ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION STUDIES FOR AFRICA (SISA)

# RETROSPECTIVE CONVERSION OF CARD CATALOGUES TO ONLINE CATALOGUE BY USING OCR TECHNOLOGY: A CASE OF ADDIS ABABA UNIVERSITY LIBRARIES

**A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION SCIENCE**

**BY**

KEBEDE HUNDIE WORDOFA

JUNE 1997

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF INFORMATION STUDIES FOR AFRICA

# RETROSPECTIVE CONVERSION OF CARD CATALOGUES TO ONLINE CATALOGUE BY USING OCR TECHNOLOGY: A CASE OF AAU LIBRARIES

By

Kebede Hundie Wordofa

**Name and Signature of Members of the Examining Board**

Ato Getachew Birru, Chairman, Examining Board _____

Dr. G.G. Chowdhury, Advisor _____

Ato Sisay Fissaha, Advisor _____

Dr. G.A. Alabi, Internal Examiner _____

Prof. A. Chatterjee, External Examiner _____

Dedicated To

***My brother Ato Beddassa Hundie***

***and***

***his wife W/o Getenesh Adugna.***

## ACKNOWLEDGEMENTS

First of all, I wish to extend my thanks to Addis Ababa University in general and the University Library System, my employer, in particular, for sanctioning me two years study leave. I am equally thankful to DAAD for the financial assistance extended to me in order to accomplish my study at SISA.

My heartfelt gratitude goes to Dr. G. G. Chowdhury, my supervisor under whose close supervision and constructive criticism the whole thesis was written. I am also equally grateful to Ato Sissay Fisseha, my second supervisor, for his fruitful comments, and ideas especially for his forbearing assistance during developing the prototype program. My special appreciations go to Ato Getachew Birru, the Dean, Dr. Taye Taddesse, and Ato Tesfaye Birru, all of SISA, and to Ato Lishan Adam of PADIS.

Thanks are also due to Ato Beddessa Hundie, W/o Getenesh Adugna, Ato Gemeda Hundie, Ato Worku Hundie, Ato Taddesse Kalbessa, W/t Alemtsehay Legesse, and Ato Agegnehu Habte who gave me moral support during my study at SISA and provided me with necessary materials required by this study.

I would like to extend my thanks to the staff of the University Library System for their assistance in providing all the necessary information demanded. Especially, I am thankful to Ato Azene Zenebe, Head of Computer Center.

Last but not least, I am appreciative to my parents, and all my friends who morally supported me throughout my study.

*K. H.*

## ABSTRACT

Conversion of card catalogues to a computerized catalogue, called retrospective conversion or RECON, is the primary hurdle of effective library automation. There are several options for RECON, but recent advances in OCR technology make it a prominent alternative. This thesis has mainly examined technical feasibility of the OCR technology for RECON with particular reference to Addis Ababa University libraries' union catalogue.

The thesis reviews the various options for RECON in the context of Addis Ababa University Library System. It also reviews technological advancements of optical scanning and OCR technology stressing on document scanning and OCR with particular reference to their application in libraries, particularly RECON. An introduction to the Addis Ababa University Library System gives an overview of the library system with particular reference to its cataloguing department and cataloguing practices followed therein, and automation plan of the library system.

Two sets of catalogue cards, each consisting 115 added up to 230, were chosen from the union catalogue for this study. The sets of cards consisted of main entries made under name of personal author (40 each), under title (40 each), and corporate body (35 each). A prototype program, called PAEBE, was written in C++ using the first set of cards. The various steps include: scanning and conversion of the cards to ASCII text files, analysing of the files, manual editing of the files, called preprocessing, if necessary, writing and running the program creating output records consisting of each bibliographic item preceded by an

appropriate field identifier.

It was noted that the success of the prototype depends much on the accuracy of OCR software, consistency of information on card catalogues, and quality of card catalogues. The prototype was tested with the second set of catalogue cards; and the results are discussed. It was found out that the performance of the prototype program is encouraging. The thesis also highlights implementation strategy for the RECON using OCR technology. Finally, some conclusions and recommendations for further studies are forwarded.

## TABLE OF CONTENTS

DECLARATION . . . . . i

DEDICATED TO . . . . . ii

ACKNOWLEDGEMENTS . . . . . iii

ABSTRACT . . . . . iv

TABLE OF CONTENTS . . . . . vi

LIST OF TABLES . . . . . xiii

LIST OF FIGURES . . . . . xiii

LIST OF ABBREVIATED TERMS . . . . . xiv

**CHAPTER ONE**

**INTRODUCTION** . . . . . 1

1.1 BACKGROUND . . . . . 1

1.2 STATEMENT OF THE PROBLEM AND JUSTIFICATION . . . . . 3

1.2.1 Statement of the Problem . . . . . 3

1.2.2 Justifications . . . . . 5

1.3. OBJECTIVES . . . . . 8

1.3.1 General Objectives . . . . . 8

1.3.2 Specific Objectives . . . . . 8

1.4 METHODOLOGY EMPLOYED . . . . . 9

1.4.1 Source of Data . . . . . 9

1.4.2 Sample Cards . . . . . 9

1.4.2.1 Personal Author . . . . . 10

1.4.2.2 Corporate Body . . . . . 10

1.4.2.3 Title . . . . . 11

1.4.3 Sampling Method . . . . . 11

1.4.4 Literature Review . . . . . 12

1.4.5 Discussions and Observations . . . . . 12

1.4.6 Prototype Development Tools . . . . . 13

1.5 SIGNIFICANCE OF THE THESIS . . . . . 14

1.6 SCOPE AND LIMITATIONS OF THE THESIS . . . . . 14

1.6.1 Scope of the Thesis . . . . . 14

1.6.2 Limitations of the Thesis . . . . . 15

1.6 ORGANIZATION OF THE THESIS . . . . . 15

**CHAPTER TWO**

**RETROSPECTIVE CONVERSION: AN OVERVIEW . . . . . 16**

2.1 INTRODUCTION . . . . . 16

2.2 REASONS FOR RECON . . . . . 16

2.3 METHODS OF CONVERSION . . . . . 17

2.3.1 Contracted Services . . . . . 18

2.3.1.1 Advantages . . . . . 19

2.3.1.2 Disadvantages . . . . . 20

2.3.2 In-house Conversion . . . . . 20

2.3.2.1 Advantages . . . . . 21

2.3.3 Bibliographic Networks . . . . . 22

2.3.3.1 Advantages . . . . . 23

2.3.3.2 Disadvantages . . . . . 23

2.3.4 Batch Cataloguing Databases . . . . . 24

2.3.4.1 Advantages . . . . . 26

2.3.4.2 Disadvantages . . . . . 26

2.3.5 Combination of Methods . . . . . 28

2.4 FACTORS FOR SELECTING RECON OPTION . . . . . 29

2.6 COSTS OF RECON . . . . . 30

2.7 GUIDELINES FOR RECON . . . . . 33

2.8 BIBLIOGRAPHIC STANDARDS FOR RECON . . . . . 34

CHAPTER THREE

OPTICAL CHARACTER RECOGNITION (OCR) AND ITS APPLICATIONS IN LIBRARIES . . . . . 38

3.1 INTRODUCTION . . . . . 38

3.2 OPTICAL SCANNING . . . . . 39

3.2.1 Mark-sense Recognition . . . . . 39

3.2.2 Barcode Label Recognition . . . . . 40

3.2.3 Document Scanning . . . . . 40

3.2.3.1 Types of Scanners . . . . . 41

3.2.3.1.1 Drum Scanners . . . . . 41

3.2.3.1.2 Desktop Scanners . . . . . 41

3.2.3.1.3 Hand-held Scanner . . . . . 43

3.2.3.1.4 Video Scanners . . . . . 43

3.2.3.2 Main Features of Scanners . . . . . 43

3.2.4 Optical Character Recognition (OCR) . . . . . 44

3.2.4.1 Types of OCR Packages . . . . . 45

3.2.4.1.1 Trainable OCR Software . . . . . 45

3.2.4.1.2 Omnifont OCR Software . . . . . 46

3.2.4.1.3 Intelligent OCR Software . . . . . 46

3.2.4.2 Features of OCR Software . . . . . 47

3.2.4.2.1 Deferred Processing . . . . . 47

3.2.4.2.2 Templates . . . . . 47

3.2.4.2.3 Spell Checking . . . . . 48

3.2.4.2.4 Other OCR Software Features . . . . . 48

3.2.4.3 Criteria for Selecting OCR Software . . . . . 48

3.2.4.3.1 Accuracy . . . . . 49

3.2.4.3.2 Integration . . . . . 50

3.2.4.3.3 Batch Processing . . . . . 50

3.2.4.3.4 Input and Output Formats . . . . . 51

3.2.4.3.5 Throughput . . . . . 51

3.2.4.3 OCR Software Products . . . . . 51

3.2.4.3.1 OmniPage Professional . . . . . 52

3.2.4.3.2 Wordscan Plus . . . . . 53

3.2.4.3.3 Recognita Plus . . . . . 54

3.2.4.3.4 TextBridge . . . . . 54

3.2.4.3.4 Merits and Demerits of the Software . . . . . 56

3.3 APPLICATION OF OCR FOR RECON . . . . . 59

**CHAPTER FOUR**

**ADDIS ABABA UNIVERSITY LIBRARY SYSTEM . . . . . 65**

4.1 INTRODUCTION . . . . . 65

4.2 BACKGROUND . . . . . 65

4.2.1 Staff of the AAULS. . . . . 67

4.2.2 Holdings . . . . . 70

4.2.3 User Population. . . . . 71

4.3 CATALOGUING DEPARTMENT . . . . . 72

4.3.1 Cataloguing Practices. . . . . 73

4.3.2 Entries in the Union Catalogue . . . . . 74

4.3.3 Arrangements of Entries . . . . . 74

4.3.4 Problems of the Department . . . . . 75

4.3.5 Levels of Catalogue Description. . . . . 75

4.4 AUTOMATION PLAN OF THE LIBRARY SYSTEM . . . . . 76

**CHAPTER FIVE**

**PROTOTYPE PROGRAM FOR AUTOMATIC EXTRACTION OF BIBLIOGRAPHIC ELEMENTS (PAEBE)79**

5.1 INTRODUCTION . . . . . 79

5.2 GENERAL PROCEDURES . . . . . 79

5.2.1 Scanning the Sample Cards . . . . . 80

5.2.2 OCR Conversion . . . . . 82

5.2.3 Understanding Formats/layouts of the Records . . . . . 84

5.2.4 Conditions Used for the Program . . . . . 84

5.3 LOGICAL STRUCTURE OF THE PROGRAM . . . . . 86

5.4 PSEUDOCODE OF ALGORITHM DEFINITION. . . . . 87

5.5 REQUIREMENTS OF THE PROGRAM . . . . . 91

**CHAPTER SIX**

**TESTING RESULTS AND IMPLEMENTATION STRATEGY OF THE PROTOTYPE PROGRAM . . . . 93**

6.1 INTRODUCTION . . . . . 93

6.2 RESULTS . . . . . 93

6.2.1 Average Errors of OCR Output Records . . . . . 94

6.2.2 Types of Errors Identified. . . . . 96

6.2.3 Preprocessing of Records. . . . . 97

6.2.4 Final Output of the Prototype Program . . . . . 99

6.2.5 Performance of the Prototype . . . . . 101

6.2.6 Factors Determining OCR for RECON . . . . . 103

6.3 IMPLEMENTATION OF THE PROTOTYPE SYSTEM . . . . . 103

6.3.1 Software Considerations . . . . . 103

6.3.2 Hardware Considerations . . . . . 106

6.3.3 Staff Considerations . . . . . 107

6.3.4 Procedural Considerations . . . . . 108

**CHAPTER SEVEN**

**CONCLUSIONS AND RECOMMENDATIONS . . . . . 110**

7.1 INTRODUCTION . . . . . 110

7.2 CONCLUSIONS . . . . . 110

7.3 RECOMMENDATIONS . . . . . 113

BIBLIOGRAPHY . . . . . 115

APPENDICES . . . . . 121

APPENDIX I

Partial sample of personal author main entry records output by OCR software . . . . . 121

APPENDIX II

Partial sample of title main entry records output by the OCR Software. . . . . 127

APPENDIX III

Partial sample of corporate body main entry records output by the OCR Software . . . . . 133

APPENDIX IV

Guide for a Discussion Held with the Cataloguing Team of the Cataloguing Department of the Library System . . . . . 135

APPENDIX V

Guide for a Discussion Held with the Head of the Computer Center of the Library System . . . 136

## LIST OF TABLES

Table 2.1 Price of leading cataloguing and RECON databases. . . . . 72

Table 3.1 Accuracy of four leading OCR software packages. . . . . .50

Table 3.2 Summary of merits vs. demerits of the major OCR software packages. . . . . .58

Table 4.1 Number of the library staff by category. . . . . .70

Table 4.2 Collections of the libraries of the university. . . . . .70

Table 4.3 Number of potential and registered or actual users of the AAULS. . . . . 71

Table 6.1 Table showing average errors per record in character.. . . . . 95

Table 6.2 Example of wrongly converted characters having similar features. . . . . 96

## LIST OF FIGURES

Figure 4.1 Organizational Chart of the Library System.. . . . . 68

Figure 5.1 Example of card catalogues output by the scanner. . . . . 81

Figure 5.2 Examples of records converted to text formats after text zones are created.83

Figure 5.3 Flowchart showing logical structure of the prototype program. . . . . 86

Figure 5.4 Pseudocode of the prototype program.. . . . . 91

Figure 6.1 An example of a record before and after preprocessing .. . . . . 99

Figure 6.2 Records output by the program.. . . . . 101

Figure 6.3 Flowchart describing steps required for implementation of the prototype RECON system for AAU Library System. . . . . .109

## LIST OF ABBREVIATED TERMS

| | |
|---|---|
| AACR2 | Anglo-American Cataloguing Rules, Second Edition |
| AAU | Addis Ababa University |
| AAULS | Addis Ababa University Library System |
| CCS | College of Social Sciences |
| FBE | Faculty of Business and Economics |
| HTML | Hypertext Mark Up Language |
| ISBD | International Standard Bibliographic Description |
| ISBN | International Standard Book Number |
| ISSN | International Standard Serial Number |
| LC | Library of Congress |
| MARC | Machine Readable Cataloguing |
| OCR | Optical Character Recognition |
| OPAC | Online Public Access Catalogue |
| PAEBE | Program for Automatic Extraction of Bibliographic Elements |
| RECON | Retrospective Conversion |

# CHAPTER 1

# INTRODUCTION

## 1.1 BACKGROUND

Library collections and user population of libraries in general are always increasing and new services are being launched to meet diversified users' information needs. This is mainly because new fields of subject specializations are emerging and large amount of information is being generated.

Consequently, libraries having huge collections and a large number of users obviously face problems. First of all, managing bibliographic information of large collections manually using conventional catalogues, involves numerous costs. Second, quality, efficiency and flexibility of library services to the users will be very difficult to achieve, for instance, access points to the library collections will be limited.

For efficient and flexible library services, implementing computer-based systems has thus, become crucial especially, for those libraries owning large collection and having a large number of users (1) to provide a service at a lesser or no great cost, (2) to give added benefits at a lesser cost, (3) to increase efficiency of services that are clerical, routine and repetitive and thus prone to human errors (Tedd 1993, 6-7), improved services to users for example, online public access catalogue (OPAC), resource sharing and library cooperation (Wedgeworth 1993, 471). In other words, library automation is a means to increase efficiency, to manage costs, to improve library

services and management, and to remedy breakdowns or shortcomings of existing manual systems (Riger 1992).

To introduce automated services to any library, there needs a machine-readable version of library catalogue records as it is probably the prime and central issue for library automation. Therefore, the availability of a machine-readable catalogue has been also long considered a necessary prerequisite for effective applications of computers in libraries (Tedd 1993, 122).

This is because library catalogues play a key role in facilitating access to collections of a library. Library catalogues are maps to library's collections. However, conventional library catalogues involve intellectual costs, production costs, maintenance costs, and other overhead costs for example, space, cabinets, furniture, etc. Moreover, there are a limited number of access points they provide.

Broadly speaking, computers especially in view of its contribution towards savings in some or all of the costs of manual catalogue, are economical to operate. Specifically speaking, computerized catalogue offers the following benefits for libraries (Tedd 1993, 124):

- the ability to produce multiple copies of the card catalogue;
- an increased number of access points for both staff and users for searching the online catalogue;
- elimination of card maintenance, and production costs;
- potential use of catalogue records obtained from elsewhere;
- the ability to produce subset catalogues for special purposes;

- an opportunity for the catalogue to interface with other house keeping modules such as circulation control, interlibrary loan, acquisitions, and management activities.

To this list, Skapura (1990) adds a point by saying that computerized catalogue makes the act of searching easier.

However, a major problem for libraries that want to automate their catalogues is the conversion of existing manual records to computerized catalogue or online catalogue that are not converted through day-to-day processing, the process commonly referred to as **retrospective conversion (RECON)**. This refers to how to go back and do the conversion efficiently, accurately, and economically. Obviously this is time consuming and quite costly especially, for libraries in Ethiopia in general and Addis Ababa University (AAU) libraries in particular.

## 1.2 STATEMENT OF THE PROBLEM AND JUSTIFICATION

### 1.2.1 Statement of the Problem

As has been mentioned earlier, in order to keep pace with the rapid developments made possible in library services and to adequately tackle the effects of the increase in, and diversification of, demands for information, the automation of all library processes has become inevitable. RECON serves as an indispensable tool to reach such an objective. Thus, the question is not whether a library should undertake RECON, but rather how it should be done (Jacobs 1990).

In line with this, libraries in Ethiopia in general, and AAU Library System, which is the biggest academic library in the country, in particular, do not have RECON option that well suits to their

resources at hand mainly financial resource, even though there exist various conversion methods. Major RECON options have been discussed in the next chapter.

Nevertheless, as advances in optical scanning and optical character recognition (OCR) have been showing rapid change in accuracy and improvement in reading different quality of source documents, its application for RECON has long been considered as an alternative. Having this in mind therefore, the problems that this study has concentrated on are as follows:

1. Is OCR technically a feasible solution of retrospective conversion, especially for libraries in Ethiopia with a particular reference to Addis Ababa University (AAU) Library System?
2. What limitations are anticipated if it is used in retrospective conversion?
3. To what extent can OCR technology help AAU libraries as far as retrospective conversion is concerned or how much is it applicable to the libraries' card catalogue?
4. How much is the average number of errors in the card catalogues of the libraries that require editing?
5. What are the factors that determine the effectiveness of OCR technology for retrospective conversion with particular reference to the library system?
6. Can the outcomes of the study about application of OCR for RECON be generalized to other libraries? What are the factors?

### 1.2.2 Justifications

There are a number of ways by which retrospective conversion can be done. First, by deriving of bibliographic descriptions from various sources or databases. These are mainly computer network systems or bibliographic utilities such as the OCLC (Online Computer Library Center), RLIN (Research Libraries Information Network), Utlas, WLN (Washington Library Network) and other online reference sources including BRS, Dialog, etc. These systems provide network members or customers, mainly libraries and information centres of the industrialized countries such as the US, UK, Canada and other European countries, with online catalogues, automated acquisitions, and interlibrary loan services.

It is possible for these libraries to use the systems to obtain bibliographic records for online catalogue and other services once they have got connected to the central databases via telecommunications. They are nationwide in scope (Tedd 1993, 146; Hunter 1985, 169), often expensive and records can be cumbersome to convert (Riger 1992), and require considerable effort in the correct identification of the required record and have ancillary problems in attaching and transferring of full local information fields (Harrison 1985). Consequently, libraries in developing countries in general cannot get the services because of too much costs involved.

Second, for libraries that do not have access to the bibliographic networks the Library of Congress (LC) and other vendors provide MARC bibliographic records on magnetic tapes and other media. The MARC records can be used to avoid a great deal of keyboarding by retrieving and matching the records against an existing catalogue.

However, searching full MARC database on magnetic tapes and downloading records that match library's holdings to an in-house database one by one and verifying and editing the downloaded records will be a routine and time consuming job on the top of all the costs involved (Burton and Petrie 1986, 119-120; Riger 1992). Due to much efforts and time needed for local processing of records, many libraries seem to be lessening their dependence on LC cataloguing services (Hunter 1985, 174). There are also a number of titles that vendors cannot find (Skapura 1990). Though this approach is relatively cheap, perhaps, many of the publications produced in developing countries in general and in Sub-Saharan Africa in particular, could not be available from the MARC records.

The third option is to have a specialized bureau do the conversion. In this case a vendor will take the whole conversion project and do the conversion without much involvement of library's staff. This approach to RECON can be the best option in terms of quality, effort, time and bibliographic standards but it is still quite costly (Drabenstott 1986).

The last option is in-house conversion. Due to the above-mentioned reasons, libraries will have to be able to do their own conversion by using whatever options they have at hand (Tedd 1993, 146), especially libraries in the developing countries. In-house conversion is usually done by keying records directly into a database. This takes very long time and requires a considerable amount of labour costs. To this effect, OCR has become an alternative way of capturing data from printed media including card catalogues.

As far as most of the Sub-Saharan African countries are concerned retrospective conversion has

still remained a major problem as the aforementioned options that are available elsewhere are neither directly applicable nor cost-effective and feasible for them.

As mentioned earlier the first three options require a huge amount of money for telecommunication charge, cost of records, expertise, hardware, and/or efforts required, and so on that many libraries of developing countries cannot afford. In countries where there is shortage of funds, like those in Africa, there are many priority areas to deal with. Thus, it makes sense for them to fully, effectively and carefully utilize what information technology (IT) systems they have at hand (Woherem 1995).

Therefore, libraries in developing countries in general and in Ethiopia in particular, will have to look for a better alternative of retrospective conversion which has to be efficient and less expensive. Technological advances have increased the range of alternatives, and the continued development of promising technologies of optical scanning and OCR may add to the conversion options (Drabenstott 1986). It is one of currently used methods by which original data is captured into computer databases and has made great advances in the last several years and recently emerged as a viable method of accurate transcription of source materials (Beutler 1995, 61-62).

To this end, as pointed out by Rice (1981), optical scanners with an associated optical character recognition (OCR) software is a breakthrough that would permit economical retrospective conversion alternative. The application of OCR technology for RECON has been reviewed in Chapter 3, Section 3.3.

## 1.3. OBJECTIVES

### 1.3.1 General Objectives

The overall objective of the thesis is to investigate the technical feasibility of OCR technology as an alternative to other retrospective conversion options and to develop a prototype program that extracts bibliographic elements from records output by OCR software, with a particular emphasis to Addis Ababa University (AAU) libraries' card catalogue.

### 1.3.2 Specific Objectives

In order to achieve the general objective, the specific objectives of the study are:

1. to look into formats, layouts, and bibliographic elements on the card catalogues of Addis Ababa University Library System (AAULS) in order to come up with the degree to which they are consistent in following uniform or standard cataloguing rules;
2. to scan card catalogues and convert to text (ASCII) files by using an appropriate OCR software to analyse formats and layouts of the converted information in order to determine the amount of errors, which helps to know the extent of preprocessing required in editing the text;
3. to develop an algorithm to be used in writing a program that identifies various bibliographic elements from text files output by OCR software, based on the results of number 2 above;
4. to write a prototype program using the algorithm to be developed;
5. to test the effectiveness of the program on a good set of sample cards scanned and converted into text files by OCR software;

6. to investigate major factors determining the effectiveness of OCR technology for RECON; and to identify limitation(s) of OCR technology for RECON as well.

## 1.4 METHODOLOGY EMPLOYED

### 1.4.1 Source of Data

To carry out the study, the union catalogue of AAU library system was used as source of data from which sample card catalogues were drawn.

### 1.4.2 Sample Cards

Two sets of **main entry** sample cards from the union catalogue were taken. The first set was used to develop, test and improve a prototype program that identifies each data element from a card catalogue once they are scanned and saved as a text file using OCR software. The second set was to test the effectiveness of the prototype on another set of data.

For each set, samples were taken according to three types of main entries. Obviously entries are made under name of a personal author, corporate body, or title. The total number of cards for each set was **115** added up to 230. To reach at this figure, first of all main entry card catalogues were examined for various local variations; and a sample had to be taken for each **major variation** of entry types. The figure is distributed for each type of main entry. Much attempts have been made to consider as much different possibilities or variations of the entry types as possible.

#### 1.4.2.1 Personal Author

Mostly, main entries are made under the name of personal author. According to the AACR2 (Ango-American Cataloguing Rules, second edition), works are mainly entered under the name(s) of personal author:

- for a work of a single personal author;
- when two or more authors have shared in creation of a work; or
- for a collective work not having collective title.

**Forty main entry cards** made under name of personal author were taken for each set. As far as the contents and details of the cards are concerned, there are lots of variations among them. Some of them are more detailed than others. Structurewise, there are also variations among the samples. These are, for example, single author item, multiple authors item, items having series title, cards having edition statement of responsibility, long notes, added entries, etc. (see **Appendix I**).

#### 1.4.2.2 Corporate Body

A work is mainly entered under a corporate body i.e., any institution, organized body or assembly of persons known by a corporate or collective name, again according to AACR2 when the work is by its nature the expression of the corporate body; or when the wording of the title or title-page clearly implies that the corporate body is collectively responsible for its content.

When compared to personal author, main entries entered under corporate body are much less. Consequently, the number of samples taken was also less. **Thirty-five** samples were taken for

each set again taking into account different variations. The formats and contents of the entries vary according to the nature of the material. For instance, monographs, reports, newsletters, conferences, constitutions, laws and treaties, etc. issued by a corporate body have different structure (see **Appendix III**).

#### 1.4.2.3 Title

Title main entries are made for works whose authors have not been ascertained, works by more than three authors, for collections of independent works or parts of works, or works known primarily or conventionally by title rather than by name of author (Maxwell 1980, 4).

**Forty** samples were taken for this type of entries. Different varieties of main entry cards were selected including for example, titles having statement of responsibility and without statement of responsibility. Usually the variations depend on the type of materials - monographs, serials, journals, reports, conferences, etc. (see **Appendix II**).

### 1.4.3 Sampling Method

A purposive/judgement sampling technique was employed to draw the two sets of sample cards because some card formats and variations might be escaped in case of random sampling. In other words, this was to include in the algorithms or the prototype as much variations, rules, and possibilities of main entries as possible, and to test the prototype accordingly.

### 1.4.4 Literature Review

Literature review was conducted to collect background information relating to the problems at hand. All the available and relevant literature have been reviewed for the following reasons:

- to explore the trends on retrospective conversion options, mainly their advantages and disadvantages;
- to study advances in OCR software packages and optical scanning technology that are available in the market, their application for retrospective conversion; and
- to see if similar studies have been done, and to analyse their outcomes in relation to the problem at hand.

The following information systems and libraries provided the much needed information in course of the literature review:

- Addis Ababa University libraries and information sources;
- PADIS, and ILRI database resources; and
- Searches on CD-ROM databases.

### 1.4.5 Discussions and Observations

Formal and informal discussions were held with some of the staff of the library system, mainly with the head of the university library's Computer Center and cataloguing team of the cataloguing department. Since the researcher is one of the employees of the library system, there was much cooperation from the staff.

A discussion was held with the head of Computer Section of AAU libraries to get information

on its activities, staff, hardware available, etc.; and about its future plans regarding automation of library services in general and retrospective conversion in particular (see **Appendix V**).

One formal and frequent informal discussions were also held with the cataloguing team consisting of head, chief cataloguers, assistant cataloguer, and verifiers of the cataloguing department of the library system. These were to learn about their cataloguing practices, procedures, problems, etc. (see **Appendix IV**).

Besides, the union catalogues were observed and examined to understand their formats, level of description, variations and up-to-datedness, etc.

### 1.4.6 Prototype Development Tools

A scanner and OCR software were used to convert the bibliographic information on card catalogues to text files for which a program was written to identify each data element of each record in the file. The resulting records were printed and analysed to know success and failure rates of the program and so on. And finally, a prototype showing how to apply OCR in retrospective conversion was developed.

Moreover, the following software packages available at SISA computer laboratory were used:

- Turbo C++ compiler to write and compile the program;
- Harvard Graphics to process graphical data;
- WordPerfect 5.1/6.1 mainly for word processing.

## 1.5 SIGNIFICANCE OF THE THESIS

Retrospective conversion at present is a practical problem for libraries in Ethiopia in general and AAU libraries in particular. Therefore, it was expected that the investigation would provide information of practical use, specifically, information that could be used in the development of an automatic identification of data elements from a machine-readable surrogate of catalogue records so that they could be used for creation of bibliographic databases.

It was expected as well that it could be used as a starting point for further development for the libraries to convert their existing card catalogues to online catalogues. It could also probably be useful for other libraries in Ethiopia who may wish to automate their library in-house activities by taking the prototype with modifications or adapting to their specific practices.

## 1.6 SCOPE AND LIMITATIONS OF THE THESIS

### 1.6.1 Scope of the Thesis

This study was based at Addis Ababa University Library System. The scope of the study was limited to the union catalogue of the library system consisting of a main library (J. F. Kennedy Memorial Library) and branch libraries over the university campuses.

### 1.6.2 Limitations of the Thesis

In this study, development of the prototype for the application of OCR for retrospective conversion was limited to software and hardware available to the researcher. There was no chance to use the state-of-the-art scanner and OCR software packages to develop and test the prototype. However, this does not mean that the outcomes of the thesis can be underestimated; all the recommendations can be implemented accordingly. This research was also limited to the technical feasibility of the OCR technology for RECON in the context of AAU Library System. Other factors such as time and cost involved in RECON using this option could not be studied due to lack of time and other resources.

## 1.6 ORGANIZATION OF THE THESIS

This thesis is divided into seven chapters. The first chapter is an introduction to the thesis. The second chapter discusses available retrospective conversion alternatives, their advantages and disadvantages so as to highlight their implications for developing countries as a whole. The third chapter is about optical scanning and optical recognition. It reviews technological advancements of optical scanning and OCR software capabilities, compares different OCR software packages on the market, and examines the applications of OCR in retrospective conversion. Chapter 4 provides information on Addis Ababa University Library System. Chapter 5 presents the prototype development procedures used, and algorithms of program developed. Chapter 6 presents test results of the prototype system and its strategies for implementation. Last not but not least, chapter 7 draws some conclusions and forwards recommendations for further research.

# CHAPTER 2

# RETROSPECTIVE CONVERSION: AN OVERVIEW

## 2.1 INTRODUCTION

Broadly speaking, retrospective conversion is the process of converting the information from a shelf list or public catalogue to a machine readable format using one or more of retrospective conversion methods available. This chapter offers background information on the problems at hand discussing different conversion methods and exploring their advantages and disadvantages from different viewpoints - costs involved, time taken, quality of the records, etc. Before that, the reasons for retrospective conversion are given in the next section.

## 2.2 REASONS FOR RECON

As has been mentioned, the primary reason for retrospective conversion is to install an automated library system that requires a machine-readable database. Most libraries do RECON for one or more of the following reasons. These are (Boss 1984): (1) to create an online catalogue, (2) to create a union list of libraries' holdings, (3) to create an automated circulation system, (4) to provide access to records that are difficult to locate in the present environment-the materials that have only been partially catalogued, and (5) to create a security backup file for the card catalogue. Bryant, et al (1995) add to this list by saying that resource sharing is an other reason which enables sensible management decisions to be made in relation to acquisition,

preservation and withdrawal of stock. The immediate need of a library may be one of these reasons. As RECON is very expensive and time consuming task, libraries should look beyond the immediate need and consider how the database will be used for many purposes in future.

The purposes behind the creation of bibliographic database, as clearly stated by Bossers and Law (1990), are twofold. The first one is to maximize access to library collections. This gives the ability to profit existing and future developments in retrieval techniques. The second is to facilitate library management. RECON provides data necessary for the automation of library administration. It can be needed in such situations as reorganization, integration of library collections, or as part of a general introduction of automation in the library.

## 2.3 METHODS OF CONVERSION

As briefly discussed in the previous chapter in section 1.2.2, there are several options available and in use by libraries. These include:

- contracted services or commercial bureaus;
- in-house conversion;
- bibliographic utilities or networks;
- batch cataloguing databases; or
- a combination of these.

These methods have been seen by many writers from different perspectives; and studies have been conducted by different people in these areas. This section reviews works relating to the

conversion methods. Reynolds (1985), has offered detailed discussions on two major retrospective conversion methods - bibliographic networks and contracted services from different angles. He has discussed at length in terms of financial resource requirements, impact on staff, required facilities, completeness and accuracy of data, amount of record editing and verification required, time and cost factors, hit rate, overhead expenses, and other related issues.

Drabenstott (1986) has discussed in an article in Library Hi Tech the methods in detail and raised different issues of RECON. The whole issue of IFLA Journal (volume 16 number 1 1990) was also devoted to retrospective conversion projects in Europe. Saffady (1994, 260-264) has described these conversion methods and examined different products. Each method is discussed below.

### 2.3.1 Contracted Services

In this method a service bureau or vendor takes care of a conversion project without bothering library staff with much extra labour. First the library intending retrospective conversion has to evaluate vendors according to local needs and available budget, and select one to sign a contract with. Detailed and clear contract is essential to specify the amount of editing and data entry required by the vendor, to identify the standards to be used, and local needs, etc. (Drabenstott 1986).

In using such a service libraries may be required to prepare and supply a microfilm or photocopy of the shelf list or sometimes the shelf list itself. This is another work that libraries will have to do before the actual conversion takes place. Instead of requiring the shelf list or photocopy, a

vendor may supply a library with a microcomputer system to prepare a database of minimal records containing search keys, local call numbers, and holdings, etc.

Then the vendor will match the received library's holdings information against its database and produce machine readable records that are revised to include local call numbers, holdings information, or other relevant information. Besides, the vendor will do all the necessary original cataloguing of records not found in the database if this is specified in the contract.

Nowadays, there are a number of vendors doing retrospective conversion on contract basis. Bibliographic utilities as well as commercial vendors contract to provide conversion. OCLC's Retrocon, Utlas, and Brodart are some of them.

#### 2.3.1.1 Advantages

Drabenstott (1986) says that in contracted services when compared to other methods, the library staff must do very little of the actual retrospective conversion work; contracting with an outside vendor will result in predefined costs and time frames for completion of the conversion; and librarians who contract for RECON services are generally very pleased with the quality of the work done, speed, experienced staff, proper facilities, and adherence to standards. There will be no need to maintain and supervise a special staff for the project (Bossers and Law 1990).

If a library has already a circulation system installed, this is usually an easier way to expand the records to full MARC using this approach. This is done by sending backup files of a brief circulation database to the selected vendor.

#### 2.3.1.2 Disadvantages

However, this approach has some drawbacks. First of all it is quite costly. Beaumont (1986) has enumerated that the costs include the preparation of conversion specifications, editing the shelf list cards, service bureau charges, quality control, and loading the resulting records into the library's automated system. The investment of library staff time should not be underestimated. Bossers and Law (1990) have pointed out that there are also other limitations: (1) the preparatory work for the project as well as the continuous monitoring during the project can prove to be very time-consuming and labour-intensive, and incur additional costs for supplies; (2) errors will be made by staff unfamiliar with the library; and in-house quality control will be required; (3) adjustments to a contract during a project are generally extremely difficult to arrange without extra time and cost; and (4) finding errors and correcting them can be time-consuming.

### 2.3.2 In-house Conversion

In-house approach to RECON is mainly done by keying data from the original catalogue entries or shelf list into a database by clerical staff and using equipment specifically hired for this purpose. It is carried out using either existing staff or contract labour. Of all the options, this is by far the most labour intensive and probably most costly, especially, for libraries having big collection. Now its application is in smaller libraries, in libraries with systems not using the MARC format, and in libraries that want to enter very brief records into an automated circulation system.

Since direct entry of records into a database takes long time, libraries have long begun

considering optical character recognition (OCR) as an alternative to in-house conversion approach. But to automatically convert the information output by OCR into a database file, first the information has to be analyzed and fields on the card have to be identified so as to import into a database. There needs an automatic format recognition software that analyses the information and identifies data elements. So far, this method has not proved to be very successful for this purpose, but developments continue to be made in the field (Bossers and Law 1990).

#### 2.3.2.1 Advantages

This approach to RECON allows for a project over which the library has a high degree of control, and is likely to provide a retrospective conversion of high quality. In terms of cash cost, it requires less than the cost of other methods. This method will incur lower up-front per item costs and allow files to remain on-site (Cohn et al 1992, 77).

#### 2.3.2.2 Disadvantages

Haddad (1990) has noted that an in-house approach to RECON is quite costly in terms of finding funds, time taken to set up the project, provisions of equipment, hiring and training of staff. Moreover, editing and verification of records will be a big task. This is often considered and desirable for some of the material due to changes of cataloguing standards, or obscurities in the original catalogue, and so on.

Cohn et al (1992, 77) have also noted some limitations of this approach. These are: impact on existing work flow, excessively long timeliness for project completion, additional space and

hardware requirements, added supervisory and quality control efforts, and increased personnel costs. In-house conversion can be impeded by insufficient experienced staff (Drabenstott, 1986).

### 2.3.3 Bibliographic Networks

The term 'bibliographic networks' is often used to describe networks that have been established mainly to support shared cataloguing by means of a central database (Tedd 1993, 78). There are many such bibliographic networks now exist in North America and Europe.

OCLC (Online Computer Library Center) in the US, BLCMP (Birmingham Libraries Co-operative Mechanization Project), LASER (London and South Eastern Library Region), PICA (Project on Integrated Catalogue Automation), RLIN (Research Libraries Information Network) that has been set for four large North American research libraries such as Colombia, Harvard, and Yale university and New York Public Library, SIBIL and REBUS in Switzerland, and UTLAS (University of Toronto Library Automation System) are the major ones. Tedd (1993, 78-82) has noted brief details of and references to these bibliographic networks in the book entitled An introduction to computer-based library systems.

Using this system a member library can search the online database to identify records that match local holdings. Once found, the records are edited and the library's holding information is added. Then all of the retrieved records are written to a magnetic tape and taken to the library to be downloaded onto the library's database.

#### 2.3.3.1 Advantages

As has been noted by Boss (1984), this approach is an attractive retrospective conversion option because of the comprehensiveness of the databases. The databases include the complete MARC database of records distributed by the Library of Congress (LC), as well as supplementary, original cataloguing records distributed by member libraries for locally held titles not in the MARC database. That means this approach provides a higher hit rate and minimum original cataloguing. In addition to the comprehensiveness of the databases, the library staff can have complete control over the amount of editing to be done on retrieved records from the shared cataloguing database (Drabenstott 1986).

#### 2.3.3.2 Disadvantages

Though the use of bibliographic networks can facilitate retrospective conversion, a conversion project may prove too expensive for a given library as of Cohn, et al (1992, 78). The cost will be much more if the following expenses are considered: one time costs for such items as the purchase or rental of terminals, profiling charges, telephone line installation, or training; on going costs for purchasing records from the database; and the costs of employees to supervise and operate the terminals.

Moreover, since the networks cover only certain geographic area and are meant to serve the member libraries, a library which is not participating, cannot generally become an online participant simply to carry out a RECON project. They therefore, do not extend RECON use to libraries that are not participating in the system (Reynolds 1985, 291). Telecommunications costs are in inherent characteristics of the bibliographic networks; and this poses potential

problems (Boss, 1984). Moreover, this technique is not invariably applicable to cataloguing records created in non-roman character sets (Saffady 1994, 263).

### 2.3.4 Batch Cataloguing Databases

Libraries are now able to obtain a fixed database that can be locally searched to retrieve machine readable records. These databases are available from various producers (Drabenstott 1986). These include:

- GRC (LaserQuest)
- The Library Corporation (BiblioFile)
- Library Systems & Services (MiniMARC)
- Utlas (ReMARC), and others.

GRC publishes **LaserQuest**, which is one of the world's largest CD-ROM database for libraries for RECON and for on-going cataloguing, having a set of six CD-ROM laser discs that contain **seven million MARC records** arranged in title order, as pointed out by Cibbarelli (1993, 270). Records for books, serials, computer files, music, visual materials, maps, and manuscripts are included.

The **BiblioFile** databases, which consist of MARC-format records in various sets, are available on three CD-ROMs. Over eight million MARC records are now available on the databases: the full LC MARC database (all LC cataloguing in English), the LC MARC Foreign database, the Catalog Card Corporation's SEARS/Dewey database, special databases such as the BiblioFile Contributed MARC discs with records contributed by BiblioFile users in school, public,

research and academic libraries, and Canadian MARC records. Moreover, there are three new databases: A/V ACCESS with MARC records for current and popular audio-visual materials, A-V ONLINE MARC with individual MARC records representing more than 340,000 titles from the National Information Center for Education media's (NICEM) A-V Online database, and DOCUFILE with government publications from a mid range of nations (Cibbarelli 1993, 262).

The **MiniMARC** system contains the total MARC database on two twelve-inch laser discs. Libraries operate this system in much the same manner as BiblioFile, although the MiniMARC equipment includes a very useful tractor feed card stock printer not available on the BiblioFile system. The productivity rate is about the same as that of BiblioFile (Drabenstott 1986).

A potential complement to either of the above systems is the Utlas **ReMARC database**, which is also provided on a twelve-inch videodisc. It contains all foreign language materials not found in the Library of Congress MARC records (Drabenstott 1986).

A user can search the database; edit records; add call numbers, holdings, and other local information; and then store records on floppy diskettes. If the CD-ROM database finds shelf list title, the amount of typing is relatively small. Since the costs of the data are fixed, the cost of each record declines as the database is more heavily used. While the hit rate varies by size and type of a library, a portion of each local collection requires original cataloguing, and so on (Drabenstott 1986). A library could search a bibliographic database and alternatively, could use a **MITINET/MARC software package** to write the data on floppy disks in MARC format

once the information is entered. It makes on-site conversion easier.

Another variant of this approach of retrospective conversion is the employment of the database of another library which has a similar collection profile; a copy of such a database can be used as a basis for one's own data conversion (Bossers and Law 1990).

#### 2.3.4.1 Advantages

It provides faster and less expensive than full keystroking of the information content of a library's shelf list. It also results in a MARC-format record (Saffady 1994, 261). Drabenstott (1986) reveals that if the library's collection is not too unusual or too large, this approach can be very cost-effective. It is the least expensive option (Cohn et al 1992, 78).

#### 2.3.4.2 Disadvantages

This approach requires additional equipment, space, personnel, and management. Moreover, it may delay the conversion until the library's computer system is installed. Obviously the **hit rate will not be as high as on online cataloguing databases**. The hit rate for most of the libraries in the developing countries is still very less when compared to the developed countries as most the materials produced by and/or in the former are not covered by the databases. Furthermore, the retrieval and modification of existing resource records is time-consuming and labour-intensive; there is no way to obtain copy-specific information from an external source; and there is much greater potential for error in this approach (Boss 1984).

Library may require a number of databases to increase the hit rate, which means the number of

matches achieved between library holdings and the databases, since it depends on the size of the databases and quality of the database. Complete conversion may not be achieved by a single database. Moreover, most of the databases rely heavily or exclusively on the **LC MARC database**, which may be segmented or offered as a series of separately priced subsets (Saffady 1994, 232). The **BiblioFile** databases that are subdivided into several segments for which separate subscriptions are required can be a typical example in this respect. Consequently, this will not be affordable by most of libraries in the developing countries where the library budgets are being drastically cut down from time-to-time. In line with this, the price for major batch conversion systems is given in **Table 2.1**.

**Table 2.1** Price of leading cataloguing and RECON databases

| DATABASE | PRICE (US$) |
|---|---|
| **BIBLIOFILE** | 2,250 plus subscription of choices |
| **Choices:** | |
| LC MARC English Weekly | 2,995 |
| LC MARC English Monthly | 1,690 |
| LC MARC English Quarterly | 1,090 |
| LC MARC Foreign Quarterly | 800 |
| Sears/Dewey Quarterly | 850 |
| Contributed Cataloguing Schools & Public Quarterly | 240 |
| Contributed Cataloguing Research & Academic Quarterly | 180 |
| Canadian MARC Quarterly | 445 |
| A/V Access Quarterly | 995 |
| **LASERQUEST** | 4,300 first year;<br>2,600 second and subsequent years |
| **MINIMARC** | 8,000 plus<br>12,000 maintenance fee per year |
| **REMARC** | Individual records cost 0.3 cents each. |

**SOUREC:** Cibbarelli (1993, 262) and Drabenstott (1986).

### 2.3.5 Combination of Methods

There is no one method that is necessarily superior to another since each method has its own limitations. The conversion methods are not mutually exclusive. Based on the following major points outlined, it is obviously possible to combine any two or more of the methods.

Libraries considering RECON need to consider the following points (Murphy 1990):

- budget and staff size;
- collection size;
- special features of the collection such as date of the material, language, and uniqueness of publisher; and
- acceptable quality of the catalogue records to be obtained.

Therefore, libraries can adapt the best features of each option to meet their local needs and budget constraints. For example, materials such as serials, nonprint, and local history may be converted on-site using either existing staff or contract labour. No one method, but a combination of several, has proved necessary to ensure a cost-effective, and high-quality database to support the automation needs of a library (Skapura 1990).

For non-matching records in-house conversion method may be required because for any library using the different conversion methods, the hit rate cannot be hundred percent. In line with this, Asher (1982) reveals that the largest bibliographic database, OCLC, has converted over three million records for public, academic, and special libraries with a hit rate of **78 percent**. This can imply that the hit rate for other databases such as BiblioFile, LaserQuest, MiniMARC, ReMARC, etc. is definitely very low.

## 2.4 FACTORS FOR SELECTING RECON OPTION

The RECON option best suited to a particular library will depend upon a number of factors. Some of the factors that should be taken into account to select a suitable option or a combination of options are noted by Drabenstott (1986), as follows:

**(i) Current cataloguing system:-** If currently cataloguing is not being done in machine-readable form, then RECON problem will continue to grow. Because records are being added from time to time as new acquisitions arrive.

**(ii) Size of the collection:-** The size of the collection determines the method of conversion to be used. For instance, for small collections, in-house conversion may be usually the best option; and libraries having large collections may probably find contracted services or bibliographic utilities to be better option than in-house one.

**(iii) Nature of the collection:-** If a collection contains a considerable proportion of unusual items for example, items of only local interest, in-house conversion may be the only practical option, although there are services that specialize in the conversion of such collections through direct keying of the catalogue entries.

**(iv) Age of the collection:-** The age of the collection also determines the conversion method. For example, if the collection contains a significant percentage of pre-1968 titles it will be necessary to find a vendor with records contributed from non-LC sources (Epstein 1990).

**(v) Language mix:-** The LC MARC files began with English titles, and added other languages, starting with the romance languages, throughout the 1970s. The files still do not contain any non-Roman vernacular records (Epstein 1990).

**(vi) Desired amount of information:-** The amount of information desired in each record also determines the conversion method. For example, since most of the vendors and utilities use MARC format libraries using other standards may not use these options.

**(vii) Costs involved:-** As a matter of fact, RECON is very expensive, labour intensive, and time-consuming. Its costs are central issue for libraries, especially for libraries in Ethiopia in general, and AAU libraries in particular which have faced acute shortage of funds. To look at costs of, for example, the least expensive RECON option, i.e., making use of the batch cataloguing databases involves costs that include database subscription, workstation, labour, etc. Costs for RECON have been discussed in the next section.

## 2.6 COSTS OF RECON

As has been mentioned, one of the major factors in undertaking RECON is the costs of the project. As noted by Epstein (1990), libraries often ignore the full costs of alternatives, focusing mainly on the cash costs. The actual money that will have to be paid someone else are sometimes the only costs considered. But the true costs of the project include much more than the costs billed by contractors or vendors. For example, the cash costs of contracting for RECON are higher than doing it in-house, but the true costs of contracting the entire project

to a vendor may be lower.

Valentine and McDonald (1986) have outlined the factors that determine the costs of RECON. These factors are: (1) the definition of conversion, (2) standards of acceptance, (3) method of conversion, (4) hit rate, (5) standards for the creation of machine-readable records for non-hits.

Peters and Butler (1984) have classified the full RECON costs into three: labour cost; equipment and supplies cost; and vendor or bibliographic utility charges. The labour cost must be based on the total compensation actually given to all employees working on the project and will depend on the salary levels of the participants.

Equipment and supplies costs are dependent upon the vendor selected. Some of these costs will be fixed and some prorated. Some utilities and vendors require the purchase of terminals, computers, manuals, or special interfacing hardware. These costs will be fixed. Certain costs, such as line charges and maintenance costs for an existing computer terminals, could be prorated by the percentage of time devoted to the project. Vendor or bibliographic utility charges will depend upon the method selected and the vendor and/or utility chosen.

Therefore, there are four steps to arrive at an approximate cost for each RECON alternative on the basis of the above classified costs. These, as outlined by Peters and Butler (1984) are:

(1) the cost of searching;

(2) verification and editing;

(3) coding and input of non-hits; and

(4) obtaining final records.

Variables which must be examined in each step include the amount of labour costs required, charges imposed by bibliographic utilities or vendors, and the cost of equipment and facilities. The resulting costs will be a function of these variables for each step. Peters and Butler (1984) also reveal that although the costs may vary drastically, the procedure used to calculate the costs will not vary with regard to in-house or vendor production.

Once costs for the individual steps for each method have been determined, the total cost for each alternative, including any of the optional costs, can be calculated. The average per-item cost is the total cost divided by the number of titles to be converted. The resulting per-item cost will reveal the method that is least expensive for the particular library (Peters and Butler 1984).

Any method can be costed following these procedures, as long as the user remembers to include all labour, supply, equipment, and vendor costs for each step. In this case of in-house conversion using manual entry of the information, the costs involve are labour cost, supplies and equipment costs.

Libraries should include these costs in any analysis or RECON into their projects. However, the cheapest method is not necessarily the most cost-beneficial method; the major factor is the set of standards set for the project depending on the purpose of the RECON (Peters and Butler 1984).

## 2.7 GUIDELINES FOR RECON

Retrospective conversion, as has been already said, has the reputation of being one of the most time and money consuming activities in library automation. It should therefore, be well planned before hand as a sound and detailed plan is a key to a successful retrospective conversion project.

Having this in mind, the LIBER Library Automation Group consisting of 15 delegates from ten European countries and six representatives of the European Communities such as the Council of Europe, IFLA, LIBER, and EFLC suggests that libraries should have guidelines, and has prepared a checklist intended to aid the management of libraries taking decisions on how to undertake RECON project (Bossers and Law 1990).

The basic things that libraries should take into account, according to the guidelines are:

- determination of the need and reasons of conversion in order to gain support of the parent institution;
- assess the order of priority in which the conversion is to take place, meaning which part of the collections to select for processing first, second, etc;
- identification of participant libraries, meaning cooperation between libraries on a national and international level can reduce the costs of RECON projects, thus allowing more rapid and economic progress;
- choosing the best conversion method according to library's local needs, nature and size of collection, and available resources;

- processing the material for RECON i.e., selection of items on the basis of the frequency of use, and the accessibility to the user;
- preparing material basis for conversion - catalogue or shelf list, or making a copy of the chosen list, etc.;
- specification of bibliographic descriptions to be used i.e., brief or full descriptions, or others, since the level of description influences, to a great extent, the rate of conversion; and
- management of the project including (1) preparatory works such as stock checking for availability of the physical item or collection, sampling the collection in order to estimate the duration of the project (2) planning and evaluation of the project including distinct specification of phases in the project; preparation of a realistic timetable; documentation of procedures, statistics and planning and every decision.

## 2.8 BIBLIOGRAPHIC STANDARDS FOR RECON

Retrospective conversion will be of more use if it is conformed to certain standards in order to make the data gained with great effort, money, and manpower usable to other libraries i.e., resource sharing. Moreover, a database is a long-term investment, a resource with the same kind of life as a card catalogue (Boss 1984). Under these circumstances, it makes sense to conform to certain standards. Even though the need for a standard is no more questionable, the amount and types of information that constitute a full bibliographic record is certainly open to debate, as revealed by Reynolds (1985, 284).

In line with this, Michael Gorman (1977) has identified nine parts of the catalogue record as constituting the minimum set of data for RECON. These are: call number, main and added entries, subject headings, uniform title, title proper, edition statement, publisher and date, brief collation, and series statement.

However, Avram (1972) says that the standard for converting retrospective records should be the same as those for current records which includes, in addition to Gorman's requirements, the subtitle, statement of responsibility, place of publication, complete collation, and notes.

Furthermore, there are the minimum requirements for RECON which are suggested in the study of Bibliographical Standards Requirements for Retrospective Catalogue Conversion prepared for the Working Party on Retrospective Cataloguing (Sule 1990). These requirements are the eight areas of ISBD(M) and (S) as a whole except few elements. These are as follows:

1. Title and statement of responsibility area
    1.1 Title proper; this includes other titles in publications without a common title as well as titles of subseries
    1.3 Parallel title
    1.4 Other title information
    1.5 Statements of responsibility
2. Edition area
    2.1 Edition statement
3. Publication, distribution etc. area

4.1 Place of publication and/or distribution

First statement

4.2 Name of publisher and/or distributor

First statement

4.4 Date of publication and/or distribution

5. Physical description area

5.1 Specific material designation and extent

6. Series

6.1 Title proper of series

6.4 Statement of responsibility relating to the series

6.6 Numbering within series

6.7 Enumeration and/or title of sub-series

6.12 Numbering within sub-series

7. Note area

8. Standard number (and alternative) and terms of availability area

8.1 Standard number

Sule (1990) points out that they can be used as the minimum when retrospective conversion is done with the book-in-hand. When converting from existing card-catalogues they have to be used just as a framework and only in as far as the information is available in the card catalogue.

However, since the late 1960s the accepted standard has been the MARC format. The format is being used by a number of libraries in the developed countries and bibliographic database

hosts. It is used by the Library of Congress (LC), the bibliographic utilities, and commercial catalogue record vendors.

Therefore, Asher (1982) strongly suggests that in order for retrospectively converted records to interface with the machine-readable records created by LC, the utilities, or vendors, it is necessary that cataloguing records be converted using this standard format for bibliographic communication, using the accepted standards of AACR2 and LC subject, name, and series headings as far as possible. Adhering to a standard help libraries to share data. It is also insurance for the future use. MARC format seems to be generally accepted standard for RECON (Sule 1990).

The data that can be contained in a MARC record include the entire spectrum of information normally represented on catalogue cards - authorship, title, edition, physical description, series statements, notes, subject headings, added entries, call number, various control numbers - plus a great deal of other potentially valuable categorizing information that can be encoded in fixed elements and elsewhere on record (Reynolds 1985, 285).

# CHAPTER 3

# OPTICAL CHARACTER RECOGNITION (OCR) AND ITS APPLICATIONS IN LIBRARIES

## 3.1 INTRODUCTION

There are several input systems that help to convert human-readable or understandable information to machine-readable forms. They can be: (1) voice recognition systems which are used to convert voice to digital forms; (2) keyboard oriented systems that are used to keypunch onto punched cards or key onto storage devices; (3) scanning and optical recognition systems; and (4) scientific instrumentation systems.

In library applications the two input systems that account for the majority of all input activities are keyboard oriented and optical scanning and recognition (Saffady 1996, 19).

This chapter offers information on scanning and optical recognition technology with much emphasis on document scanning and optical character recognition (OCR) with particular reference to their applications in automation of library functions mainly, retrospective conversion. The performance, features, and comparisons of major OCR software packages are also given.

## 3.2 OPTICAL SCANNING

Optical scanning is one of the computer data input systems converting human-readable information to machine readable form required by the central processor. This method uses reflected light to determine the information content of source documents. Once identified, the information is encoded and transferred to a computer.

Saffady (1994 19) divides scanning technologies into four: (1) mark-sense recognition, (3) bar-code label recognition, (3) document scanning or image digitization, and (4) optical character recognition (OCR). The most suitable system for any given application will depend upon the nature of the data to be input into computer systems. As mark-sense recognition and barcode label recognition are out of the scope of this study, they are therefore, only briefly discussed.

### 3.2.1 Mark-sense Recognition

This is the sensing by machine of data manually recorded in a fixed location with an electrographic pencil on a nonconductive surface such as paper (Young 1983, 141). It relies on especially designed input documents with demarcated spaces or boxes to be filled in. The completed documents are scanned by a machine, which determines the meaning of individual marks by their exact positions. The device used for this purpose is known as **an optical mark reader (OMR).**

Its use in library applications is very limited. OMR is used in surveys and multiple choice examinations, time-sheets and order forms; and it works well for standard applications where

selection from a few alternatives is possible (Rowley 1993, 13).

### 3.2.2 Barcode Label Recognition

This technique is used to read a specially designed barcode labels consisting of closely spaced lines of varying width representing binary digits (Young 1983, 20). A barcode is a pattern of thick and thin bars divided by thick and thin spaces, each bar code representing a number. In addition to the barcode itself, the label may contain a barcode number in human-readable form and other printed information. The device used to read barcodes is known as **a barcode reader**. For library applications, a barcode label recognition technique is used mainly in automated circulation system to read the ISBN or accession number of documents and a borrower's card.

### 3.2.3 Document Scanning

A document scanning technique makes electronic versions or images of a document for storage or processing purposes. As noted by Schein (1989), it converts images of text or graphics occurring on paper into digital form and transfers the digital images into a computer; and it produces a pixel-by-pixel representation of images, which is transferred to the computer's memory in the form of a bitmap, that is, addressable locations or bits that correspond to the screen coordinates defining the images.

This process of conversion is properly termed **document digitization** and the resulting electronic images are described as **digitized images**. The devices used for this purpose is know as **document scanners** or simply **scanners**. A scanner is nothing more than a device with a

sensor that detects the amount of light reflected or transmitted by a given point on the document; and all scanners use a light source, some means of moving the sensor (or mirror that reflects light to the sensor) over the surface of the document (or vice versa), and circuitry to convert the captured information to digital form (Busch 1991, 36).

#### 3.2.3.1 Types of Scanners

There are several models of scanners available depending on their geometrical orientation, scanning process, or configurations. These are (ECA/PADIS 1994, 27): floor-standing, desktop, and hand-held. There exist innumerable products of each category on the market. The floor-standing model are big in size and meant for high-volume applications and equipped with high speed automatic page feeders.

##### 3.2.3.1.1 Drum Scanners

These are higher priced, very high resolution colour separation scanners found in the graphic arts industry. With these, artwork is wrapped around a drum and rotated at very high speeds. Laser light is usually used to illuminate very tiny sections of the original. These scanners can provide highly detailed image files that can be used for sophisticated layout and page composition, electronic retouching, and colour separating (Busch 1991, 40).

##### 3.2.3.1.2 Desktop Scanners

They are meant mainly for microcomputers. They can be used for various purposes, for example desktop publishing, photocopying, database creation, image processing, etc. They can be categorized by their mode of operations as: flatbed and sheetfed varieties.

**Flatbed Scanners**: They feature a flat surface on which individual pages are positioned. These kind of scanners scan face-down, many sizes and various types of documents such as, books, magazines, a maximum-of-A4-size paper, and can be equipped with automatic document feeders (ADF) (ECA/PADIS 1994, 27). They are well suited to library applications. Diehl and Eglowstein (1991) have pointed out that a flatbed scanner with an automatic document feed (ADF) is highly recommended for serious work for libraries and office applications.

**Sheetfed Scanners:** These are sometimes described as "pass-through" or "pull-through" scanners. In this case source documents, inserted into a narrow opening, are transported by rollers across a stationery optical head assembly and light source. This kind of scanners can easily attach sheet feeders for unattended scans, handle double-sided documents in one pass, accept only single sheets and not suitable to scan books and magazines; they may affect originals; and their maximum size is legal size (8.5" X 14") (ECA/PADIS 1994, 27).

Like other hardware, the price of scanners is decreasing. The price of **Color OneScanner 1200/30**, a flatbed with enhanced image-capture capabilities is US$800; and it features 30-bit color depth and 600-by-1200 dpi optical resolution; integrates scanning, image editing, OCR, Fax, copying, and archival functions, etc. (Beale 1996). The prices of the two UMAX Technologies' color flatbed scanners, **Vista-S6E and Vista-S12**, are US$399 and US$750, respectively (Kahney 1996). **MacIRISPen** is a handheld scanner with price tag US$229 (Busch 1996).

#### 3.2.3.1.3 Hand-held Scanner

This must be manually moved, slowly and in a straight line, across a page that is positioned face-up on a desk or other flat surface. They view a document while scanning, are compact, and affordable. But only manual scanning is possible, and tedious for capturing large quantities of text. Their maximum document width is only A4 size (ECA/PADIS 1994, 27).

#### 3.2.3.1.4 Video Scanners

These make it easy to grab good quality video images in file formats compatible with RSG, QuarkXPress and PageMaker (Busch 1991, 41). They differ from other types in several ways. One key difference is that video input often consists of **a stream of still images captured as 30 different frames each second**. Another difference is that the image capturing step is distinctly separate from imaging process itself. Video imaging is done with a camera. The camera is much like photographic camera in that a lens is used to focus an image onto a light-sensitive surface, as Busch (1991: 41) notes in the book entitled The Complete Scanner Handbook for Desktop Publishing. The advantages and disadvantages of different types of scanners have been given in detail.

### 3.2.3.2 Main Features of Scanners

There are a number of different scanners being manufactured by various companies having plenty of distinguishing features and capabilities. In order to select an appropriate scanner, knowledge about their features is very essential. Some of the features are given below. These are (ECA/PADIS 1994, 27):

**Colour handling** - monochrome gray scale and colour;

**Speed** - This denotes speed of scanning characters per second (cps) or pages per minute. It ranges from 200 to 12000 cps for character recognition, and 1000 pages per minute for image scanners.

**Resolution** - A resolution of a scanner is defined by the number of pixels per horizontal and vertical inch that it employs. The different resolutions include 600 X 600, 800 X 800, 2000 X 2000 dots per inch (dpi). A resolution of at least 300 dpi is almost a standard among all non-video scanners.

**Size of document** - This denotes the size of document handled by the scanner. It is possible to scan a part or whole page of a document, to scan different sizes such as, A4, legal type, etc.

**Compatibility** - It refers to the compatibility of a scanner with different platforms such as PCs, MACs, UNIX machines, etc.

**Number of passes** - This denotes the number of passes in case of colour scanners - one pass or three passes for each of the basic colours such as red, blue, green.

**Scanning speed regulation** - This indicates whether a scanner has the mechanism of regulating scanning speed.

**Media handled** - These are papers, coloured paper, transparency, slide, etc.

**Built-in storage** - Hard disk, 3.5" floppy, 5.25" floppy.

Finally, the availability of **image compression software** is also essential.

### 3.2.4 Optical Character Recognition (OCR)

OCR is the detection, identification, and acceptance by a machine of printed characters using light-sensitive devices (Young 1983, 158). Document scanning is an essential preliminary to

OCR; the term OCR and scanning are often used interchangeably, but that usage is imprecise and misleading (Saffady 1994, 27). The former is one work step in OCR system.

OCR is the process of making computer processors understand or recognize digitized images. It uses **document scanning and image analysis to identify characters contained in documents**. Using a special program (known as OCR software), which usually comes with the scanner, makes it possible to convert digitized images to a data file that can be edited in a word processing program.

The OCR program, which may operate within the scanner itself or on a computer to which the scanner is attached, analyses the digitized images and attempts to identify the characters they contain. Most of OCR software can generate output directly into word processor formats, often preserving attributes such as font changes and boldfacing.

#### 3.2.4.1 Types of OCR Packages

There are three types of OCR software packages based on the recognition techniques they apply. These are trainable or matrix matching, Omnifont, and intelligent OCR software packages (Diehl and Eglowstein 1991).

##### 3.2.4.1.1 Trainable OCR Software

The earliest OCR packages were strictly trainable. Before they can recognize texts, they must learn each character of a particular font. User has to scan a document, preview the scanned bit map of each character, and identify the character for the program. The program builds a

database that assigns each image to its corresponding ASCII character. On the next pass, the program compares the scanned image of every character with the stored images in its database. If the program finds a reasonable match, it returns the ASCII character assigned to the matched image.

This is a less expensive recognition technique (Saffady 1994, 27). However, it can be very tedious and time-consuming. It also requires much more disk space to hold the font dictionaries that are created.

#### 3.2.4.1.2 Omnifont OCR Software

To overcome limitations of the trainable OCR the omnifont technology has come into the picture. Unlike the former, the omnifont OCR packages use **feature extraction** to recognize fonts regardless of their size. An omnifont product contains a database of shapes for example, lines and circles. It can recognize a letter by its unique combination of shapes. Such OCR products do not impose restrictions on type styles. Theoretically, an omnifont package can understand any character without training (Diehl and Eglowstein 1991).

However, omnifont is a generic term. As noted by Diehl and Eglowstein (1991), different vendors use different omnifont algorithms, and the effectiveness of these algorithms can vary widely. Some algorithms do an outstanding job of feature extraction while others are barely functional. And even the best omnifont algorithms are not perfect. Therefore, to enhance recognition, many OCR packages use omnifont algorithms and a learning facility.

#### 3.2.4.1.3 Intelligent OCR Software

These are more advanced OCR packages that use lexical context to improve accuracy of recognition. That means as part of the recognition process, the software compares its best guess to a stored dictionary; and employs contextual clues, spelling dictionaries, and other tools as supplement to feature extraction (Diehl and Eglowstein 1991).

Nowadays, the top packages have added spelling checkers to the recognition software to catch the most common errors. A few even use lexical word analysis to make sure the spelling follow common rules of the language (Eglowstein 1994).

### 3.2.4.2 Features of OCR Software

As there are innumerable products having several essential features, it is worth noting some of these features. These include (Diehl and Eglowstein 1991):

#### 3.2.4.2.1 Deferred Processing

Instead of scanning and recognizing one document after another, you can scan a batch of documents, queue them up, and instruct the software to recognize them all at a later time. Some packages can process multiple jobs with an ADF (automatic document feeder) attached to the scanners, and some let you demark jobs with a blank page between each job in the ADF. When the program encounters the blank, it creates a new text file and loads subsequent pages to it. This can remove much of the tedious work.

#### 3.2.4.2.2 Templates

These are means of simplifying the OCR process by specifying particular areas or zones to be recognized. A user can use templates for every page that follows the same structure. Some OCR packages can automatically create templates and even recognize which zones contain text and which contain graphics. You can also limit zones to the type of entry you desire i.e., numeric, alphanumeric, or user defined. This is particularly useful if you are processing spreadsheet or other documents that contain only numerical entries.

#### 3.2.4.2.3 Spell Checking

In the context of OCR, spell checking enables the software to use a dictionary during the recognition process. With a spell checker, the OCR software will first consult the dictionary to see if either of the spellings constitute a real word. With the dictionary as tiebreaker, conversion is usually more accurate.

#### 3.2.4.2.4 Other OCR Software Features

Moreover, there are other features that should be taken into account when considering purchase or selection of an OCR software package for a particular application. Diehl and Eglowstein (1991) have listed these features that include trainability, lexical recognition, largest and smallest readable point size, graphic preview, landscape, foreign language dictionaries, multiple jobs with automatic data feeder, automatic parsing of text and graphics, reorder text blocks, decolumnizing text, numeric recognition, learning of ligatures, dot-matrix support, support for fax files, retention of format (type style, indents, justification, centring, columns, tables), output file formats, and proofing tools such as built-in editor, search and replace, query on questionable

characters.

### 3.2.4.3 Criteria for Selecting OCR Software

When searching for letter-perfect OCR software, one has to keep an eye out for implementation quality in the following areas (Grunin 1996):

#### 3.2.4.3.1 Accuracy

This is the single most important characteristic to look for any OCR package; unfortunately, it's the hardest to judge without hands-on use. Most of the recent innovations in OCR software involve increased accuracy, particularly when reading poor quality documents; and prices for OCR software have come down.

Judging a package's accuracy across a variety of document types requires large-scale, statistically valid testing. Grunin (1996) has noted that there are organizations that routinely perform tests and post results on the World Wide Web (WWW). Accuracy of a package may differ based on the types of documents.

In the past, users had to spend a lot of time correcting mistakes in OCR documents, but the translation has improved significantly. Even with the latest programs, however, you can still expect to spend some time making corrections, although this varies with the program and the type of document you are converting (Kawamoto 1996).

As an example, the accuracy of four major OCR packages (OmniPage Professional 5.0,

WordScan Plus 3.0, Recognita Plus 2.0, and TextBridge 2.0) based on different kinds of documents is given below.

**Table 3.1** Accuracy of four leading OCR software packages.

| Outputs | OmniPage | Recognita | TextBridge | WordScan |
|---|---|---|---|---|
| Daisy-wheel text | 99.3% | 97.0% | 99.3% | 98.9% |
| Ink-jet text | 99.3% | 96.2% | 98.7% | 99.0% |
| Tiny text | 88.6% | 97.2% | 95.6% | 97.3% |
| Times text | 99.2% | 99.1% | 98.9% | 98.4% |
| Copied text | 88.9% | 92.3% | 96.4% | 95.3% |
| Fax | 98.8% | 87.4% | 78.1% | 98.0% |

**SOURCE:** Eglowstein (1994)

Accuracy is measured by counting the number of words that scanned correctly and dividing by the total number of words in the document.

#### 3.2.4.3.2 Integration

If OCR is needed only occasionally, or generally a single document is processed at a time, the ability to launch recognition from within the target application (or drag-and-drop operation) will save a lot of time. For example, as revealed by Grunin (1996), **TextBridge Pro 96** integrates better with other applications than any other OCR products, but it tends to be confusing when operated on its own.

#### 3.2.4.3.3 Batch Processing

If one prefers to process several documents at once, one will want a product with flexible batch processing. This requires more than simply recognizing a list of files or scanning multiple

documents. It means the product must be able to set individual parameters for each document and schedule jobs.

#### 3.2.4.3.4 Input and Output Formats

Like any conversion product, OCR is useful only if it can read and write the formats one needs. OCR packages should ideally support .DCX, multipage TIFF, .BMP, and .PCX for input, and a broad variety of word processing and spreadsheet formats for output. If format retention is needed, the program should intelligently translate input into the output application's native file type.

#### 3.2.4.3.5 Throughput

The throughput of an OCR package determines its performance. Eglowstein (1994) has measured throughput of the aforementioned OCR packages by counting the number of words the software scanned correctly (total minus errors) and dividing by the time it took to scan the document. An accurate package that runs slowly can have a better throughput measurement than a fast one that makes lots of errors (Eglowstein 1994). In general, accuracy is more important than speed.

### 3.2.4.3 OCR Software Products

There are a number of OCR software products on the market. Some of top-rated OCR software products have been reviewed by Eglowstein (1994) in Byte based on different features. These products are Caere's OmniPage Professional 5.0, Calera's WordScan Plus 3.0, Recongita's Recognita Plus 2.0, and Xerox Imaging System's TextBridge 2.0. They have been reviewed in

terms of the above criteria and other features they incorporate.

The latest versions of these leading and rival OCR software packages have been also reviewed in length in terms of accuracy, speed, interface, price, and other essential features by Gann (1996), Kawamoto (1996), EDGE: Work-Group Computing Report (1996), Mendelson, O'Malley (1996), Grunin (1996), Edward (1996), to mention a few.

#### 3.2.4.3.1 OmniPage Professional

OmniPage Professional 5.0 is available for both Windows and Mac machines. As Eglowstein (1994) has reviewed, OmniPage Professional 5.0 (Windows-based version) has got tremendous improvement in accuracy (see **Table 3.1**) and new features. Except for its performance on bad photocopies, it held its own admirably against the competition. It also offers a number of new features. It has got a one-touch automatic OCR conversion function (**Auto OCR**). It has also got a format retention (True Page) function to preserve the document's original format and reproduce it in the final output with all the text and graphics positioned exactly as they were in the original.

The **AnyFax function** of the package increases recognition by employing image enhancement on characters it perceives as broken, joined, or jagged. It also attempts, by reegineering the fax image's CCITT code, to reconstruct missing lines in faxes that have suffered from noise on the phone lines (Eglowstein 1994).

Besides, the **3D OCR** takes advantage of a scanner's gray-scale capability and uses a learning

facility for more accurate recognition. Since this analyses the depth of gray in each character's pixels, 3D OCR technology increases OmniPage's chances of recognizing faded or broken characters.

The **Language Analyst** compares the text to lists of common three-letter sequences and word groupings to determine a likely match. It also checks for common OCR errors and attempts to correct them.

However, all this slows down the recognition process but seems to greatly improve OmniPage's accuracy (Eglowstein 1994).

The most recent version of this package is **OmniPage Pro version 7.0** which is meant for Windows 95. It offers a refined user interface, ease-of-use enhancements, improved integration with word processors, and new tools for improved control over the formatting of output text (EDGE: Work-Group Computing Report 1996).

As of Kawamoto (1996), an intuitive and familiar interface makes this version the easiest OCR program to use, and offers options for scanning, importing, building zones, converting, proofing, and exporting. It supports many file formats; and its proofer is the best of the OCR bunch.

#### 3.2.4.3.2 Wordscan Plus

Calera's WordScan Plus 3.0 is also a Windows-based package. It uses neural-network and

image enhancement technology for improved dirty-document support. It can also retain page formatting and offers a one-touch OCR function being one of the top performers in terms of accuracy (Eglowstein 1994).

Moreover, it includes support for scanning stacks of two-sided documents, provides automatic deskewing of images that may be tilted on the copy glass, has excellent document-template support, and has an **OCR Aware function** to start up WordScan Plus 3.0 from within other applications.

Templates in WordScan Plus 3.0 let the user easily define regions on a page where text is likely to be, store these region definitions, and reuse them for every page in a document. The most recent version of WordScan Plus is **WordScan Plus 4.0**. Though it is a windows-based package its Windows 95 version has not been released yet.

#### 3.2.4.3.3 Recognita Plus

Recognita Plus 2.0 is available for CTOS, DOS, Windows, and OS/2 operating systems. It lacks the overall accuracy and throughput that other packages have. Its particular strength is its superb language support. The package also offers the ability to start an OCR process from within other applications, and has better handling of inconsistent spacing (Eglowstein 1994).

Its recent version is **Recognita Plus 3.0**. It is a poor performer as of the tests done by Kawamoto (1996) because of its confusing interface, low accuracy when deciphering poor quality documents and others. In terms of accuracy, Recognita Plus scans and converts

documents slightly slower than OmniPage Pro but faster than WordScan. The program also does an excellent job straightening our tilted fax. In terms of price, the product is less expensive than all but TextBridge Pro.

#### 3.2.4.3.4 TextBridge

TextBridge 2.0 is available for Mac, Power Mac, and Windows. It is the cheapest OCR package, as of Eglowstein (1994). TextBridge's recognition is fast - often faster than that of the more expensive products; but the product falls short on accuracy. According to testing done by Eglowstein (1994), it did well on clean daisy-wheel output and adequately with the clean ink-jet and Times Roman text, but poorly with faxed documents.

Its latest version, **TextBridge Pro 96**, runs on both Windows 95 and Windows NT 4.0 platforms. TextBridge Pro 96 OCR software, like its rivals, is very accurate, but suffers from an unintuitive interface (Gann 1996). TextBridge's proofer can only be used from within Word or WordPerfect.

One issue that definitely weighs in TextBridge's favour is its price (Kawamoto 1996). It will accurately scan tables and graphics (Gann 1996). Moreover, TextBride Pro 96 can recognize tightly spaced and degraded text that OmniPage cannot decipher, and it can generate true word processor tables. Using the two software packages you can save files in all standard processing formats, as well as in HTML (Mendelson 1996).

However, it can only handle as many as 50 pages, so it's not a good choice for big jobs. Still,

for recognizing most documents and forms, such as letters, memos, and newspaper or magazine articles, TextBridge Pro yielded accurate results and was easy to use (Heck 1996).

#### 3.2.4.3.4 Merits and Demerits of the Software

The summary of some of the essential features of recent versions of the leading OCR software packages such as OmniPage Pro for Windows 95 7.0, Recognita Plus 3.0, TextBride Pro 96, and WordScan Plus 4.0 is given in table 3.2. OmniPage Pro is the only OCR package that comes with a wizard to take you through the OCR process. TextBridge offers a higher level of accuracy, while OmniPage offers a better interface and better cooperation with Windows 95. OmniPage Pro however, frequently failed to detect the space between two words correctly (Mendelson 1996).

If most of your work involves importing data in tables, you will want to opt for a package that is good at this, such as WordScan Plus. However, if you want to preserve the layouts, then OmniPage Pro is the package of choice. If cost is an issue, TextBridge is the least expensive product of the group. Based on the testing done by Kawamoto (1996), Recognita Plus 3.0 just did not compare with the other products in terms of accuracy when reading poor-quality documents, and the interface was difficult to navigate and use.

TextBridge needs an interface as powerful as its recognition engine, but if one values accuracy over convenience, it is the OCR package to choose as suggested by Mendelson (1996).

Suspect words are flagged in Recognita Plus, giving users the opportunity to perform

corrections (Kawamoto, 1996). With TextBridge it is possible to add words to its internal dictionary while in the process of editing text files; and its proofreader will check converted documents from within the word processors such as Microsoft Word or Corel's Wordperfect (Kawamoto, 1996).

But WordScan Plus may lack some of the flush of the other products in this roundup, it comes with an adequate proofing editor that identifies words it does not offer suggestions for replacement (Kawamoto, 1996).

In general, Kawamoto (1996) has ascertained that OmniPage Pro and TextBridge Pro are the two leading OCR software packages in the market having good performance and still the former is better in few features. Heck (1996) and O'Malley (1996) have reviewed these leading OCR software packages. Pricewise, TextBridge (**US$260**) is the least expensive. The list prices of OminPage, and Recognita Plus are **US$499**, and **US$395**, respectively; and their upgrades price **US$129** (O'Malley 1996).

**Table 3.2** Summary of merits and demerits of the major OCR software packages.

| OCR Packages | Merits | Demerits |
|---|---|---|
| **OmniPage Pro 7.0** | Flexible interface; OCR wizard simplicity; scheduling options; good batch-processing options; recognizes standard image formats; works from within most Windows 95 applications; HTML support. | Carries a high price tag; no graphics via direct input; no enough background processing. |
| **TextBridge Pro 96** | Highly accurate; excels at reading complex documents; requires little user intervention; works within most Windows 95 applications; easy to choose areas you want to scan; instant access supports graphics; good integration with word processing software; captures original format without using frames; HTML support. | Cryptic toolbar; doesn't flag unrecognized text; could use more background processing and better batch processing; not clear markings for some interface elements; no applicable for high-volume OCR. |
| **Recognita Plus 3.0** | Supports many languages; suspect words are flagged giving the opportunity to correct; less expensive than all but TextBridge; outputs to all popular word processors & spreadsheet, and converts texts to HTML. | Confusing interface and low accuracy when deciphering poor quality documents. |
| **WordScan Plus 4.0** | It comes with an adequate proofing editor that identifies words it does not recognize; straightforward to learn and use; good at handling spreadsheet files and tables; good at retaining fonts | The slowest of the bunch; does not offer suggestions for replacement; |

## 3.3 APPLICATION OF OCR FOR RECON

The literature on applications of OCR for RECON is quite scanty from sources of citations in library and information science, but reveals the potential of this option for library automation projects, especially when in-house conversion is required. Scanning and OCR technology has matured sufficiently to arouse renewed interest in the area of applications of it for RECON (Weibel et al 1989).

Since the late 1980s, the availability of relatively inexpensive document scanners and microcomputer-based OCR programs has made OCR an attractively priced data entry option (Saffady 1994, 27). Its potential for library applications is obvious: it is a faster, more automated, and presumably less expensive alternative to the keyboard-oriented input devices.

From the early 1980s librarians started considering OCR as a new alternative for RECON, though it had not been successfully implemented due to some limitations of the technology. Attempts that have been made with this respect are briefly mentioned below.

In the late sixties the Library of Congress attempted this option to create a machine-readable records of its collection. This was terminated just after 1970. Hein (1986) reveals that there were two major parts to the project- one is optical character recognition of printed bibliographic information and one consisting of a software package for the application of MARC tagging to the strings of characters of bibliographic information i.e., turning them into full MARC records. There was reasonable success within certain limitations - for the analysing software package,

but the multi-font OCR reader to be developed by the Farrington Company failed to emerge, thus causing the termination of the project (Hein 1986).

The British Library has worked with its GK conversion (conversion of the general catalogue up to 1975). In this conversion a CIM scanning device (Computer Input Microfilm) was used, although this is now almost outdated technology. This project contains, however, only a few elements of computer controlled formatting (Hein 1986).

Another attempt in this area is the British company Libpac/Optiram claims to have a complete solution for any kind of scanning and subsequent automatic formatting using a class facsimile device and an integrated software package for interpreting the patterns read and for the formatting (Harrison 1985).

Although they might be successful, Libpac is still not offering what it was intended to do as a-large-scale commercial facility, though up till mid 1986 they have achieved the most in this respect (Hein 1986). One of the greatest assets of the Optiram system is its 'intelligence'. It uses software that learns from the past history of all previous documents processed (Harrison 1985).

Standard scanning techniques using OCR have been in use for many years. However, they were limited in scope by requiring that all the data read is in a printed form using an OCR typeface (Harrison 1985). Now as has been mentioned before, an application of sophisticated computer technology has enabled data to be read and correctly interpreted, whatever the typeface, with error rates that are typically insignificant.

A number of companies in the past three to four years have offered low-cost scanning devices in the office automation environment. In this respect Hein (1986) has stated some obstacles to optical scanning for RECON. These are:

(1) There must be a lack of interest in the dominant circles of the database business. This could be caused by the situation already mentioned that the big money is in brand new information, which will be created in machine-readable form using other techniques, such as online data entry.

(2) From the Library of Congress problems it could be observed that the analysis software made better progress than the scanning system. The present state-of-the-art is that the scanning no longer contains unsolvable problems, but the art of turning pure strings of characters into understandable information is still on a low level.

(3) The complexity of an actual application concept will be dependent on the computer system to handle the result. Old information considered desirable to store and retrieve in electronic systems could have several formats. In fact, one might wish to follow the comprehensive formats definitions used by the LC project.

Rice (1981), in an article in Library Journal, noted a breakthrough in omnifont OCR that would permit economical automation of library catalogues and other processes. Hein (1986) also supports Rice's idea by estimating the cost range of OCR for RECON to be between 10 and 25 percent of the costs for conventional conversions. Though prices will differ according to actual

technologies used.

In 1989, Grotophorst has discussed the use of OCR technology to produce a bibliographic database of dissertations at the George Mason University. A tutorial on digital scanning and OCR is also provided.

Besides, a few writers have pointed out the potential of OCR technology for retrospective conversion. It has become an emerging RECON option, as noted by Drabenstott (1986), Jacobs (1990), and Haddad (1990) who have revealed the potential of OCR as a possible new technology for performing retrospective conversions from printed cards.

However, they have not given detailed methods for format recognition of scanned documents except the few who attempted and did not achieve hundred percent success. To this end, Weibel, et al (1989), in their article in the Information Processing & Management, point out the advancements of scanning and OCR technology to matured and sufficient system. They have reviewed attempts made in automatic format recognition, especially in line with automatic cataloguing of title page; and developed a prototype of a rule-based system to explore the impediments to automating from title pages. Their study has attempted to identify the various bibliographic elements on the title page; and was able to capture a substantial part of the regularity in the title page layout in a small set of rules, i.e., 80 percent of the bibliographic fields present on a random sample of title pages. It may help as a complementary to this study that attempts to extract the various bibliographic elements on card catalogues.

Molto and Svenonius (1991), also have addressed the feasibility of developing automatic name recognition algorithms to distinguish character strings representing names from other character strings occurring on English language title pages. They have developed two algorithms, one for recognizing personal names and the other for corporate names. The algorithms involved matching title page names with names in authority files and identifying post name markers. The success rates for the corporate and personal name algorithms were 85.8 and 84.5 percent, respectively.

More recently, a brief description of work done by the RIDDLE (Rapid Information Display and Dissemination in a Library Environment) project has been given by Harrison, et al (1995). This project has used optical scanning and OCR to capture bibliographic information from journal content pages for inclusion in an online library catalogue. The system is broken down into several stages: the initial scanning of a document, the capture of the text, the identification of the parts of the text which are relevant, and the generation of commands suitable for loading the online library catalogue. The useful distinguishing features proved are journal title, ISSN number, bar-code, logo and layout by searching them in surrogate information. Similarly, this complements this study as well.

As long as text extraction is concerned there are methods commonly used. The technique of template mining, which is used in the natural language processing, can be used to extract data directly from text if either the data and/or the text surrounding the data form recognisable patterns (Lawson et al 1996). In this case, when texts match a template, then the system extracts data according to instructions associated with that template. They have pointed out that data

mining is another technique that had long been dismissed due to its narrow domain specificity that made it seen as an inelegant non-linguistic approach.

Another technique is the use of layout and content of the information (Harrison 1985). The most widely applicable for bibliographic information is the use of positional field recognition and typographic field recognition. In this respect, once various major fields have been clearly identified by positional analysis the data strings in between fall into the more limited categories of the known permutations that are likely to occur. These strings may then be divided up by their punctuation, their mix of key elements, numerics, use of brackets or parenthesis etc.

# CHAPTER 4

# ADDIS ABABA UNIVERSITY LIBRARY SYSTEM

## 4.1 INTRODUCTION

This chapter gives background information about the university library system with a view to highlighting its organizational and functional structure, collections, cataloguing practices being followed in organizing the collections, and identifying activities currently going on, future plans, and bottlenecks regarding automation of library functions.

As mentioned in Chapter 1 section 1.4.5, most of the information has been acquired through discussions with the staff of the Cataloguing Department of the library system, and with the head of the Computer Center; by reviewing annual reports and a study carried out by the university library committee in June 1996; and by taking a close look at the union catalogue and shelf list of the library system.

## 4.2 BACKGROUND

Addis Ababa University used to be a college, known as University College of Addis Ababa, that was established in 1950 at the present main (Sidist Kilo) campus with the aim of enhancing the socio-economic development of the nation by producing trained manpower, carrying out researches on various areas, and disseminating the results to the public at large.

In 1961, 11 years later, it was grown to a full fledged university and changed its name to Addis Ababa University (AAU) having different faculties and colleges (Wedgeworth 1993, 287). These are the Faculty of Medicine, Science Faculty, Technology Faculty, Building College, Faculty of Law, Faculty of Business & Economics (FBE), College of Social Sciences (main campus), and Faculty of Veterinary (about 45 kilo metres from Addis).

It is a known fact that any academic institution, whether big or small, cannot properly achieve its objectives without having library/information services. To this effect, Addis Ababa University Library System (AAULS) was founded in the same year with the university by building on the collections of the library of the then University College.

AAULS has got several branch libraries in the faculties and colleges of the university which were established around the John F. Kennedy Memorial Library (or Main Library) mainly to serve the students, academic and administrative staff, and researchers. The branch libraries include Science Library, Technology North Library for Technology Faculty, Technology South Library (Building College), Law Library, Central Medical Library, and FBE Library - all situated in Addis Ababa.

The Main Library is the central library for the whole library system; and the branch libraries depend upon it for many activities such as technical processing that includes acquisitions, cataloguing, binding, and some aspects of administrations (see **Figure 4.1**). All the branch libraries report on their activities to the university librarian, likewise the librarian reports to the university academic vice president.

Moreover, the Main Library has got its own departments including Cataloguing, Acquisitions, Binding, Media Production, Circulation, Ethiopiana Collections, Documents, Reference, and Periodicals Department. Except the last, all are located in the main library building, which was purpose-built.

User or reader services, mainly circulation and reference, are decentralized by the branch libraries while cataloguing, acquisitions, and binding are done centrally.

### 4.2.1 Staff of the AAULS

Like most libraries, the staff of AAULS is composed of employees with various levels of education and responsibility. Broadly speaking, the staff of the system are categorized as professionals, para-professionals, and support staff. This classification is mainly based on the qualifications of the staff.

Professional staff are employees having graduate degree (master's degree), and above in library and information science and other related fields. Most of them, basically work as heads of the branch libraries or departments.

To achieve its objectives the AAULS is organized as under.

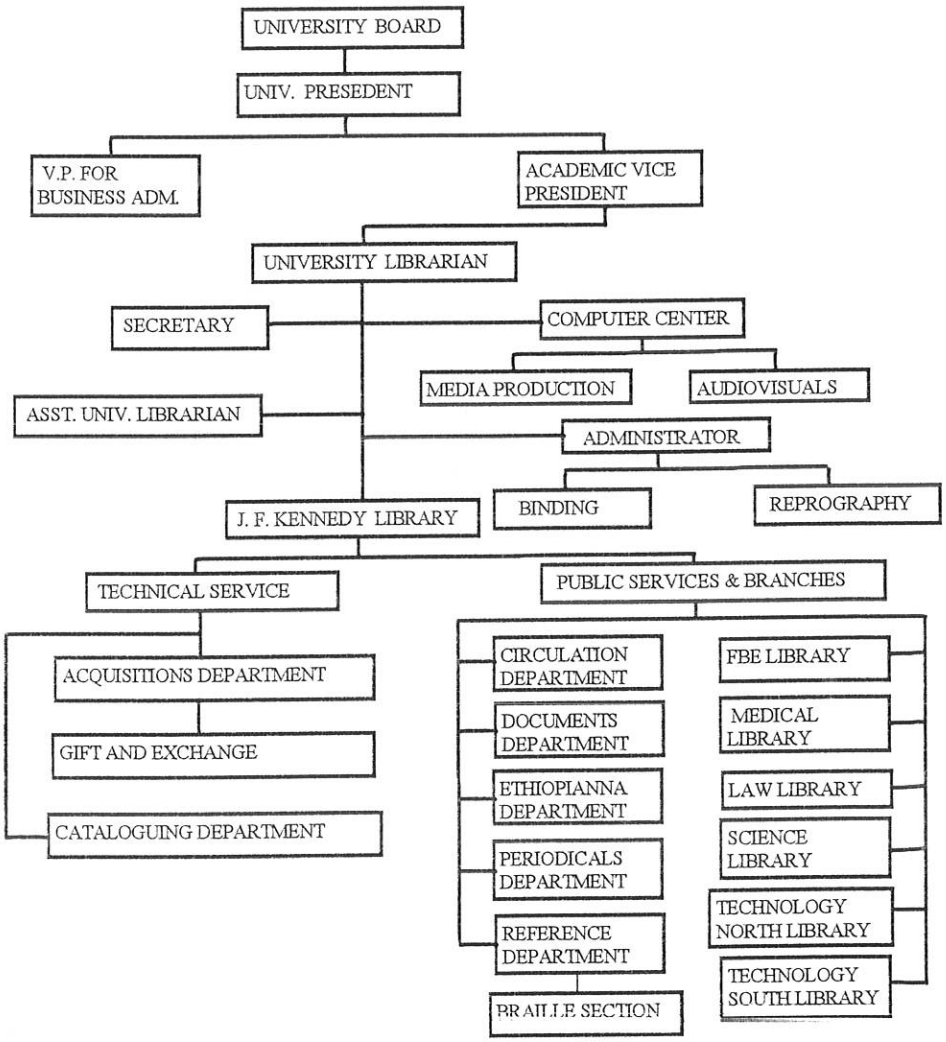


**Figure 4.1** Organizational Chart of the Library System

SOURCE: Addis Ababa University Libraries, Computer Center. 1997. Proposal for networking of the AAU Libraries for information resource sharing. Unpublished material.

Para-professional staff are employees having primary degree in library science or in related fields. Both the professional and para-professional staff are employed at the rank of library academic staff.

Support staff include the majority of the employees of the library system. Some of them are only having high school certificate and/or diploma in library science and other fields. The routine operations almost in all branches are handled by the support staff.

Besides based on the type of employment the staff of the library system are also categorized into three: library academic staff including both professionals and para-professionals who are hired on equal ranks with other academic staff of the university, administrative staff who are hired according to the rules and regulations of the Federal Civil Service Commission of the country, and contract staff who are hired on contract basis.

Totally, the AAULS has 273 staff including professional, para-professional, and support staff working in different branch libraries. The number of professional and para-professional staff, who have first degree and above, is only 27 having equal rank to teaching staff of the university. This indicates that a large proportion of the staff is on the lower end of career structure of the university.

**Table 4.1** Number of the library staff by category

| Types of staff | Number |
|---|---|
| Academic | 27 |
| Administrative | 188* |
| Contract | 58 |
| **Total** | **273** |

**NB *24** are diploma holders in **library science.**

### 4.2.2 Holdings

The total collections of the library system according to the survey carried out by the Addis Ababa University Library Committee (1996) is given in the following table.

**Table 4.2** Collections of the libraries of the university

| LIBRARY | MONOGRAPHS (in volume) | BOUND PERIODICALS (in volume) |
|---|---|---|
| KENNEDY /MAIN | 263,335 | 1,629 |
| LAW | 22,473 | 5,801 |
| FBE | 21,170 | 3,500 |
| MEDICAL | 15,792 | 11,599 |
| SCIENCE | 56,981 | 2,019 |
| TECHNOLOGY NORTH | 15,972 | 5,083 |
| TECHNOLOGY SOUTH | 8,700 | 200 |
| **TOTAL** | **404,423** | **29,831** |

**SOURCE:** Report of the committee for the study of AAU libraries, June, 1996.

The collections of AAULS were extensively developed with the assistance of the Ford Foundation and the U.S. AID (Wedgeworth 1993, 287). About 500,000 volumes of monographs and bound periodicals in the early 1990s, they include a unique library of the Institute of Ethiopian Studies (IES), where the holdings of more than 25,000 volumes are made

up of a comprehensive collection of books about Ethiopia and the Horn of Africa; 10,000 books in Amharic, printed in the country; and a collection of about 1,000 manuscripts and scrolls (Wedgeworth 1993, 287). As the Institute has become an autonomous institution, the library is no more part of the university library system.

### 4.2.3 User Population

The library system offers services to the university staff, both academic and administrative, students including graduates, undergraduates, and extension (night), and to some external readers on special cases. The number of potential and actual users is given in Table 4.3.

**Table 4.3** Number of potential and registered or actual users of the AAULS.

| COLLEGE/ FACULTY | POTENTIAL | | | ACTUAL/REGISTERED | | |
|---|---|---|---|---|---|---|
| | STAFF | STUDENT | TOTAL | STAFF | STUDENT | TOTAL |
| C.S.S | 1,906 | 4,438 | 6,344 | 1,610 | 1,728 | 3,338 |
| F.B.E | 81 | 2,584 | 2,665 | 81 | 749 | 830 |
| LAW | 32 | 694 | 726 | 32 | 254 | 286 |
| MEDICINE | 227 | 598 | 825 | 187 | 598 | 785 |
| SCIENCE | 313 | 2,823 | 3,136 | 284 | 2,177 | 2,461 |
| TECHNOLOGY (N) | 118 | 2,450 | 2,568 | 112 | 850 | 962 |
| TECHNOLOGY (S) | 86 | 750 | 836 | 76 | 450 | 526 |
| TOTAL | 2,763 | 14,337 | 17,100 | 2,382 | 6,806 | 9,188 |

**SOURCE:** Compiled from the 1995/96 annual reports of branch libraries.

## 4.3 CATALOGUING DEPARTMENT

As mentioned above, the Cataloguing Department of the library system is located at the Main (Kennedy) Library having the objectives of doing original cataloguing, producing and maintaining card catalogues.

The department maintains a copy of shelf lists for all materials centrally catalogued; and a copy goes to a respective library. The public catalogue at the main library is a union catalogue, which is also maintained by the department. It also does all cataloguing for all materials acquired by the Acquisitions Department as well as received by the branch libraries through gift, donations, exchange, etc.

Broadly speaking, the objectives of library catalogues as of Matthews (1985, 5) are to: (1) enable a person to find a book about which the author, title, or subject is known; (2) show what the library has by a given author, on a given subject, or in a given kind of literature; and (3) assist in the choice of a book as to its edition, i.e., bibliographically, and/or as to its character, i.e., literary or topical. Definitely, the purposes of the library system's catalogue do not deviate from these objectives.

The staff of the department consists of the head, three chief cataloguers, four clerical staff for filling and checking, and two typists. It was learnt that there is no adequate number of staff when compared to the volume of cataloguing being done. Even though the department uses BiblioFile CD-ROM databases for cataloguing and card production, the computer and printer

are quite outdated. Original cataloguing is also done for those materials not indexed in the database. As mentioned by the chief cataloguers, **about 50 percent** or so of materials that go to the department are not found in the database being used.

The department has quite big backlog to be catalogued that will take a long time. As found out by the survey, delays in the centrally controlled cataloguing services is one of the major problems of the AAULS (Addis Ababa University 1996). In other words, cataloguing is one of the major bottlenecks in the whole library system. The study also found out that one of the causes of the problem with this respect to library holdings is **manual record handling.**

### 4.3.1 Cataloguing Practices

Cataloguing systems, whether manual or automated, encompass two interrelated activities: descriptive cataloguing and the production of library catalogues in whatever form required - book, card, computerized, or computer output microform (COM).

The purpose of the former is to produce informative bibliographic and physical descriptions of library materials in sufficient detail to permit the conclusive identification of a given item and to differentiate it from other, possibly similar items (Saffady 1994, 220). The department does descriptive cataloguing based on the detailed set of instructions embodied in the Anglo-American Cataloguing Rules, the second edition (AACR2).

### 4.3.2 Entries in the Union Catalogue

In cataloguing there are two possible approaches to the provision of multiple entries for a work. These are (Rowley 1987, 20): main and added entries, and unit entry. Reference entries are also made. The department follows the use of main and added entries, and references. A main entry is the complete catalogue record of a document. The main entry for a work is made under a personal author's name, name of corporate author, or under a title. All other entries beside the main entry are added entries.

Moreover, references are used to refer the users to another location or entry where this information can be found. A reference is not generally as helpful to the user of a catalogue or index as an entry might be since a reference provides little direct information about a document (Rowley 1987, 27). There are two types of references used: 'see' and 'see also'.

### 4.3.3 Arrangements of Entries

Generally, there are two types of arrangements of entries in any library catalogue, viz. alphabetical and classified catalogue.

In the former case, there are two possibilities of arrangements: a dictionary catalogue in which all entries may be arranged in a single alphabet of authors, titles, subjects, etc., and a divided catalogue in which the entries are divided into two parts: an author-title catalogue and a subject catalogue. The arrangement of the union catalogue of the AAULS is divided catalogue of alphabetical catalogue, that is divided into author-title, and subject catalogues.

A classed catalogue consists of three parts: the classed catalogue, the author-title catalogue, and the alphabetical index to the classification scheme.

### 4.3.4 Problems of the Department

Obviously, there are various alternatives to library catalogues, including card catalogues, book catalogue, computer output microform (COM) catalogue, and online or computerized catalogue.

The library system's card catalogues have a number of **problems associated** with their production and maintenance: filing and maintenance costs are too high and labour-intensive; costs of cabinets and related furniture are expensive; labour costs for interfiling and maintaining the card catalogue could be even more costly; adjustment of the cards in the card trays throughout the cabinet will be difficult as the card catalogue expands in size; the cost of replacing damaged or destructed cards.

Besides, the department has encountered acute shortages of manpower, space, and lack of adequate computers and their peripherals such as printers, etc. Consequently, the department has a big backlog waiting to be catalogued.

### 4.3.5 Levels of Catalogue Description

AACR2 offers a number of options for cataloguers. One of these is the option of choice among three levels of details in catalogue description (Maxwell 1980, 11).

The first level of description is brief cataloguing, including only the title proper, edition statement, material (or type of publication) specific details statement for cartographic materials and serials, name of the first publisher, date of publication, pagination (for books), notes, and standard number, if available.

The second level of description includes all of the information given in level one plus parallel titles and other title information and statement of responsibility, edition statement together with its first statement of responsibility, first place, first publisher, and date of publication, physical description area, the series statement, notes and standard numbers.

The third level includes all the rules applicable to the item being catalogued. Third-level description is appropriate to large libraries and research collections (Maxwell 1980, 11). The department follows this level of description.

## 4.4 AUTOMATION PLAN OF THE LIBRARY SYSTEM

The library system has a computer Center located at the main library, which was established in 1988. The overall objective of the section is to automate library functions of the whole system as well as to offer IT-based information storage and retrieval services.

At present, its major activities include provision of literature search services from various CD-ROM databases, and provision of electronic communications (mainly electronic mail) services; it assists the Cataloguing Department in catalogue production; and it coordinates & supervises

IT related activities of the branch libraries. It has got connectivity to the Internet very recently.

The Computer Section has got seven computers of different types, two of them are outdated and out of order. Most of them have been acquired from donors and some bought; the library has also got a pentium machine. There are 4 printers, of which 2 are non-functional. It has also a scanner and photocopiers.

The library has also been acquiring a variety of CD-ROM databases of general, subject specific, and speciality areas. These include 14 CD-ROM titles for medicine, 7 titles for business and social sciences, 10 titles for science and technology, 1 title (ERIC) for education, 4 titles for information science, and 10 titles for general reference. Except for Medical Library that offers CD-ROM search services, for the time being all of the CD-ROM search services are rendered by the Computer Section. The databases are both bibliographic and full-text. There have been multimedia reference databases integrating texts, sound, pictures, graphs, etc. at the Medical Library.

Despite the fact that the library system has long begun considering automating its library functions, it has encountered several problems to go about it. Measures towards introduction of computers were taken in all branch libraries but the Technology South Library by the Library Automation Committee of the library system. The committee has repeatedly attempted to study means of initiating library automation, with no success so far. Major problems concerning automation indicated by all libraries, according to the recent survey, were: lack of adequate budget, lack of trained manpower; lack of adequate equipment.

Currently, the overall budget of the library system is **2,324,865 Birr**. This is a decrease of almost 90,000 Birr since last year at a time of devaluation, inflation and rapid increases in book and journals prices (Addis Ababa University 1996, 19). In light of this fact, lack of finance has highly contributed to the problem of **retrospective conversion** which is quite expensive for the university library system.

In spite of the problems, as the researcher has learned from the discussions held with the head of the Computer Center and from Annual Report (1995) of the library system, the need for automation of the library functions such as circulation, cataloguing and acquisitions has been recognized. Though there is an automation plan it has been slowed down due shortage of funds. Besides, there is also a plan to establish a local area network (LAN) the main library and to later extend the network to establish links among the branch libraries of the university. The proposal for the latter is underway. This includes costs of training, equipment, consultants, software, etc. (Addis Ababa University Library System 1996).

# CHAPTER 5

# PROTOTYPE PROGRAM FOR AUTOMATIC EXTRACTION OF BIBLIOGRAPHIC ELEMENTS (PAEBE)

## 5.1 INTRODUCTION

The objective of this chapter is to delineate the procedures complied in retrospective conversion using OCR as an alternative, and to describe the algorithm developed and used for writing a prototype program called a **Program for Automatic Extraction of Bibliographic Elements (PAEBE).** As the name implies, the program extracts bibliographic elements from a machine-readable catalogue records/files output by OCR software. **Prototype** as defined by Davis (1994, 528), is a reasonably complete, working model of a system. **PAEBE** is expected to successfully extract bibliographic elements from the catalogue records scanned and converted to machine-readable text format; it also assigns a field name to each element and writes to an output file by inserting record separators (##).

## 5.2 GENERAL PROCEDURES

In the course of analysing requirements of the program (PAEBE), there have been several stages gone through or considered before writing the actual source codes. These include the initial scanning or digitizing of the sample card catalogues that are meant for prototype development, converting the scanned images to text files, and analysing the layout and format

of the texts output by OCR software so as to develop an algorithm, and finally to write source codes of the program. Each of the steps is briefly discussed.

### 5.2.1 Scanning the Sample Cards

As said in Chapter 1, there are two sets of sample catalogue cards chosen, for prototype development, and testing. The first set of sample cards, which consist of three types of main entry, were scanned using an **HP ScanJet IIcx** scanner which also supports colour scanning, though this capability was not required for the study. At this stage much attempt has been made to increase the legibility and quality of the images produced by the scanner as this, in turn, reduces the proportion of possible errors that may occur during OCR conversion. This can easily be done by using image enhancement capabilities of the scanner software. Some of these capabilities include controlling of light intensity (brightness or darkness), resolutions, and scaling of the scanner. In addition, boxing or specifying a particular area or 'zone' that is to be scanned is also found helpful for some of the cards that bear additional information which not required.

The scanner software has also been used to preview an image and improve the legibility by making use of image enhancement capabilities required before capturing the final image that goes to the OCR software. This is especially essential in case of poor quality cards by filling in broken portions of characters or by fading out hair lines in badly-set type. Figure 5.1 depicts examples of images of card catalogues produced by the scanner.

```
RD
559
.A26      Acute hand injuries : a multispecialty
          approach / edited by Francis G.Wolfort;
          foreword by Joseph E.Murray. - 1st ed. -
          Boston : Little, Brown, c1980.

             xiv, 266 p. : ill. ; 29 cm.
             Includes bibliographical references
          and index.

             1. Hand - Wounds and injuries. 2. Hand
          Surgery. I. Wolfort, Francis G.

                                        79-90896
             RT/em
             15/R/A6
```

(a) Title Main Entry

```
RA
8
.W65      World Health Organization
1992         Basic documents : including amendments
          adopted up to 31 October 1992.-- 39th ed.
          -- Geneva : WHO, 1992.
             iv, 182 p. ; 24 cm.
             Includes index
             ISBN 92-4-165039-7

             1. World Health Organization. I. Title.
MB/ag
21/7/95
```

(b) Corporate Body Main Entry Card

```
48
.A26      Abrahams, Peter Herbert.
1975b        Clinical anatomy of practical procedures /
          /by/ Peter Abrahams & Peter Webb; illustrated
          by John Hardie. - Turnbridge Wells : Pitman
          Medical, 1975.

             vii, 119 p. : ill. (some col.) ; 26 cm.
             Bibliography : p. 118-119.
             ISBN 0-272-79343-4

             1. Medicine, Clinical. 2. Anatomy,
          Human. I. Webb, Peter John, joint author.
          II. Title.

                ZK/rt                   75-596281
                4/8/82
```

(c) Author Main Entry

**Figure 5.1** Example of card catalogues output by the Scanner.

### 5.2.2 OCR Conversion

As the images had to be converted to ASCII text files to be processed by the prototype program (PAEBE), OCR conversion is one of the stages in the prototype development. **WordScan 3.0** was used to convert the images to text files. It is to be noted here that the scanning and OCR conversion are not separate tasks since the software has an **"ACQUIRE IMAGE"** option which helps to get images from a scanner. But since the software does not have much image enhancement facilities, in case of poor quality cards, **HP DeskScan** software was used as well.

After an image is properly acquired **two text zones** are manually created: **call number zone**, and **body zone** for a card because the **'Auto Zone'** option of the OCR software normally creates a number of inconsistent text zones for varying positions of the call number, and body of the card catalogue. The position of the call number is not consistent i.e., it does not fit to a specific area on the card. On some cards it is found on the top right above the body; and on others it is found adjacent to the body. In the former case it is taken by the OCR software as the first zone, and second zone in the later case. Moreover, the body is divided into different zones. Sometimes the OCR software takes both zones as one.

Therefore, the 'Auto Zone' option could not help to maintain the consistency of the layout and/or format of the output records which are required for automatic processing. Similarly, the position of cataloguer code and date of cataloguing, and accession number is also not fixed (see Figure 5.1).

To overcome this inconsistency of the output records, manual creation of text zones has been

found to be helpful in maintaining consistent layout and format of the output records to be processed by the prototype program (PAEBE). The OCR conversion was undertaken for each zone per a card according to the following order. **The call number is always the first zone, then the body is the second zone.** Since the cataloguer code and date of cataloguing are not required for automation purpose, they are excluded. Similarly, accession number that does not appear on most of the cards is also excluded by converting only the two zone, though it may be needed in bibliographic database. If it is appearing on all of the cards, third zone could have been created for the accession number. Using the manual zoning technique the following layout/format has been produced as shown by Figure 5.2.

RA
965.3
.NSS

Humerof , Rita E.
Managing stress : a guide for health professionals / Rita E. Numerof. -- Rockville. md. : Aspen systems Corp. , 1983.
xii 350 P. : ill. ; 2 cm.
Bibfiography: p. 317-333.
Includes Index,
ISBN 0-89443-939-1

1. Health facilities--A4ministration-- Psyqhological aspects. 2. Health services administrators--Job stress. .3 Medical personnel--Job stress. - I . Title.

**(a) Personal Author Main Entry**

RC
581.2
.C44

The Cell in medical science / edited by F. Beck and J. B. Lloyd. London; New York : Academic Press, 1974-1976.
4v. : ill. ; 24 cm.
Includes bibliographies and index.
ISBN O-l2-o842o4-1

1. Cytology. 2. Pathology, Cellular. I. Beck, Felix. II. Lloyd, John Benjamin.

**(b) Title Main Entry**

RE
20
.B7
1979

British Orthoptic Society.
A glossary of terms used in orthoptic practice in the United Kinidom, - 3d ed, reve - London I British Orthoptic Society, 1979.

24p. ; 26cm.

1. Ofthoptice - Terminology. I. Title.

**(c) Corporate Body Main Entry**

**Figure 5.2** Examples of records converted to text formats after text zones are created.

It is to be noted that none of these records produced by the OCR software has been changed or edited in any way.

### 5.2.3 Understanding Formats/layouts of the Records

The study of the formats and layouts of the three types of records has helped to understand or to identify key features or phrases to be used in automatic extraction of bibliographic elements. As shown above, the texts output by the OCR software are not exactly the same as the format/layout of the information on the card catalogues. This is because of the specific area or text zones selected. When the manual zoning technique is used, the call number is always put at the first three to five lines depending on the details of the number. The body appears as it is except the removal of indents and extra spaces on both left and right margins.

As far as the contents of the body is concerned, the only difference among the three types of entries is the first line of the body of the cards (see Figure 5.2). The first line contains the name of the author, and name of the corporate body, in personal author, and corporate body main entry records, respectively. In the case of title main entry the body begins with title of the document. Otherwise, the layout of other bibliographic elements remains similar, except in the corporate body main entry where there is no statement of responsibility given explicitly.

### 5.2.4 Conditions Used for the Program

Once the cards have been scanned and converted to a computer processible text (or ASCII file format), these texts or records will be used as inputs for the PAEBE. As mention is made in Chapter 4, the library's cataloguing practice follows AACR2. This format has got its own standard in making

use of punctuation marks that separate an element from others appearing on the card. These delimiters have been used as conditions in writing the program. Furthermore, some key strings such as, ISBN, Roman and Arabic numerals have been used.

The goal of this study was to automate the process of extracting bibliographic elements from card catalogue record surrogates produced by OCR so as to finally, import them to a user specified text retrieval or library management system software packages; it is necessary to investigate ways in which the process is easily automated. To this effect, the following distinguishing features used for prototype program proved useful:

- the program has used position/layout of information in OCR output records, in case of extracting call numbers;
- the program has chiefly used punctuation marks following AACR2 format to 'recognize' an element;
- certain descriptive phrases such as "cm.", "ISBN", "1.", " 2.", "3.", "4.", etc. and "I." have been also used; and
- End-of-file mark flag was also used.

## 5.3 LOGICAL STRUCTURE OF THE PROGRAM

The logical flow of the prototype program looks like the one in figure 5.3.

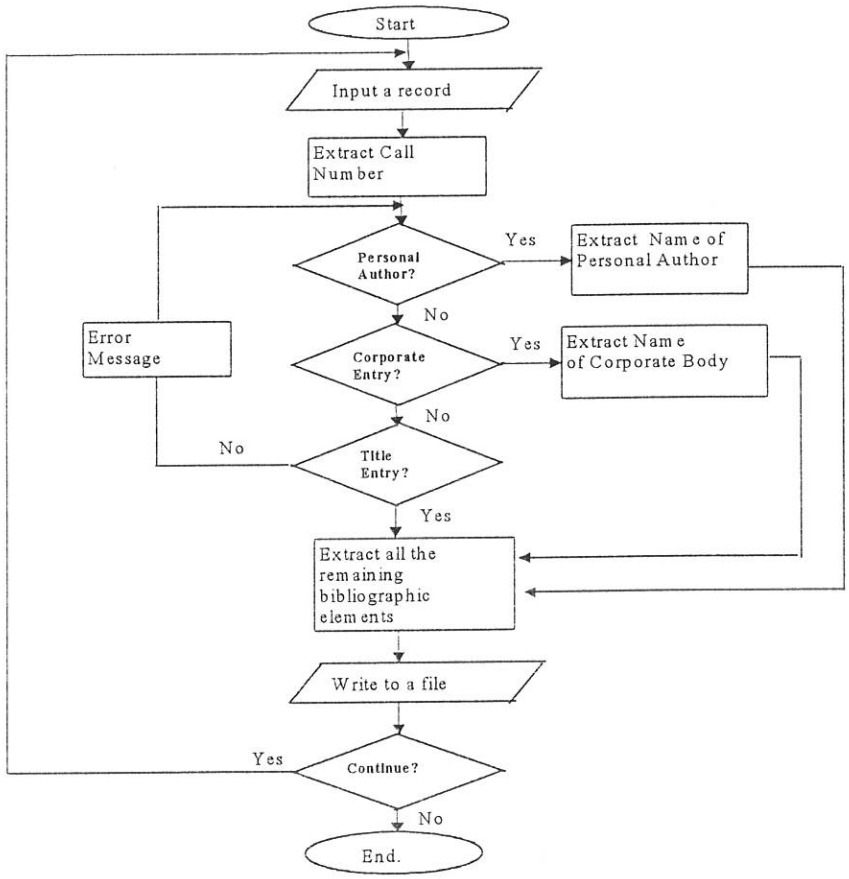


**Figure 5.3** Flowchart showing logical structure of the prototype program.

## 5.4 PSEUDOCODE OF ALGORITHM DEFINITION

As shown in Figure 5.3, extracting bibliographic elements from all types of entry make use of the same algorithm except for name of personal author, and corporate body for the differences lie only at the beginning of the body; and for minor differences, the program has incorporated appropriate conditions.

**Pseudocode** is used in defining the processes of extraction of bibliographic elements. As defined by Davis (1994, 528), Pseudocode is " ... [a] procedure or a module that resembles program code." Its statements resemble compiler source statements but the syntax is simplified. The Pseudocode of the algorithm used for the prototype program is given below. It is to be noted that each algorithm is not self-contained, but it is a continuation. The algorithm that has been used for writing the source codes of the program is given below.

**BEGIN**
GET record
REPEAT
READ line at a time FROM record
TRIM left and right spaces
STORE line in array **Call_No**
INTRODUCE single space
UNTIL empty line is encountered OR five lines are read
WRITE Call_NoTO output file

IF record is personal author THEN
REPEAT
READ character at a time FROM record
STORE character in character array **Surname**
UNTIL character = COMMA
TRIM leading spaces
WRITE Surname TO output file

REPEAT
READ character at a time FROM record

```
            STORE character in character array Forename
        UNTIL character = COMMA OR character = NEWLINE
        TRIM leading spaces
        WRITE Forename TO output file

        IF last character = COMMA THEN
            REPEAT
                READ character at a time FROM record
                STORE character in Other_Info
            UNTIL character = NEWLINE
            TRIM leading and trailing spaces
            WRITE Other_Info TO output file
        END IF

ELSE
        IF record is corporate body THEN
            REPEAT
                READ character at a time FROM record
                STORE character in Corp_Body
            UNTIL character = FULL STOP
            TRIM leading and trailing spaces
            WRITE Corp_Body TO output file
        END IF
END IF

IF record is not corporate body THEN
        REPEAT
            READ character a time FROM record
            STORE character in Title
        UNTIL character = SLASH
ELSE
        REPEAT
            READ character a time FROM record
            STORE character in Title
        UNTIL character = FULL STOP
END IF
TRIM leading spaces of Title
WRITE Title TO output file

IF record is not corporate body THEN
        REPEAT
            READ character a time FROM record
            STORE character in Stat_Responsibilty
        UNTIL character = HYPHEN
        TRIM leading spaces of Stat_Responsibility
        IF record is Personal Author THEN
```

Author(s)=Stat_Responsibility
END IF

REPEAT
    READ character a time FROM record
    STORE character in Temp character array
UNTIL character = HYPHEN OR character = COLON
TRIM leading spaces
IF last character = HYPHEN THEN
    **Edition** = Temp
    WRITE Edition TO output file
    REPEAT
        READ character at a time FROM record
        STORE character in **Place_Pub** array
    UNTIL character = COLON
    TRIM leading spaces
    WRITE Place_Pub TO output file
ELSE IF last character = COLON THEN
    **Place_Pub** = Temp
    WRITE Place_Pub TO output file
END IF
REPEAT
    READ character at a time FROM record
    STORE character array **Publisher**
UNTIL character = COMMA
TRIM leading spaces
WRITE Publisher TO output file

REPEAT
    READ character at a time FROM record
    STORE character array **Date_Pub**
UNTIL character = FULLSTOP
TRIM leading and trailing spaces
WRITE Date_Pub TO output file
REPEAT
    READ character at a time FROM record
    STORE character array **Pagination**
UNTIL character = COLON
TRIM leading spaces
WRITE Pagination TO output file
REPEAT
    READ character a time FROM record
    STORE character in Temp character array
UNTIL last three characters = "cm." OR last character = SEMICOLON
TRIM leading spaces
IF last character = SEMICOLON THEN

```
        Illust = Temp
        WRITE Illust TO output file
        REPEAT
                READ character a time FROM record
                STORE character in Physical_Desc array
        UNTIL last three character = "cm."
        TRIM leading and trailing spaces
        WRITE Physical_Desc TO output file
ELSE
        Physical_Desc = Temp
        WRITE Physical_Desc TO output file
END IF

REPEAT
        READ character a time FROM record
        STORE character in Temp character array
UNTIL last character = ')' OR last four characters = "ISBN"
 OR last two characters = "1."
TRIM leading and trailing spaces
IF last character = ')' THEN
        Series_Title = Temp
        WRITE Series_Title TO output file
        REPEAT
                READ character a time FROM record
                STORE character in Temp
        UNTIL last four chars = "ISBN" OR last two chars = "1."
        TRIM leading and trailing spaces
        IF last four characters="ISBN' THEN
                Notes=Temp
                IF Notes <> spaces THEN
                        WRITE Notes TO output file
                END IF
                REPEAT
                        READ character at a time FROM record
                        STORE character in ISBN
                UNTIL character = NEWLINE
                WRITE ISBN TO output file
        END IF
ELSE
        IF last four characters="ISBN" THEN
                Notes =Temp
                IF Notes <> spaces THEN
                        WRITE Notes TO output file
                END IF
                REPEAT
                        READ character at a time FROM record
```

```
                    STORE character in ISBN
                UNTIL character = NEWLINE
                WRITE ISBN TO output file
        END IF
ELSE
        Notes=Temp
        IF Notes <> spaces THEN
                WRITE Notes TO output file
        END IF
END IF

REPEAT
        READ character a time FROM record
        STORE character in Subj_Headings character array
UNTIL last two characters = "I. " OR End of File is reached
TRIM leading and trailing spaces
SEARCH "2.", "3.", "4.", "5." AND REPLACE WITH ","
WRITE Subj_Headings TO output file
END.
```

**Figure 5.4** Pseudocode of the prototype program

## 5.5 REQUIREMENTS OF THE PROGRAM

Even though the cards follow the same format (**AACR2**), this cannot be taken for granted that the program successfully works. Because there might be inconsistencies in use of the punctuations during card production in some cases; and while OCR conversion there might also be the possibility of converting the punctuation marks incorrectly which can result from accuracy rate of the software used. These will be discussed in next the chapter.

Therefore, in order for the program to correctly extract each element, the records are required to meet the conditions and format mentioned above. These are summarized as follows:

- There always has to be an empty line(s) in between the call number and body of a record, if the number of lines of the call number is less than five;
- Commas have to be used to separate surname from forename, and forename from other author information such as date of birth, ed., comp., etc., publisher from date of publication;
- There has to be a FULL STOP at the end of name of a corporate body, and date of publication;
- Title information has to be ended with SLASH (in case of personal author) and with a FULL STOP in the case of corporate body;
- Hyphens have to be used to separate statement of responsibility from edition, edition from place of publication, titles from edition (in case of corporate author);
- Colon has to be used to separate pagination from illustration or physical description, and place of publication from publisher;
- Semicolon has to be used to separate illustration (if exists) from physical description, and
- The key phrases or strings must be there, if they exist on the original.

# CHAPTER 6

# TESTING RESULTS AND IMPLEMENTATION STRATEGY OF THE PROTOTYPE PROGRAM

## 6.1 INTRODUCTION

By using a set of sample cards chosen for the purpose of testing (see Chapter 1, Section 1.4), the effectiveness of the prototype program, with respect to correctly extracting bibliographic elements has been tested. The records output by the OCR software have also been analysed to determine the proportion of errors occurred during the conversion process and to comprehend the possible causes of the errors accordingly. This enables to decide whether the application of OCR for retrospective conversion is viable or not. This chapter presents the findings of the test in these regards. Basic strategy for implementation of the prototype system by the university library system is also given.

## 6.2 RESULTS

The results of the analysis are organized to get answers to the following questions which have been given in Chapter 1. These are as follows:

1. What is the proportion of error rates in OCR conversion of catalogue records?
2. What are the possible causes of the errors?
3. How much preprocessing of the OCR output records is required in terms of editing their

contents for the errors, and checking and correcting those records which are not consistent in using the prescribed cataloguing format or according to the format that the prototype program requires?

4. How much consistent the catalogue of the university library system is in following uniform format, especially with respect to strictly using punctuation marks?

5. What is the success rate of the prototype program in extracting bibliographic elements correctly, and partly?

### 6.2.1 Average Errors of OCR Output Records

It is a well known fact that OCR conversion cannot by any means, be hundred percent accurate. But for practical applications at least its accuracy should be quite nearer to hundred percent. In line with the facts, there were errors encountered during conversion of the scanned catalogue cards to text files. To determine the number of the errors, each and every record was manually checked against the original information appearing on the cards. Then the errors were counted in terms of characters. This is because there are several erroneous characters occurring in a word.

According to the analysis done, the average number of errors per card in terms of characters for the three types of entry are shown in Table 6.1.

**Table 6.1** Table showing average errors per record in character.

| Type of Entry | Average number of errors/card (in character) | Average percent of errors/card |
|---|---|---|
| **Personal author** | 10.69 | 4.46 |
| **Title** | 12.36 | 4.75 |
| **Corporate body** | 9.44 | 4.54 |

The average number of characters per card was found to be **238.85, 260.10, and 208.08** for personal author, title, and corporate body entries, respectively. These figures have helped to calculate the **average percentage of errors per card for each entry type**. Accordingly, it was found out that the average number of errors per card are **10.69 characters (4.46%), 12.36 characters (4.75%), and 9.44 characters (4.54%)** for personal author, title, and corporate body main entry cards, respectively.

The figures indicate that there is no significant difference among the entry types (see **Appendices I, II, and III**). For the purpose of generalization to the whole samples, the grand average errors per card for the library catalogue that require editing is **4.58 percent.** As far as this figure is concerned, the errors can generally be said to be much less and can easily be edited by someone requiring less effort when compared to keying the whole information from keyboards as online editing is definitely faster than keying in the whole information from the keyboard.

### 6.2.2 Types of Errors Identified

Generally, there have been two types of errors identified during the analysis of the records. The first and most prevalent one was wrongly converted characters or letters. This type of errors is directly related to the accuracy and limited fonts that the OCR software supports. In most cases the above errors, the software could not differentiate among characters having similar features or resembling each other. Some of the errors of this type are given in Table 6.2 below as an example.

**Table 6.2** Example of wrongly converted characters having similar features.

| Characters on cards | Converted as |
|---|---|
| s | a, 8 |
| , | 9, l, g, t |
| 1 | I, l, I |
| 0 | o, @ |
| . | *, -, ,,e, o, a, @ |
| S | B, J |
| e | o, 6 |
| I | J, 1, l |
| Q, a, P | @, 0 |
| I, A | l, 1 |
| E, P | F, G, C |
| V | Y, U |
| a | m, @ |
| 9 | g, ;, , |
| 7 | ? |
| 2 | ?, 7 |
| u | il, ii, n, y |
| n, m | ii, in |
| E, F | X |
| h | n, li |
| &, S, 5 | @, 8 |
| : | ;, l |
| B, etc. | 8, S, etc. |

As discussed in Chapter 3, accuracy of OCR software packages is very crucial and is an essential criterion. This has been evidenced by this study as well. Almost all of the above errors could have been drastically reduced by using high-accuracy OCR software package, such as, **OmniPage Professional or TextBridge**.

The second type of errors is that characters were rarely missing or non-existing. The possible cause of these errors can be attributed to the quality of the card catalogues, that means they might be weary, old having scratches, or torn out, and illegibility of type set i.e., broken characters, in which case the OCR converted them as a pair of characters, missing, or adding extra characters to the output. In determining the types of error, each error was again checked against the sample cards.

In a nutshell, however, the OCR conversion process of the sample cards has ,to a great extent, been successful as per the analysis of the output records. It is therefore, possible to use the card catalogues of the library to convert to computer processible files by using OCR, whatsoever OCR software will be used in fact, with varying accuracy rates.

### 6.2.3 Preprocessing of Records

To test the effectiveness of the program (PAEBE) or to implement it, two types of preprocessing of the records are demanded. Firstly, contentwise the records have to be edited online using the original catalogue cards using any text editor software. In other words, this is the verification of the records for the errors discussed above. Secondly, records that are not compatible with the format and layout used by the prototype program, have to

also be preprocessed i.e., to check the existence of required punctuations or key strings or phrases used as delimiters. This is because inconsistent use of cataloguing format will have direct impact on the outcomes of the programs. Therefore, preprocessing of the records for consistencies is required. This is mainly checking and correcting of the records deviating from the prescribed format as the program has been written in this manner. Otherwise, the program cannot correctly extract the bibliographic elements. These have been discussed in the previous chapter. This type of errors is mainly attributed to typographical errors, and to some extent inconsistent use of the prescribed rules of the cataloguing format.

As has been noted above, on the average the editing required for the first type of errors is **4.58% per card**. According to the sample records, **87.14%** found out to be **consistent** in following the same format - **AACR2 format**. The remaining **12.86%** were found to be **inconsistent**, meaning the records were not using fixed punctuation marks, though the order of bibliographic elements is according to the dictated format. These are the ones that have to be adjusted to make them acceptable by the program. Even though the amount of preprocessing required seems higher, it should be noted that the preprocessing at this stage is insignificant when the amount of keystroking is considered for it requires only correcting the punctuation marks when compared to verifications done contentwise.

A great deal of these records are materials that were **published before 1975** i.e., out of the records required the second type of preprocessing or found to be inconsistent in following the same principle, **81.48%** were materials dated more than 20 years as per the date of their publication. An example of a record before and after preprocessing is shown below.

HV
697
. P45
F67

Ford, Anii Suter.
The physicians assistant :. a nitionaL and local analysis / Ann-Suter-'. Ford. - New York : Praeger Publishers, 1975.

xiv, 245p. ; 24 cm. - (Praeger special studies in U. S. economics, social, and political issues)
Bibliography : P-232-245
ISBN 0-275-28855-2

1. Physicians; assistants-United States. I. Title.

(a) Unprocessed record

HV
697
. P45
F67

Ford, Ann Suter.
The physicians assistant : a nitionaL and local analysis / Ann Suter Ford. - New York : Praeger Publishers, 1975.

xiv, 245p.: 24 cm. - (Praeger special studies in U. S. economics, social, and political issues)
Bibliography : P-232-245
ISBN 0-275-28855-2

1. Physicians; assistants-United States. I. Title.

(b) Processed record

**Figure 6.1.** An example of a record before and after preprocessing

The reason that the program's algorithm mainly depended on the punctuation marks was that sometimes the OCR software does not retain the layout of the cards since it makes text zones and converts these zones into columns which is not consistent resulting in unnecessarily complex and inefficient algorithms requiring further processing which could be tedious.

### 6.2.4 Final Output of the Prototype Program

As has been mentioned, the final output of the program is supposed to be **an ASCII text file consisting of records having bibliographic elements being labelled or named accordingly**. That is the final outcome of the program from which the records can easily be accepted by or imported to library management software packages, specifically to library house keeping software that have **desirable import and export features.**

For the outputs are in ASCII format, it is possible to import these records to any data base management software by giving delimiters according to the specific software which is going to be used. Some of the records output by the prototype program look like as under.

CALL NUMBER : 697 .P45 F67
TITLE : The physicians assistant : a nitionaL and local analysis
AUTHOR(S) : Ann Suter Ford
PLACE(S) OF PUB : New York
PUBLISHER(S) : Praeger Publishers
DATE OF PUB : 1975
PAGINATION : xiv, 245p.
SIZE : 24 cm
SERIES TITLE : Praeger special studies in U. S. economics, social, and political issues
NOTES : Bibliography : P-232-245
ISBN : 0-275-28855-2
SUBJ HEADING(S) : Physicians; assistants-United States
##

CALL NUMBER : SB 818 .w6
CORPORATE NAME : World Health Organization
TITLE : Insect and rodent control through environmental management : a community action programme
PLACE(S) OF PUB : Geneva
PUBLISHER(S) : WHO
DATE OF PUB : 1991
PAGINATION : vii, 107p
ILLUSTRATION : ill
SIZE : 30cm
NOTES : Bibliography : p. 106-107
ISBN : 92-4-154411-2
SUBJ HEADING(S) : Insect pests--Control
##

CALL NUMBER : QR 270 .C34
TITLE : Cancer diagnosis: new concepts and techniques
STAT OF RESPONS : ed.: Richard J. Steckel; A. Robert Kagan
PLACE(S) OF PUB : New York
PUBLISHER(S) : Geune and Strattong
DATE OF PUB : 1982
PAGINATION : xi, 340p.
ILLUSTRATION : ill.
SIZE : 29cm
NOTES : Includes bibliographical reference and index
ISBN : 0-8089-1451-0
SUBJ HEADING(S) : Cancer-Diagnosis
##

CALL NUMBER : BC 108 .C69 1990
TITLE : Introduction to logic
AUTHOR(S) : Irving M. Copi; Carl Cohen
EDITION : 8th ed.
PLACE(S) OF PUB : New York ; London

PUBLISHER(S) : Macmillan ; Collier Macmillan
DATE OF PUB : 1990
PAGINATION : xiv, 569 P.
ILLUSTRATION : ill.
SIZE : 25 cm
NOTES : Includes bibliographical references and index.
ISBN : 0-02-325035-6
SUBJ HEADING(S) : Logic
##

**Figure 6.2.** Records output by the program.

### 6.2.5 Performance of the Prototype

After records are verified and/or corrected, the next step is to process them by using the prototype program. Accordingly, the effectiveness or performance of the program was determined by the percentage of bibliographic elements correctly extracted (exactly as they are); one bibliographic element has error(s) out of the total number of records considered for each entry type.

A correctly extracted element is that element whose content is the same as the content of the input record, meaning a word-for-word correlation between fields in the machine-readable form of card catalogue records and the extracted bibliographic elements. On the other hand, a bibliographic element is said to have errors if there are differences in completeness, forms, etc., along with differences resulting from the occurrence of extraneous characters in the element. Analyses have been made for each entry type as follows.

**Personal Author Main Entry :-** The performance rate of the program in extracting bibliographic elements from records made under name of personal author and title is slightly less than corporate author. Recordwise, of the total 40 sample records considered, data

elements from 39 records **(97.50 %)** could be successfully extracted without a single error occurring. Only one record had an incomplete element in subject heading field.

**Title Main Entry :-** The overall performance rate is the same as for personal author entries i.e., data elements from 39 records out of 40 **(97.50%)** records were extracted correctly with no error in any of the fields. Again, only a single record had an error in the statement of responsibility field.

**Corporate Body Main Entry :-** The program was able to successfully extract **100%** (35 out of 35) of the records in the case of corporate body entry. There was no error found in all 35 records of the corporate body main entry.

To sum up, on the average the proportion of bibliographic elements correctly extracted for the taken sample catalogue records is **98.33%** for the three types of main entry samples chosen for the study. This indicates that the performance rate of the prototype program is very high; only about **2%** of the records need to be corrected. The errors were found to be that the records were not according to the format expected; otherwise, the information was there in both cases. In the case of title entry, **"role"** of the responsible person(s) was not indicated correctly due to inconsistent use of the phrases such as edited by, editors, ed., compiled, comp., etc. appearing on the cards. As mentioned earlier, in case of personal author entry the element was not complete.

### 6.2.6 Factors Determining OCR for RECON

There are some factors that should be taken into considerations when applying OCR for retrospective conversion: these are the accuracy of the OCR software, consistency of information on the card catalogues, quality of card catalogues, legibility and uniformity of characters, typesets, and fonts on card catalogues. Therefore, these are the deciding factors for application of OCR for RECON.

## 6.3 IMPLEMENTATION OF THE PROTOTYPE SYSTEM

Even though a detailed analysis of hardware and software specifications is required to come up with a document describing specifications of the hardware/software, staff, and procedures suggested in the next sub-sections, due to time constraints and the overall goal of the thesis, only their basic considerations have been highlighted for implementation of the prototype system by the Addis Ababa University Library System. Therefore, to fully implement the system for retrospective conversion **rudimentary hardware/software, personnel, and procedures** required include as follows.

### 6.3.1 Software Considerations

For the implementation of the prototype system, at least the following software packages are required.

**Operating Software:-** It is apparent that an operating system is a must to run any computer applications. To this effect, **MS-DOS 6.xx** is suggested.

**Turbo C++**:- For modifications of the source codes or addition of new features to the program may be required. **Turbo C++ version 1.01** or above is required to compile the source codes after appropriate modifications are made.

**Library Management Software**:- There are a number library management software packages available on the market. Most of them have got an OPAC module. These are for example, CARL (Colorado Alliance of Research Libraries), CITE (Computerized Information Transfer in English), CAS (Catalogue Access System), DYNIX, GEAC, LIAS (Library Information Access System), LIBERTAS, MELVYL, TINlib, BLCMP and so on. These software packages have been reviewed by Kidei in her Msc thesis (1996).

Kidei (1996, 210) has recommended **TINlib** which is an integrated library management system supplied by IME (Information Made Easy). This software has been recommended for the proposed prototype system due to various desirable features including the following:

- powerful online catalogue features;
- flexible database design features such as:
  - creation of records of variable field length and indefinitely repeatable fields,
  - integrated search and edit functions that allow editing of fields within the record creation mode, without having to quit data entry and go back in via an edit function,
  - interactive windowing facilities, whereby windows are used to display information such as authority lists without overwriting data on the current

screen being worked on;

- compatibility with MARC format;
- powerful **import/export facilities** which is highly demanded by the library to implement the prototype system; and
- full suite of library modules which are well integrated. This is desirable feature for implementation of the prototype system which might in future, have to be integrated with other library operations such as cataloguing, OPAC, circulation control, acquisition and serials control modules.

However, if the library decides to use other software, then their import/export features will have to be powerful; in other words, there should be desirable text importing facilities to the database.

**OCR software package:-** Obviously OCR software is required for conversion of scanned documents or cards in this case, to text files or ASCII format to be processed. The major OCR software packages have been discussed in Chapter 3. The library should consider OCR software with **reasonable accuracy** so as to reduce the amount of preprocessing especially online editing of records. In line with accuracy and other desirable features **OmniPage Professional** is recommended.

**Scanner software:-** This software may be required in order to increase eligibility of images while the scanning process, especially in the case of poor quality cards and typeset it is very essential.

**Other necessary software packages**:- Other software required to support implementation of the prototype system are **word processing software package** or simple text editor software for instance, **MS-DOS editor** suffices to preprocess i.e., to verify and correct records output by the OCR software package. Moreover, utility software is also required for formatting diskettes, disinfecting computer viruses and checking bad sectors of diskettes, and so on.

### 6.3.2 Hardware Considerations

The hardware required to implement the proposed system must be compatible with the recommended software packages. The basic hardware requirements include the following:

**Computer system**:- Microcomputer system, IBM or IBM compatible is recommended for implementation of the prototype system as this brand is the one which is easily available on the market of the country. To **increase efficiency** or to make the scanning and conversion process much faster computer system with **desirable processor speed** such as Pentium is suggested. In other words, a machine with **high processor** is required to increase the scanning speed which is essential for the system.

**Computer storage capacity**:- The total collection of the university library's books, bound periodicals, etc. is about **434,254**. The average length of bibliographic record in machine readable form is about 550 (calculated from the sample records). Giving the same amount of storage space for the accompanying index and auxiliary files, we can assume that a record will require about 1100 bytes of storage space in a computer. Therefore, the total amount

of storage space required will approximately be (434,254 x 1100 = 477679400 bytes or about **478 MB**. Besides allowance storage spaces needed for various software packages should also be considered.

**Scanner**:- A flatbed scanner with necessary peripherals is required to scan the card catalogues. At least, a scanner with **minimum resolution of 300 dpi** is recommended to capture better images of the cards. This and other criteria of scanners have been discussed in Chapter 3 at length.

**Printers**:- Printers are also suggested to enable the library staff or clerical personnel to print records need to be checked against the library catalogue, and to be used for other purposes.

**Other peripherals**:- Diskettes, magnetic drives and tapes are also required by the prototype system for back-up of the records converted or databases and off-line storage of some files.

### 6.3.3 Staff Considerations

The prototype system can be implemented with existing staff of the library system, for most of the professional staff already have adequate knowledge in computer applications. Especially, the staff of the **Computer Section** of the library should be able to implement the suggested system with appropriate modifications or adjustments. To this effect, however, for **scanning and on-line editing** of the records output by the OCR system, semi-professional staff having a good knowledge of cataloguing may suffice. It is believed that the library has adequate number of staff to implement this system.

### 6.3.4 Procedural Considerations

For better efficiency it has been recommended that there should be division of works at different stages and batches. A five- step process that is to be worked out in batches is suggested as follows.

1. A group of catalogue cards should be scanned and converted to text files by using OCR software package;
2. This group of converted records should be verified against the original cards and edited online by using a text editor software;
3. Then the prototype program is to be run on these records to extract bibliographic elements by giving appropriate field name or tag;
4. The records should be imported to chosen library management software or text retrieval software;
5. Finally the imported records should be rechecked by browsing and should be corrected if any errors have been encountered.

These procedural steps are shown by Figure 6.3.

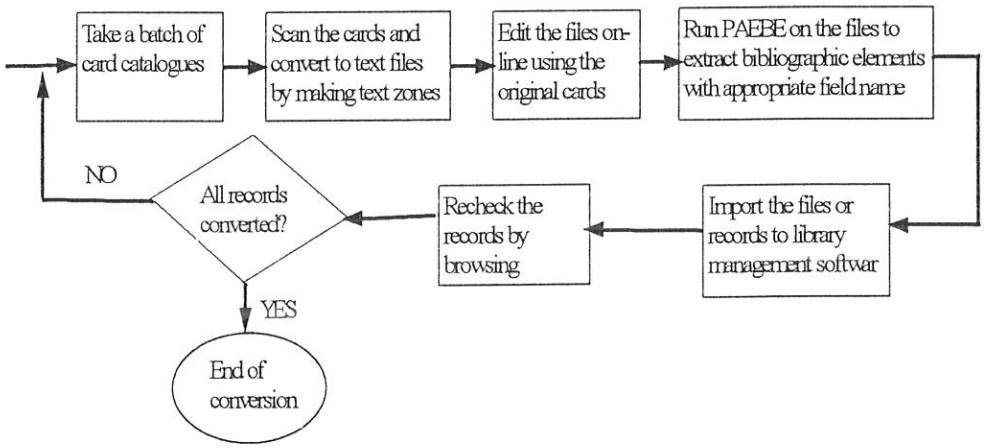


**Figure 6.3** Flowchart describing steps required for implementation of the prototype RECON system for AAU library system

# CHAPTER 7

# CONCLUSIONS AND RECOMMENDATIONS

## 7.1 INTRODUCTION

This chapter offers conclusions of the study and forwards recommendations accordingly.

## 7.2 CONCLUSIONS

The study has given insight to arrive at the following conclusions:

- Library automation is of absolute necessity for libraries particularly for those having a large collection and a large user population to reduce costs and to improve efficiency and flexibility of their services when compared to manual system.
- Addis Ababa University Library System is the biggest academic library in the country having a large collection; consequently, installation of computer-based library system is found to be necessary to overcome most of the problems related to the existing manual operations.
- RECON is the first step towards implementation of computer-based library system as a machine readable version of manual records is a necessary prerequisite.
- Though there exist various methods of RECON ranging from downloading bibliographic records from batch cataloguing databases or bibliographic networks to contracting commercial bureaus to do the conversion, they are quite expensive, time-

consuming and labour-intensive, or may not be applicable to libraries of developing countries in general and Ethiopia in particular as considerable collections or materials produced in these countries are not covered by most databases that are used for RECON. As far as these methods are concerned, their costs are the main prohibiting factor for most of libraries of the developing countries, specifically Ethiopia.

- In spite of the fact that Addis Ababa University Library System is planning to automate its library functions, lack of adequate budget that has also highly contributed to the problem of RECON impedes its implementation.
- The OCR technology has been improved from time-to-time in terms of accuracy in reading poor quality documents, throughput, batch processing, and integration with other software packages; and it has long been considered as a promising alternative to other RECON options.
- To this effect, the thesis aims at investigating the technical feasibility of OCR technology for RECON, with a particular reference to the university library system's card catalogue; and to develop a prototype program accordingly.
- It is found out that to use this technology for RECON, the formats and layouts of card catalogues have to follow uniform or consistent cataloguing rules.
- Moreover, preprocessing of records output by OCR is of absolute necessity in this respect; and this should be less in order for it to be feasible for RECON. The amount of preprocessing required depends on the accuracy of the OCR used, consistent use of cataloguing rules, and typographical errors that might be occurred during card production.
- In line with this fact, the preprocessing of records output by the OCR software used

was found to be fairly less, i.e., **4.58% per record** for the university library system's catalogue cards. Even this percentage of errors could have been reduced by using state-of-the-art OCR software packages, such as OmniPage Professional or TextBridge.

- The study has also shown that a great deal of the library catalogue cards **(87.14%)** were found to be consistent and follow uniform cataloguing rules, which is AACR2.
- The general objective addressed by the thesis was concerned with the technical feasibility of developing automated system for extraction of bibliographic elements from card catalogue surrogates output by OCR software in order to eventually, export its outputs to a text retrieval or library management software. The results of the thesis provided evidence in the affirmative that the developed algorithm has had attractive performance, i.e., **98.33%** success on the average.
- Thus, the result of the algorithm used in writing the prototype program called PAEBE, which uses punctuation marks and key phrases and strings as conditions, was also proved successful.
- Taking the above points into consideration, it can be concluded that OCR is indeed technically feasible alternative to convert card catalogues of the university library system to machine-readable records which is a prerequisite to library automation for which the AAU Library System is currently planning.
- It was found out that the factors, that determine the effectiveness of OCR for RECON application, are accuracy of OCR software to be used, and conformity of cataloguing rules to a standard format. Moreover, typographical errors occurring during card production also determine its effectiveness.
- Preprocessing of records can be considered as one of the limitations of the technology

for application to RECON since OCR conversion may not be hundred percent accurate.

- For implementation of the prototype RECON system batch processing of records is found to be faster, since the system has different implementation steps identified.
- Last but not least, the AAU library system has fairly good number of staff trained in computer applications and has relatively adequate hardware required for the implementation of the prototype system.

## 7.3 RECOMMENDATIONS

The study has forwarded the following recommendations.

- Due to time constraint, the thesis has, by giving much emphasis, investigated only the technical feasibility of OCR for RECON. To implement the results of the thesis, a feasibility study should therefore, be carried out to estimate costs and time requirements for application of OCR as a RECON option for the university library system.

- As the thesis has been carried out using only card catalogues produced in English language characters, similar study should also be conducted to convert cards catalogues produced in Amharic scripts for Ethiopianna collections. In line with this, a research is underway to develop an Amharic OCR program by one of fellow

students, Worku Alemu at SISA this year.

- The program has to be tested, and is to be improved if necessary, to process cards produced at different times (20 or more years ago); and it may be desirabe to explore the effectiveness of the proposed algorithm with more samples.

# BIBLIOGRAPHY


Addis Ababa University Library System. 1997. Proposal for networking of the AAU libraries for information sharing. Addis Ababa: Computer Section. Unpublished.

Addis Ababa University. June 1996. Addis Ababa University libraries: towards the 21st century. Report of the committee for the study of AAU libraries. Revised version. Addis Ababa: Addis Ababa University.

Annual Report of the Addis Ababa University Library System. 1995. Addis Ababa: Addis Ababa University Library System. Unpublished.

Asher, Richard E. 1982. Retrospective conversion of bibliographic records. Catholic Library World 1982(November): 155-161.

Avram, Henrietta D. 1972. RECON pilot project: final report on a project sponsored by the Library of Congress. Washington: Library of Congress. P.1. In: Asher, Richard E. 1982. Retrospective conversion of bibliographic records. Catholic Library World Nov. 1982: 155-161.

Beale, Stephen. 1996. Apple scanner sports new software. MacWorld. 13(12): 44.

Beaumont, Jane. 1986. Retrospective conversion on a micro: options for libraries. Library Software Review 1986(July/August): 213-218.

Beutler, Earl. 1995. Assuring data integrity and quality: a database producer's perspective In Electronic information delivery: ensuring quality and value. Edited by Reva Basch. Vermont: Gower. 59-68.

Boss, Richard. 1984. Retrospective conversion: investing in the future. Wilson Library Bulletin 1984(November): 173-178, 238.

Bossers, Anton and Derek Law. 1990. Guidelines for retroconversion project: foreword. IFLA Journal. 16(1): 32-36.

Bryant, Philip; Ann Chapman; Bernard Naylor. 1995. Retrospective conversion of library catalogues in institutions of higher education in the United Kingdom: a study of the justification for a national programme. Report submitted to the Follett Implementation Group on IT (FIGIT). University of Bath.

Burton, Paul F. and J. Howard Petrie. 1986. The librarian's guide to microcomputers for information management. Van Nostrand Reinhold: Berdshire.

Busch, David D. 1991. The complete scanner handbook for desktop publishing. Macintosh edition. Illinois: Business One IRWIN.

_____ . 1996. MacIRISPen: portable, accurate pen OCR scanner. MacWorld. 13(4): 75.

Cibbarelli, Pamela. ed. 1993. Directory of library automation software, systems, and services. NJ: Learned Information, Inc.

Cohn, John M; Ann L. Kelsey; Keith Michael Fiels. 1992. Planning for automation: a how-to-do-it manual for librarians. New York: Neal-Schuman Publishers, Inc.

Davis, William S. 1994. Business systems analysis and design. Belmont: Wadsworth Publishing Company.

Diehl, Stanford and Howard Eglowstein. 1991. Tame the paper tiger. Byte. April: 220-238.

Drabenstott, Jon Ed. 1986. Retrospective conversion: issues and perspectives. Library-Hi-Tech. 4(2): 105-120.

ECA/PADIS/SCH/VI/ST/3. 1994. Sixth meeting of the standing committee on harmonization and standardization of documentation and information systems in

Africa. 14-18 Nov. 1994. Addis Ababa: PADIS.

EDGE: Work-Group Computing Report. 1996. OCR: Caere announces OmniPage Pro version 7.0 for Macintosh; Caere reduces the OmniPage Pro retail upgrade price to $129. 18(November 18): 22.

Eglowstein, Howard. 1994. Due recognition for OCR. Byte October: 145-148.

Epstein, Susan B. 1990. Retrospective conversion revisited, part 1. Library Journal 1990(May): 56-58.

Gann, Roger. 1996. Accurate OCR for complex pages. PC User 292(October): 501.

Gorman, Michael. 1977. The economics of library automation, edited by J.L. Divilbiss. Urbana-Chapaign: University of Illinois Graduate School of Library Science. P.26 In: Asher, Richard E. 1982. Retrospective conversion of bibliographic records. Catholic Library World 1982(November): 155-161.

Grotophorst, Clyde W. 1989. Keyless entry: building a text database using OCR technology. Library-Hi-Tech 7(1): 7-15.

Grunin, Lori. 1996. Stealth OCR in action. Windows Source 4(6): 49.

Haddad, Peter. 1990. Retrospective conversion in national and research libraries: the Australian experience. IFLA Journal 16(1): 67-69.

Harrison, A.D.; F.A. Roos; R.E. Thomas. 1995. Semi automatic capturing of bibliographic information from journal contents pages for inclusion in online library catalogues: the RIDDLE project. The Electronic Library 13(1): 15-20.

Harrison, Martin. 1985. Retrospective conversion of card catalogues into full MARC format using sophisticated computer-controlled visual imaging techniques. Program. 19(3): 213-230.

Hein, Morten. 1986. Optical scanning for retrospective conversion of information. The Electronic Library 4(6): 328-331.

Hunter, Eric J. 1985. Computerized cataloguing. London: Clive Bingley.

Jacobs, Gijs J. 1990. Retrospective conversion. How many?... IFLA Journal 16(1): 37-40.

Kahney, Leander. 1996. Two UMAX flatbeds tricked out with bundles. MacWEEK 10(28): 6.

Kawamoto, Wayne. 1996. Reading between the lines. Computer Shopper 16(11): 558.

Kidei, Jane Grace. 1996. An online public access catalogue for the Egerton University Library System: a proposal and strategy for implementation. Msc thesis. Addis Ababa University. Addis Ababa.

Lawson, Matthew; Nick Kemp; Micheal F. Lynch; Gobinda G. Chowdhury. 1996. Automatic extraction of citations from the text of-English-language patents - an example of template mining. Journal of Information Science 22(6): 423-436.

Matthews, Joseph R. Ed. 1985. Public access to online catalogues. 2nd ed. New York: Neal-Schuman Publishers.

Maxwell, Margaret F. 1980. Handbook for AACR2: explaining and illustrating Anglo-American cataloguing rules second edition. Chicago: American Library Association.

Mendelson, Edward. Sept. 1996. Two Windows 95 programs that read the future of OCR. PC Magazine 15(16): 43.

Molto, Mavis and Elaine Svenonius. 1991. Automatic recognition of title page names. Information Processing & Management 27(1): 83-95.

Murphy, Catherine. Winter 1990. Questions to guide retrospective conversion choices for school library media centres. School Library Media Quarterly 18: 79-81. In:

Nakamura, Margaret. Fall 1991. Retrospective conversion using a combination of choices: a case study of the Hawaii School Library Network Project. School Library Media Quarterly : 24-29.

Nakamura, Margaret. 1991. Retrospective conversion using a combination of choices: a case study of the Hawaii School Library Network Project. School Library Media Quarterly Fall 1991: 24-29.

O'Malley, chris. July 1996. OCR, easier than ever. PC Computing 9(7): 78.

Peters, Stephen H. and Douglas J. Butler. 1984. A cost model for retrospective conversion alternatives. Library Resources & Technical Services: 149-162.

Reynolds, Dennis. 1985. Library automation: issues and applications. New York: R.R. Bowker Company.

Rice, James. 1981. OCR for libraries: only a few years away. Library Journal 106(15): 1603-1605.

Riger, Robert E. 1992. Retrospective catalogue conversion in mid-sized law libraries: some practical guidelines for automation. Secial libraries 10: 10-15.

Rowley, Jennifer E. 1987. Organizing knowledge: an introduction to information retrieval. Aldershot: Gower Publishing Company.

_____ 1993. Computers for libraries. 3rd ed. London: Library Association Publishing.

Saffady, William. 1994. Introduction to automation for librarians. 3rd ed. Chicago: American Library Association.

Schein, A. 1989. Optical storage and OCR - key components of automated information management systems. Optical Information Systems 9(1): 9-15. In: Molto, Mavis and

Elaine Svenonius. 1991. Automatic recognition of title page names. Information Processing & Management 27(1): 83-95.

Scisco, Peter. April 1996. Now read this. Computer Life 3(4): 80.

Skapura, Robert. 1990. A primer on automating the card catalogue. School library media quarterly 18: 75-78.

Sule, Gisela. 1990. Bibliographic standards for retrospective conversion. IFLA Journal 16(1): 58-63.

Tedd, Lucy A. 1993. An introduction to computer-based library system. 3rd ed. Chichester: John Wiley & Sons.

Valentine, Phyllis A. and David R. McDonald. June 1986. Retrospective conversion: a question of time, standards, and purpose. Information Technology and Libraries : 112-120.

Wedgeworth, Robert. Ed. 1993. World Encyclopaedia of library and information services. 3rd ed. Chicago: American Library Association.

Weibel, Stuart; et al. 1989. Automated title page cataloguing: a feasibility study. Information Processing & Management 25(2): 187-203.

Woherem, Evans E. 1995. Towards a culture of management of software systems maintenance in Africa. Information Technology for Development 6(1995): 5-14.

Young, Heartsill. Ed. 1983. The ALA glossary of library and information science. Chicago: American Library Association.

# APPENDICES

## APPENDIX I

### Partial sample of personal author main entry records output by OCR software

RD
102
. A23

Abel, Francis L
Cardiovascular function : principles and applications / Francis L. Abell Ernes P. McCutcheon. - lot ed. - Boston : Little Brown, cl979.

xix, 424p. : ill. ; 27cm.
includes bibliographies and index.
ISBN 0-316-00190-2-

1. Cardiovaacular system. 1. McCutche. Ernest P., joint author. II. Title.

RA
410
.A22

Abel-Smith, Brian*
Paying for health services; a study of the costs and Bources of inance in six countries. Geneva, World Hoalth Organization, 1963.

86 p. tables. 22 on* (Public health papers, no.17)

1. Medical care, Cost of. 2* Hygiene, Public - Finance. I. Title. II. Series

World Health Organization Public health paper, no*179

R
12 5
.A2

Abercrombie, George Francis, ed.
The encyclopaedia of general practice. Edited by G. F. Abercrombie and R. Mo S. McConaghey. London, Butterworthsio 1963-

6v. illus. 23 cm.
Include8 bibliographies*
1. Medicine - Dictionarier,. I. McConaghey, Richard Maurice Sotheron, joint ed. II. Title.

Rp
801
. A63
A2

Abraham, Edward Penley, 1913 -
Biochemistry of some poptide and aterbi antibiotics. New York, Wiley, 1957.

96p. illus, 19cm. (Ciba lectures in mi robial biochosistryg 1957)

1. Antibiotics* 2. Peptidono 3. Ster oids. I. Title.

690
.A33

Acheson, Ernest, Donald.
Medicine : an outline for the intendin student; edited by E. D. Acheson. London Routledge & K. Paul, 1970.

viii, 159 P. illus. 20 cm.
Includes bibliographies
ISBN 7100-6866-2

1. Medicine as a profe88ion - Addresse essays, lectures. I. Title.

RD
57
.A2
1981

Ackermang Lauren Vedder, 1905-
Ackerman's Surgical pathology /by/ Lauren V. Ackerman /and/ Juan Rosai. 6th ed. St. Louis, Mosby, 1981.

2v. , illus. 26 cm.
Includes bibliographies and index.
ISBN o-8ol6-oo45-6

1. Pathology, Surgical. I. Rosai, Juan, 1940- joint author. II. Title.

Q
372
. A313

Adam, Gyorgy
Interoception and behaviour ; an experimental study, by G. Adam. /Translate by R. de Chatel. Translation, rev, by H. Slucki/ Budapestg Akademial Klado, 1967.

vili, 151p, lllus* 25cm,
Bibliography : pe 141-149

1. Interoception* 1. Title.

RC
961
. A43
1963

Adams,, Alfred Robert Davies.
Tropical medicine for nurses /by/ A. R. D. Adams and B. G. Haegraith, 2d ed. Oxford
Blackwell Scientific Publications, 1963.

viii, 343p. illus., diagra. 23cm.

1. Tropics - Diseases and hygiene. I. 14Aegraithl Brain Gilmore, 1907- joint author.

RD
731
.A35
1986

Adams, John Crawford
Outline of orthopaedics / by John Crawford Adamso- 10th ed.- Edinburgh ; New

Yrok : Churchill Livingstone, 1986.

vii, 506p. : ill. ; 22cm.
Bibliography : p. /468/-494
Includes index.
ISBN o-443-03442-7

1. Ortliopedia. I. Title.

RC
185
.A3

Adams, Edward Barry
Tetanus /by/ E.B. Adams, D.R. Laurence /and/ J.W.G. Smith. Oxford. Blackwell -Scientific, 1969,

vii, 165 P. illus. 23 cm.
Bibliography : p. 142-157
SBN 632-o6O9O-5

l. Tetanus. I. Laurence, Desmond Roger, joint author, II. Smith, Joseph William Grenville, joint author. III. Title.

RC
685
.H8
A44

Akinkugbe, 010.
High blood pressure in the African /by/ O.O. Akinkugbe. Edinburgh, Churchil Livingstonot 1972o

xg 13,3 p. illue ; maps 21 cm.
Bibliography : po 119-125

ISBN o-443-oo856-6

1. Hypertension. 2. Africa-Statistics Medical, 1. Title.

RC
961.5
. A4

Ajose, Oladele A
Public health in tropical countries / Oladele A. Ajose. - London : J & A. Churchill, 1958.

viig 191p. : ill. ; 2lcm.
Includes index
1, Tropics - Disoases and hygiene.
1. Title.

57
.I35
1958

Aird, Ian, 1905-
h companion in surgical studies /by/ Ian 'Aird and the staff of the Dept. of Surg6ry. Postgraduate Medical School of London. 2d ed. F-dinburgh, Livingstone 1990.

xii, 1332 p. 26 cm,
Includes bibliographies.

1. Surgery. 2. Pathology, Surgical. I. Lond.ons University. Postgraduate Medical School of London. II. Title,

RA
44O06
. A2

Abbattl F.R.
Continuing the education of health workers:a workshop manual / F.R. Abbatt, A.Meilao - Geneva: Hold iialth'urganiza tion, 19800

X9185p. ; 24cm,,
ISBN 0-4-154220-9

1. Health education* 1. Mejial A. II. Title.

HY
5804
. A23
Ref o

Abel, Ernest L., 1943-
A dictionary of drug abuse terms and terminology / F-ruent Lo Abel. - Westpo-rt, Conn* : Greenwood Press, cl984.

xi, 187 po ; 24 cm.
Bibliography : p . /185/-187
ISBN 0-313-24095-7

1. Drug abuse-Dictionaries. I. Title.

RE
925
.D8
.1978

Duke-Elder, William Stewartl Sir, 1898-19'78
Duke-Elderis practice of refra-ction // revised by David Abrams. - gth ed. - Edinburgh ; New York : Churchill Livingstone ; New York : distributed by Longman, 19?8.

vi, 204 p. : ill. ; 26 cm.
Includes index
ISBN 0-443-ol478-7

RD
99
.F63

Forrest, Jane.
Foundations of surgical nursing / Jane Forrest. - London : Edward Arnold, 1974.

012 p. : ill., form. ; 72 cm.
Includes index.

1. Surgical nursing. I. Title.

RC
646
. F6

Foldi, Mihaly, 1920-
Diseases of lymphatics and lymph circu-lation, by Micliael Foldi. Budapest, Akade-miai Kiado, 1969.

-Xiv, 188P. illils. 24cm.
liicludes bibliographies

1. Lymphatics - Diseases. 1. Title.

BC
108
. C69
1990

copi, Irving M.
introduction to logic / Irving M. CoPi, Carl Cohen. -- 8th ed. -- New York : Macmillan ; London : Collier Macmillan,

C1990.

xiv, 569 P. : ill. ; 25 cm.

Includes bibliographical references and index.
ISBN 0-02-325035-6

1. Log ic. I. Titie.

TD
913
. F7

Franceys, R
A guide to the development of on-site sanitation / R. Franceys, J. Pickford & R. Reed.-- Geneva WHOT 1992.
viii, 237 P. ill. ; 26 cm.
Bileliography p. 196-202@
Includes index
ISBN 92-4-154443-0
1. Sanitation. 2. Waste disposal altos. I. Pickford., J. II. Reed, R. III, Title.

HD
8395
.W75

Wright, D. G.
Pop4lar radicalism : the working-class experience, 1780-1880 / D.G. Wright. -- London ; New York : Longman, 198 .

X, 211 p. ; 22 cm. -- (Studies in modern history)

Bibliography: p. 191-205.
Includes index.
ISBN 0-582-49440-0

1. Working class--Great Britain--Political activity--History--18th century. 2. Great Britain--Politics and government--18th century. I. Til-le.

RT
81.5
. D46
1992

Dempsey,.Patricia Ann.
Nursing research with basic statistical applications / Patricia Ann Dempsey, Arthur D.
Dempse@. -- 3rd ed. Boston Jones and Bartle t, cl992.

xii 324 p: - ill. 25 cm. (The Jones an rtl't series in n rsing)

Rev. ed. of: The research process in nursing.
2nd ed. 1986.
Includes bibliographical references and index.
ISBN 0-86720-449-4

RJ
50
.T86

Tunnessen, Walter W., 1939-
Signs and symptoms in pediatrics / Walter W
Tunnessen, Jr. -- 2nd ed. -- Philadelphia : Lippincott, cl988.

xvii, 702 p. ; 25 cm.

Cover title: Signs and symptoms in pediatrics
Includes bibliograpnical references and index.

ISBN 0-397-50863-8

RA
971.3
. F73

Frank, C w
Maximizing hospital cash resources /
C. W. Frank. - Germantown, Md. : Aspen
Systems Corp., 1978.

131 p. 24 cm.
Includes 4-ndex.
ISBN 0-8944-@4-076-9

1. Hospitals - Finance. 2. Hospitals -,
Business management. I. Title.

RT
86
.035
1960

Odlum, Doris X
Mental health, the nurse and the

patient. Editor : Ethel Johns, Phila,-

delphia, Lippiucbttt 1960.

192p. 21CM.

1. Nurses and nursing- 3- PsYchiatric -
nursing. 3. Mental iil. I. Title*

## APPENDIX II

### Partial sample of title main entry records output by the OCR Software

RB
37
. c18

CRC handbook series in clinical laboratory-
scieince / David Selii;soni editor-in-chief.
- Cleveland : CRC Press, 1977.

v. : ill. ; 28 cm.

Spiiie title .0 Handbook series in clinical
Iaboratory science.
Includes bibliographies and index.
ISBN 0-8493-7000-0 (set)

RA
643
.AS
1990

Control of communicable diseases in man
Abram S. Benenson, editor. -- 15th ed.
Washington, D.C. : American Public Health
1990.
xxvii, 532 p. ; 17 cm.
"official r-oport of the American
Health Association'

1. communicable diseases--Provention 2.
Communicable
diseases -- Nursing--Handbooke mafnuals, etc.
I. American Pliblic Health Assoc ation.

RC
902
@444

Nephi-ology nursing : perspectives of care /
edited [by] Francine P. Hekelman Carol A.
Ostendarp. -- New York : ticGraw-H'ill,
cl979.

xvii, 326 p. : ill. ; 24 CM.
Includes biblio raphies and index.
ISBN 0-07-027940-9

1. Kidneys--Diseases--Nursing. I.
HeKelinan, Xrancine P. 11. Ostendarp, Carol
A.

QP
376.5
.838

Basic mechanisms of the EEG / S. Zschocke, E.-
Speckmann, editors. -- Boston : Birkhauser,
cl993.

xv, 355 p. : ill. ; 24 cm. -- (Brain dynami
series)

Includes bibliographical references and
index.
ISBN 0-8176-3596-3

PC
255
.N46

Neoplasms - comparative pathology m,f growth in animaleg plants, and man / Hans X. Kainerg editor4 - Baltimore, London : Williams & Wilkinsg, 1981.

xxxll, 908 p. ; ill. ; 28 cm.
Includes bibliographies and index.
ISBN o-683-o45o.3-9

1. Tumor. 2. Growth disorder. 3. Pathologyl Comp&rative, 4. Tumors, Plant. I. Kaisert Hans Elmart 1928-

RT
120
.09
C65

Common problems in primary care //edited by/ Lynne Lesak Gorlinel Cheryl Cummings Stegbauer.- St. Louis: C,V. Mosbyl 1982.'

xiij 286 p. : ill. ; 24 cm.
Includes bibliographies and index4
ISBN o-8ol6-1939-9

1. Nursing 2. Fgmily medicine 3*.. Ambulatory medical care. 4. Nurse practitioners I. Gorline, Lynne Leske

p
572
.S7
B56
1984
Biochemistry of steroid hormones / edited by H.L.J. Makin. - 2d ed. - Oxford : Blackwell Scientific Publications, 1984.

Ili, 714 p. : ill.; 25 cm.
Includes bibliographi cal references and index.

ISBN 0-632 0 986.1

1. Steroid hormones. I* making Hugh Llewellyn John,

R
60
. B488
1974

The Biology of animal viruses /by/ Frank Fenner /and others/ 2d ed. New York, Academic Press, 1974.

xvi, 834 p. illus. 25 cm.
Edition of 1968 by F. Fenner.
Bibliography : p. 642-775.
ISBN 0-12-25304O-3

1. Virology. 2. Virus diseases. 3. Virology - Bibliography, I. Fenner, Frank John, 1914-

647
.c55
1960

Blood coagulation and haemostasix : a practical guide / edited by Joan M. Thomson. - 2d ed. - Edinburgh ; London Churchill Livingstone, 1080.

viii, 369 p, : III, ; 24 cm*
Includes bibliographies and Index.
ISBN 0-443-01813-8

1. Blood-Coagulationg Disorders.

I. Thousons Joan M. II* Practical guide to blood coagulation and haomestamis.

280
.B8
B65

The Breast /edited by H. Stephen Gallager /et al./.- Saint Louis : Mosbir, 1978.

xiv, 564p. ; /3/ leaves of plates : ill. ; 26cm.
Includes bibliographies and index.
ISBN 0-80160-1727-8

1. Breast - Cancer. I. Gallager, H. Stephen, 1922-

RC
.n 1
OC5
C29

Cancer chemotherapy 1980 / edited by H.M . Pinedo. -- Amsterdam : Excerpta Medica, 1980.
xivg 486 p. ; 24 cm. -- ( The EORTC concer chemotherapy annl.al. 2 )
Includes bibliographical references and index
ISBN 90-219-3054'-.4
1. Cancer--Chemotheraphy I. Pinedog H. M. ed.

270
.C34

Cancer diagnosis: new concepts and techniques/editors, Richard J. Steckel, A, Robert Kagan.- Now York: Geune and-Strattong ct982.

xi, 340p,: ill,; 29cme
Includes bibliographical reference and index
ISBN 0-8089-1451-0

1. Cancer-Diagnosis. 1. Steekell Richard J.

281
C4
636

Cancer in childhood/ John 0. Godden, editor. New York, Plenum Press, /cl973/
249 p. illus. 24 cm.
"Proceedings of the 17th clinical con ference of the Ontario Cancer Treatment and Research Foundation."
Includes bibliographical references.
ISBN 0-3o6-30763-4

1. Tumors in children - Congresses. Godden, John O., ed. II. Ontario Cancer Treatment and Research Foundation.

RC
681
.C313

Cardiology in the USSR / edited by E. Chazov ; translation edited by R. Oganov.- Moscow : Mir, 1982.

223p. : ill. ; 22cm.- (Advances

in science and technology in the
USSR medical series)
Includes bibliographies.

1. Cardiology - Russia. I. Chazov. E., ed.

(4m
642
.C444

Cell interactions and development: molecular mechanisms/ edited by Kenneth Ml Yamada.-- New York : Vjiley, cl983.

ixg 287p. : ill. ; 24 cm.
"A Wiley-Interacience publication.'
Includes bibliographical references and index*
ISBN o-471-07987-i
1., Cell interaction. 2. Developmental cytology. I. Yamada, Kenneth M.

581.2
.C44

The Cell in medical science / edited by F. Beck and J. B. Lloyd. London; New York : Academic Press, 1974-1976.

4v. : ill. ; 24 cm.
Includes bibliographies and index.
ISBN O-l2-o842o4-1

1. Cytology. 2. Pathology, Cellular. I. Beck, Felix. II. Lloyd, John Benjamin.

425
. C75

Community health / edited by ttune Clark, Jill Henderson ; foreword by Grace M. Owet.- Edinburgh : New York : Churchil Livingstone, 1983.

xiii, 317P. ; 24cm.
Includes bibliographical references and index.
ISBN o-443-02000-0

1. Public health. 2. Community healt services. I. Clark, June. II. Henderson Jill"

RD -
641
. C48
1983

Chronic problem wounds / Ross Rudolph, Joe M. Noe, /et al./, - lst ed. - Boston Little, Browng cl983.

xii, 250 p.9 /l/ leaf of plates : ill (some col.) : 25 cm.
Bibliography : p. 177-194
Includes index
ISBN 0-316-761lo-9

1. Ulcers-Surgery. I. Rudolph, Ross II. Noe, Joel M.

RC
ill
.C565
1982

Clinical manual of infectious diseases / edited by Stephen A. Berger. - Menlo Park, Calif. : Addison-Wesley Pub. Co., Medical / Nursing Division, 1982,

xiv, 363 P- ; 22 cm.
Includes bibliographies and index.
ISBN 0-201-10093-2

1. Communicable diseases. I. Berger. Stephen A.

182
..s4
D39

The Detection of Septicemia / editor, John A. Washington II. - West Palm Beach, Fla. : CRC Press, 1978*

155 P- : ill- ; 27cm.
Includes bibliographical references and index.
ISBN o-8493-5207-X
1. Septicemia - Diagnosis. 2. Bacteriologyg Medical - Cultures and culture media. I. Washington, John A.

PC
'$"o
. C33
1967
Ref.
Cassell'& Italian-English, Fnglish-Italian dictionary ; prepared by Piero Re'boral with the assistance of Dr. Francis Me Guercio and Arthur L. Hayward. 7th ode Londong Cassell, 1967.

xxig 1096 P. tables, 22 cm*

1, Italian language - Dictionaries - F.nglish. 2o English language - Dictionarie - Italian. I. Ro'boral Piero, 1889- con

RB
127
. C473

Chronic non-caner pain : assessment and practical management / edited by Sven Andersso
... [et al.l. -- Lancaster ; Boston : MTP Press, cl987.

207p. : ill. ; 24 cm.

ISBN 0-7462-0047-1

1. Intractable pain. 2. Chronic Disease. .5 I. Andersson, Sve,- Anders, 1927- e-d.

RB
37
. C54
1984

Clinical diagnosis and management b laboratory methods / [edited by] John Bernard Henry. -- 17th ed. -- Philadelphia Saunders 1984.
xx, 1502' : ill. (some Col.) 27 cm.
At head of-title: Todd, Sanford, Davidsohn.
Includes bibliographies and index.
ISBN 0-7216-4657-3

1. Diagnosis, Laboratory. 2. Diagnosis, Laboratory. I. Todd, James Campbell.

RC
801
. c6

Clinical gastroenterology / editors, Richard
G. Farmer, Edgar Achkar, Bertram Fleshler.
New York -.-Raven Press, cl963.
xii 614 p. : ill. ; 26 cm.
Incfudes bibliographical references and
index.
ISBN 0-89004-780-4

1. Gastrointestinal system--Diseases. 2.
Gastrointestinal diseases. 1. Farmer.
Richard, Gilbert, 1931- ed. II. Achkar,
Edgar, ed. III. Fleshler,' Bertram, ed.

## APPENDIX III

### Partial sample of corporate body main entry records output by the OCR Software

RE
20
.B7
1979

British Orthoptic Society.
A glossary of terms used in orthoptic practice in the United Kinidom, - 3d ed, reve - London : British Orthoptic Society, 1979.

24p. ; 26cm.

1. Ofthoptice - Terminology. I. Title.

RA
8
, W65
1992

world Healtk Orgasization
Basic documoatib : including ameadments adapted up to 31 October 1992,-- 39th ed. -- Goneva : WHO, 1992.
iv, 1012 p. ; 24 cm*
Includes index
ISBN 92-4-165034)-7

1. World healtk Orgaaizatioze I. Titl

RJ
21
.A4

American Pediatric Society.
Combined program and abstracts /of the 80th annual meeting of the American Pediatric Society and the 40th annual meeting of the Society for Pediatric Research held at/ Atlantic City, N.J., April 29-May 2,1970. /n.d/, 1970.
xlix, 376p. 23cm-
1. Pediatrics-Congressea* 1. Society for Pediatric Research* II, Title*

RA
407
.w6

World Health Organization
Health dimensions of economic reform.- Geneva : WHO, 1993*
ix, 68 p. : ill., maps ; 26 am.
ISBN 92-4-156146-7
1. Health status indicators 2* Econom development* I. Title.

SB
8i8
.w6

World Health Organization
Insect and rodent control through environmental management : a community action programme. -- Geneva : WHOs 1991.

vii, 107p- : ill- ; 30cm-
Bibliography : p. 106-107
ISBN 92-4-154411-2

1. Insect pests--Control. I. Title.

403
.A5

American Association of Clinical chemists.
Standard methods of clinical chemistry.

New York: Academic Press, 1953-

V, 24cm.

1. Chemistry, Medical and theoretical. I. Title.

ni
456
.D5
W?l

World Health Organization
Readings on diarrhoea : student manual- -- Geneva : WHO, 1992.

vil 147p- : ill- ; 30cm.
ISBN 92-4-154444-9

I. Diarrhea, Infantile. J. Title.

RA
644
.v4
W6

World Health Organization*
Control of sexually transmitted diseases*- Geneva: World Health Organization, 1985.
v, 110p.; 24 cm.
Includes biblio raphical references*
ISBN 92-4-154igg-9

l. Venereal diseases. I* Title.

RA
577
- E7
W6

World Health Organization
Alpha-cypermothrin. Geneva : WHO, 1992.
112 po : ills, 21 cm. -- (Enviromental health criteria 142 )
"Publioh*d under the joint sponsorship o the United Nations Environment Programmog the International Labour Organisation, and the World Health Organization.".
Bibliography : p. 719-86
ISBN 92-4-157142-x
1. Environ mental impact. I.Title.

RE
45i
OW6

World Health Organizatioa
Management of cataract im primary health car* services,-- Gen*va t WHOO 19900
43 pe : ill*, platen ; 24 cm.
ISBN 92-4-1544o8.2

1. Cataract. I. Title.

RA
8
.A275

World Health Organization.
Basic documents. lst. ed.- Geneva, 1951-

v, 24 cm.
Title varies : lst-6th ed., Handbook of basic documents.
Addenda accompany some editions,

1. World Health Organization. I. Title.

## APPENDIX IV

**Guide for a Discussion Held with the Cataloguing Team of the Cataloguing Department of the Library System**

1. What is the total number of staff of the department?
2. What are the responsibilities of the staff?
3. What are major activities in the department?
4. What is the present cataloguing system being followed in the department?
5. What are tools and facilities (computers, printers, CD-ROM databases, etc.) being used for cataloguing and card production?
6. What are major bottlenecks or problems you have had in the department with regard to cataloguing?
7. Specifically, what are the problems faced in production and maintenance of card catalogues?
8. Is there plan to automate the departmental activities?
9. What are the problems in this regard?

## APPENDIX V

**Guide for a Discussion Held with the Head of the Computer Center of the Library System**

1. What are the objectives of the Computer Section?

2. What are the services being rendered by the section to library users and staff of the library system?

3. Does the section have adequate and trained staff and facilities to achieve its objectives?

4. What are the number of hardware such as computers, printers, photocopiers, scanners, CD-ROM databases, etc. available at the Computer Section?

6. What are major bottlenecks or problems you have had in the Computer Section with respect to carrying out its activities?

7. Have you ever thought of automating the library functions?

8. What is your plan in automating the library house keeping activities? Would you make note of your future plan?

9. What are the problems in this regard?

## DECLARATION

**This thesis is my original work and has not been submitted for a degree in any other university.**

Kebede Hundie

Kebede Hundie

June 1997

**The thesis has been submitted for examination with our approval as university advisors.**

G. G. Chowdhury

June 1997

Sisay Fisseha

June 1997