Byproducts of Computer Processing

By BARBARA FISHER

University of Oregon

"All I could see from where I stood,
Were three long mountains and Mrs. Wood."

WRITING in the *American Archivist* in April 1967, Elizabeth Wood posed the dialectical question: "What are the by-products of machine-assisted information handling?"[1] In the mountained majesty of Oregon, I have tried to answer her. I have brought my own dialectics here to Ottawa and invite you to share them with me.

Before these sessions are over, I shall probably hear at least one of the following questions: Why should I use the computer? What can I use the computer for? I shall probably not hear the question: How does the computer work? There are both good and real reasons to avoid the question of how a computer functions. As archivists our primary concern is the preservation of records and their orderly description. How can machines help us in our continuing struggle against backlog, mass, diversity of record, and diversity of record conditions? I sometimes think we almost enjoy our private professional dilemmas. Like the calluses on a farmer's hand, our burdens of backlog, mass, and diversity are symbols of our professional brotherhood. If we are really professionals, however, we are seldom stymied by burdens. As archivists we are determined to produce a distinguished archives. We are determined to keep backlog from stagnating, mass from swallowing us in our archival little acre, and diversity from burying the pure wealth of our record groups in a tundra of trivia. We are concerned with records preservation. How much intellectual control we can gain through machines or our own in-depth content analysis is totally dependent on maintaining physical control.

There is another real reason why we do not often ask how the computer works. Someone is likely to tell us! He will discuss hardware in exotic terms when we have always thought of it as faucet washers and toggle bolts. He will tell us about software, system flows, interfaces, confrontations, sense switches, loads, bits, and bytes. With every word he will isolate us from the comfort of our own archival jargon like retention, disposal, chron files, record series, record groups, provenance, and *respect des fonds*.

The author, Archivist of the University of Oregon, read this paper on Oct. 1, 1968, at the 32d annual meeting of the Society of American Archivists in Ottawa, Canada.

[1] Elizabeth B. Wood, "From the Information Soapbox: Information Handling Dialectically Considered," in *American Archivist*, 30:319–320 (Apr. 1967).

Let me approach the computer despite protest, however, and enter a world where projects are finalized rather than finished, where they may be feasible but not practical, and where—between the Scylla and Charybdis of input and output—the mystical, mnemonic, hexadecimal magic of machine processing is wrought.

All I could see from where I stood after I read Elizabeth Wood's article was the bewilderment of the archivist struggling to relate to computer technology. The first problem is the term itself: computer technology. The archivist is not a technologist. His preoccupation is with content and value. Technique is only a means to an end. Other computer terms in current use—mechanization, for example—also tend to reflect what an archivist is not. The archivist is not a mechanic. He will deny it even as he unpacks 50 nailed, wired, and bolted shipping crates. Then there is the term—cybernetics. How can anyone relate to that?

Because I want to discuss some of the characteristics of the computer and its components without alienating you by use of engineering jargon, I have created a new term—computernetics, or the systematic study of the computer's system. Let Hayes, Becker, Shera, and Wiener be warned.

Mrs. Wood will approve of my study, I hope. Her approval is implied in her postulates (1) that the computer "forces us to be systematic" and (2) that it assists us to achieve continuity, the final results of which are better, more comprehensive histories. Mrs. Wood postulated well. I should like to explore her meanings with her. That being denied for the moment, I should like to explore her meanings for her. System, continuity, and comprehensiveness are, in addition to what Mrs. Wood has said of them, three terms descriptive of computer operations.

If we substitute the term logic for system, flow for continuity, and mass storage for comprehensiveness, we will open our dialog with computernetics. The computer and its components, the hardware configuration required for information processing, are logical machines. Working in an integrated, total, or partial flow system, they perform multiple tasks sequentially. If we agree at the outset that the archivist is also a logical performer of precise tasks and that he is, as a species, endowed with an extraordinary memory, we can then conclude that he really does understand what the computer technician means when he speaks of system, flow, and storage.

I have not said that the archivist is no more than a task performer with a good memory. He is far more than that because he does not perform his tasks sequentially. The archivist learns, retains, and applies his learning in a system that can best be called intuitive. The computer must be instructed anew every time it is turned on. The archivist does not need a job control statement or assembly language every time he starts a day's work. He does not need to be programed. Lord help us if he does. The sequencing and programing factors in computer process-

ing may be the most distinguishing difference between human and machine task performance. Any sequential tasks—filing, sorting, matching, listing—can be performed by means of computer programing. The archivist ought to be doing the imaginative, judgmental, and intuitive work that the computer cannot be programed to do.

Another aspect of computernetics is reflected in a comment by Thomas Condon of the American Council of Learned Societies. He writes:

> Far more significant [than the impact of the computer on information problems] has been the impact of the new technology on men's minds and dreams. New possibilities are forcing scholars to think in new ways about the information they use and need. In setting men to dream about the information world in which they would like to live, the computer has caused present practices to be submitted to critical examination and searching analysis.[2]

Condon is saying that as archivists we can, with the help of the computer, do whatever we want to do with our records. How does that sound? What would you like to be able to do? I grasp at Condon's statement just as you do because I want to resolve problems of backlog, mass, and diversity. I would emphasize, more than he does, however, the fact that the basic characteristic of computer processing is logical analysis. The computer does not set men to dream. Instead it urges a sometimes rude awakening, for it exposes to us our old habits, practices, and thoughts in an entirely new perspective. It is the computer's unlimited and uncompromising analytical system that has literally forced us to reexamine old ways and dared us to dream of new. Cumulative experience in machine manipulation of information has pressured men to undertake a searching analysis of all kinds of information and of the activities that generate information.

Computer processing relentlessly breaks information down into its constituent parts. If need be it will break paragraphs from full texts, sentences from paragraphs, words from sentences, consonants from words, syllables from words, letters from words. If information is expressed as an activity or function, that activity will of necessity be analyzed step by step. When information is thus reduced, its logic, relevance, and objective values are exposed to our critical view. The computer and its components must convert data into manageable bits. Much information when so dissected and so converted raises questions about the validity of the information itself.

A pearl necklace sometimes blinds us to the imperfections of the individual pearls. It is an assembly of many parts but we are conscious only of its total effect. If one deliberately unties the string and removes each pearl, he can evaluate each, alter or confirm its relationship to larger or smaller, finer or poorer pearls, and can then reassemble them either on the old string or on a new one, and in any order he wants

[2] Thomas Condon, "Abstracting Scholarly Literature: a View From the Sixties," in *ACLS Newsletter,* vol. 28, no. 8:3 (Dec. 1967).

them. This is what the computer does to information. It liberates all constituent parts of information from rigidity of assembly, exposes each part in context and then reassembles the parts in whatever sequence or logic we wish to impose. Computer processing thus produces something of which all of us as archivists have dreamed, freedom of unlimited choice in the kinds of information we want about our archives and unlimited choice of how we want to describe our holdings and their contents. We can be freed to exploit that which traditional procedures and clerical habits stultify, our creative professional initiative. We may call this a byproduct of computer processing, but it is in reality a means of achieving one of the ultimate goals of our profession.

There are very real byproducts produced by machine-assisted information handling, less profound in their implications but of surprising usefulness. At the completion of each step in the processing flow system, reports and summaries may be requested from the machine operator. Since the increasing development of closed-shop computer operations, these reports are sometimes bypassed. They can, however, be built into the program package. Designing step-by-step reports of computer processing procedures is important because these reports show us information behaving in ways we cannot otherwise perceive.

When we manually study our operations or accumulated data, we tend to bypass or skip over details; we slur over things that we know by assumption or presumption. Processing reports are like the slow-motion camera, time-lapse photography, or stop-action TV. They keep critical tabs on what we are doing. In addition, we can learn to build, through the processing system, supplemental programs to control what we are doing and to submit our data to continuous editorial scrutiny. When I first started processing by computer I was conscious only of the wanted end product, the print-out. When I learned to study processing reports and summaries, I found that they contained valuable documentation. In actuality, they constitute the analytics of processing. Once understood, they can teach us something of the intrinsic nature of information and of its alphabetic and numeric symbols. In my byproduct library, for example, I now have readable listings of input, edit listenings to show what I am doing to my records, update listings of new information matched against obsolete information, statistics of word frequencies that show the adequacy or inadequacy of descriptions, and statistical accumulations relating to operational costs.

An ongoing computer program that can usefully demonstrate how the computer processes is IBM–KWIC, a key-word-in-context indexing program. A rather typical entry from a manuscript collection finding aid might be: "Correspondence relating to the purchase of Alaska, the Bering Sea fishing controversy, Alaskan statehood, and the University of Alaska." The entry can be converted to KWIC format by simply punching it, just as it appears, without inversion or abbreviation, onto punchcards. This entry is only one of many we would convert from the

same finding aid. When the total card deck is complete, step-by-step processing begins. First, the cards are read by the card reader, one of the components of our hardware configuration. Output, after this first step, is a simple listing of the data on cards. It is a read-print program occasionally called a utility listing or a packaged program tagged IEBEGNER. The list serves not only as an edit print-out for the detection of punching errors but produces a text that can be edited for words not wanted in the final index—verbs, adjectives, and prepositions. Unless otherwise instructed the computer will index every term in the text. It does not use judgment. It computes.

After editing of the input text is completed, each unwanted word is keypunched into a new deck called the stopword deck. Both the textual deck and the stopword deck are then assembled for processing step two. The KWIC program is stored in the computer's memory along with the stopword deck. The text is read under a rotational program, word by word. The pearls, as it were, are removed from the string. Each word is shifted to keyword position. Correspondence is stored as a keyword, purchase is stored as a keyword, Alaska is stored as a keyword, Bering is stored as a keyword, Sea is stored as a keyword, and so on. Because KWIC is a keyword-in-context program, operating on a rotational shift basis, each of the keywords takes along with it as it shifts to key position, approximately 60 characters of the sentence in which it appears. Byproducts at this stage include statistics of word frequencies and a recount of words stopped.

The third step in processing KWIC is a sort and merge program used to alphabetize the keywords, and a print program to produce a finished index. KWIC now exists in many variant forms and may require as many as seven sequenced processing stages. The program may be terminated after any sequence is finished. Some users stop after a words-used and a words-stopped statistical analysis is printed out. KWIC can thus be adapted to linguistic or syntactical analysis. Reports, summaries, and listings can be requested after every sequence in the processing operation, depending on the options built into the program.

There are really two kinds of byproducts resulting from this kind of processing: the machine-produced and the intellectual. The machine tells us, for example, that in our sample entry we have used the term Alaska twice, the term Alaskan once, and the term Bering once. It does not tell us that these are generically or historically related terms. It simply lists both uses of the term Alaska in sequence, the term Alaskan as a discrete entry following, and the term Bering as a discrete *B* entry amputated from its coword, Sea, which of course will appear in the alphabetical listing under the letter *S*. All these discrepancies and anomalies can be corrected, not by machine but by our own intelligence. Computer-stored thesaurus control programs, which automatically wed generic terms under logical subject headings, have already been designed. The flexible program capability of the computer encourages us to do

whatever we want to do. In the meantime, byproducts resulting from computer processing can help us to decide what that is.

The computer analysis of our sample input shows us that frequency of term use implies that the correspondence is dominated by Alaskan history, that the description itself is rather broad, and that there is a distinct difference between one kind of term and another: that is, correspondence is a form term, Alaska is a subject term, and purchase is a verbal term. In archival and manuscript research the three terms do not interfile very logically nor are they adequate for a graduate or professional searcher. For that reason it might be well to code them to index in different ways. SPINDEX I and developing SPINDEX II computer indexing systems, both variants of KWIC, allow for several levels and kinds of indexing—by names, dates, series number, subject descriptors—and for retrieval in a variety of formats.

Total running time for preliminary and final processing of basic KWIC, exclusive of operating and editing intervals, has been estimated at 8 minutes for a 7,000-line text of approximately 120,000 words. As filers and indexers we cannot compete with that kind of manipulative speed. Speed is not the main goal of the archivist, but if it is speed attended by analysis, he ought to explore possible application to his indexing and management problems.

Computer processing byproducts pose a question that is of increasing importance to all of us in the archival and library professions. How does the researcher want our materials? Are we producing too much information? Are we producing the wrong kind of information? Does the researcher want a detailed index, a catalog description, a series description? We ought to design systems for maximum input so that all these elements can be stored in a central file and retrieved in a form suitable for the individual inquirer. Computer processing accepts unlimited options and variations as long as they are formatted, coded, and translated into precise program language. Do we know yet what options the researcher wants us to have?

Computer analysis has also been applied to records management and archival procedures at the University of Oregon. The analysis has resulted in byproducts that are as useful as those resulting from KWIC indexing. All the information elements relating to an integrated records management–archival accessioning system are now stored on tape. Data relating to records accumulation, records reporting, and shifting of records, as well as the consequent flow of information, have been built into the tape-stored program. By storing all of the elements of the system, manipulating output to test the validity of the information, and developing sequential job steps on a stop-action basis, the program has become a working model toward an ideal system. Byproducts in this instance are objective test results in a laboratory-controlled environment. They demonstrate painstaking steps toward realization of an ideal university information system.

Machine byproducts of the university's developing program, tagged REMARC, include print-outs of a master record, edit listings of input and update information, reports on the status of inventorying record holdings for the State Archivist, automatic signal design for flagging the transfer of noncurrent records to the archives, and listings of massive accumulations of subject information and record series descriptions. These listings incidentally demonstrate that, in addition to the traditional problems of archival backlog, mass, and diversity, universities tend to produce many duplicate records. I wonder whether this duplication would continue if a computer-stored administrative data bank with on-line and direct access terminal capability were developed. On the basis of program tests and model study, I have begun to speculate, not only about university archives and records management, but about many basic principles and practices of archival administration and records generation. Data processing has inspired me to dream of a set of workable archival principles particularly shaped to American archival experience and American corporate recordmaking practices. In the form of questions raised by Mrs. Wood back in April 1967 and with bold acceptance of Condon's invitation to create the information world in which I want to live, let me indulge in my own dialectics.

1. If archival backlog includes containers of foldered and labeled official files, why must they remain backlog? Under a computer indexing system, the information on the folder labels in current files still retained by offices and in noncurrent files in the archives can be transcribed in machine-readable form with location symbols. Sorting first on the information field, an alphabetical subject and name index can be prepared. A second sorting on the container field will produce a shelflist, a container list, or a location reference. Two basic requirements of records management and archival administration have been met: we know what is in the records, and we can find it. Additionally, rapid index control provides the archivist with overview of all records and with an index that matches up similar files even if they are in widely separated containers or widely separated offices. It also enables him to improve the quality of accession records and receipts, helps him to relate new additions to records he already has, and gives him a clue to possible gaps in the record. In my information world, I don't want to respond to inquiry by saying, "I'm sorry, it's unprocessed," or what is worse, "I'm sorry, I don't know."

2. Is diversity of form and condition of material a real problem? If sufficient descriptive data can be stored in a computerized index bank, patterns of types of records can be perceived. As a result diversity will assume a different perspective, and while it may prove to be a problem from container to container, it may be a very minor consideration in the context of the entire archives. Accumulation of data about forms and types of material tends to separate disposable records from permanent records. On the basis of these patterns, recommendations can be made in advance for discarding duplicated or irrelevant materials before the records come to the archives. Although I have considerable misgivings about secretarially oriented disposal programs, in my information world I do not want to waste my time on duplicates and trivia.

3. How can I establish overview? Overview of large accessions can be established by rapid punching of folder label information, box by box. Such overview also pro-

vides control of the condition of material. Unfoldered, unarranged, disorganized records cannot be indexed. When this fact is recorded and these containers are shelf-listed separately, the archivist has a true backlog estimate and has isolated the unknown records from the known. Overview folder indexing also indicates arrangement discrepancies and can be used to guide and instruct processors before actual rearrangement takes place. In my information world I want to know what I do not have, as well as what I do have. And I want to plan a workable processing schedule.

4. Is continuity of history dependent on the physical continuity or the order of records on the shelves? With massive storage of current and noncurrent information derived from records survey indexing and indexing of the archives, information can be centralized even though the records are not. Continuity of the facts of administrative history is thereby established far in advance of the physical possession of the records. For example, the inquirer may discover at the central data bank that documentation of a hundred years of the history of the University of Oregon School of Education exists, even though he may have to visit the Education Office, the Archives in the Library, the President's Office, and the State System of Higher Education Office to consult them. He may also discover that certain records are scheduled for transfer to the Archives in a year and may therefore better plan his research program. Since archives of a living organization are open ended, continuity can only be established by continuing records survey indexing. Records inventorying is concerned only with records at the series or group level. The indexing program, on the other hand, is an information gathering program rather than a retention-disposal management system. Inventorying serves the archivist's purpose; indexing serves the searcher. In my information world I want to serve research as well as the administration.

Why not item index by computer all unprocessed or unarranged items or units that appear to be unrelated to known record groups? One element of information—a name, a date, a description of the form of material—can be matched with similar or related data in the index bank under a retrieval program. A decision may then be made whether to interfile the record or to simply interfile the information and leave the orphaned record in its own container. Improper filing of information can generally be corrected, but a folder filed in the wrong box is sometimes lost forever. In my information world I want to excel.

5. Why not plan now for paperwork management that is free of paperwork? Convert paperwork management systems—whether Federal, State, local, or institutional—to an on-line computer service, where new reports, amendments, and follow-ups on disposals and transfers can be transmitted to a master record or entered through direct access terminals, amending the record and keeping it current. Concurrently, administrators should be encouraged to develop a master computer library of common administrative records and forms, accessible to terminal inquiry by scattered departments and offices, and thus to eliminate proliferation of duplicates, carbons, and circulars. In my information world I do not want to make paperwork; I want to reduce it.

These dialectics do not for a moment suggest that we should stop processing archival records or reduce our efforts toward the preservation of source documents. They do suggest, however, that we should explore the possibilities of managing information and descriptive programs by computer. The principle underlying the dialectics is that it is the infor-

mation in the records that the inquirer wants and that such information is what ought to be processed first. Once control has been established, the archivist can plan better, use his limited manpower with more effectiveness, and still provide access to archival records at the earliest possible moment. If he is a traditionalist, through computer indexing and management systems he can preserve original order without violating his own principles of historical continuity.

Byproducts, as I have used the term, like so many computer terms, is a misnomer. It implies that in doing something important, you do other less significant things along the way. Computernetics, my brief and superficial glimpse at the computer's processing system, implies for us as archivists a parallel, step-by-step study of our current archival processing systems. There is nothing insignificant in that. Since a system is the sum total of its parts, it is the parts themselves that we ought to isolate and subject to scrutiny. Through computer testing and analysis each element of archival management and information can be arrested and its characteristics observed. Most of us wish we had time to study our own programs and our habitual methods. There is little reason why, in due time, the computer cannot do most of this analysis and pose most of the critical questions that relate to our archival and records systems, thereby allowing us to concentrate on providing in-depth bibliographical and reference services to researchers. There is nothing insignificant about the opportunities that computer processing affords to convert our energies to the performance of exclusively professional work. I have therefore accepted Mr. Condon's invitation to dream of the information world I want. I have accepted Mrs. Wood's invitation to scrutinize the information world I am in. I am increasingly confident that, with the computer as an ally, we can move anew toward our traditional goals: to improve the quality of our archives, the quality of our service to administrators and historians, and in due course to improve the cumulative quality of our own dynamic profession.