nations Industrial Development Organization (UNIDO) under contract with the International Federation for Information and Documentation (FID).

The report was compiled by an international team of experts consisting of:

Mrs. E. El-Shooky (Egypt),
Mr. F. Gibb (United Kingdom),
Mr. P. Robosz (Hungary), and,
Mr. P. Vásárhelyi (Hungary), who acted as overall supervisor and editor of the report.
Mr. Ch.L. Citroen (Netherlands) acted as assessor to the project.

Stella Keenan
Secretary General FID

CONTENTS

PART III: ACTION RECOMMENDED

INTRODUCTION

Online searching is a tool of information retrieval which is accessible to an increasing number of developing countries. The number of databases which can be used in order to find an appropriate answer to an industrial information request can reach several dozens. Users of databases face, therefore, an important problem; how to find among the numerous potentially helpful databases those providing the most relevant answer, at the lowest possible cost. This study is aiming at identifying the characteristics which may differ from database to database and from one service provider to the other. It is intended to provide a basis for collecting appropriate data with a view to elaborating guidelines for use by developing countries when they want to select the most appropriate databases and hosts in specific fields of industrial development covered by Unido.

In order to select the most appropriate database in the case of a specific question, rather than in a broader (although limited) subject field, a great number of data should be known on each potentially useful databases and hosts. It seems, therefore, necessary to store these data in a computer and to develop gradually a knowledge based or expert system which can facilitate the identification of the database(s) to be searched in order to provide the best service at the lowest cost to a user asking a specific industrial question.

The International Federation for Information and Documentation (FID) has been subcontracted to prepare this study using international input; the team established for this purpose presents findings and recommendations on the experience gained in the field of online searching in both developing and industrialised countries in different regions.

PART I

CRITERIA FOR THE SELECTION OF THE MOST APPROPRIATE DATABASES AND HOSTS

1. ONLINE SEARCHING

Online searching has become a major tool of information retrieval in the developed industrial countries. The superiority of cost effectiveness of online searching has been proved beyond doubt in comparison with traditional manual literature searching. Scientists, medical doctors, engineers and managers involved in research and development, but also businessmen, executives, planners, teachers, economists, government officials and numerous other professionals in academia, industry and business within the reach of a computer terminal now look at online searching as their principal way of retrieving literature references, patent references or even technical, business and other data.

This is the reason why online information systems have grown in the past decades to become and industry itself. In the USA only, the online industry has been growing at least by 18% yearly, the total revenues have increased from 469 million $ in 1978 to 2.2 billion $ in 1986; the revenues in 1990 are expected to reach 4.3 billion $. In Europe, the revenues of 1.2 billion $ in 1986 can reach a figure of 4 billion $ by 1990. The number of online accessible, publicly available databases has grown from 400 in 1980 to about 550 in 1987. The approximate number of online accessible data records is 1.68 billion.

It is out of the scope of this report to introduce the reader to the basics of online searching. For beginners who wish to gain some fundamental knowledge of online information systems and retrieval, reading of textbooks is recommended, e.g. that of [1], an excellent work for beginners. It is intended here to concentrate on the database and service side of the industry, to define criteria how to make preparations for online searching and, especially, how to select the most appropriate databases for given information retrieval problems in a multiple database - multiple host environment.

The discussions will be restricted principally to scientific and technical fields with the purpose to help R&D people engaged in any branch of science. Principally, bibliographic type searches will be considered. The databases discussed here are all publicly available databases, private databases are excluded. Also, most of the databases can be accessed by anybody online.

2. TYPES OF DATABASES

The 3500 online accessible public databases available today for information retrieval can be categorized in the easiest manner according to the type of information they provide.

2.1 Bibliographic/reference databases

As their name indicates, these databases refer to source documents containing - with certain probability- the final information

needed. The output of a search in this type of database is a bibliography, in most cases, complete with abstracts, of relevant documents which must be obtained and read to acquire the final information. Therefore, these databases are usually called bibliographic databases. It is the most used type of database in online searching, especially by scientists and engineers who look for literature and/or patents for R&D purposes. Examples of bibliographic databases are: Chemical Abstracts, INSPEC, COMPENDEX etc.

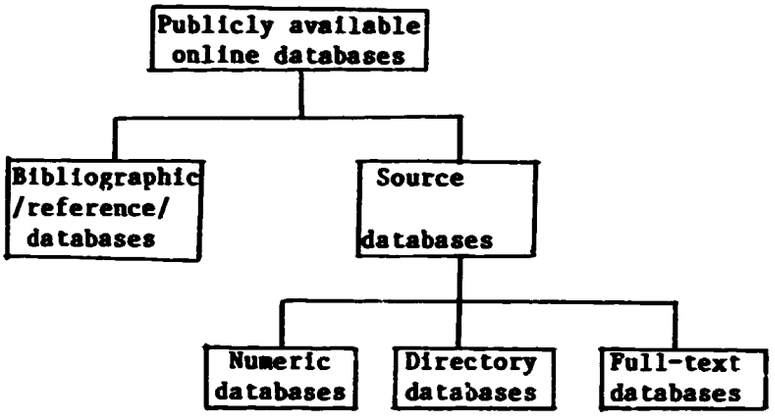


Fig.1 Types of online databases

2.2 Source databases

Contrary to bibliographic or reference databases, source databases may contain the final information sought after. Sometimes they are called also factual or factographic databases, though the name "source" database is considered to be more expressive.

Source databases can be subdivided as follows:

- Numeric databases
  contain data, figures, time series etc. on specific businesses, products or economics, like the output of a country or production data in a given year, financial data series etc. They are very specific, usually for macroeconomic utilization by well-trained economists and statisticians. A frequently available feature is that they can be used either for searching or for manipulating data in order to provide econometric analysis, results/regression analysis, forecasts etc. Numeric databases will not be discussed in this report.
- Directory databases
  present the bulk of source databases, a rapidly expanding proportion of the online industry. They are also referred to as textual-numeric databases, since they contain both text and

figures. These databases are similar in type to the telephone directory, i.e. the sources of information for names, addresses, locations, institutions etc. Also, bankers and businessmen can use them to search for current stock exchange figures, financial information, airline schedules etc. This type of database is intended for wider public use, i.e. telephone directories, entertainment and sports information, news, in viewdata type services

- Full-text databases
  A fast growing segment of the online industry, the full-text databases contain entire articles of magazines, newsletters, newspapers etc., or complete texts of encyclopedias, legal and legislation matters, recipes, biographies, rules and regulations etc. They represent the emerging electronic publication.

3. PREPARATION FOR AN ONLINE SEARCH SESSION

The online searcher - end user or intermediary - has the task of retrieving relevant literature references containing information on certain topics. The aim of the search is expressed as a query, i.e as a rigorous and precise description in natural language of the subject of the R&D job to which the information is requested.

The first task of the searcher is to decide whether the query is suitable for online searching. This issue must be considered very seriously. The searcher should be well aware of all possibilities on one hand, and of all limitations on the other, of this technique of information retrieval. The searcher should know and avoid all traps, snags and pitfalls of online systems which an average layman might easily overlook. In the author's experience, people who are not familiar with the capabilities and techniques of online searching, often regard it as an automatic question answering tool, something like "you ask, the computer responds".

The user must never forget the golden rule when preparing for an online search that, in present online systems and databases, only that information can be retrieved which has been published, and usually in that form only as it was published in documents. In other words, in general no selection, synthesis or analysis of fragment information or implicit information is possible by the present online systems. Therefore, online systems, especially those based on bibliographic databases, cannot be question answering services. Online searching is, in most cases, directed toward identifying information sources which should be subsequently obtained, then studied and evaluated. The final information can be retrieved by the classical ways of mental analysis, synthesis, learning, comparison, combination, extrapolation etc. In the future, perhaps, the situation will change.

After having established the suitability of the query for online searching, the intermediary should interview the end user in order to

convert the query into a form best suited for searching. It means that a vague query must be put within precise frames by defining what the user wants and what he or she does not want to be retrieved.

The next step is to analyze the query for the subsequent development of the search strategy, and to find the appropriate search terms. The elaboration of the search strategy should be started by the selection of databases and the host service where they are available. The rest of this report will be devoted to this problem.

4. SELECTION OF DATABASE(S) AND HOST

Following the preparational steps outlined in the previous section, one of the most crucial points of the search procedure is the identification of the most suitable database or databases, and the selection of the host system to be accessed.

The selection of database and host cannot be seen separately.

- If access to a host or hosts has been given conditions, then its or their databases can be selected as the first selection step.
- If, however, a local online service gets started, the first step is the preliminary selection of the host or hosts to be accessed.
- If the selected database is offered by more than one host, then the determination of the host where the search will be performed will follow as a second step.

First, a concise literature survey will be given, followed by a discussion of selection criteria.

4.1 A brief literature survey

There are far less papers in the literature on database selection than one would expect. In the INSPEC database from 1977 to 1987, 42 papers, from 1984 to 1987 only 15 relevant papers were found, using the search strategy of "database(2N)selection/DE,ID,TI" in the host Dialog. In the complete LISA database, only 3 additional papers were found using the same search profile. The reason can be that it is difficult to find a general methodology on the selection criteria of databases. As an author expressed: "Online searching is a personal experience, reflecting a combination of skill, luck, imagination and curiosity. There are probably as many different styles of searching as there are searchers".[2] We give a brief survey of some recent papers on the topic.

In a paper on database selection in the life sciences [3] a subject cross-tabulation of topics and availability is given for all databases with any information in the life sciences. Also, the ease of access, the indexing of secondary concepts or search modifiers, and chemical substance indexing is tabulated for 6 databases. In another

table, coverage of the most important databases are compared. In this paper many useful hints are presented on the process and criteria to find the database which might be the most suitable one in the given circumstances.

An interesting paper, though not on engineering problems [4] outlines some of the problems associated with the use of currently available databases for African studies, i.e. how online retrieval may be best utilized by Africanists. One of the two general headings of the problems explored by the author is database selection.

The paper [5] concentrates on the selection of the most appropriate databases for finding answers to business and financial questions. The primary determining factors are content, coverage, accuracy and structure, with pricing as a consideration after judging the value of the data and the likelihood of finding the data. The purpose of the presentation is to cover the foci of various business databases and some especially useful features which will help searchers to determine which databases will most likely provide the information they seek in a cost effective manner.

There are very few papers on the automation of database selection as an element in a more complex project. In [6] a microcomputer software is described which can assist in one or more of the following search functions: database selection, logging-on, uploading, downloading, post-processing and record keeping. [7] discusses front-end systems which can simplify online searching. Such systems have features including aids for database selection, search strategy reformulation and storage, simple keystroke log-on and log-off etc. The paper compares microcomputer and mainframe based front-end software which can be used for accessing medical databases.

A comparison of the effectiveness of computers and humans as search intermediaries is reported in [8]. An experimental computer intermediary system, CONIT, that assists users in accessing and searching heterogeneous retrieval systems, is described. Consideration is given to the prospects for much more advanced systems which would perform such functions as automatic database selection and the simulation of human experts.

Results of a survey of one hundred end users of online searching in a corporate research environment are analyzed in [9]. The survey was conducted to assess attitudes and expectations toward end-user searching. Potential end-user searchers expressed high levels of requirements for system aids in database selection, system menus, thesauri and cost control, but expressed less need for natural language query systems.

An interesting comparison is reported by [10], on the search costs of the hosts Dialog and ESA-IRS. Twenty-five databases available at both hosts were used in the comparison. It was found that 5 databases were less expensive to use in Dialog and 5 others were less

expensive at ESA-IRS. This was determined by free factors: online connect charges, online displays, and offline prints of citations. The remaining 15 databases represented a mixed bag. ESA-IRS is competitively priced when compared to Dialog and, despite occasionally higher telecommunication costs, it may even be more economical to use in some cases. Thus, the physical location of the computer should not be a major consideration in determining the usefulness and economic feasibility of the search services.

A theoretical paper on the topics investigated is the "Comparison guide to selecting databases and online services" [11]. It focusses on numeric and textual-numeric (defined here as directory-type) scientific databases. It suggests standards for evaluating scientific database services that can be organized under three categories: those related to the content of the database, those which describe the system used to create and access the database, and those which are determined by the management of the database and the related system. A systematic consideration of the content, accessing systems, and management supports for a database should increase the probability of a successful match between the user's needs and database capabilities.

[12] looks at database selection in an academic library. It tackles the problem whether multi-database searches are really necessary. Experience gained in a university library showed that, in most search requests, 11 databases accounted for 75.5% of the total number of online queries, and 29 databases were sufficient for 94.7% of all queries. 29 databases represent half of the total databases used. For 455 queries, 1290 database searches were used, resulting in an average of 2.84 database searches per request.

[13] discusses how searchers decide which database may be of potential relevance to their literature search. Database selection begins with the analysis of the information problem, i.e. with relating the search problem to the world of information. Selection should be made on the basis of subject/content/topic coverage, source document coverage, time period coverage, searchable and printable data elements. Relevant details are available through various promotional materials and user aids, but searchers must invest time to "dig-out" information on databases with which they are not completely familiar.

4.2 Selection of databases

A necessary condition for proper database selection is the availability of a catalogue and description of databases for the user-either a complete database catalogue or the description of the set of databases of the host. The best database directory in our practice is the "Directory of Online Databases" by Cuadra [15] appearing four times a year. Two issues are complete updated directories and the two others are issued as supplements to the main volumes. In addition to the short description of the available databases, the searcher is advised to obtain and consult database sheets and/or database chapters issued by the host (see below) before a selection decision is made.

The first consideration before the actual selection is whether a bibliographic or a source database search will be needed. It would be tempting to do a search which supplies final information or data right away, but one should be aware that source databases have their limitations: they are very specific, numerical data can be retrieved only for limited economic or business categories, most directory type databases are oriented for US affairs, full-text databases are usually expensive and contain mainly selected newsletters, magazines, legal issues etc. In most cases, factual information like specific business or statistical data, economic indicators of particular industries, trade information, product prices etc. are scarcely available or are very expensive in the present international databases. There are, however, directories to technical or industrial data but it is very unlikely to find the needed data directly.

The majority of proper queries in an industrial or academic environment can be answered through bibliographic search, i.e by using subject-field-oriented; multidisciplinary; mission-oriented; document-type-oriented reference databases. They number about 1500 and form the less growing proportion of databases, because almost all fields of science and technology have been covered by one or more databases already. This latter issue is where difficulties may arise.

The number of databases used in an online search session depends on the query and on the user. For certain queries a single-database search is adequate, if the subject field is covered by a very good database. Examples are the INSPEC in electronics and telecommunication, Chemical Abstract in all fields of chemistry (with the possible exception of business), METADEX in metallurgy etc. In these cases, there is usually no need to extend the search to another subject-oriented database, only to, perhaps, multidisciplinary databases (NTIS, COMPENDEX), or to patent-oriented databases.

There are, however, queries which cannot, or are not recommended to, be searched in one database only. These are interdisciplinary queries or those queries whose subject field is not covered by a good comprehensive database. An example of an interdisciplinary query is: "Model calculations on emission of nitric oxides by diesel engines". The topic could be covered by both environmental and motor vehicle databases. In fact, it was searched in eight databases [16] yielding a total of 48 references. Other typical interdisciplinary queries are those related with computer applications or control aspects or waste management in some specific areas of the power industry, metallurgy, chemical or mechanical engineering etc. These queries have to be searched in a minimum of two but possibly in four or five databases.

Industries which do not boast with a high quality database are, e.g. mining, transportation. For answering a query like "Techniques of convergence measurements and calculations in mines", four databases were used [17] and, in addition, a manual search was conducted to achieve a fair recall.

In the majority of cases, there is a choice of two or more databases to search. If more than one database is used for answering a query, it is called a multidatabase search. If many similar databases are considered for searching and they are all available to the searchers, the following questions should be answered:

- Does the query justify a multidatabase search?
- What is the user expectation - higher retrieval recall at the expense of precision, or a better precision, sacrificing recall/retrieval completeness?
- Are there any cost limitations involved?

After these questions have been answered and it is concluded that there is a choice of two or more databases for searching, the crucial question arises: what database or databases to choose? In the following ection, the selection criteria will be discussed.

(a) Coverage

The coverage of a database involves the extent of the scanned literature determining scope and content of the database. It implies geographical and language coverage, document selection for scanning, selection of papers within a publication scanned etc. The more international the coverage is, the more documents a database processes, the higher the chance of finding papers to solve the information problem will be. Document coverage and selection policy of the database producer should be known, though sometimes it is not an easy task. A guide can be the number and variety of source publications processed, or the numbers of records per update.

An example of considering coverage criteria is the use of various patent databases of 29 industrialized countries. CLAIMS databases cover US patens only, but their time span is longer and they contain patent claims text. JAPIO contains Japanese patents. INPADOC contains all patents and all applications but only the administrative data and class of patents rather than their content. There are patent databases on certain industries like COMPUPAT and APIPAT on computers and on petroleum industry, respectively.

(b) Indexing

The indexing structure and the type of controlled vocabularies of a database plays a vital role as a selection criterion. It is a well-known fact that the use of controlled vocabularies (descriptors, indexing terms, classification codes etc.) is preferred in the retrieval to free terms. The more thoroughly a database is indexed, the higher the relevance of the retrieved documents can be. Obviously, controlled vocabulary aids (thesaurus, classification etc.) should be available for and consulted by the user, and it is also advised to learn the indexing philosophy of the database producer.

(c) Time Span

If a searcher wants to retrieve as much background of a given subject field as possible, the database with a longer time span should be chosen. In the geosciences, for example, the time span coverage of the "Georef" and "Geoarchive" databases is from 1933 to the present and 1969 to the present respectively. This criterion should be discarded if only the literature of recent years is needed.

(d) Timeliness, update frequency

Better timeliness (shorter time lag publication) can be expected of a database with frequent updating. The figures of 2 or 3 months of average time lag claimed by several database producers could be accepted critically. Real average timeliness of databases can be far longer, but there also exist databases with exellent - e.g. weekly-updating.

(e) Language

The language of the database (to be distinguished from the language of the original documents falling into the category of Coverage) is another criterion to be considered. The vast majority of databases are in English, but some German and French language databases covering international literature also exist. An example, the Volkswagenwerk database on motor cars, though many non-German documents are also covered by it, is a German-language database. It is worthwhile to note that there are databases with bilingual descriptors. Recently, some Russian databases have started to become available. Some databases provide indexing even in 3 or 4 languages.

(f) Prices

In a multidatabase search, the prices of similar databases are not serious decisive factors, due to the SAVE capability of many search systems. However, price can be a criterion when choosing between identical databases at two hosts, see Section 4.3.

(g) Documentation

The better the printed aids of a database are, the easier the use will be. Here again, the different hosts provide different documentation of their databases, apart from the aids the database producer publishes. The producer of INSPEC for example, brings out three regularly updated publications, a newsletter and some other publications. Many database producers as well as hosts hold one day training seminars or workshops.

(h) Type of document

Several databases are document-type oriented rather than subject-oriented databases: report-, patent-, conference proceedings databases etc. Their selection should be matched with the user needs. Many subject-oriented databases do not cover all types of documents within their subject scope. Some databases e.g. COMPENDEX process only journal articles etc. It should be noted that sometimes copies of reports or of conference papers are more difficult to obtain than journal articles.

(i) Availability of abstracts

Most databases contain abstracts to each record or to part of them. Some databases provide better, more informative abstracts, others only short indicative annotations. There are still some databases without any abstracts. A criterion of selecting between databases is the existence or the quality of abstracts. The availability of printed abstract journals can overrule the problem of the absence of abstracts in the database like it is in the case of Chemical Abstracts.

(j) Familiarity of usage

Perhaps one of the most important criteria in selecting a database is how familiar the searcher is with a database. In many instances, the thorough knowledge of the coverage, of the indexing structure and philosophy of the database, extensive experience in using the database can overrule other factors, even if another database might be superior regarding other aspects. It does not mean, of course, that a searcher should not learn to use other, new databases too [13].

Generally speaking, the experience of the searcher plays the most important role in choosing the best database for his needs. Of coure, in the case of previously not used databases or if new subject fields must be searched, but also with special searches involving known databases, the above criteria should be analyzed and weighed carefully, according to the guidelines presented, even by the experienced user.

(k) Codes and Classification e.g. for Products

In some databases search and retrieval is based on determination of the product codes and not the product name. In this case selection criteria would be based on the effectiveness of the coding system used and the availability of reference tools needed whether the coding system is an international system or developed by the database producers. Another issue to be considered is the breakdown of the coding system used (number of digits) and the extent it helps in retrieving data relevant to specific type of a product.

4.3 Selection of the host service

The reader is assumed to be aware of the fact that online databases are stored and made available for remote online searching by online services (or vendors or spinners), throughout this report called hosts or host services, and by telecommunication networks. There are about 550 online host services in the world at present, some offering only one or two databases, others many (hundreds) databases for public online access. Obviously, only a fraction of hosts would be accessed on a permanent basis by any online searcher since the particular information needs can be satisfied by a relatively low number of hosts.

There are different user types. If an end user or an intermediary user is engaged in a certain field of science or engineering or in a certain industry, he or she will use only a few databases. If all databases are found at one host and access to this host exists already, the reader can skip this section and go to the next one, since there is nothing to do with selecting hosts. However, if the user works as an information broker or as an intermediary at a regional or other central service for heterogeneous type end users, the host or hosts for establishing a permanent access should be selected carefully.

The major criteria to consider in making a priori or preliminary decisions on which host or hosts must be accessed when establishing a new online search service for a country, for a university, for a company etc. can be categorized as follows:

The most important criterion is the range of databases the candidate has on its list, in conjunction with the field of interest of the would-be online service, i.e. with the anticipated topics of queries. If a good proportion of clients would use unique databases (those offered by one host only), like RAPRA, APILIT, Hoppenstedt, World Surface Coating Abstracts, EMIS, Coffeeline etc., then, obviously, that particular host must be accessed where that unique database is available.

The next and equally important criterion is the accessibility through telecommunication lines or networks of the host from a given physical location of the user. Unfortunately, there are no guidelines or rules establishing data communication links from a terminal at a given geographical point to a host located at another, perhaps on another continent, because it is beyond the control of the user. It is the responsibility and function of national PTTs that must be consulted before any decision regarding access to hosts is made. It should be emphasized that the most serious bottleneck of online searching is its telecommunication aspect, not only in developing countries but also in highly industrialized nations (with the exception of North America).

The third aspect of preliminary host selection is related with costs. If the pricing schemes of hosts are known a comparison of their pricing policies can be made. By anticipating the extent of usage, one

can determine the most advantageous price structure if the first two criteria have already been considered.

Other selection criteria of hosts are associated with their search systems. Several of them can be regarded also as a preliminary decisive aspect of selecting host for establishing access, but also as selection criteria of host when a particular search problem has to be solved in a defined database available at multiple hosts. These criteria include search language issues, mail delivery times, multidatabase search problems, cross-database search, SDI possibilities etc. They will be discussed below in association with the selection of databases offered by multiple hosts.

As it is known, the most popular databases can be searched at several hosts. Chemical Abstracts and INSPEC for example, are available at a great number of hosts. When using such a database, the user has to select the host he will access to perform the search, an equally crucial decision to the database selection. Host selection is feasible only if some conditions are met:

- The user has established access to more than one host,
- The user is familiar with the search language of these hosts,
- The search is not centered around, i.e., the primary database to be used is other than, a unique database since it would preclude any further selection,
- The user is not bound financially or by any other reason to one or another host, and is not excluded from using a host.

In summary, if the searcher is in the position to choose freely between at least two online hosts to search a given database or a range of databases, then the following criteria should be considered for the selection.

(a) Multidatabase searching

Suppose that three databases were identified for answering a given query. Two of them are found at host A, the third at host B, and all three at host C. The obvious choice would be host C, if other criteria (costs, search features etc.) are not stronger, though even then host C would be highly preferred. Example: A query related with nuclear techniques applied to foods and agricultural products must be searched in the databases CAB Abstracts, FSTA and INIS. The first two are on the list of both Dialog and DIMDI, the latter at IAEA, However, the host ESA-IRS should be the preferred choice, because here all three databases are available.

(b) Search features

Advanced searchers consider in the first place those criteria which are associated with the search features a host system can or cannot offer. The quality of the retrieval and also search costs may depend in many cases on the possibilities of the search software that

can be fully exploited by experienced users. The most important search features which can vary from host to host are:

- Proximity searching
- EXPAND feature capabilities and thesaurus resident in the computer.
- Postqualification of set numbers.
- Structure of the inverted files, basic index, additional indexes.
- Word and phrase inversion in basic index, double posting.
- Limit feature flexibility, limiting options.
- Cross-database searching. Some hosts offer a look-up possibility of the basic indexes of several databases by a single search step. This can help with the identification of the most suitable databases in a cheap way. Such aids are offered by Dialog/Dialindex/BRS and Data Star/CROS/Orbit/DBI.
- Special command features, like the ZOOM and GET type commands, STRINGSEARCH capability, REPORTing, MAPping capabilities etc.
- Split files. Large databases are split usually into two or more databases for computer reasons. Sometimes, however, it would be better to search the total database rather than to search the latest file and repeat the search strategy in backfiles. BRS and Data Star have the unique feature of Superlabels which enables the user to either do split file searches or a total database search in a single step. Dialog added its latest One Search feature, ESA-IRS the CLUSTER feature to solve this problem too.
- In searching Chemical Abstracts, the possibilities of searching for general subject headings, using substrate dictionary files, segmentation of chemical names, availability of CA abstracts, structure and substructure searching etc.

These and other features of the various search systems must be known before making a decision about the host.

(c) Online document ordering

A number of hosts have introduced the capability of ordering copies of complete documents online. Dialog, ESA-IRS, and Orbit's document delivery services are called DIALORDER, PRIMORDIAL and ORBDOC, respectively. The ORDER feature of STN International can be used for placing online orders to Chemical Abstract's Document Delivery Service.

(d) Cost factors

The cost of online searching can be broken down into three components:

- Subscription fee,
- Connect hours and print/display rate,
- Telecommunication charges.

The rates of database usage at a host are composed of the usage of host computer and of database royalties. They are proportional to the connect time (connect hour rates) and to the unit price of online

displays and offline prints of references.

What comparative prices do not reflect are the different response times and system reliabilities of various host services though both are - sometimes overlooked- cost influencing factors. With shorter response times even a higher connect hour rate can result in cheaper overall search costs than vice versa. Response times should be measured by the users by making identical searches at two or more hosts. It is interesting to note that in Europe, Dialog's response time is shorter in the morning hours than in the afternoon when the American users start working. Implicitly, various search features discussed above can also influence connect time and, hence, search costs.

An important component of search costs are telecommunication charges which usually increase with the distance of the terminal to the host computer. In industrialized countries with advanced data networks, in most cases packet-switched networks, telecommunication tariffs are comparatively low. From remote locations, the tariffs are higher if public telephone lines must be used. The rates in any case are fixed by the local PTTs which the user can hardly influence them.

Overall costs of online searching are much lower if high data speeds can be used, i.e. the lines and terminal allow them. With 1200 or 2400 Baud lines connect times and telecommunications costs are much lower. Costs can also be reduced by using an intelligent terminal or microcomputer with a suitable software for offline data preparation, uploading, downloading etc. It should be noted, however, that equipment may be more expensive for high speed.

(e) SDI services

Some hosts offer also SDI (Selective Dissemination of Information) services most of their databases. If so, all new records of documents matching a previously developed and stored search strategy will be automatically retrieved, printed and sent to the user's address at each update (monthly, two-weekly etc.). The availability, prices and quality of SDI service vary from host to host.

(f) Delivery of search results

When the retrieved records are ordered as offline prints, they will be mailed to the user's address. The distance of the user's address to the host's location determines delivery time by mail, so it is an important selection criterion. Delivery time can be shortened if the host offers an electronic mail service like Dialog's DIALMAIL which is somewhat more expensive than offline printing but much cheaper than online display of the results. By electronic delivery, search results can be printed online with DIALMAIL by the user's printer, within 24 to 36 hours of searching. An even more efficient - but not necessarily a cheaper way of delivering a high number of records is downloading and subsequent offline searching which is possible at certain hosts and from certain databases (under special contracts) if

the user is equipped with a suitable microcomputer, software and high-speed data link.

(g) User aids, training, communication with users

An important factor in choosing a host is the availability of good user aids and database documentation. As it was emphasized above, three types of user aids exist:

- Database documentation provided by the database producer,
- Online system manual published by the host (search language, description etc.),
- Database sheets published also by the host for searching the databases in the hosts systems.

The quality of the latter two presents itself as a factor to be considered for selection, since good and well maintained user's manual and database sheets/chapters can save much money and the search can be more efficient.

Another form of communication with the user is the Help Desk where the user can contact the host directly by phone to obtain assistance during the search process.

(h) Familiarity with the command language

As previously stated, one of the most important criteria for the selection of databases is how familiar the user is with a particular database. Similarly, if no other decisive factors prevail in the host to select, the user would usually choose the host system with a search language he has most experience with. Sometimes it should be preferred to search a given database at host A even if it would be perhaps more effective to search it at host B, but the user can handle the search in a freer manner at A than at 3.

(i) Online facilities

Finally, selection of host could consider the facilities that the service provider offers online as:

- Online index files
- Online help functions
- Menus helping in formulation of search strategies
- Electronic mail

(j) Selection criteria of host searching the Chemical Abstracts database

As an illustrative example of most selection or decision criteria discussed above is the case study published in [18], on the comparison of three hosts in searching Chemical Abstracts, its results will be summarized here briefly.

The main aspects of comparison were time span, splitting of Chem. Abs. database, search features including proximity searching, utilization of substance dictionary searching, segmentation of chemical names, limiting options, print formats etc., and also search times and costs, cross-database usage.

The search times of a sample search at the hosts Dialog, Data Star and CAS Online (now STN International) were found to be 0.109, 0.085 and 0.071 hours respectively. These data reflect the order of computer response times (at the time of writing of [18] since the search strategies were identical at all hosts.

CAS Online provides also the abstracts to each item, and it allows also chemical structure searching in a graphical way. At chemical substance dictionary searching, Dialog was found to be the best, and it is also the preferred host if multidatabase search is involved because of its widest choice of databases.

5. CONCLUSION

In summary, the preparatory phases for an online search session can be broken down into the following decision or analysis processes:

- Decision whether the query is appropriate for online searching.
- Analysis of the query and, if needed, its modification.
- Decision on the type of database: reference/bibliographic or source databases should be used for the query.
- Whether bibliographic, literature or patent search should be performed.
- Single database(s) and/or host(s).

This last decision is a combined process consisting of:

- Preliminary selection of host or hosts to access before the search service starts functioning,
- Selection of database or databases to retrieve information for a particular query,
- The selection of the host where the selected database(s) will be searched.

Selection criteria for consideration and guidelines for decisions are given in Table 1.

The discussions and selection guidelines presented here are valid primarily for bibliographic databases which are used in the majority of online searches involving internationally relevant information retrieval in science, technology or industry. It should be noted that it is not claimed that all criteria for selection and decision were considered and discussed in full in this report.

Search strategy formulation and online search itself can start only after these preparational and selection processes have been completed.

Table 1 Summary of criteria and guidelines for host and database selection

Criteria to consider for selection	Selection guidelines		
	1. Preliminary selection of host	2. Selection of database or databases for a query	3. Selection of a host for searching the selected database or databases /if more than one host is available/
Fields of interest, range of databases	According to anticipated usage. Unique databases	-	At multidatabase searching: Possibly all databases be available at the same host
Accessibility through telecommunication networks	Depends on local PTTs	Only databases at accessible hosts can be selected	-
No restriction of usage	Financial, geographical or political reasons	Some databases are unavailable to certain users	Some databases at some hosts are unavailable to certain users, at other hosts they are available
Costs	Pricing policy of hosts Telecomm.charges	Expensive and cheap databases	Database usage costs vary from host to host Telecomm.charges depend on geography. Computer response times vary.
Coverage	-	Content coverage, geographical and language coverage, literature scanned, document selection etc.	-
Indexing	-	Quality of controlled vocabulary and indexing /thesaurus, classification, others/	Controlled vocabulary /thesaurus etc./ resident in host computer
Time span	-	Retrospectivity, if more years' literature is needed /time coverage/	Time span of identical databases may vary from host to host
Timeliness, update frequency	-	More frequent updating means better timeliness	Updating of identical databases may vary from host to host
Database language	-	Other than English databases	-
Document type	-	Report, patent, conference paper etc. oriented databases	Document-type oriented databases as supplementary databases to subject-oriented search. For patent searching, other criteria exist
Availability of printed equivalent	-	Helps formulation of search strategy or retrieval evaluation	-
Availability of abstracts	-	Completely or partially abstracted databases. Databases without abst.	Very rarely, a distinguished host provides abstracts to a database which others do not
Search system features	See Column 3. + system reliability, billing quality	-	Cross-database searching, split-file searching, proximity operators, postqualification of set numbers, inverted indexes, data fields, limiting print/display formats, word/phrase inversion, expanding, truncation, special command capabilities, chemical search features etc.
Mail delivery time of offline prints	Geographical location dependent. Availability of electronic mail	-	Delivery time of offline prints is faster from hosts of closer location. Electronic delivery speeds up printing of search results
Online document ordering	Document copy delivery/ordering service	-	Copies of primary documents may be ordered online at certain hosts. Fast document delivery
SDI service	See Column 3.	-	Automatic printing and delivery of new references from updates against stored strategy
Training, communication with users	Training courses, training files, help desk etc.	Database training courses	-
Documentation, printed aids	System user's manual	Database printed aids and their updating	Database sheets at a host and their updating
Familiarity with usage	-	Very important aspect!	Very important aspect!

6. BIBLIOGRAPHY

1. Turpie, G.
Going online 1988. London: ASLIB, 1987, 86 p.

2. Jack, R.F.
"Meatball searching: The adversarial approach to online information retrieval". Database, vol.8, no.4, 1985, p.45-52

3. Snow, B.
"Database selection in the life sciences". Database, vol.8, no.3, 1985, p.15-44

4. Seeley, J.
"The use of bibliographic databases for African studies". In Databases and Networking in Development Seminar, Brighton, England, 4-6 Sept.1986, p.253-262
Bergen: DERAP Publications, 1986

5. McBrown, L.
"Sensing the right database for business and financial affairs". In Online '85 Conference Proceedings, New York, USA, 4-6 Nov. 1985, p. 205-209
Weston: Online, 1985

6. Fenichel C.H., and J.J. Murphy
"Using the microcomputer to communicate. II. Specialized software". In Microcomput. Inf. Manage., col.2, no.3, 1985, p. 155-170

7. Wible, J.G. Wible
"Searching made easy: front-end systems for medical databases" In Med. Ref. Serv. Q., vol.5 ,no.2 , 1986, p. 1-13

8. Marcus, R.S.
"An experimental comparison of the effectiveness o. computers or humans as search intermediaries". In Journal of the American Society for Information Science, vol.34, no.6, 1983, p.381-404

9. Tatalias, J.
"Attitudes and expectations of potential end user searchers". In National Online Meeting Proceedings, New York, USA, 30 April-2 May 1985, p. 457-462.
Medford, USA: Learned Information, 1985

10. Jack, R.F.
"Cost considerations in database selection: a comparison of Dialog and ESA/IRS". In Online, vol.8, no.4, 1984, p. 51-54

11. Swbank, W. Bruce
"Comparison guide to selection of databases and database services". In Drexel Library Quarterly, vol. 18, no.3-4, 1982, p.189-204

12. Evans, J.E.
"Database selection in an academic library: are those big multifile searches really necessary?" In Online, vol.4, no.2, 1980, p. 35-43

13. Wanger, Judith:
"Multiple database use: the challenge of the database selection process". In Online, vol.1, no.4, 1977, p. 35-41

14. "International comparative price guide to databases online".
In Online Review, vol.11, no.4, 1987, p. 255-266

15. Directory of Online Databases, vol.8, no.3, 1987.
New York, N.Y.: Cuadra/Elsevier, 1987, 581 pp.

16. Valas, G.
"Online public services in OMIKK, Hungary - The first 20 months".
Proceedings of the 8th International Online Information Meeting, London, UK, 4-6 December 1984, p. 125-138
Oxford: Learned Information, 1984

Updated version:
G. Valas:
"Retrospective information retrieval services at the National Technical Information Centre and Library on the basis of access to foreign systems". In International Forum on Information and Documentation, vol.10, no.2, 1985, p.33-39

17. Horvath, O. and P. Roboz:
"Literature search by computer and manual methods on a query in mining" In Tudomanyos es Muszaki Tajekoztatas, vol. 33, no. 9, 1985,p. 455-459
(in Hungarian, summary in English)

18. Novak, T.:
"Searching the Chemical Abstracts database: a comparison of Dialog, Data Star and STN International". In Proceedings of the 10th International Online Information Meeting, London, UK, 2-4 Dec. 1986, p. 353-364.
Oxford: Learned Information, 1986

PART II

ESTABLISHING AN EXPERT SYSTEM TO ASSIST IN THE SELECTION OF THE MOST APPROPRIATE DATABASES AND HOSTS

1. POTENTIAL EXPERT SYSTEM ROLES

Expert systems can assist information retrieval in a number of ways as described below. Existing research has demonstrated that most of these functions could be successfully embedded within an expert system.

1.1 Query Translators

The ability to take a query from a user and encode it in the search language of the appropriate database host is generally seen as being central to any intelligent interface between a user and a database. Research in this field is extensive and utilizes a number of techniques for mapping an enquiry onto the descriptors or free text terms which represent a database record. Techniques for analyzing and representing queries include the use of:

natural language interfaces
menu driven systems
graphical representations

Natural language interfaces have been developed by a number of researchers [see 12,13] and it is possible to buy of-the-shelf PC based software to tailor to specific applications [14]. However, the use of such an interface does make general assumptions about the ability of users to express their information need in appropriate terms. Belkin and Oddy [15,16] have argued that information needs vary from those which are ill-defined (often encountered at the beginning of a piece of research or with a novice)through to those which are well-defined (typically encountered with expert users). It is highly likely therefore that the information queries presented to a system will not be sufficient in themselves to ensure that satisfactory retrieval performance is achieved. In addition, users from developing countries will have varying degrees of ability in English (for instance the second language for many African countries is French) and queries may be presented in a relatively unstructured form. Indeed studies of how users expressed their information needs on the Plexus project (5] showed that the inability to give precise and grammatically correct accounts of the information sought extends to those for whom English is a first language. A natural language interface, therefore, while providing ease of access to a system, will not of itself solve the problem of query representation.

Fidel [17] sees the search process as consisting of three basic intellectual components:

(a) definition of query structure
(b) selection of search keys (terms)
(c) evaluation of feedback

and has suggested that "as intermediary expert systems are based on text analysis rather than on models of human searching, they cannot

process request-related criteria such as precision or recall requirements". Her studies have therefore concentrated on an analysis of the searching behavior of human intermediaries and have revealed a decision tree based routine for the selection of search keys which goes beyond the simple mapping algorithms which are generally used. These rules, which the author recognizes as requiring further development, could be used to structure search strategies which are more sensitive to subject-specific approaches which indicate the degree of specificity/exhaustivity or precision/recall required by a user. Using a distinction between common terms (those with multiple meanings or contexts) and single-meaning terms (those with an unambiguous subject specific meaning) and between unmatched, partially matched and fully matched descriptors, Fidel has identified fifteen basic rules for selecting free-text or controlled vocabulary terms. Such work could be extended by drawing on classical information retrieval research. Online thesauri combined with relevance feedback has been suggested by Salton [18] and others as one way of automatically generating broader or more specific search statements according to user evaluation. The Connection Machine [19] uses the concept of seed documents (a small number of highly relevant documents) to generate a new search based on the descriptors assigned to those documents. While this particular implementation in hardware-specific it might be possible to automatically generate a new search strategy by exploiting features such as Zoom on ESA-IRS and Postings On in Dialog, and the search descriptors found in relevant documents. This would be a two-stage process, with restatement of the search strategy being prepared off-line to save connection costs.

The Plexus project has shown that frame based systems can be an extremely effective method of structuring ill-defined queries presented by users. The strengths of a frame based system are that it can operate on incomplete evidence and gradually clarify a problem description through its expectations of what it should be encountered given a small number of parameters. It might, therefore, be possible to define a series of core query structures for particular industry related problem domains. These could then be used to interpret a user's information need.

A further application for expert systems is to provide a common command language interface that would allow a user to type instructions in the command language normally used or preferred. The expert system would then be able to act rather like an internet protocol by translating the command statements into those used by different databases hosts. The process of restating search strategies and submitting them to databases would of course appear transparent to the user.

1.2 User modelling

User modelling has been used by a number of researchers including Brooks [5] and Mukhopadhyay [20] to improve and personalize the retrieval process. Modelling should take account of various

characteristics that will allow the system to adapt its output to each individual user. These characteristics include:

(a) Level of experience with expert system interface
(b) Location of user
(c) Working languages
(d) Previous satisfaction with specific hosts and databases
(e) Presentation preferred (abstracts, citations only etc.)
(f) Quantities of information required
(g) Duration of research
(h) Updating requirements
(i) Ability to express information need
(j) Familiarity with problem domain
(k) Familiarity with database content
(l) Familiarity with paper-based equivalents
(m) Degree of pre-search research
(n) Expectations of hit rates.

1.3 Monitors and instructors

Meadows et al [21,22] has shown that a further role for an expert system is that of monitoring the performance of a searcher. IIDA (Individualized Instruction for Data Access) was developed to encourage end users of information retrieval systems to perform their own searches by (1) instructing them in how to search, using computer-assisted instruction, and (2) assisting with the performance of the search by providing diagnostic analyses of the user's performance as well as answering their questions about how to use the system commands. The type of errors that IIDA can detect include syntactic (e.g. inaccurately stated commands) and procedural (e.g. syntactically correct but ineffective commands) faults. Online help screens and tuition modes are also available to support the search process. Information on this performance could, of course, be passed on to a user modelling module.

1.4 Value adders

An area which has received little attention is the potential for expert systems to add value to the retrieved information either through straightforward filtering techniques or through desk top analysis of the information that has been retrieved from the database. One of the first systems to include post-retrieval processing was Paperchase [23] which included in its functions the simple, but effective, sorting of records retrieved from the database into the order in which they could be found on the shelves of the library. Other functions which might be investigated include:

(a) automatic limiting of records to particular languages
(b) division of records into those held locally, those held nationally and those which must be obtained from abroad.
(c) weighing of records using journal impact quotients
(d) division of records by country of origin
(e) post or inter-search analyses of index terms in

documents to identify other potential search keys

Option (a) would be based on the model of a user built up by the module described in section 1.2. It is possible that users in developing countries might not wish to retrieve information in languages outside those predominantly used by the information worker or his/her client. Automatic limiting of the number of records retrieved would reduce online costs and lead to less demand for expensive translations.

Option (b) would be based on knowledge held by the system about the journals, report series etc. held by the institution, national libraries, and important specialist collections. Using codens and/or ISSNs the retrieved records could be divided into those which could be satisfied immediately from the local collection; those that could be obtained at a relatively low cost from other national collections; and those which could be obtained from overseas document brokers at a higher cost. This would help the information centers to exploit national resources as fully as possible and to provide users with an instant assessment of the accessibility of information.

Option (c) would allow the process to be refined by using impact measures of journals to indicate which articles are most likely to yield high quality information. Such measures are already being used to assess the research output of academic and industrial organizations. Where resources are scarce, it is increasingly important to ensure that their effect is maximized. Impact measures could provide a means for ranking those articles which must be obtained from overseas.

Option (d) would be based on the location of the user. On the basis that similar countries will have similar problems it might be useful to cluster retrieved documents by their country of origin. This could be used to identify potential collaborative work, solutions to problems that have had to be made under similar economic and/or technological conditions; researchers who might have more empathy with a user through closer cultural and/or ideological ties; more readily accessible expertise and potential technology transfer. Beyond the simple massaging of retrieved information there are wider possibilities related to the merging, repackaging and analysis of information. This system would have knowledge about the structure of records held within databases and would therefore be able to selectively present information that has been downloaded from a database based on the analysis of the query. In addition, it would be able to extract fields from records downloaded from a number of databases and synthesize new records which match the exact requirements of a user. In this way, for instance, a patent search could be amplified with information about the company that has lodged the patent, who its local distributors are, plus relevant articles about the process. To achieve this the system would have knowledge about relations between databases and would be furnished with a core of automatically executable searches. A further extension would be the ability to extract information from selected databases in order to carry out inter company comparisons.

Option (e) would be used to automatically generate new search strategies and to improve recall and precision.

1.5 Database targetters

Identifying the best databases to use for a specific search is increasingly difficult as more and more research crosses traditional disciplinary boundaries. Bradford's classic work on the scattering of information in the primary literature can be extended to secondary information sources. For each subject there will be a core of databases which provide a large percentage of references; a secondary, and larger, group of databases which provide a similar number of references; a third, and even larger, group of databases which provide the bulk of the residue of references; and a fourth group of databases which provide negligible or zero results. Facilities such as Dialindex provide an entry point to the core databases on the basis of over fifty general topics. These ready packaged clusters of databases could form the basis of an adaptive targeting module. User specific clusters could be developed on the basis of ideas suggested by Mukhopadhyay [20]. The MINDS (Multiple Intelligent Node Document Servers) system was conceived for controlling access to file servers on a local area network but could be applied to online information retrieval services. MINDS incorporates metaknowledge about search terms, users, databases and certainty factors to identify the most appropriate database for retrieval purposes. Four-tuples express the likelihood that a given database will satisfy a perticular user on the basis of the:

(a) the breadth of knowledge pertaining to a keyword in a database
(b) how useful documents from the database have been in the past
(c) how recently the database acquired its information.

Factors (a) and (b) are directly relevant to online systems although (c) would have to be altered to take account of how often a database has been updated since it was last used.

Another factor that must be taken into account is the degree of overlap between databases. A number of databases might form part of a group of highly productive databases but much of the material in such databases is duplicated. Selection of databases must therefore be supported by functions which pass over highly productive sources where there is a high level of probability that the contribution in terms of novel references may be low. A basic figure for this overlap could be calculated on the basis of the journals represented in the citations retrieved from the first database.

EasyNet is an example of a commercial system which targets databases and hosts for the user. Its goal is "to create an effortless pathway where the searcher has only to make a series of uncomplicated selections while the system makes the more difficult choices behind the scene". [24] As well as choosing hosts and databases EasyNet will

assist in query formulation and will effect communication links and administer the appropriate logon procedures. The success of EasyNet has been demonstrated through experiments with school children and novice business and medical users. This interventionist approach is, however, based on the assumption that end users will generally be satisfied with smaller more carefully selected quantities of information. Context-dependent criteria such as precision and recall levels would have to be picked up through user modelling.

1.6 Resource managers

The costs associated with online searching can be broken down into the following categories:

(a) Local telephone costs
(b) Public data network costs
  - Volume
  - Time
(c) Host connect time costs
(d) Database access costs
(e) Royalty costs

Up to date management information about such costs is vital particularly where resources, such as hard currency, are scarce. Online users need to have accurate estimates of the potential costs of a database search in order to:

(a) Advise users of the likely cost of a search where cost recovery is operated
(b) Project total online costs for annual budgets
(c) Compare estimated costs (pre-search) with estimated costs (post-search) and actual costs to test pricing algorithms and to identify anomalies and/or time-frames where system performance is above or below par
(d) Flag users when cost ceilings for searches or budgetary periods have been reached
(e) Advise on the best time to conduct search in terms of cost
(f) Allocate search costs to nominal budgets
(g) Identify local, soft and hard currency commitments
(h) Allocate costs to specific search intermediaries
(i) Analyse search costs in terms of databases and hosts
(j) To match costs to user satisfaction levels.

2. SYSTEM SPECIFICATIONS

2.1 Selecting an expert system application

Before a specification for an expert system can be drawn up, it is necessary to ascertain whether expert system technologies are appropriate. Turner [8] provides a useful list of positive and negative indicators:

Positive indicators:

- conventional techniques not obviously appropriate
- uncertain data involved
- maintenance of system knowledge will rely on non-computer staff
- explanations of advice and conclusions required
- knowledge maps onto rules rather than onto equations

Negative indicators:

- knowledge more readily available in algorithmic form
- problem solving methods can be completely specified in advance of implementation
- when problem can be adequately and efficiently solved using conventional techniques

Using these indicators it would appear that an expert system would be an appropriate way to tackle this problem. However, one other solution should also be considered: Hypertext [25]. Hypertext, a term coined by Ted Nelson, allows the creator of a knowledge base to provide explicit linkages between the individual items within a collection of records. Using these linkages the user can be taken on a tour through a knowledge base and can refer directly to the full text of the relevant portion of any supporting record. In addition, users can develop their own tours by creating new linkages and by adding new information. Windowing environments can then allow these records to be viewed side by side and for further, related information to be retrieved. An extension of hypertext - hypermedia, allows linkages to be made between related text, voice, data, graphics and video. A number of experimental systems exploiting these techniques were developed by researchers at institutes such as Brown University and Xerox Parc. Until recently, however, the cost of hardware and software which would support hypertext has been prohibitive and has ensured that wide access to hypertext based information systems has been severely restricted. The emergence of Hypercard for Macintosh, and similar products announced for the IBM PC range of machines, has meant that relatively low cost technology is now available. Hypercard is being shipped free with every new Macintosh and is available at a nominal cost for existing users.

For some application hypertext has distinct advantages over information retrieval and expert systems. Information retrieval systems do not in general provide information that can be immediately

used for decision support. Secondary systems (i.e. those providing bibliographic citations rather than source documents) place the user several steps back from the point at which a decision can be made. Primary systems (i.e. those providing the full text of documents or source data) bring the user closer, but are still based on one-dimensional, flat-file structures and require the end-user to invest considerable effort in analyzing the information to make it usable.

It has also become clear that expert systems may not always be the most suitable route to follow despite the temptation to encode any rule-based system using such technology. Where the problem domain is primarily concerned with navigating through an assemblage of information which is linked by formal relationships then an expert system may prove to be an expensive and inflexible solution.

Hypertext can offer a cost-effective alternative to expert systems, combined with an extremely user-friendly interface. Recent figures suggest that familiarization times can be reduced by a factor of ten using Macintoshes. Hypertext will allow primary information to be linked to commentary, related documentation and graphics. In addition, users are released from the requirement to learn a complex command language by providing control through the use of a mouse and menu based instructions. Finally, the user is able to add new documentation and personal commentary to the core knowledge base. The ability to enhance and augment the knowledge base by adding local information, interpretation etc. will ensure that the user can maintain an information system which is both personalized and up to date. This may be important where, for instance, there is a shift in research priorities or when users sign up to use new hosts. As hypertext can offer a cheap and flexible method of storing knowledge, it should be borne in mind as a possible technological solution if studies show that a decision tree approach with well defined boundaries is needed.

2.2 User requirements

It is often tempting to produce a system specification that is entirely technology-driven. Given adequate resourcing a seamless multi-function, high performance system can always be developed. It may nor however be what the users actually need or what is financially possible. Any decision on system specification must be prefaced by an extensive survey of user requirements. This should include novice as well as experienced users as any system will have to make allowances for the differing approaches of each of these classes of users. Information on common research strategies, the types of information sought, user preferences, reactions to interfaces, common problems encountered with paper-based and online systems etc., must be collected before development work starts.

3. BIBLIOGRAPHY

1. Van Rijsbergen, C.J.
Information retrieval, 2nd ed. London: Butterworths, 1979

2. Barr, A. and Feigenbaum, E.A. (eds)
The Handbook of artificial intelligence. Volume 1. London: Pitman, 1981

3. Michaelson, R.H., Michie, D. and Boulanger, A.
"The technology of expert systems". In Byte, 1985, Apr., p.303-512

4. De Jong, G.
"Prediction and substantiation". In Cognitive Science, 1979, 3, p.251-273

5. Brooks, H.M.
"Expert systems in reference work". In: Gibb, F. (ed). Expert systems in libraries. London: Taylor Graham, 1986

6. Duda, R.O. and Gasching, J.G.
"Knowledge-based expert systems come of age". In Byte, 1981, Sep., p.238-280

7. Gordon, J.
"Expert system development routes". In ITEMS, 1984 Dec., p.9-11

8. Turner, M.
Expert systems: a manager's guide. London: PACTEL, s.d.

9. Waterman, D.A.
A guide to expert systems. Reading, Mass: Addison-Wesley, 1985

10. EUSIDIC survey of public data networks - 1987. London: EUSIDIC, 1987

11. "An overview of the new business intellegence center". In Online Access, 1987, Sept/Oct., p. 16-17

12. Bisaws, G. Bizdek, J.C., Marques, M. and Subramanian, V.
"Knowledge-assisted document retrieval: 1. The natural language interface". In Journal of the American Society for Information Science, 1987, 38 (2), p.83-96

13. Bisaws, G., Bizdek, J.C., Marques, M. and Subramanian, V.
"Knowledge-based document retrieval: 2. The retrieval process". In Journal of the American Society for Information Science, 1987, 38(2), p.97-110

14. Hendrix, G.G. and Walter, B.A.
"The intelligent assistant". In Byte, 1987, 14(12), p.251-258

15. Belkin, N.J., Oddy, R.N. and Brooks, H.M.
"ASK for information retrieval. Part 1. Background and theory". In Journal of Documentation, 1982, 38(2), p.61-71

16. Belkin, N.J., Oddy, R.N. and Brokks, H.M.
"ASK for information retrieval, Part 2. Results of a design study". In Journal of Documentation, 1982, 38(3), p.146-164

17. Fidel, R.
"Towards expert systems for the selection of search keys". In Journal of the American Society for Information Science, 1986, 37(1), p.37-44

18. Salton, G. and McGill, M.J.
Introduction to modern information retrieval. Tokyo: McGraw-Hill, 1983

19. Waltz, D.L.
"Applications of the connection machine". In Computer, 1987, Jan.,p.85-97

20. Mukhopadhyay, U., Stephens, L.M., Huhn, M.N. and Bonnell, R.D.
"An intelligent system for document retrieval in distributed office environments". Journal of the American Society for Information Science, 1986, 37(3), p.123-135

21. Meadows, C.T., Hewett, T.T. and Aversa, E.S.
"A computer intermediary for interactive database searching. 1. Design". In Journal of the American Society for Information Science, 1986, 33(5), p.325-332

22. Meadows, C.T., Hewett, T.T. and Aversa, E.S.
"A computer intermediary for interactive database searching. 2. Evaluation". In Journal of the American Society for Information Science, 1986, 33(6), P.357-364

23. Horowitz, G.L. and Bleich, H.L.
"Paperchase: a computer program to search the medical literature", In New England Journal of Medicine, 1981, 305(16), p.924-930

24. O'Leary, M.
"EasyNet: doing it all for the end user". In Online, 1985, Jul., p. 106-113

25. Conklin, J.
"Hypertext: an introduction and survey". In Computer, 1987, Sep. p.17-41

For a comprehensive listing of relevant articles see: Gibb, F. and Sharif, C. "Expert systems bibliography". In: Gibb, F. (ed). Expert systems in libraries. London: Taylor Graham, 1986, p.83-97.
Updates to this bibliography appear in Expertise, a newsletter on expert systems for information management.

PART III

ACTION RECOMMENDED

41 SELECTING DATABASES FOR INDUSTRIAL INFORMATION REQUESTS

1. COLLECTING INFORMATION ON DATABASES AND HOSTS

In order to facilitate selection decisions described in detail in Part I of this report, information on databases and hosts should be acquired and analyzed. The information must be supplied directly by database producers and host services.

Before planning any action to collect information of the above type, it should be considered first, what data should be asked for and who to acquire the information from. This is the purpose of this section.

1.1 Databases

1.1.1 What data?

The following information should be acquired for further analysis for each database:

- Database producer
- Subject field coverage
- Printed equivalents
- Time span of database
- Updating frequency
- Number of records per update
- Present total of records
- Online hosts where it is available
- Magnetic tape/CD-ROM version
- Prices (royalties, downloading, online search costs, CD-ROM subscription, allowances etc.)

In addition to this data which is readily available, some further information on the database service policy should be asked for. Some producers might be reluctant to supply this information. Producers should be contacted by explaining them the Project and to ask for their corporation. (However, part of the information is usually available in printed documents). They are for example:

- A detailed coverage of subject fields, priorities if any,
- Language coverage
- Type of documents, lists of journals
- Timeliness (a crucial problem)
- Selection of documents/papers for processing
- Indexing tools, guidelines, indexing policy
- Printed aids
- User education and communication, price reduction
- Other relevant information that might influence selection

1.1.2 Database producers to contact

The most important databases for industrial information retrieval on international level are:

Chemical Abstracts (CA Search), USA
INSPEC, UK
World Patents Index (WPI/WPIL), UK
COMPENDEX, USA
NTIS (National Technical Information Service), USA
IMIS, International Atomic Energy Agency, Austria
BIOSIS, USA
METADEX, USA
Sci Search, USA
CAB Abstracts, UK
Predicasts PROMT, USA
FSTA, UK-FRG
ISMEC, UK
Enviroline, USA
Chemical Industry Notes

At a later stage more databases can be selected for information acquisition and analysis.

1.2 Hosts (online services)

1.2.1 What data?

The following information should be acquired from hosts directly for subsequent analysis:

- Databases available
- Accessibility through telecommunication networks
- Restrictions of usage, if any
- Prices, pricing structure
- Time span and updating frequencies of databases
- Search system features: cross-database searching, split-file searching, proximity operators, postqualification by set numbers, inverted indexes, word/phrase indexing, print/display formats, expanding features, truncation, thesauri online, special command capabilities (like GET, ZOOM, STRINGSEARCH, REPORT, CLUSTER, MAP etc.) chemical search features (graphic search, substructure search, Registry Number crossover, availability of abstract), data fields, limiting capabilities, saving and executing strategies etc.
- System reliability
- Billing quality and delay
- Availability and features of SDI
- Mail delivery time of offline prints
- Electronic mailbox
- Downloading
- Online document ordering

- Training, communication with users, training files, help desk
- Documentation, printed aids

In addition to the above relatively readily available data, some other information should be asked for. They include expansion/specialization policy, gateways, possible price reduction, regional availability for distant locations, special educations aids (free usage, reduced price for students) regional offices, regional training courses and any other information that might influence selection.

1.2.2 Hosts to contact

The most important hosts providing databases for information searching in science, engineering and medicine fields for international utilization are:

Dialog Information Services, USA
Bibliographic Retrieval Services (BRS), USA
Data Star, UK and Switzerland
ESA-IRS, Italy
Pergamon Orbit InfoLine, UK and USA
STN International, USA, FRG and Japan
Telesystemes-Questel, France
DIMDI, FRG
International Atomic Energy Agency (IAEA), Austria

2. PROCESSING THE INFORMATION

It should be emphasized that the information asked from database producers and hosts be extracted and submitted to the above aspects by the producers and hosts, respectively, rather than analyzed and synthesized by us from regular printed documentation supplied routinely for their users. An extraction of data as indicated in Sections 1.1.1 and 1.2.1 can be the basis of the comparison of databases and hosts, respectively, for objective selection decisions.

The information obtained is recommend for preparing several tables and charts, with the objective to facilitate further efforts toward a more sophisticated selection guide system. They would be:

i) Table of data on the most important databases, containing many information listed under 1.1.1

ii) Table of most important hosts containing location, search system, number of databases online.

iii) A summary chart of decision guidelines comparing criteria to be considered for selection in:

- Preliminary selection of host(s)
- Selection of database(s) for a given query
- Selection of a host for searching the selected database(s) if more than one host is available.

Criteria are those listed under 1.2.1 except Search system features.

iv) A comparative evaluation chart of Search system features of various hosts as listed under 1.2.2 except proximity operators.

v) A comparative chart of proximity operators os some selected hosts.

vi) Comparative guidelines of database search prices at different hosts and their pricing structures.

3. EXTENDING THE GUIDELINES

3.1 Gateways

A brief description of the gateway service providers would be very useful for developing countries, as this service could solve many of their selection problems which were a result of the enormous number of databases available for online access. Very few people in developing countries are aware of the number of gateways operating in the world today and how can they make use of them.

Evaluation and analysis to these gateways would be needed some of the current gateways could be very useful for end users who have simple enquiries, i.e. simple search strategies. But in cases of sophisticated or complicated searches, gateways will not be useful and in these cases access to databases directly through the hosts would be more appropriate. Advantages and disadvantages of gateways are to be clear to database searchers.

3.2 Software packages

Guidelines to software packages available that could be useful in accessing, retrieving and manipulating data would be useful.

Examples of packages to be described could be Micro-Disclosure, Pro-Search, Dialog link, Personal Bibliographic System (PBS), Head-Form etc. Other programs that could be used for data handling as DBASE III, Wordstar, Infostar, Lotus etc. could be listed with a brief description of their usage.

3.3 Infrastructure Requirements

Guidelines to the infrastructure requirements for accessing databases would be of great help to professionals who are planning for online database access activity.

A brief listing for requirements could cover:

1. Hardware requirements, PC's. Modems, Printers
2. Software requirements
3. Communication requirements (networks, protocols)
4. Manpower
5. Agreements, or host subscriptions

3.4 Procedures and Forms

Guidelines to steps and procedures related to all activities concerned with the access and retrieval process would be useful.

Examples that could be covered in the manual are:

- Process for defining users needs
- Process for selection of databases to be accessed
- Setting up the search strategy
- Use of supporting tools
- Search and retrieval
- Data manipulating and storage
- Evaluation of results

Suggestions for samples of forms to be used for inhouse activities would be very useful for the administration of:

- User enquiry form
- Enquiry Analysis form (to be used by the information specialist while studying and analyzing the enquiry and selecting the relevant databases)
- Evaluation forms.

3.5 UN Databases

Guidelines to UN databases related to the needs of the industrial users would be useful.

What are the terms and conditions for accessing or acquiring UN database in machine readable formats for internal use.

3.6 Databases on CD-ROM

A brief description on the use of databases available on CD-ROM would be useful. What is the economy of its usage and on what basis a center could take a decision to use CD-ROM rather than

going online. This technology is quite new to developing countries.

4. ESTABLISHING THE FEASIBILITY OF AN EXPERT SYSTEM ON DATABASE AND HOST SELECTION

Once the suitability of an expert system solution has been confirmed, the feasibility of the system must also be established. Three major constraints apply to applications for which expert systems are suitable:

(a) The source of expertise
(b) The problem characteristics
(c) The state of available technology

Constraint (a) will depend on the availability of cooperative experts and the extent to which they are able to explain their knowledge. It is not anticipated that locating suitable experts will be a problem. However, some areas of expertise, specifically, those related to subject representation and natural language understanding, will prove to be difficult to pin down. Others, such as those related to communications, database coverage and content, command languages etc. will, on the other hand, be relatively easy to ascertain. A modular approach is therefore recommended utilizing a matrix of functions ranked in terms of achievability, necessity, investment and availability of alternative solutions. This matrix should be drawn up by the working group using knowledge of technological solutions and user requirements. The extent to which conventional solutions using existing software (communications packages, windowing environments, natural language front ends etc.) can be interfaced to reduce development times and costs should also be investigated.

Constraint (b) is related to how well bounded the problem domain is, how easily problems can be solved and hence explained by a human expert, and how reliable or time-dependent the data stored in the expert system is. This is a more significant constraint as problem domains in information retrieval are characterized by fuzzy boundaries, ambiguities, context-dependent variables and cross disciplinary solutions. The EasyNet solution has been to take a generalist approach and to guide the user to an appropriate database using a hierarchy of options. This leads to a low level of recall. Careful analysis of the population of queries from industrial users will be essential to ensure that the appropriate balance between recall and precision can be achieved. Fortunately industrial enquiries are likely to contain highly concrete concepts which will make this task less difficult than it would be for disciplines which utilize soft vocabularies.

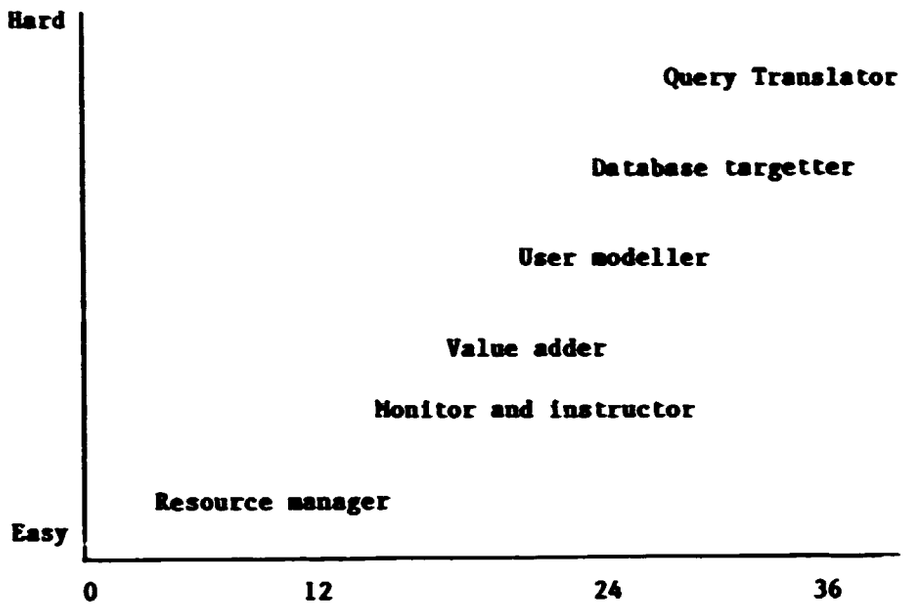
Constraint (c) relates to the limitations of hardware and software but also to the availability of technology to end users. It is the latter of these two factors which is of prime importance for this project and underlines the importance of carrying out user surveys before embarking on a technological journey for which there will be few paying passengers. The size of the knowledge bases within the proposed system will almost certainly require hard disk based PCs and, ideally 80286 or 80386 based system units. It will be possible to run expert systems on less advanced systems but the slower speed of operation (and hence higher communications costs) may make them unsuitable for field use.

5. PROTOTYPING AN EXPERT SYSTEM AND BEYOND

Prototyping an expert system is the precursor to developing a complete system. Paper-based systems utilizing flow charts and decision trees are a cheap and effective method of dealing with a few basic rules and parameters. However, once the system grows beyond this stage it will have to be properly encoded within a suitable software package. Development can be done on PCs but for full flexibility the workstation approach is advised. Experience at the University of Strathclyde has shown that continual updating and compiling of object code on a PC becomes increasingly expensive: what takes minutes on a PC can be achieved within seconds on a workstation.

When a prototype has been developed it must be tested in order to validate the rules and resolve any conflicts or ambiguities. The validation phase will require further dialogue with both information experts and end users. Amendments to the system will usually be necessary and it is therefore important to maintain a paper-based map of the knowledge base to ensure that patches and alterations do not cause short-circuits or other new conflicts in the existing system. If substantial changes are required it may be necessary to re-implement the system from the bottom up to avoid an inefficient and tortuous network of rules. Once the prototype has been tested the next stage will be to gradually refine this system by adding new knowledge on a modular basis. Field testing will then need to take place at strategic points in the growth path.

Timescales for non-trivial expert systems are notoriously difficult to establish as they will depend on how well defined the problem domain is found to be. As has been noted above most commentators suggest a minimum of three to four man years of effort over perhaps 12 to 18 months. The roles suggested in section 5 can be ranked in terms of difficulty/effort as follows:



6. Conclusion

It is recommended to start with the collection of the necessary data from database owners and hosts, as proposed in Chapter 1 of Part III. Guidelines could then be prepared for the selection of the most appropriate databases in specific fields of interest to Unido, using the methods and preparing the proposed in Chapter 2. This effort could be harmonized with Unido's programme regarding the development of specific industrial sectors, such as electronics, fine chemicals etc.

When one or two such concrete guidelines are ready, supplements could be prepared as proposed in Chapter 3.

In a third phase detailed information could be stored in a microcomputer an experimental knowledge based system could be established regarding selected, high priority fields of industrial information.