Design of an OPAC Database to Permit Different Subject Searching Accesses in a Multi-Disciplines Universities Library Catalogue Database

Maristella Agosti* and Maurizio Masotti+

* Department of Electronics and Informatics, University of Padua
Via Gradenigo 6/A, 35131 Padua, Italy
e-mail: agosti@ipdunivx.unipd.it
+ Computer Centre, University of Padua
Via San Francesco, 35121 Padua, Italy
e-mail: masotti@ipduni1.bitnet

Abstract

This paper presents searching approaches and user interface capabilities of DUO, an Online Public Access Catalogue (OPAC) designed to permit the users of three Universities of the North-East of Italy different subject searching accesses to the co-operative multi-disciplines library catalogue database.

The co-operative catalogue database is managed by one of the software systems developed under the Italian National Project for library automation: the SBN project. Since the SBN database has not been designed to be efficiently accessed for end-user searches, the DUO database has been designed to avoid duplication of the SBN database data and to be usable for making efficient subjects accesses to the catalogue documents. The DUO design choices are presented, in particular the main choice of designing a "virtual" document that corresponds to each SBN document and that has unstructured data usable for subject search purposes.

The paper presents a new kind of user-OPAC dialogue that makes available to the user different search approaches and on-line dictionaries. In particular the user during the interaction with the search tool can represent his information needs with the support of interface capabilities that are based on retrieval path history, and words and codes on-line dictionaries.

DUO is the first Italian OPAC that has been made openly available to users of universities and research institutions. For this reason, it is also the first time that OPAC log data is going to be collected in Italy. This work mainly intends to make a modern OPAC available to the users of a SBN catalogue database, but it is going to permit also to build up a knowledge on OPAC usage in Italy.

1. The Multi-Disciplines Library Catalogue Database

This paper presents the database design choices and characteristics of DUO, an OPAC [Hildreth, 1985] designed to permit the users of the Universities of Padua, Venice, and Verona different subject searching accesses to the co-operative multi-

disciplines library catalogue database.

At present the co-operative multi-disciplines library catalogue database of these three universities, all in the Veneto Region that is situated in the North-East of Italy, is set up by forty-three university libraries. These libraries are Faculty, Department and Institute libraries of the different subjects of study that are addressed in these universities: humanities, economics, sciences, engineering, statistics, languages, law, medicine, pharmacy, etc. The library system of the three universities is really complex with more that 150 libraries, 2,000,000 books, and 20,000 serials.

The co-operative catalogue database has been created and is currently managed by one of the software systems which have been designed and developed under the Italian National Project for library automation: the SBN project (SBN stands for "Servizio Bibliotecario Nazionale") [ICCU, 1985; ICCU, 1987]. SBN software systems have been developed to implement the same specifications in order to give the same library automation functions to all librarians of the libraries participating in the project.

The SBN software systems have been primarily designed to interact with a professional user: the librarian. Due to this, the functions which have been made available are those usually carried out by the librarian himself, e.g. cataloguing, acquisition, serials control, and subject indexing.

In mid-1989 a two-year project for planning and development of an experimental prototype of an OPAC for university users was set up. At the beginning of June 1991 the DUO prototype was available to university users for experimental use. At present log and other types of data concerning the interaction between user and system are being collected in a systematic manner in order to monitor the prototype and to consolidate it into a tool satisfying end-user search requirements.

2. The Design of the DUO Database

2.1 Design highlights of DUO

The DUO project was organised in three main parts [Agosti et al., 1989]: 1) identification of end-user categories and analysis and design of end-user re-

quirements [Moressa, 1989]; 2) design of functional specifications of an OPAC for university library end-users; 3) design and development of the DUO prototype. General design and functions of DUO have been presented in [Agosti & Masotti, 1992]. The architecture of the DUO database that is central to this paper was focused on making OPAC information retrieval [Salton & McGill, 1983; Van Rijsbergen, 1979] capabilities as efficient as possible. For this reason it was necessary to program the database and the software to work efficiently for words and terms access operations on the textual data [Blair, 1984] of the SBN database.

It is important to bear in mind that one important target of the DUO database design was to avoid data duplication logically integrating SBN and DUO databases. This is an innovative aspect of this project, because at present it is quite common to see OPAC implementations operating on specific databases which duplicate that data of the cataloguing database which is considered informative for the end-users.

It is imperative to recall here that the main feature of the SBN database is to contain data for administrative and cataloguing purposes. Because of this, all query access operations that have been implemented by the SBN software permit a pre-coordinated access only. This means that it is not possible to obtain textual attributes through single words or of single word expressions connected by Boolean operators (post-coordinate access [Lancaster, 1979]), as in the information retrieval (IR) process depicted in Figure 1.

The SBN catalogue database contains catalogue representations (administrative and cataloguing data) of textual documents, the construction process of this database is represented in Figure 2. Records in the SBN database lack of abstracts and the subject indexing is sparse, because it is made by different trained librarians in libraries of a wide variety of different disciplines and subject areas. Since document abstracts, indexes, or any other form of semantic representation of document contents are not included in a systematic way in the SBN database, it was not possible to base access operations on a "by content" principle designed for
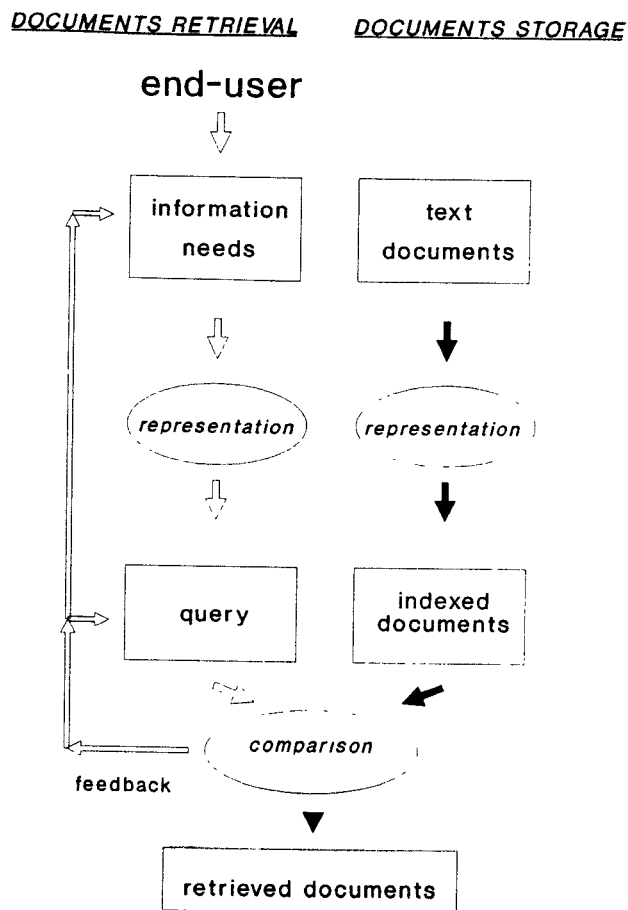
the OPAC on this data.

level of specification of the words of the query.

DOCUMENTS RETRIEVAL   **DOCUMENTS STORAGE**

Figure 1: Overview of Information Retrieval Process
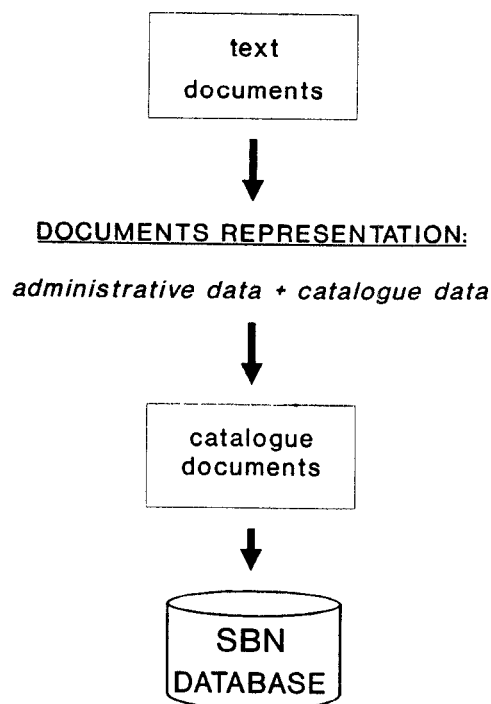
Figure 2: SBN Catalogue Database Construction

Moreover, the SBN database is not only multi-disciplines and multi-lingual, but it also contains cataloguing data of different specification levels. Because the SBN cataloguing approach supports a range of four different cataloguing levels: from a very poor to a very rich one also in content representation of the document. This means that the database content is not homogeneous. A consequence of this fact is that the design of the search mechanisms that are based on browsing and navigation is more difficult than for a database with all data of the same level of specification. The vocabulary contains words of different levels of specification and it is not possible to construct the system answers in a way that takes into account the exact

Among the items available in the SBN schema, only these have been deemed useful for construction of the DUO database:

- the title;

- the "nature" (or type) of title;

- the terms of the controlled vocabulary, developed by the National Library of Florence, directly related to the document title;

- classification subject headings directly related to the title;

- name(s) of document author(s);

- document language;

- the country in which the document was published;

- date type and date of publication;

- the type of publication; and

- the library which owns the document.

2.2 The "virtual" document

The data corresponding to the items of the SBN schema are typical structured data managed for administrative and library automation operations. The information items are connected through a network of connections that permits:

- the construction of sets of items corresponding to each document of the database, different lists (e.g. authors list), different types of catalogues (e.g. card catalogues), and

- the use of the data by the librarians for many other library automation operations.

Thus, the SBN database has not been designed to be efficiently accessed for end-user searches. Since the SBN database does not contain specific data for word searching in the textual document attributes, it was necessary to decide how to make the necessary data available in the DUO database.

A new solution has been proposed: the design of a "virtual" document that corresponds to each SBN document and contains the structured SBN document items transformed into unstructured data usable for subject search purposes. Each virtual document is represented as a flat record containing these unstructured data items and the connection with the original complete SBN document.

The unstructured data items are organised in five segments for each document: title/s, author/s, publisher/s, subject heading/s, and classification heading/s. Each part of the segments can be used to implement general and specific searches based on Boolean operators, proximity operators, and any other information retrieval technique that makes use of words in sentences. The integrated collection of virtual documents constitutes the DUO database that needs to be accessed for search purposes together with the SBN database that contains the complete data on documents of the collection.

2.3 The DUO database

We decided to avoid a total duplication of the SBN files and we physically designed the DUO database to contain only the vocabulary and the access lists that are necessary to make efficient access

to the textual attributes of the SBN (or cataloguing) database, as depicted in Figure 3. SBN system and DUO prototype make use of three software development tools that are property of the SAG Company: the ADABAS database management system, the NATURAL language, and the TRS product.
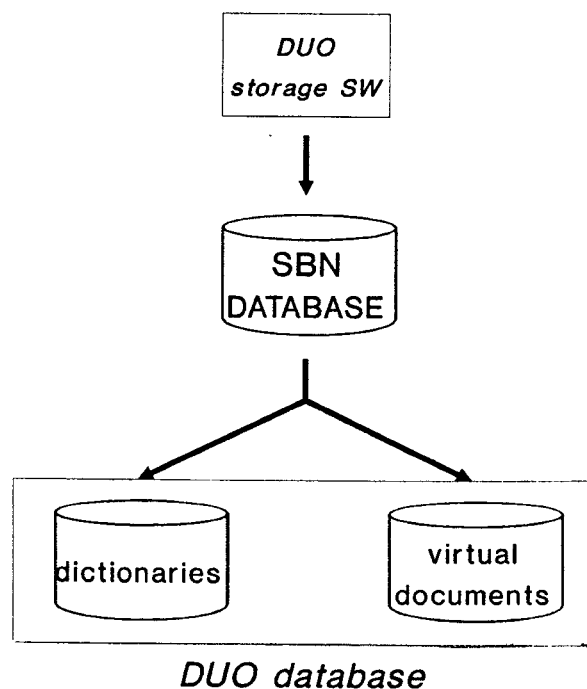

Figure 3: Construction of the DUO Database

One important engineering problem to solve was that concerning updating of the DUO database which needs to be done in sequence with the SBN database. We adopted the solution of accessing the DUO database for search purposes only; after a query has been formulated and solved, the DUO prototype accesses the SBN database in order to make the bibliographic data available to the end-user; in this way the bibliographic data displayed to the end-user is always consistent with the most recent SBN database updates. Figure 4 gives a schematic representation of the DUO process of

querying.

query
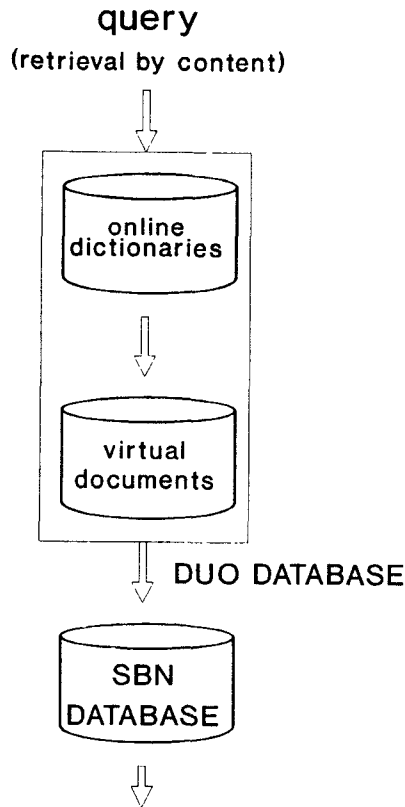### (retrieval by content)

Figure 4: DUO Query Processing

The DUO vocabulary contains all the words which occur in the virtual documents. The information on the diverse segments where a word occurs is associated to the word in the vocabulary together with the word language code and related synonym(s). The synonyms must be manually defined by a really skilled librarian, because when defining a synonym it is necessary to take note of the fact that the prototype works in a multi-lingual and multi-disciplines reality.

3. Searching Approaches and User Interface Capabilities

3.1 Information Retrieval Process and Auxiliary Data

Figure 1 contains the usually used schematic representation of the information retrieval process. See for example [Croft, 1990].

An information retrieval tool is able to operate and to answer an user query when the documents of a real collection have been represented through an indexing language and their representations have been stored as "indexed documents". Indexing languages, also called auxiliary data, consist of a *semantic structure* filled with a *collection of words* or terms which depends upon the domain concerning the collection of documents managed by the information retrieval application.

The auxiliary data structure identifies the semantic relationships between words or terms: it is this structure which permits implementation of data retrieval operations *by content* in contrast to the usual data retrieval by value, supported by traditional data processing applications. Many different paradigms and techniques can be used for the design and management of the structure of auxiliary data: object-oriented, knowledge-based, etc. All the approaches have the same basic purpose: to implement a structure which can be used to find semantically related terms; by using related terms which directly concern the user's specific information needs it is possible to launch a query to the database of the collection so as to retrieve all relevant documents.

The auxiliary data items represent the vocabulary to be consulted by the information retrieval tool. The auxiliary data items are associated to each document by means of an indexing process in order to represent its semantic content in the database. Each auxiliary data item is assumed to describe the document content only to a certain extent, neither completely nor uniquely.

Different structures have been proposed to manage the collection of terms that are used by an information retrieval system to represent the domain of pertinence of the collection of documents represented and managed together with the terms that populate the structure. The simplest structure that is used to represent and manage the collection of terms is an alphabetical list of significative words or terms; a more complex structure can be a classification scheme, semantic network and thesaurus.

3.2 Differences between Cataloguing and Indexing Process

When the collection of real documents is represented with a *cataloguing process* instead of a *indexing process*, it is not possible to have the output of the representation anymore as rich and equal to that of the information retrieval process depicted in Figure 1. During the cataloguing process the subject indexing is sparse and any other form of semantic representation of document contents is not included in a systematic way in the catalogue database. The representation of each document is poor for retrieval by content purposes and the auxiliary data structures that can be used are alphabetical list of significative words or terms only.

We have considered imperative to make explicit and available to the user for browsing and direct extraction all the alphabetical lists of words that are produced during the construction of the DUO database. Different approaches are under investigation to support in a better way the user during the representation of his information needs. One of these approaches manages the semantic structure as a schema of concepts which can be used by the final user as a frame of reference in the query formulation process [Agosti et al., 1991a&b]. By now this approach has been shown as a valid one when the collection of documents is pertinent to only one discipline. In the case of our OPAC project the collection of documents is a multi-disciplines ones and it is not possible to envisage a specific representation of the content of each document using one semantic structure populated by only one vocabulary.

3.3 A New Kind of User-OPAC Dialogue

We have decided to propose a new kind of user-OPAC dialogue, that makes usable and available to the user different search approaches and on-line dictionaries.

During the interaction with DUO, the user can represent his/her information needs in the same way as he does to interact with a traditional Boolean retrieval language, or with the support of interface capabilities that make transparent and usable by him in any moment of the construction of the query (i.e. the information needs representation process) different on-line dictionaries. DUO supports the concurrent use of these dictionaries to satisfy different information needs.

The DUO prototype makes explicit to the user two different techniques for end-user interaction: the first way is intended for inexperienced end-users who have never used the tool before or who will use the system only occasionally; the second type of interaction has been designed for end-users who use the system regularly or already have experience in on-line database query operations. In both ways the end-users have direct availability to innovative retrieval aids and help facilities of the DUO prototype.

In the first interaction method the end-user is guided through his or her query formulation and searching by a sequence of self-explaining menus with fill-in masks, and pop-up menus that contextually appear when they can be helpful for the user. The second method lets the user employ Boolean operators for query formulation and searching together with the use of the on-line dictionaries facilities. The separation between the two interaction techniques has not made sharp and definitive, so the user can mix the two techniques to reach the mixture he prefers. Contestual help is available in the two different interaction techniques together with the possibility for the user to use "functional" keys instead of commands that need to be remembered.

The innovative kind of user-OPAC interaction is based on new retrieval aids that can be used in the query formulation process. Before introducing this innovative interaction it is necessary and useful to introduce the way the retrieval process has been implemented.

3.4 DUO Retrieval Process

The DUO retrieval process consists of two different steps: the initial search and subsequent search operations. While the initial search asks the end-user to make a "blind" query formulation, the subsequent search operations start after the end-user has examined the documents retrieved in response to the previous query.

The initial query formulation reveals three pos-

sible searching techniques, depending on the kind of information which is needed by the end-user and also on the data he or she has concerning the documents he/she wants to find:

a) searching through a specific document attribute;

b) searching for a set of documents which all have the same value of a specific document attribute;

c) subject searching.

One of the scopes of our prototype is to investigate new solutions for subject searching in Italian University libraries. With this aim in mind, in the prototype's current version, though it is possible to perform search operations on a specific document attribute, the particular structure of the DUO database allows search operations using all the document attributes of the cataloguing description concurrently, and the cataloguing description can contain any semantic data.

The end-user can ask the system to find one or more words anywhere in the textual segments within the DUO database: this particular retrieval behaviour is typical of someone who has a subject in mind on which he or she needs information, but who does not know the specific title of an actual book or publication dealing with that subject.

The logical textual segments which have been created for subject searching using all the document attributes of the cataloguing description concurrently are:

- the document's author's name;

- title;

- the controlled vocabulary developed by the National Library of Florence which is used in subject indexing; in the foregoing the expression: "subject headings" will be used to refer to this controlled vocabulary;

- classification subject headings, and

- the publisher's name.

The *document author's name* allows free text searching. The option for word searching on the author's name is particularly important for author's corporate names. The cataloguing rules defined in the SBN database consider the pseudonyms and the acronyms of an author's name in a special way: they are linked to the author's name in the authority file but they are not linked with the documents related to the author. The OPAC database puts together in the same textual segment all the words which form the original author's name, the pseudonyms and the acronyms. The DUO prototype is able to find all the author's works even if a query formulation contains a pseudonym or an acronym.

In the SBN database the *titles* are related to each other in a hierarchical structure or at an equivalent level (i.e. titles which are related to each other), depending on the nature of the title. Let us give an example: a monograph can be related at a higher level due to the monograph title's relation to the title of the collection it belongs to; the same monograph can also be related to a title in its original language version. Among all these titles only the monograph title is linked to the physical book it represents and to the circulation data (data concerning its availability, etc.). The prototype again gathers all the words which form the titles related to the one linked to circulation system and puts them in the same textual segment. In this manner the hierarchical order is lost but the set of words which could directly lead to data in relation to availability of the actual book is enriched. This operation is justified also by the fact that these hierarchical links which have some meaning for the library staff are generally unknown to the end-users of a library.

The *subject headings* also have hierarchical links. As for titles, these links set on different levels are smoothed down onto same root level of the tree of relationships. The same operation has also been done for the *classification subject headings*. These links differ from those of the titles, because they mainly depend on the subject matter of the document and the descriptions relating to contents in the hierarchy become more and more specific as we move down to the lower levels.

The library staff of the library system at the three universities is not forced to adopt a common, sole, subject-headings and classification scheme which would be unique for all the libraries, but there is sufficient freedom of choice of a method which is semantically closer to the specialised documents of each library. This can be an advantage for the end-user who is searching in a specific library collection, but it becomes a disadvantage when an end-user wants to run a subject search throughout the entire collections of one or all of the universities in the Veneto Region. The solution which has been implemented in our prototype involves use of the classification subject heading terms just for their very meaning, but providing the DUO database architecture with multiple logical segments in order to distinguish different classification systems.

The *initial search* can be constructed specifying in the query some precise values of document attributes, such as data concerning publication, language of publication, or the particular library in which the end-user wants to look for the documents.

The initial query can also be constructed in an entirely different way in order to make it possible to set up the database for subject searching operations combined with a deterministic query on one or more specific document attributes. This retrieval method is not normally used by end-users in initial search operations, owing to the fact that it supposes that the end-user knows exactly what he or she is looking for. This possibility can however become very useful in *subsequent search* operations: in fact, the structured attributes of cataloguing data, language and library can be used by the end-user as "filters" to reduce the number of documents retrieved by the initial or previous query.

If the *initial query* retrieves one or more books, the prototype lists the results in two steps:

1. concise descriptions of all retrieved documents;

2. extended description of each document in the set.

The end-user can select one or more documents from the concise list (step 1) and obtain the screen display of all pertaining data (step 2): information concerning the complete bibliographic description and the description of related documents, the subject headings and the classification subject headings, the names of all the libraries which have the book or that have just ordered it can therefore be obtained.

The prototype gives many different print options, in particular a concise data printout of all or some of the retrieved documents which is made following specific bibliography construction rules.

3.5 User-DUO Interaction

The overview of the innovative user-DUO interaction is depicted in Figure 5. The user has available in a concurrent way different search approaches together with the possibility to import in the query data of different on-line dictionaries directly extracted from the dictionaries and by the DUO tool inserted in the user's query. This means that the user during the interaction with the search tool can represent his information needs with the support of interface capabilities like the on-line dictionaries, and retrieval path history.

The *on-line dictionary* feature has been provided for in order to support the end-user in the formulation of the query and during browsing, because one *on-line dictionary* gives a further possibility of finding exact access points within the vocabulary of words and in the list of pre-defined codes.

Five on-line dictionaries of words are at present available to the user, the alphabetical list of words extracted from:

- document titles,

- author's names (person names, e.g. John Smith, and organisation names, e.g. ACM and/or Association for Computing Machinery),

- subject headings,

- classification subject headings, and

- publisher's names.

Two on-line dictionaries of codes are also available:

- the alphabetical list of the names of languages of the documents of the collection, and

252

- the alphabetical list of names of libraries participating in the project together with their codes.
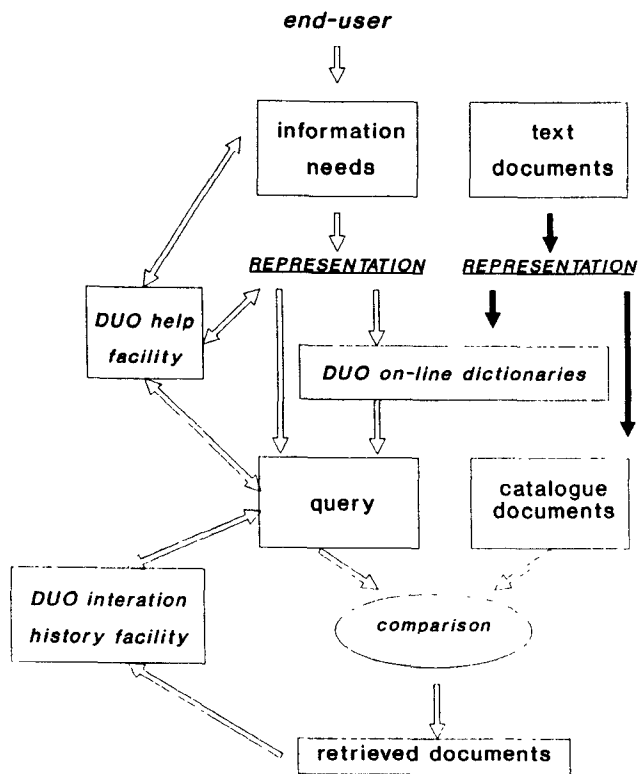

Figure 5: User-DUO Interaction

The language dictionary makes available the selection of the documents that have been written in one or some specific languages, and the libraries catalogue can be used to select only one or some of the libraries participating in the project. The user can use the on-line dictionaries having the possibility to import terms, codes or names of libraries.

When an end-user selects a word in one specific dictionary, the word is first searched for within the selected segment: if it is not found in that dictionary, the DUO prototype automatically checks if it exists in any other textual segment, if so, the relevant dictionary is then displayed to the end-user.

When the word is entered by the user of the system, a stemming algorithm performs an automatic truncation of the word which is sought and lists the set of words beginning with that stem, each one with the number of occurrences in the titles. If the end-user selects one or more words from the dictio-

nary, the system formulates the equivalent query in the syntax accepted by the system.

The *retrieval path history* feature provides information on the end-user's individual interaction history with the library catalogue database. This feature is available at any moment during the interaction: it gives a text display of previous queries and the number of corresponding titles which have been retrieved. The set of documents that has been retrieved in response to a query can be used in the construction of subsequent queries.

The prototype design tends to facilitate the user-DUO interaction; due to this feature, except for the difficulty of understanding the meaning of Boolean expressions in the query formulation, all the options are easy to use and they are always followed by help messages.

As it has been previously underlined, the DUO system must be made directly usable by untrained and inexperienced users. Help facilities are therefore very important, because they have to teach end-users how to use the system while they are using it.

Two types of on-line help options have been provided for in the prototype:

a) specific contextual help;

b) general help on prototype use and features.

Many help messages also give examples of use of the prototype's facilities. Also some sort of monitoring on incorrect attempts of the end-user has been implemented in the prototype; in fact, after a series of subsequent errors made by the end-user during interaction with the system, the help facility gives advice for an alternative search strategy.

4. Conclusions and Further Developments

The prototype has been made available to users starting from June 1991. DUO is openly available on Internet Network through a public login. Using Telnet (TCP/IP) facility, host number 147.162.1.2, public login "OPAC" (free of charge), it is possible to use and query DUO. At present the user/system interaction is using the Italian language: query language operators, contestual help, and messages.

253

The possibility for the user to choose the interaction language between Italian and English is going to be released and made available to users during 1992.

General impressions and initial logging data are currently being collected for subsequent analysis and refinement/development of the prototype. Evaluation tools of the suitability of the features provided in the prototype and efficiency of the chosen software tools are presently under way.

In parallel with the planning of the prototype evaluation tools, we are at present designing new features to implement in the next version of the prototype. One of these new features will be automatic selection of alternative retrieval paths when the set of retrieved documents is empty. Also for "subsequent search" operations it will be possible to get the system to find other documents similar to that (or those) just retrieved. Another feature is a ranking algorithm that operates using information on a "interest area" specified by the end-user and/or the main discipline or subject area of the library of reference for an end-user.

Acknowledgements

References

[Agosti & Masotti, 1992] M. Agosti, M. Masotti. "Design and functions of DUO: the first Italian academic OPAC". In: H. Berghel, E. Deaton, G. Hedric, D. Roach and R. Wainwright (Eds.), *Proc. of the 1992 ACM/SIGAPP Symposium on Applied Computing*, Vol. I, 308–313.

[Agosti et al., 1989] M. Agosti, M. Masotti, A.M. Moressa. *Linee di progetto*. Progetto di un prototipo di catalogo in linea per l'utente finale (OPAC) per l'Università - Base dati del Polo SBN Universitario Veneto, novembre 1989, pp. 13.

[Agosti et al., 1991a] M. Agosti, G. Gradenigo, P.G. Marchetti. "Architecture and functions for a conceptual interface to very large online bibliographic collections". *Proc. of RIAO '91 Conf. "Intelligent Text and Image Handling"*, Barcelona, Spain, 2–5 April 1991, Vol. 1, 2–24.

[Agosti et al., 1991b] M. Agosti, R. Colotti, G. Gradenigo. "A two-level hypertext retrieval model for legal data". In: A. Bookstein, Y. Chiaramella, G. Salton and V.V. Raghavan (Eds), *Proc. 14th ACM-SIGIR Int. Conf. on Research and Development Information Retrieval*, Chicago, USA, 1991, 316–325.

[Blair, 1984] D.C. Blair. "The data-document distinction in information retrieval". *Communications of the ACM*, 1984, 27 (4), 369–374.

[Croft, 1990] W.B. Croft. "Lecture on 'Survey of basic retrieval models' ". Notes on *Proc. of 1st European Summer School in Information Retrieval*, AICA, Bressanone, Italy, 1990.

[Hildreth, 1985] C.R. Hildreth. "Online public access catalogs". *Annual Review of Information Science and Technology (ARIST)*, 1985, Vol. 20, 233–285.

[ICCU, 1985] Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU). *Il Servizio Bibliotecario Nazionale*. Roma, 1985.

[ICCU, 1987] Istituto Centrale per il Catalogo Unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU). *Guida alla catalogazione nell'ambito del Servizio Bibliotecario Nazionale*, Vol. 1, Roma, 1987.

[Lancaster, 1979] F.W. Lancaster. *Information Retrieval Systems: Characteristics, Testing and Evaluation* (2nd Ed). New York, 1979.

[Moressa, 1989] A.M. Moressa. *La fase di analisi dei requisiti utente - Relazione sulle interviste raccolte*. Progetto di un prototipo di catalogo in linea per l'utente finale (OPAC) per l'Università - Base dati del Polo SBN Universitario Veneto, settembre 1989, pp. 15.

[Salton & McGill, 1983] G. Salton, M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[Van Rijsbergen, 1979] C.J. Van Rijsbergen. *Information Retrieval* (2nd Ed). Butterworth, 1979.