DOCUMENT RESUME

ED 096 997 IR 001 178

AUTHOR Carmon, James L.
TITLE "SDI--Where are We? The Challenge of the Future." The Infotmation Dissemination Center View.
INSTITUTION Georgia Univ., Athens.
PUB DATE Oct 74
NOTE 19p.; Paper presented at the Annual Meeting of the American Society for Information Science (37th, Atlanta, Georgia, October 1974)

EDRS PRICE MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS Bibliographic Coupling; *Computer Oriented Programs; Data Bases; *Information Centers; Information Dissemination; Information Networks; *Information Retrieval; Information Scientists; Library Reference Services; Library Science; Man Machine Systems; On Line Systems; *Search Strategies; Speeches; Use Studies
IDENTIFIERS SDI; Selective Dissemination Of Information

ABSTRACT

The historical and current status of information dissemination centers and the problem of user interface are reviewed. During the past decade, the problems of technical data processing have been conquered; information dissemination has evolved from a loosely knit group of experimental centers to an organization of established centers, many operating multiple data bases. Competitive data bases are becoming available in a number of subject fields, putting the centers in a better bargaining position with the data base producers. However, on-line retrieval, resource sharing, and networking must solve the common problem of user interface before anyone or any combination of these operating modes can be really effective. Interactions between the user with his question, the intermediary (the profile code processor), and the search system with its data base are critical to continued evolution of information centers. The intermediaries will, for some time, be the most effective bridge between the users and the computer-based retrieval services. The breakthrough needed for both on-line and batch retrieval systems is the understanding, modelling, and simulation of the man-machine interfaces which are now handled by the intermediaries. (WCM)

"SDI -- Where are We? The Challenge of the Future"

The Information Dissemination Center View

J. L. Carmon

Introduction

With a topic as broad as this one, and with free license from our session chairman to explore within it, the problem is not what to discuss but rather what aspect can be covered in the 20 minutes alotted. I would like to address the topic of the future challenge for SDI from the point of view of the information dissemination center -- the organizational entity which has evolved over the past decade to handle the retrieval processing of the computer-readable bibliographic data bases. More particularly, I would like to address the problem which we, in our center, see as the next major research and development hurdle to be bridged if SDI services are to continue to develop in the future as they have in the past. After briefly reviewing the historical evolution of information dissemination centers in general and a survey of the current status, I'll turn attention to the problem which I'll refer to as the User Interface, and, I hope, convince you that it is indeed of greater magnitude and complexity than has generally been recognized and that it will require concentrated attention by researchers and practitioners in the Information Science and allied fields if we are to ever realize the blue-sky dreams of general and widespread access to and use of bibliographic retrieval services through some network utility.

Definitions

Before going any further, I want to define some terms the way I will be using them since they may differ with some of the other panelists. I'm not sure how SDI was defined in setting up this SIG, but its use in the literature has varied. Most authors limit its scope to current awareness searches, but some give it a broader scope. I will be using SDI in its broadest possible context -- that is, the selection of information for dissemination in response to a request. No time frame is implied in the words themselves, and I choose to include such types of retrieval as have been labeled current awareness, retrospective, demand, customized, special, mission-oriented, and so fortn.

Other terms which require clarification include "center", "intermediary", "user", and "data base producer or vendor". The "center" is the organizational entity or group which processes one or more computer-readable bibliographic data bases for the purpose of distributing bibliographic citations in response to individual queries. Thus, centers may be for-profit, or not-for-profit; located in a library or a computer center, or may be set up as an independent organization, as part of a government agency, or as part of a data base producer's services. My point is that the term "center", will be used in its broadest context and should not be equated to any particular type of center or operating mode. Another term which was mentioned was "intermediary" -- or "profiler".

By these terms, which will be used synonymously, I refer to the human being who interacts in anyway with the user or his question and the search system, including such components of the search system as the data bases. These intermediaries are known by many names, -- e.g., information specialist, reference librarian, information analyst, and profile analyst. Again, the broadest possible scope should be associated with my use of the general term "intermediary" even though specific functions may vary from centers and in all functions may be performed in any given center. A "user", in my frame of reference, is the person with the information need -- the person who wants an answer to a question. A user may interact directly with a search system on his own, but more often he is one member of the team -- the other being an intermediary -- who interacts with the system. The last term to be defined is "data base producer or vendor" -- the organizational entity that creates the machine-readable bibliographic data base. Like centers, they may be for-profit or not-for-profit, located in a government agency or with a professional society, or there may be any of a number of other possibilities. If a given organization both produces and searches its own data base, then it is both a data base producer and a center.

So much for definitions. Let me turn now to a brief history of the development of information dissemination centers as a means of providing perspective for where we are, where I think we are going, and what it will take to get there.

History

Information Dissemination Centers using machine readable data bases had their beginning back in the early 1960s -- just a little over a decade ago -- with the establishment of the Medlars and RDC centers by the National Library of Medicine and NASA, respectively. They were mission-oriented and heavily subsidized by the federal government, and these two data bases were limited to processing by the agency-sponsored centers. In the not-for-profit sector, Chemical Abstracts Service led the way with publicly available data bases, first with Chemical Titles about 1962, and a few years later with CBAC and POST. In these early years, user groups tended to build up around individual data bases -- the Medlars centers got together to discuss common problems, as did the NASA centers and the CAS tape users. During those first few years, our user groups struggled with such problems as debugging search programs (which were often supplied with the data base), arguing the pros and cons of various search techniques, teaching each other how to prepare profiles, and persuading users to do their searches by computer. Retrieval systems, as a concept, did not exist at that time -- we still spoke in terms of search programs. And the file structures reflected their unit record heritage -- card image records, with fixed length fields, numerically encoded index terms, and print-oriented data representation.

Several significant changes have come about during the past decade -- changes which not only reflect the rapid maturing of an infant industry (we've been diapered and burped publicly on a number of occasions), but also reflect major changes in what centers do, the user communities they serve, and relationships between centers and data base producers. On the technical side, we've moved from the single processing shops of 1401s and 7094s to third generation computer hardware with its versatile operating systems, applications software, and multiprocessing environment with telecommunications access. The self-defining, directory-oriented, variable length file structures, such as defined by the ANSI standard for bibliographic information interchange on magnetic tape, are now state-of-the-art and are being adopted by more and more data base producers as they convert their data processing operations to integrated computer-based production operations. Search programs have evolved to large and relatively sophisticated retrieval systems, capable of handling multiple data bases with varying content and format, often with many of the processing operations under user or intermediary control (e.g., format, content, location, and media in which the search results are delivered). Computer programming, profile construction, and data base conversion are state-of-the-art and part of the routine operations of all but the youngest of information dissemination centers. The ASIDIC meetings, which now attract as many as 80 attendees from among 30 full members and 50 associate members, are now devoted to topics which reflect the interactions of centers with their environment. With data base producers, the hot topics are lease and license provisons, royalty payments, usage restrictions, and networking implications. With libraries, two areas of interaction are drawing attention: one concerning the interface with reference librarians and the incorporation of the intermediary functions into reference librarianship, and the other dealing with the location and delivery of documents which are identified through the computer-based retrieval services.

In summary, during the past decade we have largely conquered the technical data processing problems; we have evolved from a loosely knit group of experimental centers serving small parochial user groups to an organization of established centers, many of whom operate multiple data bases and serve a nation-wide user community in a competitive environment which provides shopping choices to those users. Competitive data bases are now becoming available in a number of subject fields, putting the centers in a better bargaining position with the data base producers and, indirectly at least, providing motivation for improved data base quality and serious consideration of unjustified incompatibilities between data bases.

This brings us to the present. What about the future?

The hue and cry now is on-line retrieval, resource sharing, and networking. These three concepts are by no means the same thing -- on-line retrieval may be done via a telecommunications utility but need not necessarily be part of a network, in the sense of having anything in common with other users of the utility. There are several centers which make their on-line retrieval services accessible via the Tymshare communications system yet have no relationships -- in fact are competitive --

with each other. Similarly, several centers may agree to share resources, thus constituting a network, without using telecommunications. The NASA RDC centers, for example, comprise such a network of centers without telecommunications links. However, on-line retrieval, resource sharing, and networking do have one very important problem in common which must be solved before a· · one or any combination of these operating modes can be ·eally effective, and that is the users' interface to the search system.

The User Interface Problem

I can practically hear the shrugs -- "What's the big deal about user int·rface? You prepare some good profile coding manuals, run a training session, and the problem is solved." And I might add that if we had been told the same thing a few years ago, we would probably have shrugged with the same answer. However, over seven years experience as a center, some 20 different data bases, and over 6 million document records in the retrospective collection have taught us differently. And I hope to convince you that understanding the interactions between the user with his question, the intermediary (if one is imposed), and the search system with its data bases is critical to continued evolution of information dissemination centers. It is the major block to effective use of on-line search services and to the sharing of data base resources, regardless of whether networking per se comes about.

I emphasize the word effective, because it is certainly true that on-line searching and profile exchange are going on. But experience in our center raises serious concerns which we, as information science professionals, should have about the quality of the results being obtained. (For those of you who may not know, the University of Georgia Computer Center operates a center which has remote input and output terminals located in New York, Ohio, and Atlanta, as well as several terminals on site in Athens.)

slide 1

FACE proposal

Does this look familiar? It should, because this diagram or a similar one appears in almost every profile coding manual or textbook on reference librarianship. Different names have been applied and the various sources may differ somewhat on the descriptions of the functions, but most of them present steps which are similar to those given in Figure 1. Descriptions normally concentrate on "what" is to be done with little or no attention on "how". The librarian or profiler is exhorted to discuss or negotiate the user's question until it is clearly defined, but there is little guidance as to what constitutes a clear question or what techniques can be used to arrive at it. The same situation applies to other steps in the process, some more than others, of course. Identify the concepts -- parenthetically, the "important" concepts -- but what constitutes important concepts? The next step may be something like expand the concept, which means to add the vocabulary appropriate to the data bases -- or what Lancaster calls "indexing the query". This profile coding process is often more art than science. In spite of the importance profile construction plays in the effectiveness of the retrieval, we know virtually nothing about the decision-making processes and the sources and characteristics of the information used to make these decisions for creating good profiles.

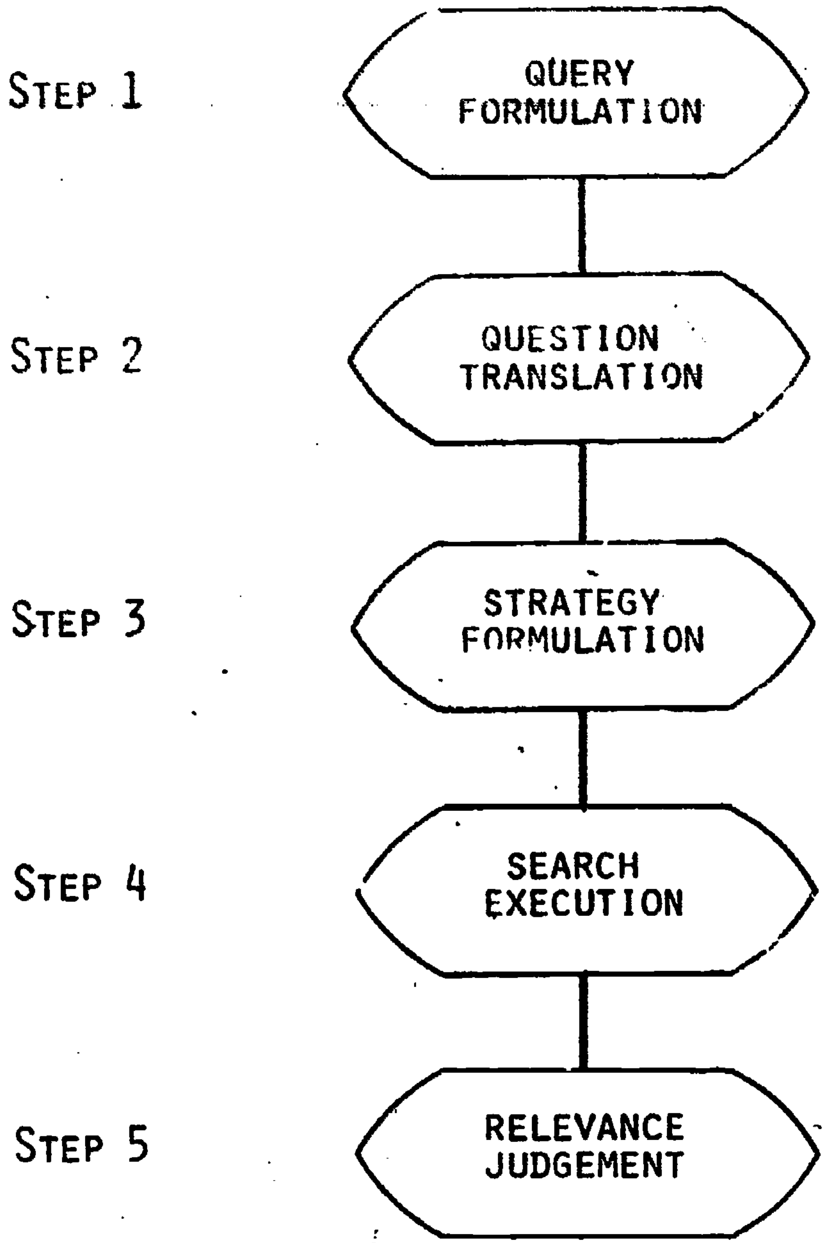
Last year, the dissemination centers at UCLA and at Georgia launched a joint study to investigate the functions, processes, and roles which take place in the interface between user and system -- what we call our "interface" study. This joint study has two major phases, the first of which is to develop a model of the interface process as it now exists. This has been called the Manual Model since most of the functions are performed manually by trained intermediaries. The fact that there are two centers involved is important, because we are concerned not only about processes within a given center but also in differences which exist between centers. Thus, the study has proceeded independently in each center but in parallel through the use of jointly defined measuring instruments so the data can be compared. The second phase of the study, which will follow development of the Manual Model, is the creation of one (or perhaps more than one) model based on a networking environment (this has been dubbed the "Network Model"). It should be clearly understood that we are looking at networks involving multiple dissemination centers, rather than a single, central dissemination center servicing a distributed user population through a communications utility, although the results may be applicable to both.

Over the past 10 months we have collected data on many different characteristics of the interface process and from several points of view. Analysis of these data for development of the models is not yet complete, but the findings already indicate that the interface process is far more
slide 2 complex than we anticipated. As shown in slide 2, the major variables being investigated are related to the user, the question, the data bases, the intermediary, and the search
slide 3 system. Typical characteristics of the user which are being considered (slide 3) include the purpose for which the search is being done (e.g., a class project or term paper, a dissertation, instruction or teaching, a research project, a patent search, etc.), familiarity with the topic being searched (e.g., is it a new project about which the user knows little or nothing, is it final wrap up on a journal article or dissertation to be sure nothing has been missed, or is it perhaps grist for a review article or book?), familiarity with literature resources in the field (e.g., can the user select the appropriate data bases?), prior experience with computer-based search services (that is, a new user or one with prior experience?), and others, as you see listed. For
slide 4 the question, (slide 4) we are looking at such things as the clarity with which it is expressed (i.e., how well-formulated is the question?), the completeness with which the initial question is presented (information on this can be obtained by comparing the user's initial question with the negotiated question), and the scope of the question (that is, is it a

broad question intended or expected to retrieve a large number of answers or is it a narrow, precise question which can be answered with a single, relevant document?). To the extent that the profile is a surrogate of the question, we are also interested in characteristics of the profiles slide 5 and their relationships to the initial question. In the area of data bases (slide 5) we are investigating such characteristics as the size (in terms of both the number of records per some fixed unit of time, such as a year, and also the size of the retrospective collection as a whole). Two other factors believed to be very critical in terms of the roles which intermediaries now play in preparing profiles are related to the vocabulary characteristics of the various data bases (that is, controlled versus uncontrolled, classification versus indexing, and various combinations of these and other attributes) and also the data content of the data bases. When, for example, is it appropriate to search the abstract, and when is it better to stick with assigned index terms or codes? Should the search strategy, hence the profile, differ depending on whether or not the abstract is being searched? Those of you who have done a great deal of profile preparation will know that this is not a simple yes-no decision. It depends on how much you expect to be retrieved, how good the index vocabulary is relative to the particular question at hand, how large the data base is and how much its coverage overlaps the subject matter of the question, and so on. I won't go into characteristics of the other major variables -- the search system and its logic and retrieval features, the background and training of the intermediaries, etc. -- but I hope I have illustrated even briefly how complex the process is when all the combinations and their associated interactions are considered. Several different data collection approaches have been used in this study -- questionnaires filled out independently by the users and the intermediaries, and tape recorded interviews which have been transcribed and analyzed for the presence or absence of over 60 characteristics and have been described in terms of event time series. Data has also been collected on the data bases, one subtask of which is the creation of a merged vocabulary file of an estimated half-million terms or term-pairs for about 13 of the data bases used in our center. This master vocabulary file, which is designed around a thesaurus-like structure, forms the basis for study of the similarities and differences in indexing terminology between the various data bases. There has also been a detailed linguistic analysis of the transformations which occur in going from the narrative form of the user's question to the formalized profile representation as prepared for search against one or more of our data bases. Transformations which are data base dependent are of particular interest in this phase of the study.

As I mentioned earlier, we have collected most of the information needed for development of the Manual Model, but are still working on the statistical analysis and interpretation of the data. Based on our preliminary findings, I would have to say that we have only scratched the surface of the problem and will undoubtedly raise far more questions to be investigated further than we will be able to answer. As Saracevic has pointed out, "The human factor, the variations introduced by human decision-making, seems to be the overwhelming variable, the major influencing factor affecting the performance of every and all components of an information retrieval (IR) system". However, I believe we cannot simply rest on the matter by acknowledging its complexity. We must devote at least as much attention and effort to this critical area of computer-based retrieval as has been poured into building the data bases in the first place, comparing indexing techniques, and programming complex retrieval systems, if for no other reason than to understand the functions and techniques of profile preparation in sufficient detail to effectively train our reference librarians and information specialists. These intermediaries will for some time be the most effective bridge between the users and the computer-based retrieval services offered by information dissemination centers like ourselves.

For my collegues who say that on-line is the only way to go I might respons that there is considerable evidence that both on-line and batch retrieval systems are presently being used in essentially the same mode. It is true that the on-line systems complete the search itself faster that do most batch-oriented shops in terms of elapsed time, but this is the only significant difference at the present time between the two types. At the ASIDIC meeting a couple of weeks ago, one of the data base vendor representatives who uses his own data base in on-line mode reported an average of 40 minutes for construction of the profile (off-line by a user-intermediary team), 18 minutes of terminal connect time to enter and search the profile, and 30 minutes to review the results for relevance. These are almost identical timings to those we get in our center where we use an on-line data entry system for input to batch search. The on-line systems have certainly shortened the elapsed turn-around time for the search, but they have not changed the process significantly, and in fact those on-line centers who started out trying to peddle terminals directly to users have rediscovered what we learned back in 1965 -- the majority of the users don't have any aspirations toward being information specialists; they just want the results. At the present time, on-line search systems look like the early days of computer-assisted instruction -- very expensive page turners with little or no advantage being taken of the interactive potentials of the computer. The breakthrough needed for both on-line and batch retrieval systems is in the understanding, modelling, and simulation of the man-machine interfaces which are now handled by those artists, the intermediaries.



MAJOR PROCESSING FUNCTIONS OF THE REFERENCE PROCESS

MAJOR VARIABLES

- USER
- QUESTION
- DATA BASES
- INTERMEDIARY
- SEARCH SYSTEM

CHARACTERISTICS OF THE QUESTION

- CLARITY WITH WHICH IT IS EXPRESSED
- COMPLETENESS WITH WHICH IT IS PRESENTED
- SCOPE OF THE QUESTION
- CHARACTERISTICS OF THE PROFILES AND THEIR RELATIONSHIPS TO THE INITIAL QUESTION

USER CHARACTERISTICS

- PURPOSE OF SEARCH
- FAMILIARITY WITH THE TOPIC
- FAMILIARITY WITH LITERATURE RESOURCES
- PRIOR EXPERIENCE WITH COMPUTER-BASED SEARCH SERVICES

CHARACTERISTICS OF THE DATA BASES

1.

. SIZE

. VOCABULARY CHARACTERISTICS

. DATA CONTENT