DOCUMENT RESUME

ED 047 718 LI 002 578

AUTHOR	Leimkuhler, Ferdinand F.
TITLE	Systems Analysis in Libraries.
PUB DATE	68
NOTE	18p.; Paper prepared for the annual meeting of the American Society for Engineering Education, University of California, Los Angeles, June 17-20, 1968
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Engineering, Information Services, Information Systems, *Libraries, Library Planning, *Library Technical Processes, *Models, Operations Research, *Systems Analysis
IDENTIFIERS	*American Society for Engineering Education

ABSTRACT

A comprehensive engineering approach to the analysis of libraries as information service systems is described in which operations research type models are developed and tested for optimizing the design and management of document storage and retrieval processes. Extension of this approach to the development of more comprehensive system models is outlined in which the interaction of the librarians with the users and the funders is explicitly taken into account in assessing system performance. (Author)

Event No. 233

American Society for Engineering Education

Annual Meeting, June 17-20, 1968

University of California, Los Angeles

Los Angeles, California

SYSTEMS ANALYSIS IN LIBRARIES

Ferdinand F. Leimkuhler
Professor of Industrial Engineering
Purdue University
Lafayette, Indiana

Abstract

A comprehensive engineering approach to the analysis of libraries as information service systems is described in which operations research type models are developed and tested for optimizing the design and management of document storage and retrieval processes. Extension of this approach to the development of more comprehensive system models is outlined in which the interaction of the librarians with the users and the funders is explicitly taken into account in assessing system performance.

Systems Analysis in Libraries *

Like the person who was surprised to learn that he had been speaking prose all his life, some librarians may be surprised to learn that to some people a library is an information service system. It is from that viewpoint that we can subject the library to modern engineering analysis and program its orderly evolution into the computerized information center of the future. The importance of the "system" viewpoint in planning library innovations is seen best perhaps in the work of Fremont Rider, a librarian, who was among the first to grapple with the information explosion in a serious manner. He became certain that "no emendations in present library methods alone were going to provide a sufficient solution to our growth problem". He sought a totally new synthesis.

With remarkable insight and anticipation of the systems analysis viewpoint, Rider said:

> And the reason for our failure to integrate what were really facets of one single problem was that we were blinded by the status quo. We insisted on continuing to accept as library axioms, unalterable and unquestionable, certain assumptions which no longer had validity, such dicta, for instance, as: Libraries are collections of books; books are stored on shelves; library materials have to be cataloged; catalogs have to be made on cards; books must be arranged by their call numbers, etc., etc., etc.
>
> It is not until we have looked behind, and beyond, every one of these -- and many other -- supposedly basic axioms of library method, and have seriously questioned their validity as axioms, that we begin to make any real progress. For, when we do this, we are suddenly amazed to find all the mismatched bits of our research library growth puzzle falling, almost of themselves, into a quite astonishingly new synthesis.

Rider's new synthesis was the microtext catalog card which he there and then invented.

This recognition of the pervasive role of the printed book in conventional libraries, was taken up more recently by J. C. R. Licklider in his book on the bookless Libraries of the Future. Marshall McLuhan has observed how printing from movable type was the first mechanization of a complex handicraft and became the archetype of

---

* [Bas]ed on work supported by the National Science Foundation, Grant GN-519.

all subsequent mechanization. The book was the first teaching machine and also the first mass-produced commodity. Originally it was regarded as only a cheap way of reproducing manuscripts and it was not uncommon to have printed books recopied into manuscript form. Eventually the book lost this "horseless carriage" character and its enslavement to an older culture. So, too, today we force computers and microphotography to serve the older book culture at a considerable sacrifice of their potential efficiency in storing and processing information.

In order that we may proceed with the design and development of new kinds of libraries, it is necessary to achieve the same degree of objectivity evidenced by Rider or Licklider and at the same time be prepared to eventually cope with the full complexity of the information exchange environment. It is here that systems analysis and operations research can be of great service by providing analytic models and quantitative methods for the better understanding, design, and management of library-type systems. The mathematical simplicity of the models may make them appear overly abstract and irrelevant to the manifold complexities of real library systems, but it is precisely this quality that (1) permits experimental validation in any library, (2) provides a valuable connection with the scientific literature, (3) makes it possible to build subsystem models that can be assembled later, and (4) indicates the feasibility of eventually automating parts of the system.

For the past several years, a small operations research group has been engaged in the study of problems in the design and operation of research libraries at Purdue University through the joint sponsorship of the Libraries and the School of Industrial Engineering. From the beginning, a continuing seminar, attended by the interested faculty, students, and members of the library staff, has met regularly to review proposals and reports on research activities both within and outside the university and to foster a common understanding of the nature of the problems to be considered and the methods of analysis employed. Over the years a considerable amount of mutual education has taken place, an excellent environment for the conducting of research has been created, and some major research efforts have been completed.

Operations Research Models

At the heart of most operations research work is the development of analytic models which both describe in a meaningful way the behavior of man-machine systems operating in natural environments and predict how the system can be regulated or improved upon through the manipulation of certain control variables. A wide variety of models can be employed for these purposes, ranging from the physical analog to the highly abstract. Among the easier models to visualize are such things as photographs, maps, drawings, scale models, pilot plants, and field tests. In short, any representative of reality is a model. Hydraulic and electronic circuit analogs of network flows lend themselves well to problems of continuous multi-dimensional measurement and can be applied to information flow problems. Digital simulation on computers can be used for the same kinds of problems and generally allows for more flexibility in the configurations being studied and ease in the processing of test results. With analog and digital simulation the assumptions about the operations of a system being studied can be kept quite simple, and these operations can be multiplied and iterated almost indefinitely through machine manipulations. In this way, complex systems can be broken down into many simple elements which can be programmed to operate in parallel or sequence as needed to generate a measure of the cumulative, over-all effectiveness of the total system.

A higher level of elegance and sophistication can be reached with mathematical models at the cost of greater precision in the statement of assumptions and some loss in the ease of enumeration. As with literary language, mathematics can be made to convey an enormous amount of meaning in a very concise way to those who are prepared to understand the fullness of the language. On the other hand, admittedly, it can be a prohibitive barrier to the layman, it can be used to disguise the true facts in a study, and can lead to sterile reworkings of old ideas. But when properly used, quantitative models can provide the most general insights into the behavior of a system, allowing for experimental validation and meaningful comparisons with other systems. Furthermore, it provides a valuable connection between a specific application and the general scientific literature so that important discoveries in other fields can be brought to bear on problems in a different field.

Among the classical mathematical methods that have proven useful in library studies is that of marginal analysis or cost minimization. With marginal analysis, for example, one identifies the various cost components which are dependent on the activity level of one or more controllable factors which tend to increase costs in some cases and decrease them in others. An optimum mix of factor levels is then sought where the total combined cost to the system is minimized or where small changes in the level of any factor would produce a net cost effect which is undesirable. When such problems can be held to a few highly important variables and where the cost functions can be approximated by well-behaved mathematical relationships, this method is extremely useful and powerful.

In large-scale systems involving many variables and the satisfaction of many system specifications, it is often possible to formulate an approximate mathematical program which can be solved on a computer to yield optimal or near optimal system configurations or sequential decision-rules for a piece-wise solution to the problem. For example, the location of branch distribution points over a network of population centers, the allocation of budgeted funds over the many items in a supplies inventory, or the long-range planning for the budgeted development of departmental libraries in a university system; these are the kinds of problems that lend themselves well to the modern techniques of linear, quadratic, integer, and dynamic programming.

Another classical method that should prove to be of greater value to libraries than it has been is that of statistical inference, whereby one can make reliable statements about the content and behavior of large book collections on the basis of relatively small samples and without recourse to the chore of total enumeration that is often used. At Purdue, statistical sampling methods have been developed and used extensively to test various hypothesis about library characteristics. Among the newer developments in the field of statistics is the subject of stochastic processes and its application to systems which exhibit random behavior patterns over time and can only be described in a probabilistic sense. Both its earlier application to telephone systems and waiting lines and its newer application to traffic streams and switching networks should be transferable to the library scene and helpful in the planning of library facilities and staff requirements as well as in the establishment of operating policies.

Again, where such dynamic systems involve complex networks of many random and partially random components with a high degree of interdependence it may not be practical to attempt a completely analytic description, but better to simulate the system's behavior on a computer. However, a considerable amount of analytic work is generally required before making best use of computer simulation. Just as linear programming has extended so greatly the ability to handle problems having many variables, digital simulation has greatly extended the ability to study dynamic systems over long periods of time. Where particular problems lend themselves well to these techniques it is often better to "explode" a problem artificially into a many-variable or many-period problem rather than simplify the problem as one with a few variables over a single smoothed horizon.

One of the "quackisms" that operations researchers have tried to avoid is that of forcing real world problems into mathematical forms which have known solution techniques. Preferably one should take the world at face value and work out the analytic structure which best fits the empirical evidence. In practice, however, there is a certain amount of compromise in order to make headway and gain a few early results. One way operations researchers have maintained a certain amount of empirical validity is through the small multi-disciplined group approach. The mixed team consisting of both theoreticians and practitioners has a better chance of avoiding the "single technique" approach and coming up with a truly authentic analysis of the problems at hand. In any event it must be recognized that opsearchers have been least successful with regard to the implimentation of their findings and are somewhat inept at the quick answer and stop-gap measure. It is the task of astute executives to make themselves aware of what is known and exploit it for all that it is worth.

Shelving Models

One of the first problems the Purdue library O.R. group studied was that of the efficient storage of library collections. An early study sought to develop a precise measure of the space requirements for various shelving arrangements and to devise an algorithm for finding that arrangement which would minimize the wasted space. This problem was not chosen because it seemed to be the most pressing issue facing libraries at the time, but for the less noble reasons that it was interesting, seemed amenable to analysis, and was a good place to start. In choosing to optimize one piece of any complex system, one always faces the bug-a-boo of narrow suboptimization to the detriment of other parts of the system. By dealing with a purely physical problem in two and three dimensions with rather obvious measurements, it seemed plausable to pursue this problem in something of a vacuum and make the necessary qualifications later.

It was only after some useful mathematical results were obtained and some barbed jibes were directed at our attempts to measure knowledge with a ruler, that the true practical importance of the work has begun to reveal itself. Fremont Rider, a true pioneer in library systems analysis, has argued convincingly that space is the most costly aspect of a library's business but because of the methods used in measuring and reporting costs it is usually a hidden item that doesn't get the attention it deserves. He devoted a great deal of his time in the Wesleyan University library to experiments with space reducing techniques and reported some very impressive results. As with all methods studies of this kind and at that time, it was difficult to generalize the results to another library and thus provide a truly convincing argument for others to follow suit. The best solution for one library is not likely to be best for another, but a good method of arriving at a solution should have general validity in other libraries.

Melvil Dewey became world famous for his relatively simple mathematical model for shelving books according to a decimal classification. It has the important mathematical property that it is theoretically unlimited in its capacity to accept new additions and preserve a unique ordering among the items. It does this at the expense of such spatial and temporal factors as random book sizes, the need for gaps and reshuffling along the shelves, and the increasing complexity of assigning unique catalog numbers. Dewey was aware to these inefficiencies and recommended subdivisions by size and age, and all libraries practice such departures from strict subject classification to some degree. The question arises that: if it is profitable to divide a collection at least once into say an oversize and a remaining group, how profitable is it to do this and what is the optimal size to use for this division point? Furthermore, if it is profitable to do this with one division, what about using two, three, or more divisions and where would these cuts be made?

The models developed at Purdue follow this line of questionning. Given a collection of books whose dimensions are known, divide the collection once at each possible point and measure the gain in space at each point. Choose the point that is best. Now do this for all possible pairs of subdivision points, measuring the net gain, and recording the best pair. By continuing this process until you have as many subclasses as there are book sizes, there is generated a whole set of possible shelving patterns with their related spatial "cost", and for any number of subdivisions the best cutting points are noted. It is readily seen that anyone with a little arithmetic ability and a great deal of time on their hands could work this sort of thing out for any library.

What was needed was an efficient technique for collecting the empirical data on sizes, and grinding out the cutting points and space savings, not just to make the chore easier, but to be able to get to more generalizeable results that should have validity for many libraries, and also to permit the introduction of constraints and dependencies which would enhance the realness of the model and its solutions. This was accomplished in part by using statistical sampling techniques to collect the data. A relatively efficient iterative technique was devised for computer evaluation of the cutting points, and later the problem was reformulated as one in dynamic programming which greatly reduced the computation of solutions. Further efforts have been made to obtain closed-form analytic solutions by making assumptions about the distributions of book sizes and the correlation of various dimensions and age. Such real world considerations as stack widths, ceiling heights, and shelf thicknesses have been introduced into some models to see what effect they would have on the unconstrained solutions.

Table 1 gives the results for shelving a representative sample of the Purdue collection at optimal shelf heights. It can be seen that the additional gain in capacity decreases quite rapidly as the number of size classes increases. The Purdue studies indicate that no more than three or four classes would ever be needed in a compact storage area. Further studies of shelving systems are needed especially with regard to their cost and compatibility in the total library context.

Table 1

EFFICIENCY OF STORAGE BY SIZE

Number of size Classes	Potential Capacity Increase	
	Storage by Height	Storage by Width
1	0 percent	27 percent
2	38 percent	94 percent
3	47 percent	111 percent
4	51 percent	116 percent
5	53 percent	121 percent
10	58 percent	128 percent

Recently, the compact shelving models were applied to three of Purdue's branch libraries: chemistry, physics and pharmacy. Twenty percent samples of the collections were taken to get the size distributions in one inch increments. It was found that there was little justification for making more than three divisions by height, as shown in Table 2. The percentage gain in space was based on the present space utilization where books are stored flat and on their spines, and shelves are set for the tallest book, all of which conserve space in a somewhat haphazard manner. As a variation on the original model, the option of storing all books over 13 inches on their spine was considered for chemistry and the results are shown in Table 2 also.

Table 2

Compact Shelving vs. Conventional Shelving

A. Chemistry (All up-right)

Heights	Efficiency
17 in.	73%
17, 11	105
17, 13, 11	110
17, 13, 11, 10	112

B. Chemistry (Over 13" on spine)

Heights	Efficiency
13 in.	96%
13, 11	111
13, 11, 10	114
13, 12, 11, 10	115

C. Physics

Heights	Efficiency
15 in.	89%
15, 11	114
15, 12, 11	118
15, 12, 11, 10	122

D. Pharmacy

Heights	Efficiency
14	95%
14, 11	115
14, 12, 11	117
14, 12, 11, 10	119

Storage and Usage Models

While the shelving models could demonstrate the availability of space savings through compact storage by height, they do not take into account the disadvantages to users of having to consult several locations for a particular need. However, this was taken up in subsequent studies, one of which considered the situation where a large collection was to be divided into two parts: one part housed in a high cost manner at the time of acquisition and the remaining portion in a low cost manner such as in a depository. There are costs associated with moving items from the prime to the secondary area, and a user penalty for obtaining a book from the depository.

Two selection policies were considered, one based on age and the other on inactivity. In each case, the questions to be answered were: what are the optimal cut-off points, what proportion of the collection should be stored, what are the savings in total cost, and how sensitive are the results to the parameter estimates? Cost models were developed for both policies, and algorithms were devised for minimizing total cost at the optimal cut-off points.

Estimates of the pertinent costs were obtained from the literature and from interviews with librarians. Data on growth, usage related to age, and on inactivity patterns were obtained from three libraries. Evaluations of both policies for each of the three libraries were made and the results were compared over fairly long planning horizons.

An important conclusion from this study is that a significant savings may only be realized if the added cost to circulate a book from storage is small; and with this condition somewhat paradoxically a large portion of the collection must be stored to achieve maximum savings. Specifically, if the user penalty is negligible then as much as 60% of the collections should be stored to gain a total cost reduction of about 10%. However, if the user delay penalty is of the order of a dollar per event and this cost is to be borne by the library, then it is better to maintain a single collection and not have a depository at all.

This analysis indicated also that selection policies based on book inactivity may be more efficient than those based on age. In either case, the quality of the storage selection activity would depend on the ability to predict future usage, and this was the subject of another important study at Purdue which is just being completed. This investigation began with a thorough review of the many usage studies (perhaps 700) which have been conducted in libraries over the past 50 or 60 years.

For the most part the past studies have failed to come up with a definitive technique for evaluating and predicting book use. One of the contributions in the present study is a method for combining current usage with carefully planned collection samples so as to come up with a measure of relative use which allows for meaningful comparisons among different classes of material and between libraries. Furthermore, there is no longer the need to rely on questionable book records.

The emphasis in this study was a comparison of the various mathematical models for describing book use as related to age, which have been advocated in recent years. By analyzing the models, it was possible to identify a few major classes of models which contain all the others. In addition, a new model was developed and tested which has the capability of separating inactivity patterns from activity patterns for any particular item or group of items. By isolating the probability of no use, it is possible to predict item inactivity as a function of age. This allows for a more critical evaluation of the policy differences mentioned earlier.

In general, this recent usage study appears to vindicate the use of age as a more important factor affecting usage. It may take considerably more research and application studies before we are able to fully exploit these results in the operation of depository schemes, but a very good start has been made.

Search and File Organization Models

In a strict sense all of the models discussed thus far have been concerned with the problem of file organization which is central to all information systems. In a library context, there are two main classes of files: document files and the bibliographic files. With bibliographic files we generally have homogeneity of form such as with cards and tapes so that the problems of space are reduced to one of simple linear growth. With document files, there is bulk, randomness in size, problems of access, handling, deterioration, loss, and obsolescence. In short, all those messy details with which librarians must cope.

By working at a level of abstraction, once removed from the physical realities of document handling, documentalists have been able to focus their attention on the production and distribution of information. While libraries are designed on the book module, as the standard for batching information, documentation services have built their systems around the journal article and the research report, while more recently the still smaller abstract and data entry are becoming the working units of specialized information centers. All of these are better suited than books to microform reproduction.

Through this process of explosive subdivision, it is possible to speed up the flow of information and exercise finer control over the admission of new knowledge to the record. This advances the quest for exhaustiveness through specialization as opposed to the quest for comprehensiveness through substitutability that marks the general library.

Large research libraries are in a sense crude realizations and rough approximations of the more refined and abstract information networks of the world of scholarly ideas and facts. They also grow explosively through subdivision and are able to exploit and economize through consolidation. Thus, there should be close analogies between the information research of the documentalist and the problems faced by libraries.

The current revolution in computer-aided, micro-analytic techniques of information retrieval was largely justified by earlier studies of the literature used in scientific research. These studies showed that researcher interest was widely scattered among formal titles and classifications and much more extensive than the coverage provided by existing bibliographic services. The survey of work prior to 1950 by R. E. Stevens identifies "title dispersion" as the most important statistical characteristic to emerge from the many empirical studies that were made. Title dispersion is defined as the degree to which the useful literature of a given subject area is scattered through a number of different books and journals.

Of particular interest is the work of S. C. Bradfor, who carefully studied the title dispersion of useful papers in two areas: applied geophysics and lubrication. By arranging source titles in order of productivity and then dividing them into three approximately equal groups, Bradford concluded that the ratios of the titles in successive zones followed a common pattern, and proposed the following "law of scattering":

> If scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject, and several groups or zones containing the same number of articles as the nucleus, where the numbers of periodicals in the nucleus and succeeding zones will be as $1:n:n^2$.

In a recent paper it was shown that Bradford's verbal statement could be expressed more usefully in a mathematical model, called the Bradford distribution. There are several advantages in doing this. It can be shown that Bradford's law is closely related to Zipf's law and to a whole family of functions proposed by various statisticians for the study of human behavior. It makes possible efficient computerized techniques for digesting the data and drawing more precise conclusions about the differences in the literature between fields.

Of operational importance is its adaptability to file organization models which can relate file subdivision to search efficiency. The rationale of these models starts with the notion that in searching for information one would prefer to examine the potential sources in decreasing order of their likelihood of yielding the data sought. This requires a pre-knowledge of the file through experience or through such indicators as key words, citations, etc.

A certain amount of the cost of pre-structuring a file could be avoided by using cruder divisions, as for example, by just dividing the file into two groups of more and less likely sources and making no further discrimination within the two groups. With less discrimination, there is a corresponding increase in the expected search time, but by how much? And, given the number of subdivisions to make, where should these be made so as to minimize search effort? The answers to these questions have been worked out for a file with a Bradford distribution of item likelihood, and evaluated for a wide range of parametric vairations.

The general result is that a "two bin" system of file organization provides a very marked improvement over a single undifferentiated file. It appears that an initial search should cover only that 15 to 20 percent of the collection having the greatest likelihood of yielding the target information. This should satisfy about two-thirds of all searches, while the remaining third would have to go on to a secondary search over the remaining 80 to 85 percent of the file. If a "three bin" system is used, the gains in search efficiency are not as profound but could be substantial for some situations.

In general, these studies have pressed the case that uniformity in document storage is generally inefficient with respect to certain costs and that simple bifurcation of large collections may be used to yield modest but substantial economics. However, it remains to be shown that these separate ventures in sub-optimization can be combined for the joint improvement of a library system. There is need for more applications studies to validate the theoretical results and to extend their usefulness.

Final Remarks

These studies do not in any way exhaust the possibilities for quantitative analysis in libraries, rather they suggest only one line of thought which has been followed over several years by a small group of opsearchers. Another group would probably follow a completely different tact, and indeed at Purdue there are other researchers doing just that. To the extent that good communications can be maintained these various approaches should ultimately converge in a general theory of research libraries.

In order to reach a truly comprehensive understanding of library-type systems, it is necessary to pursue research on two levels simultaneously. At one level are the microanalytic models of the kind described above in which one aspect of a system is studied exhaustively. At the other level are the macroanalytic models which seek to simulate more of the totality of the system without worrying about a too precise refinement of all of the details. This latter approach can serve to reveal the long run patterns of system behavior, which microanalysis often fails to achieve. Furthermore, macroanalysis is perhaps the surest way to begin to identify the overall objectives of the system, the true decision-makers, and the importance of their roles in shaping system performance.

In the final analysis, operations research is concerned with decision-making as it occurs in complex man-machine systems operating in natural environments. In order to study such activity in an analytic manner it is necessary to postulate a theoretical framework which describes the behavior of the decision-making participants in the organization, and relates their behavior to some well-defined and accepted measures of effectiveness and/or efficiency. In a library, we can identify three major classes of participants: users, funders, and librarians. How these groups interact in the establishment of "acceptable" behavior on the part of the library is the focus of a current macroanalytic research study at Purdue which compliments the microanalytic studies of the technical options available to the library in meeting user needs and satisfying funder constraints.

References

1. Baker, N. R. (1966) A Descriptive Model of Library/User/Funder Behavior in a University Environment, Purdue University, U. S. Clearinghouse Report PB-174-439.

2. Cox, J. G. (1964) Optimum Storage of Library Materials, Purdue University Libraries, Ph.D. Thesis.

3. Jain, A. K. (1967) A Statistical Study of Book Use, Ph.D. Thesis, Purdue University.

4. Leimkuhler, F. F., and Cox, J. G. (1964) Compact Book Storage in Libraries, Operations Research, 12, 3, 419-427.

5. Leimkuhler, F. F. (1966) Systems Analysis in University Libraries, College and Research Libraries, 27, 1, 13-18.

6. Leimkuhler, F. F. (1967) The Bradford Distribution, Journal of Documentation, (to appear in Fall of 1967).

7. Leimkuhler, F. F. (1967) A Literature Search Model, Purdue University, U. S. Clearinghouse Report PB-174-439. 31st Meeting Oper. Res. Soc. of America, New York, June 1967.

8. Lister, W. C. (1967) Least Cost Decision Rules for the Selection of Library Materials for Compact Storage, Ph.D. Thesis, Purdue University, U. S. Clearinghouse Report: PB-174-441.

9. Nance, R. E. (1967) Systems Analysis and the Study of Information Systems, Nat'l. Meet. Amer. Doc. Inst., New York, October 1967.

10. Nance, R. E., and Baker, N. R. (1967) The Use of Simulation in Studying Information Storage and Retrieval Systems, 32nd Meet. Oper. Res. Soc. of America, Chicago, November 1967.

11. Popovich, J. D. (1967) Compact Book Storage, MSIE Project, Purdue University.

12. Raffel, L. J. (1965) Compact Book Storage Models, MSIE Thesis, Purdue University.