COMPUTER SIMULATIONS AT THE COLUMBIA UNIVERSITY LIBRARIES

Warren J. Haas

The woods are full of those who have harnessed machines—the professional literature of the past few years reports many breakthroughs of far-reaching importance and is filled with one success story after another in such diverse fields as auto-abstracting, file loading, machine translation, and successful search strategies for electronically massaging huge masses of stored bibliographic data.

For some reason, evidence of difficulties or of failures does not seem to float to the top as easily. While we have not been at this business long enough to be counted failures, we can certainly lend perspective as far as difficulties go. To characterize Columbia's libraries will help put the description of both our activities and our problems in context.

First, a few notes on size. The general cataloged collections include about 3.2 million volumes, and they grow by about 100,000 volumes each year, only about half of which are in English. Nearly 50,000 serial titles (including documents) are acquired on a current basis. A million or more manuscripts, and an adequate number of items in other typical categories such as technical reports, maps, scores, and microtext are also on hand. The full- and part-time staff, in full-time equivalent, now numbers over 400. Over a million and a half books are charged for outside use each year, and a hundred thousand overdue notices are written to get them back. On an average day during the academic year, readers enter one or another of the thirty-two library doors on the campus 16,000 times—a figure about equal to the full-time enrollment in the university proper.

The annual operating budget, supplemented by special funds, is close to $3,000,000. It is estimated that over 60 per cent of this amount goes to support research while something less than 40 per cent goes to support the instructional program of the university. Roughly one-fourth of this sum goes for books, journals, and binding;

Warren J. Haas is Associate Director, The Libraries, Columbia University, New York City.

4 per cent goes for expendable supplies, leaving approximately 70 per cent for direct wage and salary payments—about one-third for technical service departments and two-thirds for reader service departments.

Organizationally, all library units, including law and medicine, are administered by the Director of Libraries. Two library units are operated on contract—one for the National Aeronautics and Space Administration and one for the National Institutes of Health.

While still not as large as Harvard University or the University of Illinois, Columbia is out of the special library class, and it is already apparent that leading a library of this size down the path of automation is a difficult task.

How do we begin to automate significant portions of this somewhat unwieldy organization? One thing that seems certain is that the adage "the bigger they are, the harder they fall," is valid beyond question. A library with sharply limited subject responsibilities, or with a small (or at least a homogeneous) group of readers, or one without a catalog of both rational and irrational procedures and practices shaped by years of history has a far easier path to follow in instituting radical changes than does a general research library. It took a team of experts two years to decide if it was feasible to begin to plan how to automate the bibliographic processes of the Library of Congress. It will probably, and unavoidably, take two more years before it is decided whether or not to take the next step—that of planning (or more accurately, inventing) the system required to accomplish automation. Without dwelling further on this fact, it may be asserted that large general research libraries, unlike specialized libraries, are not transformed overnight.

But we cannot sit back and do nothing simply because what needs to be done is difficult and slow. Columbia, like many other institutions, has been dipping its toes in the water of automation in recent months, principally to test the temperature before actually committing itself to taking the bath.

A related project to the Yale-Harvard-Columbia Medical Catalog project is centered in Columbia's engineering and physical science group of libraries. A detailed systems analysis has been under way for a short time. The object is to create a record system that will take up at the point an item is selected for the collections (whether before or after acquisition) and be used for all subsequent transactions and processing activities. As a first step (and as evidence that Columbia is serious), those science units not using Library of Congress (LC) classification were switched July 1, 1963, to help implement a concept of collection mobility judged to be an inseparable part of automatic record generation. A draft of a universal process form has been developed, and during recent weeks it has been walked through the various phases of processing to eliminate some of the

more obvious "bugs." Among the things aspired to are printed book catalogs, a weekly printout report called "status of selections," perhaps a Selective Dissemination of Information (SDI) system, and a number of other output products. Two specifications for this project are that it be compatible with the medical program and that it be flexible enough to be extended to other subject fields. We are not walking here yet, but we do seem to be beginning to crawl.

Another example of Columbia activity, and one in which progress might come quite quickly, is in what might be called one of our special libraries. In January 1964, Columbia contracted with the National Institutes of Health to develop and operate a national information center on Parkinsonism and related diseases of the basal ganglia. Without going into detail, it may be noted that the services of the center, which is a possible prototype for other disease-oriented research and information centers, are to include on-demand searches of literature to produce bibliographies as well as substantive data, publication of critical reviews of reports of work done in pertinent subjects throughout the world, organization of symposia, creation and maintenance of a "who knows what" type of file, etc. Work on a thesaurus of terms is under way as a first step towards creation of a machineable file of bibliographic information, and initial planning for a comprehensive information system has started. A distinctive characteristic of this project is the provision that a portion of the salary of each doctor and scientist attached to the Parkinson Research Center is charged to the information center contract—a device designed to stimulate participation of the scientific staff in the work of the information center.

A fair amount of spade work in other areas has been done in recent months—for the most part, it has been directed towards learning more about what is already known. For example, Columbia has a descriptive inventory of all currently maintained records—bibliographic, personnel, process, statistical, etc.—in the library system. They total about 1,000. Much information about the flow of material through the system by the use of log sheets inserted in several hundred sample items as they were unwrapped in the shipping room has been gathered. As a matter of fact, about 10 per cent of these forms have not yet returned to home base—but it has been only a year.

These examples, along with several others that might be noted, suggest perhaps that a crash program to automate Columbia's libraries is gaining momentum. Such is not the case. Columbia's objective is not automation. It is rather to provide effective support for each of the many and diverse instruction and research programs that constitute the work of a complex university. The library services required must be appropriate in type, in quantity, and in quality. They must be flexible to meet changing needs, and they must at the same time offer continuity and incorporate perspective.

Automation will certainly help us achieve service with these characteristics, but at the moment, we are more concerned with what we do, rather than how we do it. Neither Columbia nor any other library can fulfill its obligations by doing better and better what need not be done at all.

As libraries grow in size, the process of program development and performance evaluation becomes more complex and less subject to critical administrative review. In itself, this is not necessarily bad because responsibility for this kind of review can be shared on a wider base. But this same element of size makes it difficult for the larger group of operating policy makers (40 or 50 people at Columbia) to be aware of all the facts pertinent to the problem at hand.

The problem alluded to here is deceptively simple, and can be stated in many ways, but essentially it is this: How can we make certain that we select a proper course of action from among a number of alternatives to achieve an objective that is itself related to a whole complex of other objectives?

Several months ago, Columbia embarked on a type of operations research program known as Simulation of the Columbia University Libraries (SCUL), in an effort to see if a way could be devised to give insight into this fundamental problem of library operation.

Briefly stated, the specific objective of SCUL is to study the comprehensive research library as an economic system. This approach has been successful in some business applications, and the fact that much work has been done in the study of economic systems using computer simulation and mathematical modeling techniques has enabled the SCUL study to capitalize on the experience of others in the field.

At this point, only a part of the first phase of the project, essentially a limited feasibility study, has been completed. The product of this initial effort includes a computer program that simulates the interaction of readers with materials in Columbia's Engineering Library, an outline of proposed mathematical approaches to the task of creating an economic model, a distinctive questionnaire designed for the collection of some of the required data, and a fuller realizations of the magnitude of the job we have proposed for ourselves. At the moment, funds required to get on with the main job are being sought.

The form that SCUL finally takes is certain to differ from its present state as general concepts are molded to fit library application. In brief outline, the project incorporates development of a probabilistic simulation model of a library in the form of a computer program that will be used to game with patron sets to study the nature of the interaction in varying situations between categories of patrons on the one hand and categories of library materials and library facilities on the other. The output from the simulation model will be a measure of the "satisfaction" experienced by each patron category for any given

mode of library operation (actual or hypothetical). This "satisfaction function" for a single category of readers will be adjusted in the context of the total patron population using the technique of multiple regression and will be associated with relevant cost information. Comparable information for each alternative course of action will be similarly developed. This information will be used as input to an economic model yet to be devised, and will be analyzed using linear programming to determine the mix of alternatives that best satisfy some stated goal.

In the sections that follow, the major parts of the SCUL project are described as they have been developed thus far.

I. The simulation model.—Most SCUL project time has been devoted to the development of a computer program that will serve as a prototype for a general library simulator model. In essence, the program that has been written is used to create a "computer duplicate" of the public service side of Columbia's Engineering Library.

A dynamic replica of the library is created by playing library patrons, library facilities, and library stock against each other to analyze the complex relationships that exist between these three elements to learn more about the demand on stock and to establish the satisfaction of patron groups in any given mode of library utilization. The model can be operated under different conditions in order to (1) analyze in detail the real-life library, (2) to determine the effect on this library of a shift in the composition of the patron group using it, and (3) to investigate the effect on service (patron satisfaction) of changes in management policies affecting facilities or stock.

As a first step in formulating the simulation, each of the three operating elements in the model were categorized in the following manner.

I. Patrons, or the population using the library

Major Categories	Minor Categories
Undergraduates	Chemical engineering
Graduate students	Civil engineering
Teaching staff	Mechanical engineering
Research staff	etc.
etc.	

II. Facilities

A. Those provided for the comfort and convenience of patrons:

Furniture
Microtext readers
Photocopy equipment
etc.

B. Library intermediaries between patrons and stock, including:

Card catalogs
Indexes and abstract journals
Reference librarians
Clerks
etc.

III. Stock, or objects in the library containing information used by the patron population:

Major Categories	Minor Categories
Books	Format
Journals	Full size
Technical reports	Microform
Theses	
Miscellaneous	Date
	pre-1951
	1951-1960
	1960-
	Language
	English
	Romance
	etc.
	Type of loan
	Non-circulating
	Overnight
	etc.
	Use
	Reference
	Reserve
	etc.
	Subject
	etc.

The second step in developing the model was the construction of detailed flow charts tracing the paths of patrons entering the library, performing one or a number of possible functions, and then leaving. Following completion of the charts, the program which translated the flow charts into computer code was written.

A deck of punched cards, representing a set of patrons, is processed through the simulator program, duplicating the flow of a set of real patrons through a real library. The route each "patron" takes through the simulator model is established by a gaming process. At

each decision point, the program compares a known probability that the specific patron will perform a specific function with a random number generated by a subprogram within the simulator. If the random number is equal to or smaller than the known probability, the decision is "yes"; otherwise, it is "no." Step by step through the program, courses of action are determined by probability tables tied to each decision point.

In another mode of operation, the simulator can process "patrons" in a nonprobabilistic manner—in effect, specifying that a patron will follow a specific path or will use a specified facility.

For each run of a set of patrons, summary reports of patron action and library performance for each patron category are prepared. From these reports, the "satisfaction function" already referred to is calculated for use as input into the economic model. Because the required data has not yet been collected, runs so far have been limited to small sample sets, and the probability tables have been artificially generated.

While simulator output is generated primarily for use in the economic model, it is hoped that it will be useful in itself, since the model produces an analogous account of how the library's facilities are being utilized by the patrons and how well the demands of the various patron categories are met. The model will also predict the changes in stock demand resulting from a change in the proportions of patron categories utilizing the library. Further, the simulation model is also a laboratory library, because it makes possible tests of alternative management decisions and thus provides a way to assess changes before they are actually made.

II. Data gathering.—There are two types of probabilities involved in the library simulation. The first describes the order, or sequence, of patron activities, and the second describes the patron's probability of success. To gather those facts about present library operation that are required to develop the probability tables, a questionnaire in the format of a flow-chart has been developed. The questionnaire has been tested in the Engineering Library, but is has not yet been put to large scale use. It is also possible that the results of work being carried out at the Massachusetts Institute of Technology will provide probability information that can be used in this phase of the SCUL project.

III. The economic model.—The second model implied but not yet developed for the project is an economic model that will hopefully provide insight into a wide range of administrative problems by answering questions of the following type. Given a set of alternatives in library service to various patron groups and a specific allocation of funds to the library (the library budget) and given a set of requirements imposed on the library (service goals), what is the optimum distribution of the allocated funds to satisfy the requirements set?

In brief, this model will characterize the library operation as an economic system. The output of the simulation model (e.g. the derived "satisfaction function" for any or all alternative methods of operation) is coupled with cost information and analyzed by a linear program to determine the mix of alternatives that maximizes the effectiveness of the library for every dollar spent.

Only tentative approaches to the construction of the economic model have been taken. The entire process promises to be an undertaking of great complexity. The determination of cost information for existing modes of operation requires extensive and imaginative study; to establish meaningful costs for projected or hypothetical changes makes the task even more difficult. Areas of operation that seem particularly fruitful will have to be identified. Establishing relationships between, and constraints on, variables and expressing objectives and policies in quantitative terms will require a kind of analysis and a point of view that is new to library administration. The actual formulation of the problem will present complexities of many kinds, but this is to be expected simply because the nature of a library is itself complex.

From this brief description of the SCUL project it is evident that we have far to go before we can determine the utility of this approach, but thus far the promise of the project is such that we hope to continue what we have begun.

Benefits of many kinds will inevitably come from this kind of intensive research into library operations, even if the final results differ from those looked for at the beginning of the project.

It is already obvious that any significant success of this project implies major administrative and operational changes. For example, program objectives of the library will have to be carefully related to every segment of the university program and stated with more precision than has been the case in the past. The mission of the library will have to be reviewed and understood by all concerned parties in the university as well as within the library. Because a university library is in many ways a microcosm of its parent body, this very process might have interesting and useful supplementary effects.

Second, it is evident that a management team of a type new to libraries will have to be developed to employ effectively and utilize fully the results of management techniques of the kind contemplated.

Finally, because success of the SCUL concept is dependent on a continuing flow of data to make the models honestly reflect the real-life situation, it is evident that an integrated and automatic system to generate information as a by-product of every important library operation will have to be devised. Planning for an output of useful information should be an important part of every system component designed to carry out library operations.

Thus far we have described by example some of the Columbia projects that have already, or will soon, involve us in the use of data processing equipment. In the course of the next two or three years, several of these activities now in the formative stage will be fully operational. But as we have moved along in recent months, we have been reminded again and again that we are not coming to grips with some of the basic problems that must be solved if the promise of data processing machines applied to library operations on a nation-wide scale is to measure up to the visions we have been induced to accept.

First, it seems unlikely that most members of a staff of a large research library—a staff already responsible for carrying on a substantial load of day-to-day operations—can put their regular work aside for the time required to become conversant with machine techniques, and then devise, install, and operate a new system. A mountain of undone work would quickly grow and bury them. For example, we have done some detailed work in flow charting serial processing, but so far no one on the serials acquisitions or cataloging staff has found a way to create the new world while coping with the old—the simple process of handling the half-million items that come their way each year dominates time and energy. How do we surmount this dilemma? Do we have to create a parallel system, including a duplicate staff, to move from the world of the 3" x 5" card and the visible record to that of magnetic tape and printed holdings lists? Or should we break up our central serials acquisitions system into smaller subject-oriented units and revamp them one by one? Is it possible that a very large library must break up into a federation of libraries before changes of the magnitude we envision can be accomplished? Is there a limit to how big or old a dog can be if he is to learn new tricks?

A second, and related, question concerns the amount of what might be called "risk capital" that an academic library should spend to assure a progressive program of operational evolution. The SCUL project, just described, much of which was done by a private firm, has cost about $15,000 already, not counting substantial amounts of library staff time or computer time—and this is for pure research devised simply to test a methodology, with no guarantee of a pay-off. My question—are academic institutions too conservative generally in investing capital on a planned basis to improve operations? Higher education, judged on the basis of dollar expenditures, is big business and is growing bigger. Perhaps both libraries and the institutions of which they are a part need to provide in their regular budgets for more research into their way of operating.

Next, is it reasonable or even rational for every library to go off on its own to establish a type font and design a format for what should be generally useful and useable bibliographic information? Johns Hopkins is now hunting a way to convert its shelf list to tape.

The Library of Congress might one day go to work producing machine-able records for current publications and might ultimately find itself involved in converting The National Union Catalog (NUC) and the Union List. The New York Public Library has major catalog problems and might have to get into converting some of its records into machineable form. The Yale-Harvard-Columbia medical group is working on format and establishing requirements for type fonts. One group in the government is out to establish a nation-wide information network for scientific material published in journals. This list could go on and on. Many other organizations and individuals are involved. It will be a little short of tragic if some sort of national machinery is not soon created for coordinated development of a generally acceptable format for basic bibliographic information and a standard font of characters for printing with data processing equipment. I shall go all the way and suggest that perhaps the President of the United States might properly create a permanent Commission on Access to Recorded Knowledge to tie together the multitude of activities in this area, in and out of government. The acronym ARK is itself symbolic of the flood of uncoordinated (and therefore both competitive and redundant) solutions to bibliographic control. When this fundamental problem is solved, I can suggest others just as basic to keep this commission active for some time to come. In the long run a high level official organization responsible for optimizing access to recorded information might prove as important to us all as the Atomic Energy Commission or the Fish and Game Commission. After all, when we are dealing with recorded knowledge, we are dealing with the cumulated product of the brain power of the human race. I cannot think of anything that deserves more care.

In short, individual institutions can and should move to try new methods of operations and analysis—everyone can learn from such efforts. They can, and should, seek new ways to handle administrative and operating services such as circulation control, collection maintenance and inventory records, and business and fiscal aspects of acquisitions. But in the field of bibliographic control, to say nothing of text storage, it seems both impossible and unrealistic for any large general research library to step out on its own. The research resources of this country need to be linked by more than transitory ad hoc committees or a complex of professional associations.

A final problem, one that must be solved if coordinated cataloging is ever to be achieved on a massive scale, involves the matching of a book in hand to a remote bibliographic record. How can I pick up a book in Hungarian (one of the several languages I do not read), on a subject I know nothing about, and locate the descriptive and analytical bibliographic information for that book? At present, one has practically to catalog the book before he can begin to search for this information. We need an internationally understood and

automatically derived means to describe a printed item. For example, if imprint information were noted as day, month, year, instead of year alone, it would be easy to devise a number made up of the imprint, the number of the last numbered page, and a code for the language of the title page. Thus anyone, anywhere in the world, could pick up a book, and describe it by the same number that would be established by any other person working with the same book. This number could be attached to a bibliographic record, wherever and whenever this record was created—and we would be on the way to tying the item in hand to a remote, machine-stored record. A little work on the probability of generating duplicate numbers for different items would need to go into the composition of the code. Some duplication would seem acceptable, because this would mean that one would simply select the right information for the book in hand from two or three records produced out of the system.

The problems I have isolated here are large ones—they are not concerned with machine configuration or programming shortcuts, or even such important questions as how are enough people to be trained to meet the demand for the special skills required. But the solutions to these larger questions and others of similar magnitude are required, or must at least be on the way, before large, general, research libraries can join fully and without reservations in this revolution in the methodology of librarianship.