CHAPTER II

REVIEW OF THE RELATED LITERATURE

This chapter is focused on literature concerning index and indexing in general, archive works and archive automation. Different types of indexing and archive automation are presented together with detail explanations and comments on each, as follows:

A. Purposes and Needs for Index

An index is a tool that reduces the number of documents we need to read through in order to find one or more items that will satisfy our need. Its purpose is to let through wanted documents and to hold back unwanted one (Lancaster 1968: 1-5). It permits us to go directly to the information needed. Without it, we would have to read through or do some sequential scanning, in order to find the items sought (Foskett 1974: 24-25), i.e. the index in a telephone directory. The index guides us to subjects of interest, provides us with an integrated picture of a subject field (Borko 1978: 1-3). In short, we can say that with an index the users can find the subjects they want, without having to read through the whole text.

B. Definition of Index

Experts have given different definitions to index. For example, Foskett (1974: 39) says that:

> An index is a directory which list the names of authors and subject, and the addresses of their literature, that is, the name of the book or journal in which the information appears; and if

> the index is made for a particular library, it also contains the addresses of the place in the library where the document is stored.

The American National Standard Institute (1968: 2) gives the following definition:

> An index is a systematic guide to items contained in, or concepts derived from, a collection. These items or derived concepts are represented by entries arranged in a known or stated searchable order, such as alphabetical, chronological, or numerical.

A more general definition which may be considered as a conclusion of index is found in the American Heritage Dictionary (1969: 669) :

> An index is an alphabetized listing of names, places, and subject included in a printed work that gives for each item the page on which it may be found.

C. Methods of Subject Indexing

There are many kinds of indexes, but the one preferred for information retrieval system is the subject index, since we primarily think of a system capable of retrieving documents in answer to particular subject requests. Subject indexing needs more intellectual effort than the more general one of authors, report numbers, issuing agencies, or languages (Lancaster 1968: 1-3). It required more time and money if in-depth analysis is called for (Doudnikoff 1973: 59).

To create the subject index, one needs first to decide on the subject matter then translate this conceptual analysis into index terms according to the terms in the created authority list or controlled vocabulary of terminology. The same documents may be indexed under different terms in different organizations because of the requirements and characteristics of users in each group (Lancaster 1968: 3-6).

Some documents may contain more than one specific subject. Each subject must be treated separately (Foskett 1982: 32).

The major problem of subject indexing is the system itself, in that it cannot cope with the fast growth of knowledge, especially in the field of science and technology. Another problem is that the same document must be indexed under different specific headings under different circumstances (Lancaster 1978: 1-3). The minor problem is its production. Lack of support from those who authorized resources for index is the main problem in index production. Quality control and delay in production are both to be solved (Berner 1968: 378-379). Subject indexing may be divided into two methods, assigned indexing and automated indexing.

1. Assigned Indexing

Assigned indexing or concept indexing is a way to identify the concepts involved in each document. In so doing, we use the same list of words both to help us in searching and to encode the document at the input stage. Thus, we assign the appropriate words to each document rather than rely on the author's choice. In assigning terms to indexed documents, we rely mostly on classification scheme which gives us a systematic listing of subjects and shows their relationship (Foskett 1982: 68-79). We assign class number or index terms to a document according to the classification scheme, in order to present its subject matter, instead of the natural language terms (Lancaster 1968: 8). It works well in a situation where users are not sure of their approach and need help from the arrangement. It is most useful for an information search on a broad subject,

and for browsing. With its helpful order and its simple search strategy, a classified index is most effective. The arrangement is designed to bring related subjects together in order to help the users broaden their searches by looking at the same part of the index (Foskett 1982: 79-80).

There are some weak points to classified indexing. Firstly, it cannot be expected to list all subject complexes dealt with in the document. For example, in Dewey Decimal Classification System (DC), it is impossible to synthesize the terms 'boilers' and 'corrosion' to express the precise concept 'corrosion of boiler'. To file a new document dealing with this topic, we have to enter it under both 'boiler' and 'corrosion' (Lancaster 1968: 9-10). Secondly, problems may arise if the users are not familiar with classification system, subject guides are then required in this situation (Berner 1968: 389). Thirdly, to evaluate the content of a document and adopt these essential parts to the conceptual framework of the classification, the indexer consistently faces the difficulty of interpreting the classification concepts. Interpreting from the natural language to the more general of classification can cause a change of some specific context (Highcock 1968: 85-96).

Another method of assigning terms to indexed documents is multifacet indexing. It is a permutation of all the elements in a composite subject to increase the numbers of terms. It allows the indexing of subject matter of any degree of complexity and is capable of accommodating document on new subject fields by combining of notations. The process is done through the means of rotation of notational elements, permutation of notational elements, citation or preferred order, and chain procedure (Lancaster 1968: 11-16).

2. Automated Indexing

The weak point of manual indexing led to the development of mechanical indexing. Earlier, punched cards were used for indexing purpose. Now we are in the computer age and computer devices such as machine readable punched cards, punched tape and magnetic tape can help us in indexing. The computer needs a notation which will transfer natural language terms into symbols that the machine can read. This can be done at a speed of up to many thousands of items per second. It can both perform remarkable feats of data correlation and information retrieval. With the increase of mass publication, the computers are more widely used (Foskett 1974: 53-55).

There are many kinds of computerized indexing. One of the most well known is the title based indexing, i.e. Keyword In Context (KWIC) indexing and Keyword Out of Context (KWOC) indexing. Others are indexing from the full text, and citation indexing.

a) Title-based Indexing

1) KWIC Indexing

The most primitive form of title-based indexing is KWIC indexing. The method was first used by Crestarodo for the purpose of compiling a catalog of the Manchester Public Library during the nineteenth century. The system was later applied to computer manipulation by H.P. Luhn of I.B.M. in 1959 (Foskett 1982: 39-46).

The title used in KWIC indexing can be that of an article, a book, a chapter, or a paper. Titles in the foreign language must be translated. The original language may be given in parentheses (Doudnikoff 1973: 53-55). To put items into the system needs litle

or no intellectual effort. We only transform the titles into machine readable form and the rest is done by computer. The computer will be programmed to reject non-significant words such as articles, and words that are unlikely to be useful in a particular subject area. Words remaining in the titles become index entries. The appropriate entries will be compiled and printed out readily in alphabetical order. Each title can be found under any of the significant words that it contains. The keywords appear in the middle of the line and the other words in the same title will be on both sides. Long titles may be truncated. A typical index consists of an accession file, which is an item's identifying number or a page number for a periodical, the keyword index based on the title, and an author index (Foskett 1982: 39-46).

KWIC indexing can be less effective if the titles to be indexed are not informative enough, or not retrievable. In such case, they need to be edited, either by rewriting the titles or by adding suitable descriptors. Another weak point to this type of indexing is truncation of titles to accomodate a 132-character line printer. Searchers have to refer to the bibliography section to get the full titles (Highcock 1968: 90-92). Also, a KWIC index with its unfamiliar appearance and its small sized printing letters is not very favorable. (See Fig. 1)

2) KWOC Indexing

KWOC indexing is KWIC in a developed form. Instead of the unfamiliar appearance of the keyword in the middle, a KWOC index has the keyword at the beginning of the line, followed by the complete title.

OBJECTIVE	KNOWLEDGE	LSRL111
RELIABLE	KNOWLEDGE	LSRL113
PUBLIC	KNOWLEDGE = THE SOCIAL DIMENSION OF SCIENCE	LSRL 31
VE COMPARISON OF THE SUCCESS AND FEASIBILITY OF A	KWOC CATALOGUE.. ATALOGUES = AN INDICATI	LSRL 47
SUCCESS AND FEASIBILITY O CLASSIFIED OR	KWOC CATALOGUES = AN INDICATIVE COMPARISON OF THE	LSRL 47
OF WISDOM.. OR ALL YOU EVER WANTED TO KNOW ABOUT	LAA COURSE RECOGNITION THE GETTING	LSRL 87
ION FOR LIBRARIANSHIP = THE CHANGING ROLE [OF THE	LAA] EDUCAT	LSRL 83
BASIC STATISTICS FOR	LIBRARIANS	LSRL 24
	LIBRARIANS AND ONLINE SERVICES	LSRL128
AN INTRODUCTION TO SCIENTIFIC RESEARCH IN	LIBRARIANSHIP	LSRL 9
IP = REPORT OF THE WORKING PARTY ON EDUCATION FOR	LIBRARIANSHIP EDUCATION FOR LIBRARIANSH	LSRL 78
THE FOUNDATIONS OF EDUCATION FOR	LIBRARIANSHIP	LSRL 71
HIGHER DEGREES IN	LIBRARIANSHIP	LSRL 81
EXAMINATION AND ASSESSMENT WITHIN EDUCATION FOR	LIBRARIANSHIP	LSRL 88
SOME THOUGHTS ON A FIRST DEGREE IN	LIBRARIANSHIP	LSRL 94
CATION FOR LIBRARIANSHIP EDUCATION FOR	LIBRARIANSHIP = REPORT OF THE WORKING PARTY ON EDU	LSRL 78
EDUCATION FOR	LIBRARIANSHIP = SOME THOUGHTS ON DEVELOPMENTS IN T	LSRL 77
HE UNITED KINGDOM EDUCATION FOR	LIBRARIANSHIP = THE CHANGING ROLE [OF THE LAA]	LSRL 83
HANDBOOK OF SPECIAL	LIBRARIANSHIP AND INFORMATION WORK	LSRL 2
BASIC UNDERGRADUATE EDUCATION FOR	LIBRARIANSHIP AND INFORMATION SCIENCE	LSRL 93
EDUCATION FOR	LIBRARIANSHIP AND THE MANPOWER PROBLEM	LSRL 86
IN THE MAINSTREAM = THE GRADUATE SCHOOL OF	LIBRARIANSHIP AT MONASH UNIVERSITY	LSRL 78
WHY DO PEOPLE ENROL IN	LIBRARIANSHIP COURSES? A SURVEY	LSRL 79
SYSTEMS ANALYSIS IN	LIBRARIES	LSRL 21
READER IN OPERATIONS RESEARCH FOR	LIBRARIES	LSRL 97
SYSTEMS ANALYSIS IN	LIBRARIES = A QUESTION AND ANSWER APPROACH	LSRL120
	LIBRARIES AND THE ORGANIZATION OF KNOWLEDGE	LSRL 63
THE TRAINING PROGRAMME AT THE BARR SMITH	LIBRARY	LSRL 76
A HISTORY OF	LIBRARY EDUCATION	LSRL 95
	LIBRARY EDUCATION 1977-1987	LSRL 72
TASK ANALYSIS OF	LIBRARY OPERATIONS	LSRL 27
SUCCESS IN	LIBRARY,SCHOOL = A STUDY OF ADMISSION VARIABLES	LSRL 91
MUMMY, WHAT ARE	LIBRARY SCHOOLS FOR?	LSRL 73
FOR THE PROFESSION THE ROLE OF	LIBRARY SCHOOLS IN PROVIDING CONTINUING EDUCATION	LSRL 89
TION PAPER ADOPTED BY THE ASSOCIATION OF AMERICAN	LIBRARY SCHOOLS, JANUARY, 1979 $ = A POSI	LSRL 98
1975-1 THE PROLIFERATION OF COURSES IN	LIBRARY SCIENCE AND AN ESTIMATE OF STUDENT OUTPUT,	LSRL 75
	LIBRARY SURVEYS	LSRL 16
	LIBRARY SURVEYS	LSRL 61
EDUCATION OF	LIBRARY TECHNICIANS IN AUSTRALIA	LSRL 85
ABSTRACTING SCIENTIFIC AND TECHNICAL	LITERATURE	LSRL 17
BIG MEDIA,	LITTLE MEDIA =TOOLS AND TECHNOLOGY FOR INSTRUCTION	LSRL125
AN INTRODUCTION TO	LOGIC	LSRL 66
EMPHASIS '70...ON	MANPOWER PLANNING	LSRL 88
EDUCATION FOR LIBRARIANSHIP AND THE	MANPOWER PROBLEM	LSRL 86
BIG MEDIA, LITTLE	MEDIA =TOOLS AND TECHNOLOGY FOR INSTRUCTION	LSRL125
DUCTION BIG	MEDIA, LITTLE MEDIA =TOOLS AND TECHNOLOGY FOR INST	LSRL125
THE ROLE OF SYSTEMS	METHODOLOGY IN SOCIAL SCIENCE RESEARCH	LSRL133
	MODELLING	LSRL 67
	MODELLING	LSRL132
NSTREAM = THE GRADUATE SCHOOL OF LIBRARIANSHIP AT	MONASH UNIVERSITY IN THE MAI	LSRL 78
	MULTILINGUAL ASPECTS OF PRECIS	LSRL 48
EXTRA-	MURAL STUDY IN CONTINUING PROFESSIONAL EDUCATION	LSRL 82
SCIENTISTS	MUST WRITE	LSRL117
THE VITAL	NETWORK = A THEORY OF COMMUNICATION AND SOCIETY	LSRL131
CES	NON-PARAMETRIC STATISTICS FOR THE BEHAVIORAL SCIEN	LSRL 22
UESTIONNAIRE EFFICIENCY = CONTROLLED REDUCTION OF	NON-RESPONSE MAIL Q	LSRL 51
	OBJECTIVE KNOWLEDGE	LSRL111
IRWTBLOR, CHEMFILE, STAIRS AND DIALOG SYSTEMS FOR	ONLINE SEARCHING OF ITS INDEXES AND THE S	LSRL134
LIBRARIANS AND	ONLINE SERVICES	LSRL128
TASK ANALYSIS OF LIBRARY	OPERATIONS	LSRL 27
READER IN	OPERATIONS RESEARCH FOR LIBRARIES	LSRL 97
LIBRARIES AND THE	ORGANIZATION OF KNOWLEDGE	LSRL 63
SES IN LIBRARY SCIENCE AND AN ESTIMATE OF STUDENT	OUTPUT, 1975-1980 E PROLIFERATION OF COUR	LSRL 75
THE ACCREDITATION PROCESS = A POSITION	PAPER ADOPTED BY THE ASSOCIATION OF AMERICAN LIBRA	LSRL 98
TOWARD	PAPERLESS INFORMATION SYSTEMS	LSRL 46
OPF CURRICULUM	PAPERS PRESENTED AT A WORKSHOP ON THE INTEGRATED C	LSRL 74

Fig. 1 Computer Print-out of a KWIC index

Additional terms are inserted in the titles of KWIC and KWOC index to overcome the problems of exhaustivity and specificity. It is called enriched KWIC or KWOC index. This method involves intellectual effort in selecting the additional terms. Both KWIC and KWOC indexes can retrieve subject information as well as document, provided that users can identify a document by its title and that the index entries are linked to appropriate descriptions of the document in question (Foskett 1982: 41-46).

Title-based indexes are suitable for current notifications and cover a small number of documents. They are not effective for retrospective searching among large collections. A good example is a computer produced KWIC index of 'Chemical Titles'.

b) Citation Indexing

Citation indexing is a simple method assisted by computer. It results from the concept that most scientific papers contain citations. If we collect a large number of papers, we then can build up a much larger number of original items cited. By knowing one that has been cited, we can find dozens of additional papers, every one of which provides a list of new citations that enable us to continue our searches. Citation indexing shows the linkage among papers that have particular points in common. The citation is indexed generally by title, author, date, place and name of publisher of those documents that support, provide precedent for, illustrate, or elaborate on what the author has to say (Garfield 1979: 1). A good example of citation index is the 'Science Citation Index'. (See Fig. 2-6)

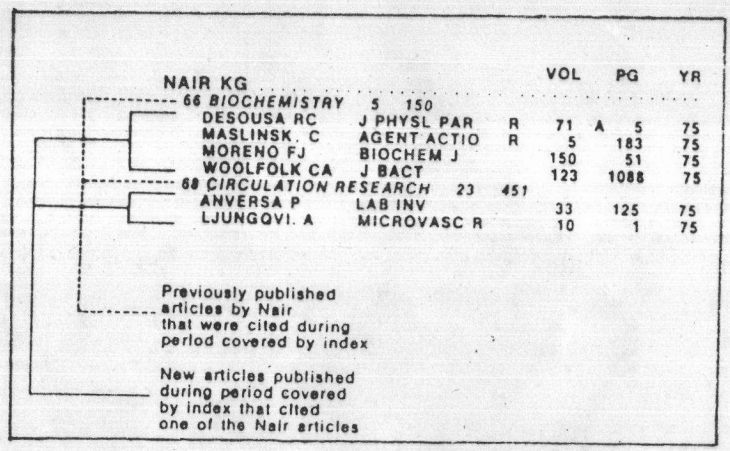

Fig. 2 Typical Entry from the Citation Index Section of SCI

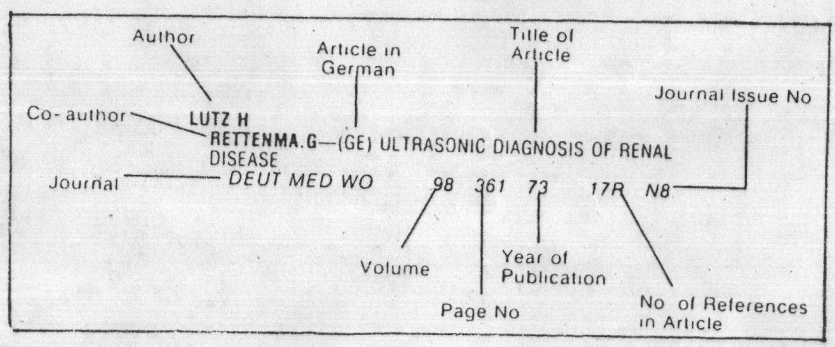

Fig. 3 Typical Entry from the Source Index of SCI

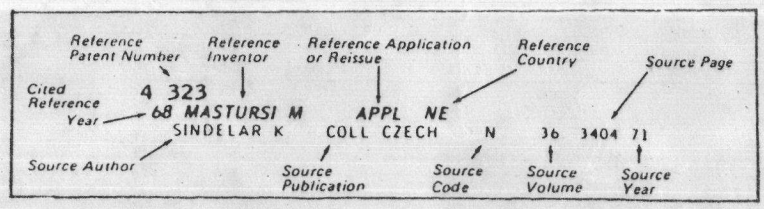

Fig. 4 Typical Entry from the Patent Index of SCI

		VOL	PG	YR
MAX PLANCK INST BIOL,TUBINGEN,WEST GERMANY				
BISSWANG H	BIOC BIOP A	321	143	73
BRAUN V	J BACT	114	1264	73
ENGELRAE M	BIOC BIOP R	53	812	73
HENNING U	FOL MICROB	18	268	73
	P NAS US	70	2033	73
SORSA V	NATURE-BIOL	245	34	73
TICHY H	GENETICS	74	S276	73
ZARYBNIC V	VIROLOGY	54	318	73

Fig. 5 Typical Entry from the Corporate Index Section of SCI

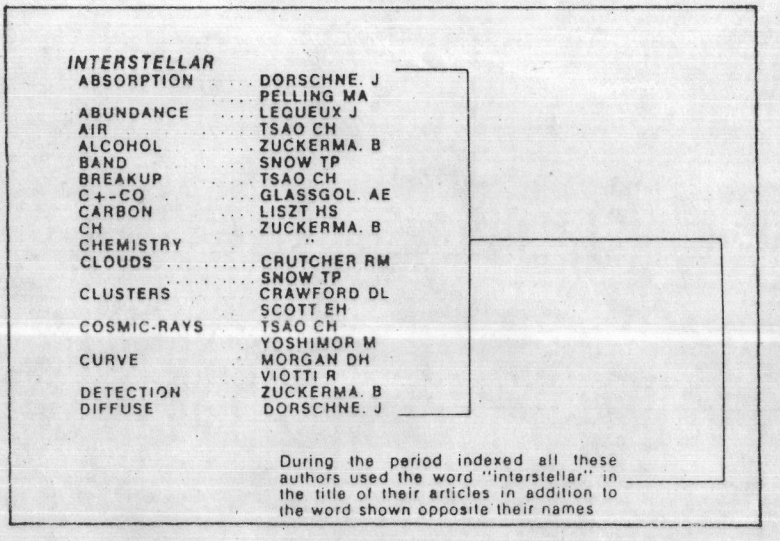

Fig. 6 Typical Entry from the Permuterm Subject Index Section of SCI

c) Indexing from the Full Text

This method of indexing is possible if we have the text of a document in a machine readable form, that is to put it in a computer data base. The computer is then programmed to derive suitable indexing terms to produce a conventional type of index or to select sentences from a document in order to write up an abstract. The computer performs its function by words count, counting the number of occurrences of a word in relation to the expected number, and by measuring the co-occurrence of terms, and by weighing each sentence in order to extract the sentences with weights above a pre-determined level. For the users who wish to find a document containing the terms they need, the computer can also be programmed to 'match' the users' 'term' with the one kept in the data base at high speed (Foskett 1982: 49-51).

D. Indexing and Archives

To present the proper view of archive works, definitions of two important terms concerned are given by Schellenberg (1971: 16) as follows.

archives:

> Those records of any public or private institutions which are adjudged worthy of permanent reservation for references and research purpose and which have been deposited or have been selected for deposit in an archival institution.

records:

> All books, papers, maps, photographs or other documentary materials, regardless of physical forms or characteristics, made or received by any public or private institution in pursuance of its legal obligations or in connection with the transaction of its proper business and preserved or appropriated for preservation by that institution or its legitimate successor as evidence of its functions, or other activities, or because of the informational value of the data contained therein.

We may say that archives are records produced or received by various groups or organizations, public or private, or by families and individuals. They are worthy of preservation for reference and research purposes. Any users can make use of them. Their physical forms vary, e.g. correspondence, reports, memoranda, directions, authorizations, awards, bids, certificates, deeds, manifests, notifications, payrolls, questionnaires, receipts, returns, schedules, specifications, vouchers, and warrants (Schellenberg 1971: 204-205).

1. Life Cycle of Records

Population increase leads to the increase of government and private activities. A lot of records of these activities are made and received. These records rapidly increase in amount as well as in complexity. The owners cannot keep all of them for administrative purposes, so parts of them have to be disposed of. Some are kept for administrative use in the office and some as archives for future administrative and educational research. Most organizations have their own systems in preserving and destroying their records. They set up a committee to draw guidelines for the destruction of non-valuable records and the preservation of records of sufficient public value as archives. The objectives of the system are to discard the valueless records and to preserve those of some value, be it administrative, historical, legal, political, or otherwise which information provide not to be obtained elsewhere (Hodson 1972: 48-51). (See Fig. 7)

2. Archive Works

Different organizations may have different methods of maintaining their archives, but the principal ones are as follows:

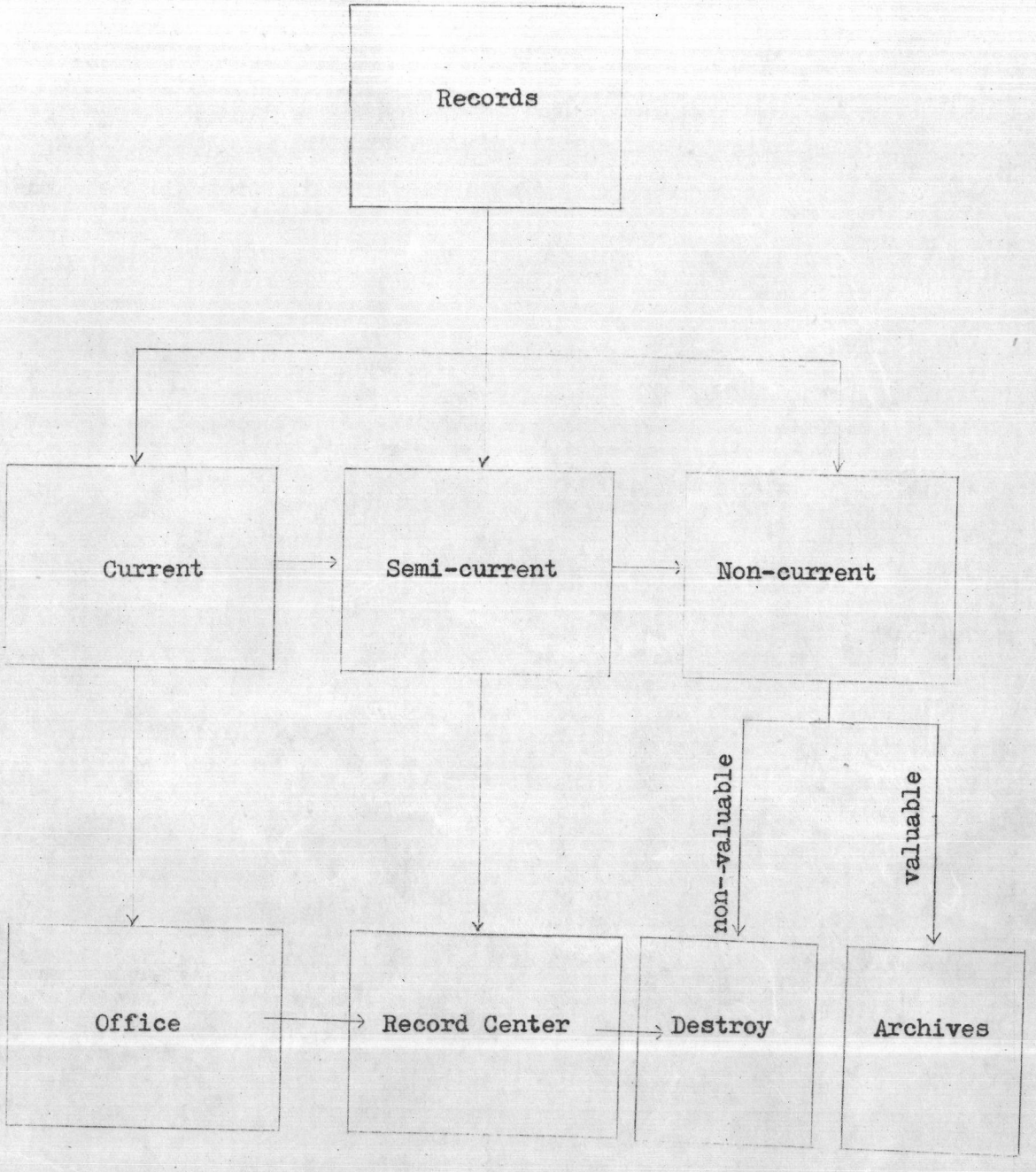

Fig. 7 Life cycle of Records

a) Arrangement

Arrangement of archives involves two activities, namely, accession register and classification.

Accession register is a methodical, chronological way to keep record of receipt of materials in archives. An account book can be used for this purpose. An accession register is to be filled in with accession number, date of accession, sources, description of records, quantity, and location or classification (Hodson 1972: 119-124). Accession register provides the archivists with an immediate, brief, and permanent record of how materials came into a repository (Schellenberg, cited by Hodson 1972: 120).

Classification is the method in which one classifies the records received in to group of principal subject. The archivists scan the whole records received and group them together according to subject and types. The purpose is to arrange the records in a classified order, that is to give them a permanent number and storage (Hodson 1972: 124-129).

b) Description

This activity of archive works has to do with the preparation of finding aids, such as guides, inventories, calendars, cataloges, and lists (Schellenberg 1971: 204-214). Index is an efficient finding aid for archives.* It identifies archives in relation to subject, that is to indicate the users to the location of subject in a particular archives (Schellenberg 1965: 272-273).

---

*More detail on archives indexing are presented in 'E. Automation and Automated Indexing of Archives'.

c) Storage

This step consists of three activities, namely, permanent storage, repair and rehabilitation, and reprography.

Permanent storage refers to practical ways in which archives are kept in the archive building. Repair and rehabilitation are done before or after the damage of archives occurs. Reprography is an activity to satisfy both internal and external needs, e.g. by photographing, xerography, statfile processing for copying of large items such as maps, and microfilming (Hodson 1972: 107-170).

d) Services

This step involves any forms of service provided by archivists within the limitation set by the owner of archives' repository policy. Most archives are open to the public for research purpose, some are reserved for a certain purpose, e.g. for administrative use, and some are private. Services include searching for information needed and reprography of the original copy.

E. Automation and Automated Indexing of Archives

The first system applied to archive works was the Termatrex Information Retrieval System, utilized in the Joseph Down Manuscript Library of the Henry Francis du Pont Winterthur Museum in the early 1960's. It was a mechanized, not computerized, system, developed by Jonker Business Machines, Inc. In this system, a card was provided for one keyword selected from the manuscript descriptions. If an item indexed had a certain keyword then a hole was drilled in the position for that item on the card. For searching procedures, the

cards were passed over a light-giving source. Any cards with holes that transmitted light were relevant to the search. The system made use of the preprinted indexing forms and made it easier to prepare input data. It worked well with a small collection of closely related documents though it provided only one type of retrieval process without recording the results of specific searches (Ingerman 1962: 331-340).

In 1958, the Presidential Papers Section of the Library of Congress Manuscript Division was set up. It started work on the arrangement, microfilming, and indexing of twenty-three Presidential Collections in the Library while using mechanized approach. The papers were arranged chronologically and each item was indexed by correspondence. The machines used were keypunched machines, card sorters, and a tabulator. An index entry was created for each document and then punched on a 80 column punch card. The entry consisted of seven fields, namely, number identifying the appropriate Presidential Collection, writer-recipient, date, series number, page count, additional information, and card count. The writer-recipient and date fields were sorted in order to arrange the entries in the printed index. In 1969, the Library acquired a computer and used three programs in its work: a 'sort' program, an 'edit' program, and a program for producing printed copy. With the computer's assistance, an editor could check errors in the original listing, add new data, and delete the irrelevants, through the change cards. This system provided chronological arrangement and item indexing (Smith 1967: 295-302).

In 1965, the Public Archives of Canada utilized item indexing to index the Papers of Canadian Prime Ministers. The system used resulted in various forms of finding aids, i.e. an author index which secondarily sorted by subject and then by data; a subject index,

secondarily sorted chronologically and then by author; and a chronological index, secondarily sorted by author. These finding aids provided precise information retrieval but preparation for input data was rather complex (Atherton 1967: 303-309).

Also in 1965, the Hover Institution on War, Revolution and Peace, began its project to apply computer usage to all collections in the Institution. The records of the American Relief Administration were chosen to fulfil this project. The method used was KWIC. The individual chose appropriate keywords from an authority list which included document forms, subject keywords, geographic place names, and personal names. The keywords were limited to twenty-four characters, and the total description for each set of documents being indexed was limited to 320 characters. Identification number and inclusive dates were also contained in the description. Each set was composed of five to ten documents, concerning the same general subject, which were drawn together by the indexes from the roughly sorted records. when print-outs were produced, the full 320 characters description was printed under each keyword (Campbell 1966: 298-302).

In 1966, the Library of Congress developed the MRMC I (Master Record of Manuscript Collection) System. It dealt with the administration and control of archives holdings but not with the information retrieval as needed by researchers (Knowlton 1975).

In October 1969, the Manuscript Division of the Library of Congress developed the computer package program known as SPINDEX II (Selective Permutation Indexing). The program was jointly developed by many other universities and archives in the United States of America. It was designed to handle information as formated in

archives or manuscript inventories, registers, or guides. The purpose of the program was to rearrange the information from such guides into permuted index of terms contained in the guides. The system combined terms, proper names, and personal names into one index. It was capable of indexing titles of varied lengths up to two thousand characters, providing a cross reference, indexing dates by year, updating the file through single corrections, formating the final print-out in a hierarchical structure used by archivists and manuscript curators, and of other sophistications that provided other sub-programs to suit the archivists' needs (Burk 1970: 19-23).

In 1979, the University of Guelph created the University Document System 'CODOC' (Co-operative Documents) for the Library Collection of government documents. The system provided means of assigning a mnemonic county code number and other codes to various area of bibliographic data in order to facilitate quick sorting. The print-out included shelflist, personal and corporate author catalogs, title list, keywords index, identification of language, source location, and document type. It is hoped that this system can be applied to archive collections with minimal change of computer programs. The system provided quick sorting capability and good level of subject approach through KWOC with the development of an effective stopword list and selective word linking and subject enrichment terms (Sadek 1982: 75-77).