INTERNATIONAL AFRICAN INSTITUTE / INSTITUT INTERNATIONAL AFRICAIN

10/11, Fetter Lane, London, E.C.4.

University College Nairobi

December - 1967 - Décembre

INTERNATIONAL CONFERENCE ON AFRICAN BIBLIOGRAPHY

CONFERENCE INTERNATIONALE SUR LA BIBLIOGRAPHIE AFRICAINE

Title: COMPUTER AUTOMATION AND AFRICAN ARCHIVES.

Sujet: TRAITEMENT AUTOMATIQUE EN ORDINATEUR ET ARCHIVES AFRICAINES.

Author/Auteur: P. Duignan.

COMPUTER AUTOMATION AND AFRICAN ARCHIVES

by Peter Duignan

The application of computer programming to African archives has now become a practical possibility. Just as in some cases African states have had to leapfrog over certain traditional stages of development, going directly to the air age before they had developed their road or railroad network, so in archival science African archives may have a unique opportunity to break loose from the dead hand of customary archival procedures. The application of computers to records management, archives organization and production of finding aids is just beginning. The future beckons the bold and promises cheaper, more efficient services than are at present provided by older archival institutions.

The Council on Library Resources, Inc. (CLR), of the United States has made a grant to the National Archives and Records Service for a two-year project to develop and apply a computer program for indexing finding aids to archival and manuscript materials. The CLR grant to the National Archives and Records Service will allow the National Archives to make a major contribution to the solution of the control of basic research materials. The project aims to develop a computer-indexing program which could be used for all archival and manuscript collections in the United States and yet allow enough variation to cover the unique problems of each.

The director of this research program is Frank G. Burke, Special Assistant for Information Retrieval in the National Archives and Records Service. Mr. Burke was instrumental in developing an automated system for the management and control of collections in the Library of Congress Manuscript Division, where he was head of the Preparation Section until joining the National Archives early this year.

It is hoped that once this program has been developed for the United States it can be adapted for world use. Individual countries could apply it to their archival finding aids and collections, and tapes could be exchanged between countries. The logical end product would be a centralized world data bank with indexes to finding aids for archival and manuscript materials as well as indexes to the collections. London, Nairobi or Bombay could query the national computers as well as the centre for world archival and manuscript data and receive printouts of the information they desired.

Although archives and libraries have lagged behind in using computers to help solve their problems - the physical and medical sciences in the United States have been more far-sighted and have developed numerous indexing, abstracting, and selective dissemination of information systems - a few institutions have tried to innovate in this field.

In 1967, the Library of Congress produced a computerized "Master Record of Manuscript Collections," from which more specialized lists may be derived. An effort is being made to transfer subject information from the Manuscript Division's finding aids to a master automated index. At present the Master Record tells you only what collections the Library of Congress has and in what quantity. It does not tell you what is in the collections or in what box, file, or folder you can find the information you want.

In the attempt of the new program undertaken by the National Archives to index finding aids to archival and manuscript materials, the promise is held out of a computer-indexing program for all archival and manuscript collections. Such a system has actually been developed by the Hoover Institution. The Hoover system could, in my opinion, be adopted by African archives in the near future and would place them years ahead of American and European archives which are imprisoned by old and expensive methods of preparing registers, calendars, etc.

The paper explosion of the twentieth century poses twin problems for all archives: how to organize these mountains of papers and how to meet the increasing demands of researchers for use of this material. The way out of this dilemma may be by intensive machine indexing of archival materials. For many African archives in the first phase of archival management, i.e., collecting and organizing their material, machine indexing may offer unique opportunities.*

Arrangement of archives without any indexing, whether by provenance, by date, 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 indexing; however, manual detailed indexing is too expensive, especially where large archives are involved.

Archival material, in whatever arrangement it happens to be, or even non-arrangement, can be machine indexed. Rearrangement from an existing arrangement is unnecessary and expensive. Some arrangement rather than no arrangement at all permits browsing, a most desirable feature worth its nominal cost.

An inexpensive way to arrange an archive is by provenance - an arrangement of archives "according to their origins in an organic body or an organic activity." 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. Although it is generally easier to find material by subject in those collections arranged by provenance, there remain difficulties, since 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. But in addition to its cost, the process is almost self-defeating since 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 is the process. Indexing, not arrangement, is the key to information retrieval.

Machine indexing gives subject clues far beyond the descriptive registers and broad subject indexing of traditional archival handling. Because of the cost of manual indexing, no clues are generally 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. However, the register 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 omission of subjects which are not central or important within a collection, means burying of material covering subjects peripheral to the known subjects of interest in the collection.

A "controlled vocabulary" dictionary is used as opposed to no dictionary control or use of an "uncontrolled vocabulary." With the latter, the machine accepts all words used in a title (key word in context, KWIC system) or document except for a brief, predetermined list of invalid terms, usually prepositions, articles, and other similar non-substantive words. Since archival materials do not usually have titles which can act 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 following pages are a summary of Dr. Rita Campbell's longer article, "Automation and Information Retrieval in Archives: The Broad Concepts," in The American Archivist of April 1967.

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 also would create a very large and unwieldy dictionary.

Machine, detailed indexing permits "browsing" in all directions. 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 which allows rapid browsing for data to answer known information needs and, even more difficult, to answer sometimes unforeseen information needs. The aim of retrieval is to place all material pertinent to the scholar's research at his disposal. The technology has been developed; however, its use because of costs is not common.

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 a deceptive conclusion because it omits any discussion of the differences in the degree of indexing obtained and success of retrieval under the two systems. In comparing the costs of a Rolls Royce and a Morris Minor one makes allowances for the difference in their performances and design. Only by comparing manual and machine costs of identical indexing of a given collection can the true relative costs be ascertained.

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 as indexers subject specialists, but rather use as indexers high school graduates. On a college campus, students and wives of students, relatively inexpensive sources of labour, may be used.

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 limited experience suggest that it is wasteful even beyond the difference in salary levels to use, for indexing, individuals who have a relatively high level of education.

The greater the subject knowledge of the individual, the more he is likely to read into material a significance or meaning which may not be there and the more likely 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 a "folder" containing several items, not each single item.

In a social science archive most users are not subject to the same degree of pressures from time, or deadlines, as users of material in the pure and applied sciences. If it is feasible for archives in the social sciences to accumulate retrieval requests for, say, one or two weeks and feed them as a block into the machine, retrieval costs would be lower than if twenty-four-hour service were given.

The archives of the Hoover Institution have been developing a general technique to index by computer. This project is still experimental, but we feel that it will prove successful in demonstrating machine techniques to search archival materials and to prepare subject bibliographies developed from these materials. A brief description of the Hoover Institution's system follows.

If the collection has some arrangement, which is the usual case, the indexer 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 do not appear to be arranged in any fashion, the indexer will group papers into whatever loose groupings they fall, place them in a folder, and again in accordance with an authority list write a descriptio 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, etc. Of course, unsorted items such as bills, etc., are not numbered. The identity number begins with a mnemonic letter, e.g., 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 are left, and then follows a number equivalent to an acquisition number, e.g., 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;

2) the substantive descriptors - currently there are some 650 words in this category;

3) geographic place names - 165 are listed, and we have 160 'see' references;

4) corporate names - the count here is over 600;

5) the names of person.

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 is used - an addition of these is discouraged - it and its cousin are carefully defined.

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

Our program is primarily based on invention of artificial titles by the indexer, 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 fed them into the machine. Likewise, frequently mis-spelled words will be corrected. The machine will print out and flag all unauthorized terms which 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 keywords must be designated by the archivist as a "common" or "precise" descriptor. A common descriptor is one which appears frequently in the particular collection being indexed and as such may not be used in a search without at least one precise descriptor. A precise descriptor 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.

The program will yield a printout of the dictionary which is of course larger than the control or authority list since it contains unauthorized synonyms and "see also" suggestions in the form of subdescriptors and also "scopenotes." The latter are really the archivist's definitions of terms to help a user and the indexer.

The anticipated, most usual type of search is generally termed "Boolean" search, which is a term derived from the English logician, George Boole, and 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, if requested, all titles containing specific keywords but not containing another keyword, e.g. "France and coal and transportation and not Czechoslovakia."

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

There is a need for archivists 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 centres, may be anticipated.

Machine retrieval in archives permits the researcher to place on the machine the monumental tasks of search and of memorization of quanitites of information while he devotes himself to the far more creative task of searching for relationships among facts and data. It also provides him with more information more quickly than the old system of calendars and registers.

Machine retrieval in African archives would permit the archivist to organize and to place on the machine more quickly the masses of data he has collected, to use less skilled help, and to provide researchers with more information more rapidly and more cheaply.

RESUME

TRAITEMENT AUTOMATIQUE EN ORDINATEUR ET ARCHIVES AFRICAINES, par P. Duignan

L'utilisation d'ordinateurs pour le traitement des archives africaines est maintenant entrée dans le domaine des choses réalisables. Le 'Council on Library Resources, Inc.' des Etats-Unis (C.L.R.) a accordé au 'National Archives and Records Service' une subvention pour deux ans en vue de créer et de développer un programme de traitement en ordinateur pour l'indexage des guides aux matériaux d'archives et manuscrits. Cette subvention du C.L.R. permettra aux Archives Nationales d'apporter une important contribution à la solution du problème que pose le contrôle des matériaux de recherche de base. Ce projet vise également à développer un programme d'indexation par ordinateur qui soit utilisable par toutes les collections d'archives et de manuscrits des Etats-Unis, et soit néanmoins assez souple pour répondre aux spécifications de chacune d'entre elles.

Une fois développé aux Etats-Unis, ce programme pourrait être étendu au monde entier. Chaque pays pourrait l'appliquer à ses guides et à ses collections d'archives, et des bandes magnétiques pourraient faire l'objet d'échanges entre pays. Le résultat final de l'opération serait alors une sorte de centre d'information mondial comportant des catalogues des guides aux matériaux d'archives et manuscrits, en plus des catalogues des collections.

La Hoover Institution a créé un programme de traitement automatique par ordinateur. Ce système pourrait être appliqué par les archives africaines dans un proche avenir. Le projet en est encore au stade expérimental mais il semble qu'il doive réussir à prouver l'aptitude des techniques automatiques au traitement des matériaux d'archives et à la préparation des bibliographies spécifiques qui en sont tirées. Suit une briève description du système de la Hoover Institution.

Si (comme c'est généralement le cas) la collection est déjà classifiée, on choisit un dossier existant de documents, et on en élabore la description en appliquant un code reconnu de mots-clés et de règles. Si les documents n'ont fait l'objet d'aucune classification, il faut les traiter suivant l'ordre dans lequel on les trouve, les grouper par dossiers, et ensuite, de la même façon, en donner une description codée.

On assigne abrs au dossier un numéro d'identification unique. Les documents individuels qui le composent ont un numéro secondaire. Par exemple, si le numéro du dossier est 100, les documents individuels portent les numéros 100-1, 100-2, 100-3 etc. Naturellement les documents non triés comme les factures etc. ne sont pas numérotés. Le numéro d'identification commence par une lettre mnémonique, ainsi A majuscule désigne l'American Relief Association. Cette lettre peut être suivie d'un chiffre arabe qui désigne le pays envisagé par les documents. Ainsi le chiffre 4 désigne la Tchécoslovaquie. Après le numéro mnémonique, quelques intervalles sont laissés en blanc, puis suit un numéro correspondant à un numéro d'acquisition, par exemple, 00826, 00827 ou 00828.

Une fois assigné le numéro d'identification il faut sélectionner, dans le code, les critères de description ou mots-clés aptes à décrire les documents.

Pour le moment, le code se subdivise en 5 sections:

1) 1) Types de documents - lettres, journaux etc. Cette liste comporte 70 mots.

2) Mots-clés - 650 mots entrent couramment dans cette catégorie.

3) Noms de lieux géographiques - la liste en comporte 165 et nous avons 160 renvois.

4) Noms d'organismes - cette liste comporte plus de 600 noms.

5) Noms de personnes.

Selon que les documents l'exigent, des mots-clés sont ajoutés au code en accord avec diverses règles. Une des règles de base stipule qu'un mot-clé nouveau ne doit pas être le synonyme d'un mot déjà compris dans le code. S'il s'agit d'un presque-synonyme - et de telles additions ne sont guère encouragées - chacun des deux mots fait l'objet d'une définition très précise.

Le programme de la Hoover Institution est une adaptation du programme de bibliothèque IBM 1401. Le modèle 1401 est une machine IBM en usage depuis plusieurs années; elle est relativement moins chère que le modèle 7090 ou le modèles récents 'on-line'. Nous pensons quant à nous ne pas être victimes de l'avance technologique - dans ce cas précis l'abandon éventuel du modèle 1401 - car le programme IBM de bibliothèque que nous sommes en train de modifier est étudié en sorte d'être adapté au modèle de machine IBM le plus récent.

Notre programme a pour base l'invention de titres artificiels, composés à partir des mots-clés mentionnés ci-dessus. Le programme de la machine comporte un système de contrôle des termes incorporés ou code de mots-clés, auquel de nouveaux mots peuvent être ajoutés. La machine peut remplacer un mot non codé par son synonyme accepté par le code, dans la mesure où l'archiviste a prévu tous les synonymes possibles et les a insérés dans la machine. La machine peut imprimer et signaler sous les termes non autorisés qui ont été utilisés. Les mots-clés doivent être completés de la mention 'général' ou 'précis'. Un terme descriptif général est un mot qui apparaît fréquemment dans la collection envisagée et ne peut donc être utilisé pour la recherche sans l'addition d'un autre terme descriptif précis. Un terme descriptif précis circonscrit le champ de recherche et des termes descriptifs secondaires généraux le rétrécissent encore par la suite. Ainsi sont éliminés dès le début des recherches des matériaux qui ne s'y rapportent pas, et une plus grande efficacité est atteinte.

Le programme produira une liste imprimé du glossaire, liste évidemment plus longue que le code dans la mesure où elle comporte les synonymes non autorisés et les renvois sous forme de termes descriptifs secondaires et également de 'scope notes'. Ces dernières sont en fait les définitions de termes faites par l'archiviste à l'usage du chercheur ou du programmateur.

Le type de recherche prévu, et le plus courant, est celui que l'on appelle 'Boolean': très sommairement on peut dire que les termes en sont 'et', 'ou bien', et 'non'. Ainsi une demande peut être faite pour obtenir tous les documents comportant certains mots-clés spécifiques comme 'France et charbon et transport' ou tous les titres comportant des mots-clés spécifiques mais ne comportant pas un autre mot-clé, par exemple: 'France et charbon et transport, non Tchécoslovaquie'.

Afin d'obtenir un indexage en profondeur, le programmateur doit utiliser au moins un terme descriptif spécifique et un terme large, par exemple pour le sujet nourriture, il pourrait ajouter produits laitiers et lait.

Il est souhaitable que les archivistes soient au courant des techniques automatiques, des possibilités qu'elles offrent, des frais entrainés par l'usage des ordinateurs, et des perspectives qu'elles laissent entrevoir pour l'avenir, tels des centres d'information reliés entr'eux.

L'extraction automatique dans le domaine des archives africaines permettrait à l'archiviste de confier à la machine toute la masse de matériaux qu'il récolte, d'employer un personnel moins qualifié, et de fournir au chercheur plus d'information, dans des délais plus courts et à moindres frais.