# Automation and Information Retrieval in Archives—the Broad Concepts

By RITA R. CAMPBELL

*Hoover Institution*

BECAUSE OF the continuing proliferation of "paper" (which does not need to be documented in the *American Archivist*) and its impact on archives, the need for more accurate and quicker access to archival holdings is becoming urgent. The impact is felt not only because of the great increase in the number of collections accumulated by individuals, corporate bodies, government agencies, and other record-gathering groups but also because of the ballooning size of each collection or archive. Before there were dictaphones, typewriters, and the like, when correspondence was limited by the physical fatigue of handwriting, the papers of a President of the United States numbered a few thousand pieces. Today they may number in the millions. This accumulation of vast quantities of paper is one symptom of the copymania of our society.

Archives are subject to the opposing pressures of mounds of paper and continuously increasing research demands. The way out of the dilemma may be found in more intensive indexing of archival materials by machine.

Although it is now almost 15 years since the first marketing of a commercial modern computer, archivists, especially in the social sciences and humanities, have only lately begun to explore the computer's potentialities for lightening the task of search and the preparation of bibliographies. More intensive machine indexing is making more readily available to scholars the known subject material of an archive and, possibly even more important, the unknown and frequently valuable peripheral material buried in hundreds of manuscript boxes.

Arrangement of archives without any indexing—arrangement whether by provenance, by data, by geographical area, by individual donor, or by some combination of these—does not in a large collection meet today's research demands. The need is for detailed

Dr. Campbell is Archivist of the Herbert Hoover Archives and Research Associate of the Hoover Institution on War, Revolution, and Peace, Stanford University. Her paper, prepared especially for this issue of the *American Archivist,* explores her subject more thoroughly than did a paper she read before the Society of American Archivists in Atlanta, Ga., on Oct. 7, 1966.

indexing. Manual detailed indexing, however, is too expensive, especially where large archives are involved.

Archival material, in whatever arrangement it happens to be, or even when not arranged, can be machine indexed. Rearrangement is unnecessary and expensive. Some arrangement rather than no arrangement at all, however, permits browsing and is a most desirable feature worth, it seems to me, its nominal cost.

An inexpensive way to arrange an archive is by provenance, which has been defined by T. H. Schellenberg as an arrangement of archives "according to their origins in an organic body or an organic activity."[1] Another common and relatively inexpensive method of arrangement is by date. From the point of view of the research scholar who is subject oriented, arrangement by date without any index is frustrating. Often correspondence is chronologically arranged with a subarrangement by name, or vice versa. Although it is generally easier to find material by subject in those collections arranged by provenance, there remain difficulties; for the research scholar's choice of a subject may not be easy to find within the lines of organization created by provenance.

Archives are sometimes arranged by subject; this, however, is not only costly but almost self-defeating. A single item can be put in only one place, and a single item may cover several subjects. In large collections there may be subarrangements by subject, but the more complicated the arrangements become the more self-defeating the process. Any businessman with extensive correspondence files knows the frustration created when his staff is unable to find a particular letter, which he remembers by what *he* thought was its most important subject matter and which his secretary, because of a different frame of reference, filed under a different subject—the one *she* felt was most important. Making several copies of an item and filing them in several places is a partial solution for such a problem but adds to the bulk of paper. Indexing, not arrangement, is the key to information retrieval.

Machine indexing gives subject clues far more specific than those provided by the descriptive registers and broad subject indexing of traditional archival handling. Because of the cost of manual indexing, clues are seldom offered to the subject content of dozens of manuscript boxes or file drawers of letters except those in brief descriptive registers. Dependence is placed on arrangement. A descriptive register may indicate that a collection has material on a broad subject or refers to a particular person, but the register

[1] T. R. Schellenberg, *The Management of Archives,* p. 41–42 (New York, 1965).

does not usually indicate in which one of several dozens or even several hundreds of manuscript boxes the references occur. Search time is thus often extensive. The register's use of broad subject terms and its omission of subjects that are not central or important within a collection mean the burying of material covering subjects peripheral to the known subjects of interest in the collection.

Some people object to machine indexing because it depends on the selection of subject tags or keywords to describe items rather than having the scholar either scan each item to find what he wants or else formulate a search request in his own words without resort to a formalized dictionary. It is sometimes said that the listing of hundreds of subject keywords for potential search somehow directs the individual's research and that this is especially true when the machine program has a controlled vocabulary.

A "controlled vocabulary" dictionary, which I believe is a necessity in machine retrieval of archival information, is used as opposed to no dictionary control or to the use of an "uncontrolled vocabulary." With the latter, the machine accepts all words used in a title (the KWIC system) or document except for a brief, predetermined list of invalid terms—usually prepositions, articles, and other similar nonsubstantive words. Since archival materials do not usually have titles to serve as a ready source for the machine to scan for vocabulary, it would be necessary for the machine to scan for vocabulary or keywords the whole or at least a large part of a document. This would be expensive, especially as the physical format of archival material varies, and it also would create a very large and unwieldy dictionary.

Machine retrieval in archives is in a way forced to use a controlled vocabulary, and this does mean that somebody must select the keywords. To object that this also means direction of an individual's research, I find to be a strange criticism, especially when some of the objectors at the same time approve of subject headings in card catalogs and subject indexes to books. Offering a particular list of subjects to a research scholar does not direct his scholarship; it just gives him clues to the subject content of material in archives. A similar objection, as indicated above, can be made against all indexing systems; it is not relevant only to machine indexing. Indexing is not, nor can it be, a substitute for reading an article or manuscript.

Likewise browsing is not an adequate substitute for indexing. Browsing may give research leads to a scholar, but it is often an inefficient method to find specific data. Browsing in a particular collection is limited by the arrangement of that collection. If the

material is arranged chronologically, subject browsing is indeed difficult, and if the material is arranged by subject, browsing to find all the material about a particular person may also be difficult. Browsing is limited not by machine indexing but by the physical arrangement of material. There is nothing in the method of machine indexing that will prevent the scholar from reading all the documents in an archive—if he has the time and the desire to do so.

Detailed machine indexing permits "browsing" in all directions. The term "browsing" in this sense seems peculiarly apt. An extensive, in-depth, keyword list may suggest new relationships to the research scholar. The new on-line computers, linked to individual consoles, permit man-machine communication of a nature that allows rapid browsing for data to answer known information needs and even to answer sometimes unforeseen information needs.

> The difficult problem in information storage and retrieval will always be the intellectual job of matching the information that is known and stored with the information that is wanted at a particular moment. No universal solution exists for this problem because each of us has a unique set of needs for information. Often the very process of trying to answer a question changes our needs. In a search through categories of references and through possibly relevant data, we discover unforeseen aspects that change our concepts of what we seek. Reference data that so modify the course of our search also become information.[2]

The aim of retrieval is to place all the material pertinent to the scholar's research at his disposal; then it is up to him to accept or reject it, to rearrange and develop his own schema and conclusions.

Although the intellectual interaction of man and machine of the sort described above may in practice be largely still in the future, its anticipated contribution cannot be ignored. The technology has been developed, but its use is not common, because of costs.

Comparative cost evaluation of a machine retrieval program is difficult. The impulsive conclusion that machine retrieval is "too costly compared to present manual indexing" is deceptive because it omits any discussion of the differences in the degree of indexing obtained and in the success of retrieval under the two systems. In comparing the costs of a Cadillac and a Ford one makes allowances for the difference in their performance and design. Only by comparing manual and machine costs of identical indexing of a given collection can the true relative costs be ascertained.

[2] Herbert Evans, "Information Transfer," reprinted in *Congressional Record,* Apr. 7, 1966, p. 7656.

Costs of machine indexing can be kept within a reasonable range if no rearrangement of material is required, if relatively untrained personnel can be used, and if folder indexing rather than item indexing is used.

A machine retrieval program should not require subject specialists as indexers but should rather use as indexers high school graduates. On a college campus students and their wives, relatively inexpensive sources of labor, may be used as indexers. Controls can be built into a program so that it is possible to use nonprofessional indexers under professional supervision. Some of the literature on machine retrieval and also the Hoover Institution's very limited experience suggest that it is wasteful, even beyond the difference in salary levels, to use as indexers people with a relatively high level of education. The greater subject knowledge the individual has, the more likely he is to read into material a significance or meaning that may not be there and the more he will be tempted to read material in order to educate himself rather than to index quickly.

The third factor important in keeping expenses in line with the returns gained is a permissive elasticity in handling the indexing. One way to retain elasticity is to index to a folder containing several items, not to each single item within it. A "folder" varies in size. I think I can best explain "permissive elasticity" by a sample illustration taken from experience with the papers of the American Relief Administration. This archive in the Hoover Institution has several hundred, perhaps a thousand, pieces of paper that are telegrams requesting the U.S. Army Signal Corps to repair breaks in communication lines and a similar number of pieces of paper that report repair of the breaks. Our elasticity in defining the size of a folder enables us to index all these pieces of paper, which fill an entire manuscript box, in a unit with a single set of descriptors as U.S. Signal Corps, unsorted, 100's (of items), communications, ARA, 1919. Although a "folder" may contain several hundred items or one, in general it averages 10 to 30 pieces that are related in some way depending upon how the material has been arranged.

It is also anticipated that search costs in some archives may be kept at a minimum. In a social science archives most users are not subject to the pressures of time or of deadlines that apply to users of materials in the pure and applied sciences. Many scholars write ahead to inquire about materials, sometimes several weeks or months before they appear in person to use the records. If it is feasible for archives in the social sciences to accumulate retrieval requests for, say, 1 or 2 weeks and to feed them as a block into

the machine, retrieval costs would be lower than if 24-hour service were given.

My thoughts about machine information retrieval have evolved while trying to develop a general technique of indexing by computer the archives of the Hoover Institution on War, Revolution, and Peace at Stanford University. Although this project is still experimental, we believe that it will prove successful in demonstrating machine techniques for searching archival materials and to prepare subject bibliographies developed from these materials.

There follows a brief description of the Hoover Institution's system.[3]

The indexer works as follows. If the collection has some arrangement, which is the usual case, he takes an existing folder of papers and then writes, in accordance with an authority list of keywords and rules, a description of the material. If the papers have no apparent arrangement, and sometimes this occurs, the indexer will group papers into whatever loose groupings they easily fall, will place them in a folder, and again in accordance with an authority list will write a description of the material.

The indexer then assigns a unique identity number to the folder and its contents. Individual items in the folder have a subscript number. For example, if the identity or folder number were 100, individual items would be numbered 100–1, 100–2, 100–3, and so on. Of course, unsorted items, such as the Signal Corps telegrams already mentioned, or accounting items such as bills, are not numbered. The identity number begins with a mnemonic letter; for instance, capital "A" stands for the American Relief Administration. This may be followed by a mnemonic Arabic number, which stands for the country with which the papers are concerned. For example, 4 stands for Czechoslovakia. After the mnemonic number, several blank spaces follow, and then follows a number equivalent to an acquisition number, for example, 00826, 00827, or 00828.

After assigning the identity or call number, the indexer selects from the authority list the descriptors or keywords describing the material.

At present, the authority list is subdivided into five sections:

1. The forms of documents—letter, diary, etc. This list has about 70 words.

[3] A description of the Hoover Institution's pilot project on machine information retrieval appeared in *American Archivist,* 29:298–302, Apr. 1966. Since that date the program has been revised, which accounts for the differences between the two articles.

2. The substantive descriptors. Currently there are some 650 words in this category.

3. Geographic place names. This list has 165 names, with 160 *see* references, a high ratio that indicates our problems here.

4. Corporate names. The count here is over 600.

5. Names of persons. The archives has no count for the latter, although we have noted the names of such important people as Georges Clemenceau, John Foster Dulles, David Lloyd George, Christian Herter, Robert A. Taft, and Woodrow Wilson. We estimate that there will be 2,000 or more.

As the material warrants, keywords are added to the authorized list in accordance with various rules. A major rule is that a new keyword must not be a synonym for a word already on the list. If a near synonym, an analogous word is used—and such an addition is discouraged—it and its cousin are carefully defined.

The Hoover Institution program is adapted from an IBM 1401 library program. The 1401, an IBM machine that has been in use for several years, is relatively less expensive than the 7090 or, of course, than the newer on-line models. We believe that we shall not be caught by technological advance—in this case, eventual abandonment of the 1401—because the IBM library program, which we are modifying, is being adapted for IBM's latest model machine.

Our program is primarily based on the indexer's invention of artificial titles, made up of keywords as explained above. The machine program has a built-in dictionary control or authority list of keywords, to which new words may be added. The machine will convert an unauthorized word to its accepted synonym if the archivist has foreseen all the possible synonyms and has fed them into the machine. Likewise it will correct frequently misspelled words. The machine will print out and flag any unauthorized term that an indexer may have used. It will also print out the frequency of use of the keywords in the various descriptions of archival materials. The keyword must be designated by the archivist as either a "common" or a "precise" descriptor. A common descriptor is one that appears so frequently in the particular collection being indexed that it may not be used in a search without at least one precise descriptor. A precise decriptor initially acts to narrow down the searchable items, and then common subdescriptors are used to further narrow down the field. Thus irrelevant items are eliminated as early as possible in the search, making the search technique very efficient.

In the ARA archives the keyword "Herbert Hoover," for example, would never be used as a single search item since we would then receive thousands of print-out items. Even if one wanted a

complete list of items in the archives about or by Herbert Hoover, it would be more useful to have it subdivided by subjects. The program will also yield a print-out of the dictionary; this is, of course, larger than the control or authority list since it contains unauthorized synonyms, "see also" suggestions in the form of sub-descriptors, and "scope notes." The last are really definitions of terms that the archivist has made to help user and indexer.

The anticipated, most usual type of search is generally termed a "Boolean" search. This is a term derived from the English logician, George Boole; it can be very simply described as an "and"-"or"-"and not" search. For example, a request can be made for all items containing certain specific keywords such as "France *and* coal *and* transportation" or for all titles containing specific keywords but not containing another keyword, for example, "France *and* coal *and* transportation *and not* Czechoslovakia."

In order to get depth indexing, the indexer is instructed to use one broad and at least one specific descriptor; for example, for the broad subject area of food, he might add specifically dairy products and milk.

Archivists now need to become knowledgeable about computer technology, the opportunities it may create, the true costs of using computers, and what future gains (as from interlinked information centers) may be anticipated. Machine retrieval in archives permits the researcher to turn over to the machine the monumental tasks of search and of memorization of quantities of information while he devotes himself to the far more creative task of searching for relationships among facts and data.

---

***Communication***

Sir Just as my messenger was about to start for the Capitol with several communications enclosing my reply to the resolution of the Senate adopted on the 12th inst I recd their resolution of this day asking why the information then asked for has not been communicated. In reply I have the honor to state that the only reason I have to give why it was not before communicated is, that it was not ready.

—AMOS KENDALL to R. M. Johnson, Vice President and President of the Senate, Feb. 27, 1839, in Letters Sent by the Postmaster General, Record Group 28, Records of the Post Office Department, National Archives.