Division of Research
Graduate School of Business Administration
The University of Michigan

January 1984

THE MANAGEMENT OF INFORMATION: BASIC DISTINCTIONS

Working Paper No. 357

David C. Blair

Graduate School of Business Administration
The University of Michigan

FOR DISCUSSION PURPOSES ONLY

THE MANAGEMENT OF INFORMATION: BASIC DISTINCTIONS

by

David C. Blair
Assistant Professor of Computer
and Information Systems
Graduate School of Business Administration
The University of Michigan
Ann Arbor, Michigan 48109

Abstract

Many of the problems of the management and use of information arise from the inability of the system designer or system user to understand certain fundamental conceptual distinctions in information management. The failure to make these distinctions can lead to certain unfortunate consequences. These distinctions (storage vs. access, physical access vs. logical access, and data vs. documents) are described, as well as the possible consequences of not making these distinctions.

Information management is a complex and highly diverse endeavor. Not only does it concern the management of a wide variety of types of information, from operational data to high level analysis, but it must also provide many different types of access to the same information (from the simple compilation of data to complex, selective ad hoc data inquiry). Such diversity can make the information manager's job extraordinarily difficult. Fortunately, though, there are several important logical distinctions which can be made in the management and design of information systems. These distinctions are critically important for both the system designer and user to understand in order for the design and use of the system is be optimal. As Hegel said, "Hell is truth seen too late."

Storage vs. Access

The first important distinction to make is between the storage of information and the access to that information. The task of information management is frequently seen as a problem of information storage. This point of view perceives the information problem as a problem of physical space. It tries to answer the question, "Where can we put (or, what will we do with) all this information?" From this perspective the solution to the "information problem" often lies in the technologies of microforms and computer-accessed mass storage devices. The contents of an entire filing cabinet can be put on one 16mm. film cartridge. The problem of physical space is solved (or at least mitigated) by putting reams of paper onto microfilm, magnetic tapes, or data cartridges, but has the information problem been helped by doing this? Not necessarily. In fact, the problem may have been made worse. This is because the solution to the information problem lies not just in the storage of information, but in the conscious

provision for access to that information.

How can the use of technologies which reduce the cost and physical space of information storage exacerbate the problem of access to information? It does so in two ways.

1. The transfer of information from paper to microform, or mass storage, with no concern for access to that information frequently reduces the access points which an inquirer may have to that information. One company, drastically reduced the floor space (and, consequently, the cost of office space rental) taken up by filing cabinets full of documents, by urging employees to put all but the most recent files on 16mm. film cartridges. The only instructions given to effect this change were to package up the files and send them to the micrographics department. In a few days the individual who sent his files to be microfilmed got one or two 16mm. film cartridges back with a label like "John Erik: personal file, 1976 - May 1978". Why was this bad? Because the nature of microfilm had changed the labeled file folders to which John Erik had direct access to, rapid scanning ability, and ease of file rearrangement, to a strip of film with one comprehensive (and, thus, undetailed) label which could only be searched sequentially and could not be weeded or rearranged without refilming the entire file. True, John Erik knew pretty well what information was on the cartridge, but if he left the company the individual who inherited that information would not have that comprehension, and would be forced into a laborious, sequential search of the film each time he wanted to look for something he thought might be in the file. Computerized mass-storage devices often make direct access files possible, but this is only an advantage over microforms if such

direct access is planned for. Such planning, unfortunately, takes time away from an employee's principal responsibilities, so even if he sees the need to provide access to computerized information he may not have or take the time to index all his information that goes onto a data base. The result, then, is likely to be the same as with microfilming, namely, the computerized information will be sequentially ordered and inadequately described.

Sometimes the concern for microfilming or mass-storing large numbers of hard-copy files is so strong in organizations that individuals in charge of the files often, by necessity or ignorance, delegate the labeling of files to the microfilming or data-entry clerks. One engineering company delegated this labeling responsibility almost entirely to the microfilming clerks. Files were usually grouped by contract numbers and these naturally became the labels for the microfile cartridges. Two problems occurred: First, if information from one contract went down to micrographics in more than one batch (or at different times) this information might be filmed on several cartridges interspersed with information from other contracts. Second, some information was not clearly concerned with one contract or another, so the micrographics clerk merely labeled the cartridge containing that information "Dept. A: miscellaneous." Within a relatively short time the company had 35 cartridges (6000+ pages of information per cartridge) labeled "miscellaneous".

2. The second difficulty with the use of low cost information storage technologies is that they encourage an unfortunate, but common, human weakness. Namely, as the cost and ease of storing information decreases,

the amount of information considered worth keeping increases to a relatively easily reached limit where nothing is thrown away. Ours is not so much the Information Age, as the Information Retention Age. The rationale given for this tendency to keep everything is that since the cost of keeping everything is often negligible, then in these cases we should keep everything because any of this information might prove useful at some time. The human mind is marvelously creative, and it's easy to imagine possible scenarios where the most trivial information would prove absolutely crucial. But this tendency to keep everything is a natural consequence of the view that information management is a problem of information storage. In other words, if you keep everything you'll always have the one unforeseen crucial item of information should you need it in the future. But do you really "have" that information? In terms of "possession" you do, but in terms of "access" you may not. Wilson pointed out that what is important is not how much information you have available to you, but how much information you have control over.[1] In other words, what is important is how much information you have reasonable access to. This is why the problem of information management is a problem of access to information, not storage of information. Even a well-designed information system suffers from the tendency to keep everything because an inquirer will have to search through the useless information on a data base (or in an archives) in order to find the relatively small body of useful information. The larger the amount of useless information (that is, useless to a particular inquirer) on an information system, the more difficult it will be to retrieve the useful information. In the vernacular of systems theory, information with a

low likelihood of utility is just "noise" in the information system. This "noise" seriously degrades the quality of information searches done on a system, and increases the cost of those searches. Thus the "negligible" cost of keeping all information results in a significant increase in the cost of accessing or "controlling" that information. A document or collection of data can be just as easily lost on a large data base as it can be lost in any other way.

Physical Access vs. Logical Access

Now that we are clear that the problem of information management is a problem of access to information and not storage of information, we can look more closely at the problem of information access. Here we can distinguish two kinds of access: physical access, and logical access. The problem of physical access concerns how actual desired information is retrieved and represented physically to the inquirer. It concerns how an information retrieval system (computerized or manual) finds the desired information (or gives directions to the inquirer on how to find the information) once the address of the information is known. Logical access is concerned with how to find the address of the desired information. Consider a library. Discovering where the book with call sign QA76.A1A84 is in the library is a problem of physical access. The discovery of which book in the library will be likely to satisfy a particular information need is a problem of logical access. Or, in terms of a modern data base: determining where on the data base the monthly sales figures are is a problem of physical access; determining whether the monthly sales figures are what is likely to be needed by a particular inquirer is a problem of logical access. Physical access is concerned with the techniques of retrieving actual information,

while logical access is concerned with determining which information (or kind of information) answers a particular question. For this reason, the problems of logical access must be resolved before addressing the problems of physical access. But this is often not the order in which these two problems are approached.

The greatest confusion between physical and logical access occurs in computerized information systems. Computers can dramatically speed up the physical access to records, but, in themselves, they do not necessarily improve the logical access to records. The incredible speed of computers has seduced many information managers into believing that such speed alone can make the largest data base into an efficient information system. But the speed of an information system is directly related to the number of logical decisions the inquirer must make in the course of his search, not the number of physical records or record representations the system can search through in a given time. For example, if an inquirer does not know the logical addresses (such as, report number, exact title, author or date) of the report he wants he will be forced into a situation where he would have to examine every document produced by the organization in a fairly broad time frame. If there were 10,000 total documents produced in this period then, ceteris paribus, the inquirer would have to expect to look at about 5,000 of them to find the report he wanted. This means his information search would consist of about 5,000 logical decisions. The fastest computer in the world would not improve his search rate much, and might not even be as fast as browsing through the reports on the shelves of an archives. The information problem is not a computer problem any more than cabinet making is a chisel or saw problem. To see the information problem

as a computer problem is to confuse physical access with logical access, or to confuse the tool with the job. Physical access in information management is concerned with tool selection: selection of microfilm techniques, mass storage devices, dedicated vs. time-shared computer systems, etc. Logical access in information management is concerned with reducing the number of logical decisions the inquirer must make in searching for information. It is concerned with developing logical access points to a body of information. To continue our example, the system would have been more useful to the inquirer if he could have searched by key words in titles, or size of report, or by minor subject headings, or some Boolean combination of these with other fields of logical access (e.g., "Retrieve all reports greater than 100 pages, with 'marketing' and 'products' in the title, within time frame 1978-79 (inclusive), and with the assigned indexing term 'Alpha'").

The physical access/logical access distinction is similar to the physical structure/logical structure distinction which is frequently made in one area of information management: data base management systems (DBMS). Although DBMS design is convinced of the importance of the independence of logical and physical structure, these structures are rarely totally independent. Nevertheless, the more independent a DBMS's logical and physical structures are, the more flexible and responsive to change will the system be. The same holds true for information systems in general. The more independent the methods of logical and physical access are the more flexible and adaptive the system will be to the inquirer's needs. Unfortunately, such logical/physical independence is much harder to achieve in information systems in general than it is in DBMS design. This means that the kinds of logical access which can be designed for an information system are often

severely constrained by the method(s) of physical access available. For example, if an organization puts a great deal of information on microfiche, they have effectively eliminated any easy direct access to that information. Thus, a kind of logical design which relies on direct access to information (or the sorting of information) is not easily available to information systems which store information on microfiche. This means that the logical design of an information system must be done before the tools of physical access (microforms, tapes, hardcopy, printouts, mass-storage devices, etc.) to that system are selected. The greatest mistake an organization can make is to acquire equipment first and then set out to design the information system.

Data vs. Documents

So far we have said that information management is a problem of access to information, and that we can distinguish two kinds of information access: physical and logical. We also argued that the logical design of an information system must precede the physical design of that system. Let's look more closely at the logical access to information. Here we can distinguish two kinds of logical access: access to data and access to documents (see Figure 1). This is an important distinction for both the system designer and the system user to make. We can delineate three principal areas in which access to data and access to documents differ (summarized in Table 2):[2]

1. Method of answering an inquirer. A data retrieval system operates by directly answering the inquirer's question. It retrieves the actual information desired. A typical data query is specific, it asks, "What is Murphy's salary?" "What are the total commissions earned last month

for salespersons in region 4?", etc. A document retrieval system is more indirect in its operation. It does not retrieve the specific desired information, but usually provides an inquirer with references to a document or set of documents which will likely contain what he wants. A typical document query is more general than a data query, it asks, "What reports do we have which discuss our competitors' marketing posture?", "What documents do we have which analyze Middle-East investment prospects?", "What company correspondence is supportive of our legal defense?", etc. Such queries are indirect because they usually cannot be easily or directly translated into a formal information query, and, as a result, may retrieve a varying amount of irrelevant information (irrelevant documents) along with the relevant information. Even if the inquirer wants a specific document there is still often a probabilistic relation between the retrieved document(s) and the likelihood of inquirer satisfaction. This is because although there may be a one-to-one correspondence between a document descriptor and a particular document (e.g., a unique report number that is only assigned to one document) there is frequently only a relatively low probability that the inquirer would remember something that specific. There are two probabilities which contribute to the indeterminacy of document retrieval: the likelihood that a given document descriptor (e.g., date, author, title, subject term, accession number, etc.) will be assigned to a document and used as a logical access point to that document; and, the likelihood that an inquirer will know or remember the descriptor or descriptors which are logical access points to the document(s) he wants.[3] There is usually a trade-off between these two probabilities whereby the

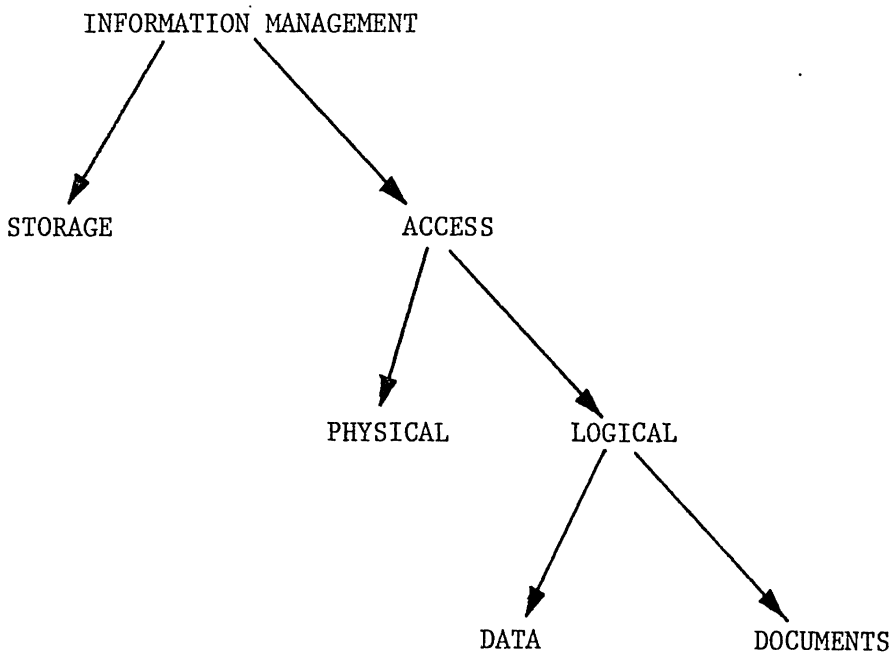
INFORMATION MANAGEMENT
STORAGE
ACCESS
PHYSICAL
LOGICAL
DATA
DOCUMENTS

Figure 1

DATA (VS.) DOCUMENT RETRIEVAL

DATA RETRIEVAL	DOCUMENT RETRIEVAL
* Direct (Answers the inquirer's question)	* Indirect (Provides or refers to a set of documents which will likely contain what the inquirer wants)
* Typical Query is Specific ("I want to know X")	* Typical Query is General or Topical ("I want to know about X")
* Necessary Relation Between the Request and the Correct Answer (Hence, Data Retrieval systems are Deterministic)	* Probabilistic Relation Between the Request and a Satisfactory Answer (Hence, Document Retrieval Systems are Non-deterministic)
* Criterion of Successful Retrieval: Correctness (objective) ("Does the system answer the inquirer's question correctly?")	* Criterion of Successful Retrieval Utility (Subjective) ("Does the system satisfy the inquirer's need?")

Table 1

more specific a document descriptor is (a unique accession number being the most specific) the less likely it will be that an inquirer who wants the document will know or remember that descriptor exactly. In one case, a document retrieval system was developed to keep track of the substantial number of documents (engineering drawings, purchase orders, sub-contracts, correspondence, receipts, etc.) which were generated during the course of a large construction project. Since the major documents (drawings, orders, receipts and sub-contracts) all had unique numbers associated with each of them, the system designers felt that these numbers should be the primary access points to the documents on the data base. Unfortunately, after the system was built it was discovered that the users could only rarely remember (or find) the exact number associated with a desired drawing, order, receipt or sub-contract. In fact, over 80% of the searches were based on subject descriptions--an access point not well developed on the system.

2. The relation between the formal system request and the satisfaction of the inquirer. In data retrieval there is a necessary relationship between the formal request and the correct answer to that request. A formal request for Fredenburg's address must be answered with Fredenburg's correct address. No other address will do. This necessary relation between the request and the correct answer (which, by definition, must satisfy the inquirer) means that data retrieval systems are, from a logical point of view, deductive. For document retrieval, there is usually a probabilistic relation between the formal request and the likelihood that an inquirer will be satisfied. If an inquirer requests all documents that have certain subject descriptions, or authors, or

dates, he does so because he believes that documents with those descriptors will likely contain the information he wants. This likelihood may be very high, but it is a likelihood nonetheless. The probabilistic relation between the formal request and the likelihood of inquirer satisfaction means that, from the logical point of view, a document retrieval system is essentially inductive.

3. Criterion of successful retrieval. For data retrieval the criterion of success is relatively straightforward. One need only ask, "Does the system answer the inquirers' questions correctly?" Hence, the criterion of success in access to data is correctness. Such an objective criterion makes the problem of system evaluation (in terms of retrieval effectiveness) much simpler than it might be since one needs only set up procedures for verifying the accuracy of the answers given by the system. The effectiveness of a document retrieval system is not quite so easy to determine. One needs to ask, "Does the system satisfy the inquirers' needs?" (or, "Did the inquirers find the documents useful?") Hence, the criterion of success in access to documents is utility, a much more subjective criterion than correctness, and, consequently, a much more difficult standard to measure.

Consequences of the Data/Document Distinction

The distinction between access to data and access to documents has a profound effect on both the design and use of information systems. From a design point of view, a data retrieval system is relatively straightforward in the sense that an item of data has only one access point. For example, if the number 23,400 represents "Johanson's salary" on a data base, it cannot simultaneously also represent something else like, say, the number

of products sold last quarter. Neither can the physical data "23,400" also represent someone else's (other than Johanson's) salary (even if they are the same value) because of possible update problems. Hence, to retrieve the value of Johanson's salary from the data base, one must know that that number is called "Johanson's Salary" and nothing else. The description "Johanson's Salary" is the only access point to that specific item of information. (One could get access to that number through "Employee's Salaries", but this is only a more general version of "Johanson's Salary" and does not retrieve only the specific information desired.) Documents, though, typically have many distinct access points. A document can be retrieved by using such descriptions as author, title, date, document number, document type, subject(s), length, addressee(s) (where appropriate), etc. This list can frequently be extended depending on the type of documents and the type of access required. What is important here is that because of the large number of access points that documents have (as opposed to data) the design of a document retrieval system is not as straightforward as the logical design of a data retrieval system. For document retrieval to be optimal it must allow the inquirer to find documents with as few logical decisions as possible, and to do this it must permit queries which allow for ad hoc combinations of access points. This means that a document retrieval system must not only provide a large number of individual access points to a single document, but must also provide access to that document via a large number of ad hoc Boolean combinations of these access points. A system designer who failed to perceive the importance of the data/document distinction in information access, would likely treat document retrieval design problems with data retrieval solutions. This would likely result in a document

retrieval system with too few access points, or an inability to combine access fields in an ad hoc way (such as retrieving by date period and document length for the inquirer who wanted a particular document but could only remember that it was written in 1979 and was exceptionally large).

The data/document distinction has important consequences for the users of information retrieval systems. The clearest point of difference between data retrieval and document retrieval lies in the inquirers' expectations of retrieval effectiveness. For data retrieval, because there is only one access-point to individual items of data, the inquirer can expect to either (1) retrieve exactly the data he wants (e.g., "Johanson's Salary"), or, (2) know that the data he wants does not exist on the data base (e.g., a request for "Johanson's Salary" retrieves nothing).

For document retrieval, though, there are many access points to individual documents, and these access points are usually not unique (that is, there may be more than one document with a specific date, document type, subject description, or a particular author's name (especially in a correspondence file)). This means that, unlike data retrieval, in document retrieval the inquirer can rarely retrieve exactly the information he wants. Requests for documents having particular dates, authors or subject descriptions usually retrieve a significant number of irrelevant documents along with those the inquirer wants. The document inquirer must therefore be resolved to doing a certain amount of browsing through useless documents to find the ones he wants.[4] It is simply unrealistic for the document inquirer to expect the same direct, specific retrieval usually available in a data retrieval context. A good document retrieval system is as specific as possible, but never as specific as a typical data retrieval system.

There is another consequence of the lack of specificity in document description. This consequence is that the document inquirer frequently may not know for certain after he has conducted his search whether there remain important unretrieved documents on the data base. The most difficult situation for a document inquirer is where he does not know how many useful documents there are on the data base, and he needs them all (or most of them). If he retrieves a set of useful documents it is very difficult for him to infer whether he has all the useful documents he could retrieve, or even whether he has the most useful ones available to him. A group of lawyers used a computerized document retrieval system for support in the defense of a large corporate lawsuit. In a study of retrieval effectiveness the lawyers were asked to retrieve all the documents necessary for their defense of the suit. They did so, and estimated that they had retrieved over 80% of the documents germane to the defense. A detailed analysis, though, showed that the lawyers had in fact retrieved less than 20% of the relevant documents.[5] Such mistakes are naive, but not stupid. The unavoidable indeterminacy of document retrieval is not widely understood by inquirers (or system designers). They are not suspicious enough of the sets of documents they receive from document retrieval systems.[6] A document search has two parts: (1) Find the information you need; and (2) Make sure that that information is all that is available to you.

A Functional Taxonomy of Information Management

We have already pointed out that the task of the information manager is not the relatively static job of information storage, but is a much more active and dynamic endeavor. Information attains its significance through its use. The information manager is fundamentally concerned with two

kinds of information; data and documents; and he is also concerned with two processes which may affect this information: retrieval only, and retrieval with modification. When we consider the two logical types of information and the two fundamental processes which effect that information we can see that there are four primary areas of information management (see Figure 2):

Data retrieved only: Data retrieval
Data retrieved and modified: Data processing
Documents retrieved only: Document retrieval
Documents retrieved and modified: Word processing

These are the four fundamental functional areas of information management. They are important areas to keep distinct from an organizational point of view, because they frequently serve different functions within an organization. Data processing serves the operational level of an organization, while word processing serves the strategic or policy levels of management. Data retrieval finds its use at all levels of management from the simple non-selective data aggregation of the operational level, to the highly selective ad hoc inquiries that may be made by the highest levels of management using sophisticated data base management systems. Document retrieval, on the other hand, finds it greatest application at the strategic and policy-making levels. This four-fold taxonomy of information management helps the information manager to identify and better serve the specific users he serves. But this is not the only use for this taxonomy. The manager can use it as a means of avoiding information system design mistakes. In the first place this taxonomy graphically illustrates the data/document distinction we have discussed before, and suggests that while hybrid systems may be designed that combine vertical (Figure 2) functions (data processing plus sophisticated data retrieval (an area of development in DBMS design);

Four Principal Areas of Information Management

	Data	Text
Information Retrieved Only	Data ("Fact") Retrieval	Text ("Document") Retrieval
Information Changed in Some Way	Data Processing	Word Processing

Figure 2

DISTINCTION	CONSEQUENCES OF NOT MAKING THE DISTINCTION
Storage (vs.) Access	Overconcern with cheap, easy storage. Reliance on individual's personal knowledge of available information. Tendency to blame searchers for failure to find information. Tendency to keep all information when financially feasible.
Logical (vs.) Physical Access	Concern for fast retrieval of all records rather than fast identification of important records. Tendency to fit information system to available equipment, or to purchase hardware before designing the logical structure of the information system.
Data (vs.) Document Access	Tendency to apply known data retrieval technology to document retrieval. Tendency to oversimplify the logic of document retrieval. Tendency to believe the specificity of data retrieval is attainable in document retrieval.

Table 2

or word processing plus sophisticated document retrieval (an area of growing concern, but no real development)) hybrid systems which combine horizontal (Figure 2) functions (data plus document retrieval; or, data plus word processing) do not have good prospects. This is because any such system would have to be designed to deal with and resolve the data/document distinction. While it may be possible to do this it would never be easy, and, as Samuel Johnson once wrote, on another matter, " . . . . it is never done well, and one is surprised to see it done at all."

What Implications Do These Distinctions Have for Information Management?

The problem of information management is everybody's problem, from the data base administrator to the individual who sends his own records down to microgaphics to be put on film. All should recognize the fundamental distinctions in information management and the consequences of not making these disctinctions (see Table 2). Since information management is everyone's problem, and, by inference, everyone's responsibility, managers at all levels should understand that their subordinates should be allowed time to properly select, store and provide access to the information they use. Such time is not wasted if the individuals involved are properly advised. This advisory role can be provided for either formally (by establishing an advisory committee) or informally (by publicity or voluntary training).

The clearest implication of the four-fold taxonomy of information management (Figure 2) is that there exists no single solution to the information management problem. The idea of one comprehensive data base holding all of an organization's information (data and documents) is simply not a realistic goal. Information systems, to be optimal, must be tailored to the kinds of information they provide access to, and to the kind of

individual who will use them.

Some of the most powerful policies an organization can implement to deal with the problems of information access are not technological at all in nature. The retrieval effectiveness of any information system can be significantly, often dramatically, improved by instituting strict information retention policies and providing time for them to be carried out. One can still keep as much of the information as before. It is only necessary to remove the marginally useful information from the up to date operational information systems, and put it either in the archives or under the management of a less critically efficient information system. Of course it will always be better to dispose of useless information entirely, but in some situations it may not be easy to identify the useless information.

The management of information is a concern for all individuals regardless of what their primary duties may be. To successfully address the problems of information management an individual must understand what some of the fundamental distinctions and issues of information access are. But this understanding is only a beginning. The manager of today must have the courage and determination to use this understanding to make his own information a resource and not allow it to become an impediment to his normal duties.

Acknowledgement: The author would like to acknowledge the helpful comments of Michael Gordon, Alan Merten and Dennis Severance of the University of Michigan, and M. E. Maron and Patrick Wilson of the University of California, Berkeley. This paper was the result of research conducted under a faculty research grant provided by the Graduate School of Business Administration, Ann Arbor, Michigan.

FOOTNOTE

1. Wilson, Patrick. Two Kinds of Power: An Essay on Bibliographical Control, University of California Press, Berkeley, 1968.

2. Blair, David C. "The Data-Document Distinction in Information Retrieval," Communications of the ACM, forthcoming.

3. Blair, David C. "Indeterminacy in the Subject Access to Documents," unpublished, May 1983.

4. Swanson, Don R. "Information Retrieval as a Trial-and-Error Process," Library Quarterly, v. 47:2, pp. 128-148.

5. Blair, David C. and M. E. Maron. "An Evaluation of Retrieval Effectiveness for Full-Test Document Retrieval," unpublished, June 1983.

6. Blair, David C. "Searching Biases in Large Interactive Document Retrieval Systems," Journal of the American Society for Information Science, v. 31:4, July 1980, pp. 271-277.