Information & Management 20 (1991) 239–246
North-Holland

Research

The management and control of written information

Growing concern amid the failure of traditional methods

David C. Blair and Michael D. Gordon *
University of Michigan, Graduate School of Business, Ann Arbor, MI 48109-1234, USA

The control and management of written information in businesses is growing in importance, and the consequences of its mismanagement are coming more clearly into focus. Written information which cannot be found impairs decision making, planning, evaluation, and organizational growth. The problem is reflected in the costly duplication of effort when available information is not or cannot be shared. Managers are frequently overwhelmed by the rising tide of useless information that crosses their desks and often hides the occasional truly significant document. The control of written information is often based implicitly on the Library Model; that is, the control of written information in businesses is much like the management of books in a library. This article discusses what the difficulties are in applying the Library Model to business information systems and offers a more appropriate model for businesses to manage their written information.

Keywords: Information retrieval, Text retrieval, Document retrieval, Information resource management, Logical structure, Guidelines.

David Blair is an Associate Professor of Computer and Information Systems at the Graduate School of Business, University of Michigan. Professor Blair is interested in a variety of information retrieval problems, including problems of language and the management of information in corporate lawsuits. Blair is author of the book Language and Representation in Information Retrieval.

* The authors acknowledge the comments of Dr. E. Sibley and the anonymous referees in improving the quality of this article.

1. Introduction

> "A man's judgement cannot be better than the information on which he has based it." – Arthur Sulzberger
>
> "If we managed our money like we managed our information, we'd have been broke a long time ago." – Attributed to Delano Sloot

Organizations depend on textual information for their survival. But the problem of dealing with such information is becoming as important to companies as the decisions surrounding their main line business. Professional workers lose an estimated ten hours a week storing and locating information (Yourdon, 1986). Information which cannot be found impairs decision making, planning, evaluation, and organizational growth, often causing costly duplication of effort when available information is not or cannot be shared. Managers are frequently overwhelmed by the amount of information which crosses their desks and often hides significant documents.

American business deals with 400 billion paper documents a year, and this number is growing by

Michael Gordon is an Assistant Professor of Computer and Information Systems at the Graduate School of Business, University of Michigan. Professor Gordon's research interests include using artificial intelligence methods for information retrieval, information retrieval theory, and information retrieval in business.

70 billion documents annually. One firm discovered it was housing enough paper to make a stack eleven miles high[1], and electronic creation, reproduction, and dissemination of information is only increasing this problem. It has been estimated that "...the processing of documents will be the primary application of personal computers in the mid-1990s...it is conceivable that every desktop eventually will require a textual retrieval program" (Dataquest, 1988).

When a large firm decides to settle a $10 million lawsuit out of court because it cannot find a specific document vital to its defense, we must ask how long we can avoid paying proper attention to the management of textual information. It is time to look for effective techniques for the storage and retrieval of text.

2. Inapplicability of the Library Model for business' information management

Most efforts to conceive and design business information retrieval systems are based on a familiar model: storage and retrieval of textual information is like the storage and retrieval of books in a library. We call this the "Library Model". But a business that adheres to the Library Model will fail to manage its information effectively.

The first characteristic of the Library Model (see *Table 1*) is that all books are of equal value. With unlimited funds, libraries would obtain as complete collections as possible and, as a consequence, library collections would never diminish. As an archive, a library houses both old and new information. Even dated information provides historical context, being a basis of comparison or for indulging historical interests.

Though businesses, too, must look backwards as well as forwards, useless information can stand in the way of gaining access to day to day information. A business that only adds to its information will find it increasingly difficult to find the information it needs. The difficulty and cost of retrieving needed documents increases with the amount of information the firm manages. Thus, the exponential growth of print, text, and image information argues for increasing the rate with which firms *discard* unneeded information. Information important to business differs widely in significance, timeliness, quality, scope, and authority. Businesses need to know a lot, but they do not need to know everything.

The second characteristic of the Library Model is lack of cost associated with the mismanagement of information. At worst, a library may hear complaints from disappointed students or frustrated scholars, but it usually incurs no cost for lost information or for providing patrons irrelevant material they must wade through to find what they really want.

Conversely, information is integral to a business' survival. A company that cannot react to market changes, deals ineffectively with capturing and disseminating information, or cannot defend itself against criminal or civil lawsuits will find itself at a competitive disadvantage. A large company that was a co-defendant in a lawsuit had a surprisingly small settlement against it compared to its *smaller* co-defendants. Head counsel attributed this savings to the company's better ability to locate textual evidence to support its defense. Organizations which fail in their information responsibilities may ultimately fail as companies. Businesses pay a real financial price when they cannot locate needed information, retrieve and act upon information that is incorrect or outdated, or spend inordinate effort attempting to locate important information.

In the Library Model, patrons often look for books only for their edification or personal enjoyment. The librarian has no stake in the outcome: its constituents' success in finding what they want rarely affects the library itself. This is not the case for business. A manager may need information to incorporate into a report, use in developing a decision model, etc. Her unsatisfied need can affect the entire company. If she writes a report or designs a computer decision model (like a spreadsheet) based on insufficient or inaccurate information, the consequences will affect everyone who uses the report or makes a decision based on the model. Stronger efforts must be made to make information a true corporate resource. Textual information "warehouses" are being explored in some progressive companies today: these shared information bases support litigation, research, as

[1] Examples in this article of problems confronting given firms are from the authors' personal experiences. Many of these examples are detailed in (Gordon 1990).

Table 1
Comparison of the Library and Business Model of Information Management.

The Library Model	The Business Model
All stored information is of equal importance or significance.	Stored information varies in significance from essential to useless.
No cost or penalty for stored information which cannot be found by interested patrons.	Significant costs and penalties result from not being able to find essential stored information.
An unsatisfied information need affects only the patron seeking the information.	An unsatisfied information need may have repeated effects if it affects a memo, report, or computer program repeatedly used by the company.
Relatively convenient, uniform access to stored information.	Different information systems within the same organization may require different types and standards of access.
Centralized control of all information management processes.	Some kinds of information (e.g. financial) require central control, while others can be decentralized.
Centralized inventory of all stored information.	Centralized inventory of all stored information is desirable but rarely implemented.

well as day to day operations. The prevailing view must shift from regarding such efforts as "skunk works" to understanding them as "vital projects."

A fourth characteristic of the Library Model is uniform, centralized access to information. A centralized inventory of all the items is maintained by a comprehensive catalogue, and all the information stored is arranged according to a uniform logical structure such as the Dewey Decimal System or the Library of Congress Classification System. Such an inventory will serve business well, but such a logical structure for information is not appropriate.

The Business Model prescribes maintaining a centralized inventory of all information. Unfortunately, this is rarely done, and much valuable information remains "hidden" and known to only a few individuals or departments. The information which a business collects or generates is usually under the control of a single department or individual, but it should be a corporate resource available to other members of the organization. While the total corpus of information may be sufficient to conduct the daily affairs of the firm, if it is not accessible to all firm members who need it, it is not truly a *corporate* resource. Unfortunately, such hidden bodies of useful information will often be duplicated by those who need but who do not have access to it. One firm described conducting two duplicate text retrieval evaluation studies – at significant cost – because neither of the groups knew of the other's work. In another firm, a Vice President lamented that none of the divisional computing departments knew what the others were doing. "Hidden" information causes the firm to incur both the original cost of generating and maintaining certain information as well as the cost for multiple copies, multiple systems, and additional personnel.

While a centralized catalog is a good idea, a uniform logical structure is unrealistic for most businesses, and it may not even be desirable. Information may be generated and maintained by different departments, and each may have widely differing commitments to its management as well as a great variety of tools (automated and manual) with which to implement its information policy. As a result, there is seldom a uniform method to access all the information of a firm.

Organizations using the Library Model will often see the centralized control of all information as a realistic goal, but with rare exceptions, this is not only unrealistic it may be counter-productive. Different classes of information require different methods of structure and access. For example, the personal information on which a manager bases decisions may be conveniently stored in an ordinary filing cabinet, while the information needed to make rapid, informed marketing decisions may need to be continuously updated in a computerized system with a "friendly," non-technical interface; financial information would probably need to be maintained in a relatively well validated and secure information system to insure that it is private, timely, and accurate. Such a variety of standards would mean that a uniform,

centralized information system would be less than ideal for some classes of information and unnecessarily good (too costly) for others.

But this lack of homogeneity in a firm's information management policy does not preclude the need for some centralization of control and uniformity of access. The Business Model, like the Library Model, does need to maintain certain standards. For example, financial data located on different systems should be accurate to the same number of significant figures; all references to individuals should be made with unique names; sensitive or confidential information should be handled in a uniform and reliable manner; etc.

Another distinction between the Library and Business Models lies in the essentially passive attitude of libraries towards their patrons. Library patrons are free to search for any information in the public domain, but libraries generally do not actively attempt to determine patrons' needs and recommend appropriate information or information sources (although library policy is becoming more proactive with, for example, current awareness systems). Such a passive posture is arguably adequate for the typical library patron. But, with timing critical and the quantity of information business faces staggering, passivity in information management can mean disaster. It is vital that companies anticipate and prepare to receive the information they need. Current awareness systems, distribution lists on electronic mail networks, and user information profiles repeatedly applied against commercially available information retrieval databases can help one become better armed with information. Equally important are efforts to ensure that information reaches its appropriate destination and is screened from those who do not need it. Technologies for digitally scanning or converting printed sources into machine readable form can be combined with software and procedures that manage information flows by filtering and automatically routing information throughout an organization. An aggressive approach to dealing with information can involve in-house development of novel information systems or can come from applying pressure on the industry's large information system providers to build more functionality into their systems.

Convincing a firm to abandon the Library Model will not be easy. Nor will it be enough to ensure its success in storing and managing information. After a firm is convinced that documents vary in quality and significance, that failed information retrieval is costly, and that an aggressive posture toward assembling information and developing information management technology should be taken, it must come to grips with the logical design of an information system that will adequately support its needs.

3. Taxonomy of logical structure

The "logical structure of information" refers to the way in which information is conceptually organized; what information is easy to access and what is not. For instance, the logical structure of a phone book makes it quite easy to find Frank Cioch's phone number but difficult to determine the number of houses with listed phones on Fairview Circle (even though that information is contained in the phone book).

The most familiar logical structure for storing and retrieving documents is categorizing them by author, title, and subject. This, too, is a legacy of the Library Model. But, experiments conducted over the years have shown that such a logical structure is only partially successful in providing searchers with useful documents (VanRijsbergen, 1979). Too often, desired documents are missed and useless ones retrieved.

What alternative logical structure would be more appropriate? There is no simple answer to this because the appropriate structure must be based on the use of the documents by individuals who need them. There are four major types of information that can be used to represent documents:

1. Information contained in the document itself. This is generally the easiest information to gather, and could include data such as:

i. Author(s): the individual(s) or organization responsible for producing the document. In documents such as the minutes of meetings, the authors could be the individuals present at the discussion.
ii. Title of the document.
iii. Addressee(s): including both "action addressee(s)" and "information addressee(s)", if specified.
iv. Document type: memorandum, directive,

policy statement, minutes of a meeting, internal or external correspondence, product or market analysis, etc.

v. Date: the time when the document was created or received, an action (such as a response) should be taken, a reminder is to be sent etc.

vi. Medium: the format of the document, such as: hardcopy, microfilm/microfiche, or machine readable (data base or mass storage).

vii. Subject: a description of the intellectual content (what it is "about").

2. The context of the document:

i. Place of origin: where the document came from, such as an organizational position (the name of the department) a geographic location (the name of a branch office), or a "corporate position" (the name of another company or organization).

ii. Present position: where to find the document, and the owner with responsibility for maintaining a copy of it. It might specify a data base or computer file name if the document is stored in machine-readable form.

iii. Routing: who has received a document, from whom these recipients received it, time of receipt, action requested of the recipient, and action the recipient took in response. Such routing information addresses questions of assigned responsibilities and dissemination of the information.

3. The "value" of the document: some documents are immediately seen (or come to be seen) to be important or significant. Examples include corporate policy statements, strategic analysis of competing markets, formal or informal statements of commitment to clients, and evidence in a lawsuit. Very often these documents have implicit, or even explicit, security classifications. Significant documents may require concern for their *physical* storage and access: they should be stored in a physically controlled place that guarantees their survival during natural calamities, on a medium appropriate for the length and conditions of the required storage (e.g., "archival-quality" microfilm), and according to their anticipated use (e.g. some machine-readable formats are not admissible as evidence in a court of law (King and Stanley, 1985)). Finally, access to sensitive documents must be controlled to prevent unauthorized reading or copying of them.

4. The relation of this document to others: Many documents are implicitly or explicitly linked to others in an organizational activity. Requests for products or services require acknowledgments of receipt or commitments to provide what is requested. These should be linked. Correspondence or memos discussing the same organizational activity should be organized into a set of documents of mutual concern. Flores et al. (1988) take a "language/action" perspective on organizational communication. This theory is based on the performative nature of language, as exemplified in the writings of Austin (1962) and Searle (1969). It concerns itself with the observation that language is often used to accomplish things, such as when we promise, give our consent, vote for, declare our intention, appoint, dismiss, warn, advise, recommend, diagnose, estimate, analyze, evaluate, etc. What a computerized information system needs to manage is not information, per se, but linguistic activity that is directed towards completing some managerial task. For example, if an individual wishes to make a request, the message system should include the address of the recipient and a "respond-by" date, "completion" date, and "alert" date. The original request can then be answered either with a commitment or some kind of negotiation to provide something other than what was requested. In this way, one can ask queries such as, "Show me all of the requests that I have received but have not answered". Thus, messages are linked together according to how they are used to complete certain prototypical linguistic activities such as requesting, promising, etc.

Another way of linking documents is by their use in completing certain well-defined tasks. For example, a bank loan request requires several specific documents be completed by both the bank and the applicant. FileNet's (1989) "WorkFlo" software is a good example of an automated system which facilitates such activity. When a loan request is made, a computer image of an application form is generated. Next, Workflo generates images for all the necessary support documents to be filled out before the loan approval. It also will prevent completion of the request if the supporting documents have not been filled out and stored on the system. Similarly, the creation or receipt of

a document will automatically bring it – and all others necessary to complete the task – to the attention of the individual needing to complete it.

Sometimes the "links" are too ad hoc to be handled automatically. For example, documents which support applications for different loan requests by different individuals may need to be linked as, "bad" loans; or, documents may need to be linked because they discuss similar kinds of subjects (such as reports on new products). There is nothing unusual about a document being linked to other documents in several different ways: it originally may be dealing with making and satisfying a request, but later it may be desirable to link it to other documents that comprise a chain of evidence in a lawsuit.

4. Prescriptions for inducing useful logical structure

Clearly it is neither necessary nor desirable to keep all possible information describing a given document. But which descriptions should be used? Some general guidelines can be given:

1. The logical structure of a document data base must support the activity/activities in which documents are used. It is surprising how often document management is considered to be a problem of document storage – what is the cheapest and easiest way to store a given set of documents (Blair, 1984). But documents are kept because they might be of some further use. By anticipating how a document might be used, one can select the appropriate logical structure. For example, if a database of internal and external correspondence were designed to serve primarily the individuals who wrote the letters, then the logical structure of the data base would include, at a minimum, the names of the senders, the addressees, and the date of the correspondence. If the same data base were to be used by other individuals it may be necessary to provide access based on the "content" of the letters: the subject of each letter would have to be described. One way to discover how documents should be represented is to identify potential users and ask for examples of the kinds of queries they would like the new system to answer. For example, if a potential user would ask the question, "Give me the minutes of all the meetings which Stan Joyce attended last year" then it is important that the full names of all participants of recorded meetings be included as access points. It might also be important to be able to retrieve documents based on a range of dates (people often have difficulty remembering the precise dates).

2. Descriptions of the subject of documents should not be their primary means of access. In the majority of cases it is impossible to describe the subject clearly and unequivocally. Natural language is marvelously creative and there are innumerable ways to describe the same subject. Even trained indexers rarely achieve higher than a 75 per cent consistency in selecting subject terms to describe the intellectual content of a document; on average, it is much lower (Zunde and Dexter, 1969). Because of this indeterminacy, subject terms should not be the primary access points for document retrieval, unless the searcher does not mind missing a good number of the documents which deal with the desired subject (Blair, 1986). Further, subject descriptions should not be used until the data base has been partitioned along more certain lines. It may not be good to search a reasonably large correspondence data base for a known letter by asking for documents dealing with "marketing". First, the number of such documents might be excessive (perhaps in the tens of thousands). Second, the desired document might not have the subject term "marketing" assigned to it. The person who assigned the subject term might have thought the terms "advertising" or "promotion" described the letter best. Thus, to reduce the effect of this indeterminacy, the searcher should first be able to reduce the number of documents by specifying more precise descriptions such as a time frame, specific author(s), or department where the document was produced, before specifying a document subject (e.g., "Give me documents written between June 1986 and August 1987, by Boylan, Mulligan or Deasy, and having to do with 'marketing'"). By restricting the subject search in this way, the response will include fewer documents.

3. The number of documents in the data base affects its logical structure. If a retrieval system manages only a few hundred documents (e.g., for a small word processing system) then the authors, addressees and dates may be sufficient informa-

tion for retrieval. But if the data base maintains thousands of documents it may be necessary to add more information about the documents to insure later retrieval. One might add department codes to distinguish documents written by individuals with the same name; also it may be necessary to retrieve documents by intellectual content or subject if there are individuals who have authored a large number of documents, or if the authors are no longer with the organization (e.g., "Find any letters on the Dublin project that Molly Bloom wrote a couple of years ago").

The richer logical structure of the larger data base can reduce the number of "possibly relevant" documents that the searcher must examine – a problem with large data bases (Blair, Maron, 1985). As a result, a searcher can still attempt to retrieve a significant number of relevant documents without being overwhelmed by the number of documents that the system retrieves.

4. Information used to represent documents must be standardized. Problems arising from multiple descriptions of the same topic can be mitigated by a controlled vocabulary of subject terms. This mandates that a given subject must be described using a fairly precise terminology. For example, it could specify that documents concerning "records management" be described in that way exactly and not by semantically similar descriptions (such as "record management," "management, records," "records mgmt.," "information management", or "file processing"). By enforcing such uniformity, the set of potential subject designations to be used as search terms is greatly reduced.

Similarly, the names of authors of documents, the addressees of correspondence, etc., must be written in a standarized form. References to individuals within the document should be made using the full name, in case such references become important in the future (e.g., "retrieve all letters in which I discuss Kathleen Kearney's proposal").

It is also important to standardize the use of dates, written either numerically or using the proper name of the month. When dates are given numerically, it is important that it be clear whether the first number refers to the day or the month. This is especially important in multi-national firms since, in Canada and Europe, the first number is the day, while in the United States it represents the month. Such seemingly trivial differences can become major difficulties if you must find documents in a particular time frame.

5. Remove unneeded documents from frequently used systems to help maintain a useful logical structure. Organizations that look at the storage of documents from a purely cost perspective may argue that it is more costly to search for and weed out unneeded documents than it is to keep them. With the steady decrease in computerized storage costs, this is a very attractive and convincing argument. But it overlooks another important cost – the cost of finding needed documents. When a large number of useless documents is kept on a retrieval system, they become just so much "noise" in the system – impediments to the effective retrieval of important information (Blair, 1984).

Computerized word processing systems have increased the number of useless documents being stored on retrieval systems. Letters and other documents written on word processing systems often go through many revisions before they reach final form. These preliminary copies of the final documents are frequently retained as unique, useless documents. In fact, unless the final copy of a letter is annotated in some way, it may be difficult to remember which version is the final one; you may think that you wrote something in a letter which, in actuality, was not included in the version that was mailed. This problem is mitigated somewhat by systems which maintain version numbers or keep track of the exact time and date each copy is written. But, unless these copies are linked together, one still might find it difficult to tell which document was the final version.

6. It must be possible to change both the logical structure of a document collection as well as the particular information that is maintained to represent a given document. A document, referenced today by the author, may be needed later because it refutes an argument in another document. In this case, the information which describes documents must be adjusted to link them together. No combination of subject, date, author, document type, linkages etc. will be sufficient for all time. Thus, a flexible logical structure is necessary as new uses of existing information spring up. In this regard, adaptive retrieval systems hold great promise for the future. Such systems are capable of learning about the ways in which a document is

useful by noting how searchers ask for it (Gordon, 1988). As a result, such systems automatically modify a document's logical structure or the information describing it on a retrieval system. Until such systems are widely available, we must make provision for allowing *manual* changes to improve the descriptions and logical structure of documents.

References

Austin, J.L. *How to do Things With Words*, Oxford, 1962.

Blair, David C., "The Management of Information: Basic Distinctions," *Sloan Management Review*, 26(1), Fall 1984, pp. 13–23.

Blair, David C. "Indeterminacy in the Subject-Access to Documents," *Information Processing and Management*, v. 22:2, 1986, pp. 229–241.

Blair, David C., and Maron, M.E. "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System," *Communications of the ACM*, 28(3), March 1985, pp. 289–299.

Dataquest: Research Newsletter (1988-13) (Dun and Bradstreet Corp. 1290 Ridder Park Dr., San Jose, CA 95131-2398).

FileNet. FileNet Corporation. Costa Mesa, CA 1989.

Flores, Fernando, Michael Graves, Brad Hartfield and Terry Winograd. "Computer Systems and the Design of Organizational Interaction," *ACM Transactions on Office Information Systems*, v. 6:2, April 1988, pp. 153–172.

Gordon, Michael D. "Information Retrieval in Business: An Unmet Challenge." Working paper. Graduate School of Business. University of Michigan. 1990.

Gordon, Michael. "Probabilistic and Genetic Algorithms for Document Retrieval." *Communications of the ACM*, 31(10), pp. 1209–1218, 1988.

King, Roger, and Stanley, Carolyn. "Ensuring the Court Admissibility of Computer-Generated Records," *ACM Transactions on Office Information Systems*, 3(4), October 1985, pp. 398–412.

Searle, J.R. *Speech Acts: An Essay in the Philosophy of Language*, Cambridge University Press, 1969.

VanRijsbergen, C.J. *Information Retrieval*, Second Edition. Butterworths, London, 1979.

Yourdon, Edward. "Paper Chase: Keeping up with Office Productivity." *Computer World*. July 21, 1986. pp. 53–58.

Zunde, Pranas, and Dexter, Margaret. "Indexing Consistency and Quality," *American Documentation*, July, 1969.