



Número 1, juny 1998

Improving subject retrieval: user-friendly interfaces and effectiveness¹

Constança Espelt

Escola Universitària *Jordi Rubió i Balaguer* de Biblioteconomia i Documentació
Barcelona

espelt@fbid.ub.es

1. Introduction

During the past decade our society has experienced a revolutionary increase in the extent to which the general public has access to electronic resources. This phenomenon has been characterised by:

- International trends leading to unification.
- Increased availability of personal computers at home in all developed countries.
- Wide-spread use of electronic media in our daily life, and not just in the work environment: cash dispensers, cinema and theatre tickets, computer games, electronic mail, ...
- Diversity of multimedia products: encyclopaedia and dictionaries, cook books, children's books, etc.
- Specific environments dominate the market: WWW browsers, notably Netscape and Microsoft Explorer, have effectively become standard interfaces.

Library and information professionals have welcomed this trend with satisfaction and high expectations, as a chance to enlarge the community of information users. The efforts involved in incorporating these new facilities into traditional libraries have been considerable. Librarians are exploring, evaluating and categorising Internet resources in specific fields in order to extend their holdings coverage.

The users' expectations of information centres and services have also changed: they increasingly expect to access library services from their usual work place (office, laboratory) or from home and they are aware that technology has made a great deal of information available to them. But user perceptions do not necessarily correspond with those of professionals.

2. User satisfaction versus information literacy

With all this technology permeating the society, the basic goal must be that all those who need information should be able to access it directly for themselves.

We know that increasing numbers of people are using electronic sources, but we do not have sufficient data on how successfully they are using it and on how well they understand the systems.

We can assume that there is a high level of user satisfaction in searches for previously known documents: author/title searches in OPACs; updated documentation retrieved from the producing institution; product demos obtained from direct contact with companies; symposium programmes available on the Internet, etc. But the ease of obtaining this information contrasts greatly with the difficulty of getting good results in subject searches.

If we take into consideration how well users understand the system, the question will lead us to the concept of information literacy. Users' information literacy--defined as the ability to know when one has a need for information, to identify information needed to address a given problem and to locate and evaluate information ([Snaveley, 1997](#))-- will be a decisive factor in the evolution of electronic information. As regards information literacy, there is no uniformity between countries, groups or individual users. In Spain, although very few studies of users in networked environments have been undertaken, the general impression is that the availability of access underscores the lack of information culture on the part of users, even in academia. We would agree with general studies indicating that users are enthusiastic about the increased access that provides them with a greater degree of independence and are pleased with the results they obtain and with the easy-to-use systems at their disposal. And if we look no further, the information world would appear to be a happy place indeed. However, more detailed studies reveal particular tendencies that should serve to alert us. The users' understanding of the systems they are using is really quite low: the results of a recent user survey in a Catalan academic library indicate that users do not understand the language of our catalogues. For example, 75% do not know the meaning of "keyword" ([Andreu, 1996](#)).

The enhancement of available resources and searching tools has to be based upon evaluations of the quality and use of these resources, with special attention to user studies. But the contributions that information professionals can make must take into consideration the negative consequences of their users' low level of information literacy.

3. Elements affecting user success in information retrieval

The need for locating information by subject goes back a long way, but has grown even more with the automation of library catalogues ([Larson, 1991](#)) and with the explosion of Internet-based resources as well, as people are performing high volumes of subject searching ([Taylor, 1995](#)). The above-

mentioned problems concerning subject retrieval start in the very first stage: the information request. Users have a tendency towards limiting themselves in their requests. They are frequently conditioned by their current knowledge, asking only for the kind of information they already know. On the other hand, they often confuse what they need with what they want, and they want much more than they really need, but request nothing about what they do not already know.

Leaving aside the problems derived from the differences between real information needs, the level of consciousness of these needs, and their expression in a given language, we will examine the elements which have a direct influence on the effectiveness of subject retrieval: database context, indexing procedures and user interfaces.

Database context

The database context greatly influences the effectiveness of search results. We can differentiate two aspects: source selection and searching context.

Database selection is a decisive factor in the success of the search. Even with the best search strategy we cannot obtain good results in a database that is not suitable for a topic. The selection of the appropriate search context for full coverage of a topic is being simplified by companies which provide access to databases from different producers, journals and relevant web sites for a community of users. Systems like Engineering Information Village, Chemport, Chemweb, and Health-Gate are intended to integrate in an easy-to-use manner, the information sources that engineers, chemists or physicians respectively may need.

The uniformity of the context simplifies the search process. The common characteristics of the specialised language within a narrow subject area allow for quite good results even in free text searching. Similarly, the fact that a database limits the types of document formats covered, contributes to a coherence in the indexing.

To a large extent user searches take place in networked environments that allow for different options for different search contexts: CD-ROMs, OPAC, union catalogues and selected World Wide Web resources. These systems differ in content, structure and capabilities. Many questions arise from this fact: Are users aware of the differences? And even more, do they know which is the best option for a given search? How can we effectively guide our users in the selection of the appropriate search context?

The coherence provided by uniformity is generally limited to CD-ROMs and traditional online services. OPACs have extended their coverage, with sound and video recordings being typical materials now in all types of libraries. And it is not rare for default screens to lead the user to search the library network holdings, and not just the holdings of the particular library where he or she is physically at that moment. But the widest variety of document types is found in the Internet. Search engines such as Altavista provide equal and uniform access to all kinds of information: books, journals, web pages of institutions, commercial entities and even individuals, bibliographic information, art and fiction, restaurant menus, etc.

Indexing procedures

The debate concerning controlled and natural indexing languages has generated a considerable amount of literature since the 60s. Jennifer Rowley ([Rowley](#), 1994) divides the history of this debate into four eras but, she concludes, "the debate still rages". Mainly for economic reasons, the question of whether or not subject heading lists, thesaurus and classification systems have any meaning is often posed. Along the same lines, much research is focussing on improving the quality of retrieval systems based on automatic indexing and the application of knowledge-based expert systems. But free text and controlled vocabulary should not be seen as antagonistic retrieval techniques. For the professional searcher, and also for end users who have received some training, the best way to achieve effective retrieval continues to be a combination of both approaches--natural and controlled—while taking into account the nature of the search and the searching environment.

Traditional databases are maintaining their thesaurus and classification systems, such as MeSH, AGRICOLA, ERIC or INSPEC, created for a specific abstracting service. Searches in CD-ROMs or traditional online services allow the controlled language to be enriched by free text in identifiers and abstracts fields, and user interfaces and command search languages are designed to take advantage of controlled language characteristics. We have mentioned the growth of combined access to databases from different producers. What effect does the incompatibility of controlled indexing languages have? It may be difficult to benefit from the specific characteristics of each controlled language, so search strategies will gain in complexity and synonyms should be included to ensure recall. In fact, some of the new interfaces do not allow users to browse the existing thesaurus (this is the case of Cambridge Scientific Abstracts).

Studies on catalogue use show that subject retrieval failures are frequently due to the controlled vocabulary used in OPACs. The defects attributed to Library of Congress subject headings stem from subject cataloguing policies, inconsistent forms of entries, and terminologies not widely agreed upon ([Aluri](#), 1991). Frost ([Frost](#), 1997) reports on the predominant solutions suggested by research in recent years: bibliographic record enhancement with the addition of abstracts and tables of contents, the use of classification schedules as browsing tools and as a source of additional subject access points using index terms, and the application of natural language processing techniques to transform users' queries into subject headings.

The situation of OPACs has been affected by the economic constraints we are experiencing. Subject analysis and representation is the most costly operation in terms of time consumed by developing classification and indexing criteria and by maintaining the systems. Moreover, due to the frequently subjective nature of these tasks and the variety of criteria used by individual cataloguers or indexers, some critics question the appropriateness of the results. So the consequences are clear: subject analysis has been the first element to be reduced in our library catalogues, and also in the US ([Williamsom](#), 1997), classification is now used solely as a shelf arrangement device, many subject authority records are incomplete and subject indexing policies have shown a tendency toward simplification and toward the replacement of the subject headings list by a general post-coordinated thesaurus.

The constantly growing and changing body of information that constitutes Internet resources has been made accessible by search engines. The early and well known engines such as WebCrawler, Lycos, Altavista,

Excite, Infoseek, etc., are continually adding new features, improving automated indexing and permitting the use of retrieval techniques characteristic of traditional online services, such as the three basic Boolean operators, nesting with parentheses, phrase searching and proximity operators. Looking at the new features added, we can see that they are mainly focused on allowing the searcher to limit and refine search results. Two examples of this dominant trend are the elimination of automatic truncation of search terms and the use of proximity and order operators in Lycos Beta Custom Search, and suggestions of alternative search terms based upon word occurrences within pages retrieved from an original search argument in Altavista LiveTopics.

Subject directories and guides, developed at the same time, are a means for classifying this vast amount of information. They are selective tools: human teams review the Internet locating worthwhile resources according to quality criteria. The initial broad subject categories are subdivided by hierarchical lists which enable the user to find resources within a topic area. A number of directories are organised using library classification schemes, such as Dewey Decimal Classification, Universal Decimal Classification or Library of Congress Classification. A list of World Wide Web sites that have adopted classification or controlled languages to organise Internet resources can be found in the clearinghouse *Beyond bookmarks: Schemes for organising the Web* ([Beyond bookmarks...](#)). A search engine is added for searching within a category. The success of Yahoo has led many search engines to introduce a classification scheme under the search box.

The current practices employed by users for searching seem to produce poor results. This general impression is confirmed by the observations of Keily ([Keily, 1997](#)), which include: single word searches were very usual while phrases using double quotation marks were used only in 10% of the searches, Boolean operators in 12%, and the combination of phrases and Boolean operators in 0,5%. And so the search results are typically very irregular. The users of Internet are too often frustrated by large recall and poor precision, and they become confused by duplicate records which frequently occur, and by the difference in results for the same request performed under different search engines.

In individual libraries, information professionals are also trying to apply some sort of organisation onto the chaotic world of Internet. In this context, there are different kinds of approaches for organising Internet resources. Some libraries have created catalogue records of Internet resources of interest to their users, with the application of adapted cataloguing standards and formats ([Snavely, 1997](#)), while other libraries have organised a parallel, virtual library accessible via their web site. But the benefits of these initiatives are restricted to a limited community of users. International projects have been undertaken with the aim of providing a standard for document access in the Internet. Since these offer a wider scope than individual initiatives, they may be able to affect the evolution of search engine performance in the near future.

The TEI (Text Encoding Initiative) is one of these international projects. The TEI header is a standard for SGML encoding of electronic texts and is divided into four parts: File, Encoding, Profile and Revision descriptions. The subject is included in Profile description where positions for keywords and class codes are provided, as well as the identification of thesaurus or classification systems used. The Dublin Core issued in 1995 following the Metadata Workshop, is much simpler. Within the 13 elements of the label, the first one listed is the subject, as the broad discipline or a group of

descriptors that can represent the content. Armstrong ([Armstrong](#), 1997) suggests the use of PICS, developed as filtering devices in response to the US Communications Decency Act 1996, for labelling the resources. The system would record qualitative and factual data, including subject and geographical coverage, but emphasis is placed on quality assurance. The use of a "standardised vocabulary and scales" would make explicit quality judgements, so "users would no longer have to interpret the meaning behind a site designated as cool or guess how current they could expect a three-star site to be" ([Armstrong](#), 1997).

Due to the broad diversity of information in the Internet, indexing must be performed while bearing in mind the public to whom the documents are addressed. The formal aspects acquire importance in the sense they can be very useful in simplifying the retrieval of relevant information, as suggested by Line ([Line](#), 1997): "what is really needed is an indication of the nature of the item or its intended audience ..., and of whether it is intended to be ephemeral or of more lasting interest". But traditional subject indexing systems still have no definitive proposals for representing formal aspects, so additional research is required in this area.

User interfaces

At present, the initial appearance of most user interfaces is quite similar. Uniformity has thrived in all contexts. Netscape and Microsoft Explorer are effectively standard interfaces that enable the user to access a great variety of information: Usenet newsgroups, electronic journals, bibliographic databases, and Opacs. The use of intranets within companies and institutions has become widespread, and the recent Microsoft Internet Explorer 4, which is intended to integrate Web navigation into the desktop, will probably be an extension of this uniformity.

CD-ROMs have migrated to Windows environments, some with an enhanced screen design, like Ovid Technologies. Online hosts have developed Web interfaces, Knight Ridder has created DIALOG Web in response to the feedback from online searchers from different types of libraries: "they saw it as a natural fit with the regard growth of corporate intranets" ([Klopper](#), 1997).

User interfaces offered by search engines allow searches to be narrowed or broadened; allow parts of the database to be selected and searched according to different criteria; allow limits for language, location, type of media, etc. But the differences in descriptive details and in indexing procedures lead to inconsistent results for the same search arguments performed under different search engines. On the other hand, even if the user interfaces look easy to use, users may feel confused with the lack of transparency of the search box and, to a large extent, they are not making good use of all of the search possibilities. All search engines provide help pages, but no one seems to want to press the help button, and the help information included on the search form page is frequently poor. Simple and precise information on how to search, with illustrative examples, could be added beside the search box.

4. How can subject retrieval be improved?

The terminology we use is not keeping pace with reality. Gilchrist ([Gilchrist, 1996](#)) makes a comment about the term, "information retrieval", suggesting that in the 60s "computer-assisted bibliographic reference retrieval" would have been a more realistic term. Today we are speaking about "Intelligent agents", the word "Knowledge" has replaced "Information" in many contexts, and "Metadata" refers to a process that cataloguers have been performing for years. But the new words serve solely for attracting attention. All search engines fail in understanding the way human beings look for information and there is a long way to go before reaching the ideas expressed by Shera in 1957: "The great promise of these machines [for organizing and searching graphic records] lies in the fact that the system they use will make possible the conceptualization of recorded information in patterns approaching the process of conceptualization in the human brain" ([Gilchrist, 1996](#)).

In the last forty years the evolution of computer systems has been significant but we have to admit that they do not provide an intuitive search context. The user's thought patterns remain unclear, so the challenge may lie much more in the cognitive field than in the technological one. Meanwhile, if users are to be able to do their own searching with an acceptable degree of success, systems should fully integrate the function of the intermediary, from the very first step --assisting users in analysing their information needs-- to the very last one --delivering the search results in such a manner that content assimilation becomes easy.

Both reference and indexing librarians, in their role as intermediaries, make use of conceptual structures to link information needs (real or potential) with the knowledge contained in documents. Here we are not making reference to specific tools such as UDC, DDC or LCSH, but to the categorisation of concepts, as Soergel ([Soergel, 1997](#)) outlines "Based on our knowledge of conceptual structure principles and on analysis of users' problems and users' thinking, we can come up with structures that take users a step further."

Keeping this aim in mind we can give an economic and standards perspective. There is evidence that economy is what makes change possible. The inclusion of advertising in information services can be a solution to individual problems generated by current budget restrictions, but in a global approach the proposals must be economically affordable. Thinking globally, the present proliferation of subject directories and guides is a wasted effort, even though the amount of time dedicated to maintaining and updating them is substantial and, as such, has significant worth. Cooperation is needed to save us time which could then be invested in providing users with value-added products.

Cooperation requires standards. The use of standards now really does matter because end users are accessing the systems directly. The "standardised vocabulary", mentioned by Armstrong in regards to quality, should be applied also to subject indexing. If subject terms are not controlled, the limitations of free text searching will persist. We should learn from the history of cooperative cataloguing and not reproduce the same mistakes, the delay of subject representation.

On the other hand, reaching a widely accepted, standard controlled language for indexing can be too difficult and costly to afford. We must look toward a realistic standard, since it is apparent that we cannot attempt to impose a highly detailed classification or indexing language, whether it be traditional or innovative. A tool requiring extensive development in order to represent highly specific concepts would not fit in with the knowledge

structures of users coming from different fields of activity. Moreover, natural language has proved to be competent for retrieving specific concepts.

Content metadata could cover two types of data:

- concept metadata, addressed to conceptual categorisation
- form metadata, including formal aspects concerning language, geographic location, type of media, audience, and content quality criteria (validity, completeness, accuracy..).

Authority files are needed to allow unique content representation in both concept and form metadata. Conceptual categorisation will require hierarchical authority files to be maintained, with a minimum of three levels of subdivision under each discipline.

Search engines may directly benefit from this sort of standard. PICS is already supported by Microsoft Internet Explorer 4, but we should keep in mind that users must be provided with searching tips on how to use conceptual categories in combination with natural language, and user interfaces should provide features for limiting searches by formal aspects.

Although this sort of content metadata would not provide adequate conceptual access to information, it can be worthwhile as an intermediate tool. The provision of information in companies and research institutions will require additional tools designed according to the users' particular needs and covering the specific field of knowledge with much more detail. The application of Java programming language can allow for integrating modules with other applications. Java "widgets", generic pieces of code designed to perform specific functions such as the representation of a hierarchical vocabulary tree, can be integrated into products that can be customised for particular data and applications ([Johnson](#), 1997).

5. Conclusions

Ensuring the information competency of users is one of our challenges. The knowledge of electronic information sources must be acquired during the secondary education level. This will involve supplying schools with the appropriate equipment and preparing teachers. Libraries will also have a significant role. Information professionals must share their information seeking skills with users, and the intermediary function must be replaced by a training function in order to help users to become effective finders of information. But technology is a means, solely the vehicle for delivering information. The training tasks must be further extended to ensure the acquisition of skills required for analysis, synthesis and evaluation of information resources.

Database context and indexing procedures, as the main factors influencing retrieval, have led us to a more complicated environment that needs to be reoriented in order to improve the relevance of present search results. Although the unification of user interfaces has been a great advance, the enhancement of help information is required in order for users to be able to enter more complex search statements.

A first step towards improving subject retrieval in the World Wide Web can


be achieved through content metadata. The automated indexing performed by search engines should be complemented by the inclusion of standard subject categorisation in headers. Recording of formal and quality data can contribute toward effective delimiting in searches. Further technological developments, like push technologies, cannot be relevant if the application of conceptual structures is not implemented.

References

- 1 Aluri, Rao; D.Alsdair Kemp and John J. Boll. *Subject analysis in online catalogs*. Englewood: Libraries Unlimited, 1991
- 2 Andreu, Cristina et al. "El catàleg de la UAB: una enquesta d'ús". *Item* 19 (juliol-desembre 1996): 79-93
- 3 Armstrong, C.J. "Metadata, PICS and quality". *Online & CDROM Review* 21,4 (1997): 217-221
- 4 *Beyond bookmarks: Schemes for organizing the Web* [en línia] <<http://www.public.iastate.edu/~CTW.htm>>
- 5 Diaz, Karen R. "User success in a networked environment". *Reference Quarterly* 36, 3 (Spring 1997): 393-404
- 6 Estivill, Assmpció. "Organització dels recursos Internet". *Item* 18 (gener-juny 1996): 42-74
- 7 Frost, Carolyn O. "Next-generation online public access catalogs: Redefining territory and roles". *Advances in Library Automation and Networking* 5 (1994): 1-37
- 8 Gilchrist, Alan. "Blissful spiders: Forty years of information retrieval". *Managing Information* 5, (May 1996): 35-36
- 9 Johnson, Dana. "Read it and reap: the benefits of Java search applications". In: International Online Meeting (21s: 1997: London). *Proceedings*. Oxford: Learned Information, 1997, p. 187-192
- 10 Keily, Lyn. "Improving resource discovery on the Internet: the user perspective". In: International Online Meeting (21s: 1997: London). *Proceedings*. Oxford: Learned Information, 1997, p. 205-212
- 11 Klopper, Susan M. Testing stretching, pushing and pulling the Dialog Web. *Online* 21, 5 (1997), 27-32
- 12 Larson, Ray R. "The decline of subject searching: Long-term trends and patterns of index use in an online catalog". *Journal of the American Society for Information Science* 42, 4 (1991): 197-215
- 13 Line, Maurice B. "Electronic information: use and users". In: Jornades catalanes de documentació (6es: 1997: Barcelona). *Cap a una societat digital: un món en contínua transformació*. Barcelona: SOCADI: COBDC, 1997, p. 25-36
- 14 Notess, Greg R. "New features of the Web indexes". *Online* 43, 2 (Sept.-Oct. 1997): 52-56
- 15 Rowley, Jennifer. "The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research". *Journal of the American Society for Information Science* 20, 2 (1994): 108-119
- 16 Snavey, Loanne and Natasha Cooper. "The Information literacy debate". *Journal of Academic Librarianship* 23, 1 (1997): 9-13
- 17 Soergel, Dagobert. *An Information Science Manifesto* [en línia] <<http://www.asis.org/Bulletin-Dec97/Soergel.htm>>
- 18 Taylor, Arlene G. "On the subject of subjects". *Journal of Academic Librarianship* 21, 6 (1995): 484-491
- 19 Williamson, Nancy J. "The importance of subject analysis in library and information science education". *Technical services quarterly*

15, 1-2 (1997): 67-87

1 Comunicació enviada al BOBCATSSS Symposium. (Budapest, 1998)

 guardando log...

Articles similars a BiD

- [RCLIS : cap a una biblioteca digital de biblioteconomia i documentació](#). Barrueco Cruz, José Manuel; Subirats Coll, Imma. (2003)
- [Implementació d'una eina de metacerca : MetaLib i SFX](#). Váñez, Mari; Benítez, Beatriz; Leg, Mireia. (2009)
- [Directrius per a l'accessibilitat al contingut web, versió 2.0](#). World Wide Web Consortium. (2009)
- [Instruments bàsics per planificar estratègicament el servei de biblioteca pública](#). Omella i Claparols, Ester; Permanyer i Bastardas, Jordi; Vilagrosa Alquézar, Enric. (2009)
- [Emmagatzematge distribuït i preservació digital : una panoràmica d'alternatives](#). Castillo, José Manuel; Jorba, Ferran. (2008)

Articles similars a Temària

- [Control o caos bibliogràfic : un programa para los servicios bibliográficos nacionales del siglo XXI](#). Gorman, Michael. (2003)
- [Catalogació/organització de documents digitals : estat de la qüestió, tendències i perspectives d'Espanya](#). Méndez Rodríguez, Eva María. (2003)
- [Amazon, Bol y Diversia : tres ejemplos de librería virtual](#). Nuño Moral, María Victoria; López de la Iglesia, M^a Concepción. (2000)
- [Recursos electrónicos y catálogo](#). Estivill Rius, Assumpció. (2000)
- [Una nueva concepción de la documentación en los medios electrónicos : retos y nuevas tareas profesionales](#). Marcos Recio, Juan Carlos. (1998)

Articles del mateix autor a Temària

[Espelt, Constança](#)

[[més informació](#)]

Escola Universitària de Biblioteconomia i Documentació
Universitat de Barcelona
Barcelona, juny de 1998
<http://www.ub.edu/biblio> •  [Comentaris](#)


[Citació recomanada](#) • [Metadades](#)
[UB](#) • [Facultat](#) • [BiD](#)