# Digitizing Library Card Catalogue Retrospectively on Exploring Scanning and OCR Technique at University of Mumbai – Some Lessons.

***Nalini A Raja***

***Abstract***

*Retrospective conversion though the old and well-studied subject requires research in finding out an automated data capturing method. In this paper an attempted has been made to explore the digitization technique of scanning and converting the title and verso of the title page into optically recognized character for capturing bibliographic details of a document. The result of the digitization exercise done in this regard clearly indicate the inherent limitations and thereby caution the library professionals to explore this technique with alternative approaches till either change and consistency in publication style is appealing to the publishers or changes in OCR software takes place. The same technique however had been explored for specific subject search and was successful.*

**Keywords:** Digitization, Scanning, OCR, Retrospective conversion

## 1. Introduction

The University of Mumbai Library since its inception in the year 1880 has kept pace with the knowledge growth published in various forms as per the technological innovations in storing and publishing. Over the period of 134 years the Library has spread over two campuses – Fort and Vidyanagari, serves as the house of cultural heritage to the society by preserving all the forms of information. Thus library has variety of documents starting from MSS to magnetic disks viz.; Books (6,76,330), Bound Volumes of Journals(85,452), Theses (25,797), Microfilms (4,835), MSS (9,986), Non-Book material(293) and Pamphlets totaling to 8,02,693 in the year 2006. Some of the collections in different forms are designated as rare collection. Along with providing access to more than 7.6 lakhs documents the Library has taken measures to safe guard these rare collection. For the purpose the Library opted to implement 3M Library security system using 3M B2 Tattle Tape for books and continue to serve as a cultural heritage for the generations to come.

The access to either of the campus collection through traditional card catalogue was delaying information access. It was of utmost importance and necessity to provide the Online Public Access Catalogue to its Faculty, Students and all potential readers. This could be achieved by digitizing the Library Card Catalogue. It was a herculean task to convert the existing card catalogue into digital form giving access to all the bibliographic fields in MARC 21 format. The Library has used predefined data entry sheet given by the INFLIBNET Centre in SOUL 2.0.

The Library had an option of digitizing (computerizing) the bibliographic details either from each library document or Accession Register or Catalogue Cards. This can be achieved:

1. Either by first scanning the title page and verso of the title page of each Library document; then converting it into optically recognized characters

and storing as it is in the note area for searching using OPAC/WebOPAC module of SOUL 2.0 Library management software;

2. Or by manually entering the bibliographic information in the predefined datasheet in cataloguing module of SOUL 2.0 giving access to document/s through OPAC/webOPAC locally/ globally.

The main objective of the second option was to avoid all possible manual data entry errors such as typographical mistakes, leaving one of the bibliographic field blank and/or incomplete information. Over and above it is difficult to cope up with checking of data entered by data entry operators which was about 10000-15000 records a day!

## 2. Objectives

The objectives of the present study is:

1. To build capacity of Cataloguers in serving readers with sought information contained in the library documents; also by accessing live bibliographic database from their home or globally with actual status, library location (Churchgate or Santacruz) and shelving location.
2. To enhance the capacity of cataloguers in creating computerized library catalogue maintaining consistency and uniformity in rendering the bibliographic information enabling readers to access exhaustive list of all the works containing information sought by a reader.
3. To find out an efficient and effective method for retrospective conversion of bibliographic description of all the Library documents irrespective of their form and type;

## 3. Methodology

As described earlier the library had options of two methodologies:

1. Either scan the title and verso of the title page of a book; or
2. Manually key-in bibliographic description from three options available – Accession Register, Catalogue Cards or the document itself as a source of bibliographic information using predefined datasheet.

### 3.1 Method 1

The Library carried out pilot study of digitizing title and verso of the title page of 100 books from 25 publishers in various subjects. An attempt was made to convert them in structured textual data which then can be tagged in MARC21 format. The bibliographic information presented on title and verso of the title pages were observed and tabulated. These pages were also assessed for preceding and succeeding texts identifying and/or separating one bibliographic field from another. These pages were also converted into Searchable PDF, Word document and Excel Worksheet. They were assessed for whether the text was completely optically converted or partially and find out the probable causes of distortion.

The search option of OPAC as well as Web OPAC of Soul 2.0 offers search on total of fifteen fields viz.;

1. Title
2. Title + Subject
3. Title + Series.
4. Author
5. Accession Number
6. Corporate Name
7. Meeting Name

8. Uniform Title
9. Subject
10. Class Number
11. ISBN/ISSN
12. Publisher
13. Note
14. Series
15. Year of Publication

Amongst these fifteen fields No.2 and No.3 are combination of three individual fields enabling one to narrow down or broaden the search on maximum of six fields using only two Boolean operators i.e. three level search. The present study is restricted to digitize the title page and verso of the title page as the value of 60% of the searchable fields is obtained from the title or verso of the title page. Class No. and Subject are assigned by the classifier whereas the Accession no. is assigned by the acquisition section.

Thus the study was carried out exploring the conversion of PDF file into OCR using eCopy PDF Pro Office 6.2 software. This software enables converting PDF files into:

1. XPS Document;
2. Word document;
3. Word Form;
4. Excel Spreadsheet;
5. PowerPoint Presentation;
6. WordPerfect Document;
7. MRC PDF;
8. Searchable PDF; and
9. Unicode Text.

For the study three options were explored viz; converting the scanned (PDF) files into searchable PDF, Word Document and Excel Spreadsheet. The first attempt was made to copy from searchable PDF. But as the name suggests it does not allow to select and copy. Therefore the searchable PDF was saved as text file i.e.in Notepad. To avoid this process scanned file was directly converted into Word document. But SOUL 2.0 does not support directly copying from this Word document file. This exercise of converting scanned document into OCR was further explored to convert into Excel spreadsheet to check whether the data further can be converted into some structured form to directly map into MARC 21.

The option of entering the data manually from catalogue Cards or Accession Register was also studied for the advantages as regards adhering to standards for data exchange and retrieval in a structured format. The methodology for this option is described in the following section.

### 3.2 Method 2

In method 2 where in manual keying in from subject catalogue card was opted because:

1. The data entry operators were not library professionals having no cataloguing knowledge;
2. Entry from Accession register was omitted as the purpose of the accession register is to record

   i. Each library document arrived either on gratis/donation or purchase;

   ii. The bibliographic details in brief, the document suggested by, budget, source details (supplier name, place); the invoice /bill/cash memo No. and date; price with original currency and amount in Rs. with the conversion rate. Thus

the accession register was falling short for actual and complete bibliographic information resulting into inability to adhere to various standards like ISBDs, AACRII and MARC21;

iii. Remarks pertaining to document either transferred to department or lost or replaced or withdrawn from circulation.

iv. Collation (size, no of pages, binding) details and ISBN.

Under the circumstances the Library decided to capture bibliographic details from the subject cards so that the books entered are taken from the shelf, tattle taped and barcode label pasted. The Collation, ISBN and Remarks are taken from the Accession Register.

The collection of 16 Department Libraries merged into University Library was reaccessioned to identify these collections uniquely by using prefixes/suffixes as given in Annexure-1.

The multivolume documents like gazetteers, Census reports, publications under one series, publications under one uniform title have been given same Accession No. even though having different author and title. These Accession Nos. were suffixed with V1-n P1-n; where V stands for Volume No. and P for Part No. The Gazetteers accession numbers were suffixed with gz1-n. or simply suffixing by a-z.

The tag generated was pasted on the title page also to incorporate the reaccessioned numbers. In absence of this a list of these Accession Nos. with Class No. was to be provided to the respective Section for making corrections on books (accession no.). This was mandatory before bar-code generation and tattle tapping the books with respective labels.

## 4. Observations

4.1 Method 1

4.1.1 The Digitized title page and its verso were checked for the presence of the bibliographic fields viz.;

4.1.1.1 Title/Uniform Title

4.1.1.1.1 Title appears first in 88% of the books scanned whereas remaining 12% of books have title in second position.

4.1.1.2 Author/Corporate Name/Meeting Name

4.1.1.2.1 Author in 12% of books published by Springer precedes over all other bibliographic description.

4.1.1.2.2 In 88% of the books Author occupies second or third position.

4.1.1.3 Publisher. All books have publisher on the title page

4.1.1.4 Year of Publication. Except four books year of publication appear on the verso of the title page. The year of Publication constitutes 16% of the searchable bibliographic fields. This field value is dispersed throughout the page and rest is again of no value in retrieving the library document as far as this field is concerned.

4.1.1.5 Series

4.1.1.5.1 Series statement appears on the fly leaf of the book i.e. only 16% of the bibliographic fields taken into consideration. It constitutes first one or two lines only. The rest is of no value in retrieval.

4.1.1.5.2 Only one book published by Cambridge University Press has series on title page and precedes over all other bibliographic description.

4.1.1.5.3 Presence of Series adds number of pages to be scanned. However Series constitutes only 13% of the searchable bibliographic fields.

4.1.1.6 ISBN Except for four books ISBN appears on the verso of the title page.

4.1.1.6.1 Except for thirty nine books more than one ISBN numbers are given. These pertain to either different binding or format e.g. Hb/e-book, electronic/ eBook, ebook, (HB), (PB), (cloth), blank/(pbk), ISBN (US-Hb, US-Pb, India Hb, India Pb), (hbk), (ebk), (hb)/(pb)/(eb), (pbk. : acid-free paper), Print ISBN, hardback/paperback. This leaves no scope to pick up the ISBN automatically as the decision regarding format and type of binding can only be taken by the person cataloguing and not from the scanned or OCR page.

4.1.2 The scanned pages were examined and it was observed that different publishers have different style and it changes with time. The observation regarding searchable PDF are as follows:

4.1.2.1 All books were searchable, but the Library stamp, book tags pasted on the title page were not searchable either partly or totally;

4.1.2.2 Title page of one book was inverted;

4.1.2.3 Two titles were not OCRed because of the printing fonts used; (Figure 1 and Figure 2);

4.1.2.4 This format of file did not support pasting the content in SOUL 2.0 worksheet therefore conversion in word was explored;

4.1.3 In Word document the library stamp, call no. and accession number appear as image on the title page and at times they distort the succeeding text or text opposite to stamp. Copying in SOUL 2.0 worksheet is also not supported. This format has distorted total of 64 books. These 64 books were examined to check whether the title page was completely distorted or partially. The observations are as follows:

4.1.3.1 Twenty five titles have been distorted;

4.1.3.2 Three subtitles have been distorted;

4.1.3.3 Twenty five publishers have been distorted;

4.1.3.4 Eleven authors have been distorted;

4.1.3.5 This format also is not supported by SOUL 2.0 therefore Excel spreadsheet was also tried.

4.1.4 The Excel spreadsheet file gives result in a structured format i.e. one row for each line and blank rows for the blank spaces with distortion;

4.1.4.1 The text not converted into OCR appear as image and on copying in SOUL 2.0 leaves blank spaces for that portion;

4.1.4.2 The two pages are stored as different sheet in one workbook;

4.1.4.3 The rest of observation is just the same as word viz.; 4.1.3.1 to 4.1.3.4

### 4.2 Method 2

Shortfalls of method 1 lead to opt for the second Method (Method 2) i.e. manually keying from catalogue card.

4.2.1 The total no. of Names (surname of the author/Corporate name/Meeting name), publisher and class no. were searched and the unique values for each field was also searched using SQL query as SOUL 2.0 uses SQL as its database. Thiese values are recorded in Annexure 2.

4.2.2 The alphabetical list of Author surname count and unique surname count was again carried out using SQL query to highlight the importance of creating authority files.

From the above observations following conclusions were drawn.

## 5. Conclusion

5.1 From the above observations it is concluded that the digitization technique – Method 1 has limited scope in retrospective conversion.

5.2 Data entered in Method 2 requires checking for accuracy, however exporting and importing from Union catalogues or Library of congress is one of the option which can be tried or insisted while outsourcing. In large Libraries only sample testing is possible. Under the circumstances the following steps are of help in reducing the manual data entry errors.

5.3 Data entry directly in the library management software used should be insisted

5.4 The authority files for Author, Publisher and Class No as listed in Annexure 2 be created/updated first to:

5.4.1 Speed up the data entry work by avoiding the redundancy of the values which is significant as listed in annexure 2;

5.4.2 To maintain the consistency and uniformity in rendering name or class numbers of multiple works of either of the fields mentioned above; as well as retrieving the documents exhaustively from the field searched.

5.4.3 The authority files for Name/Corporate name/ Meeting, Publisher and Class number is obvious from the result of Annexure 2 and Annexure 3. This not only reduces checking of retrospectively converted records to only remaining fields but also reduces the keying in of author's name to half as tabulated in Annexure 3. The Publisher and Class no. counts have not been taken but this result clearly shows that it is worth doing. On creating/updating these authority files the checking of data is reduced to half i.e. 50% of the searchable fields are error free!

5.4.4 The strength of digitizing the bibliographic fields using scanning and converting them into optically recognized characters cannot be ruled out mainly due to error free data capturing. For the reason the same type of study was carried out in digitizing the content and index pages for enhancing the subject search in terms of precision1.

5.4.5 The scanning and OCR technique can be used as complement to the bibliographic database by copying these pages in the note area and not as a substitute. The main reason being limitation of database exchange and generating bibliography. The subject bibliography can be generated by the class number. In addition it requires the use the multiple sources of bibliographic description like catalogue cards for call number. Accession register required for accession number, collation, budget, price, location etc.

From the above observations and conclusions it is summarized that the objective of finding out an efficient and effective method for retrospective conversion of bibliographic description of all the Library documents requires more than one technique and sources.

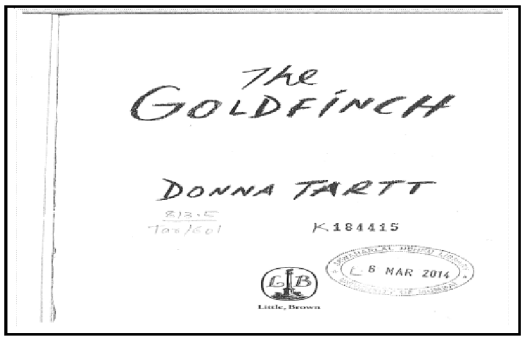


**Figure: 1**

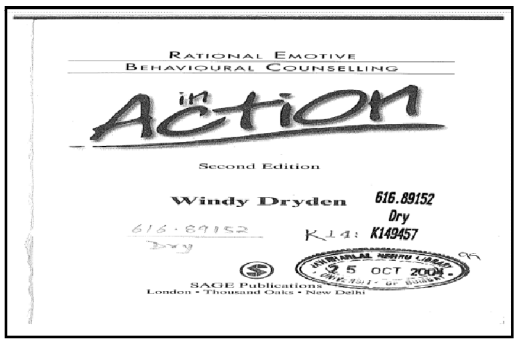


**Figure: 2**

Annexure - 1

**Accession Registers with description of Department**

| Department Description | Accession No. From | Accession No. To | Total No.of books | Acc Nos Prefixed as | Acc Nos suffixed as |
|---|---|---|---|---|---|
| KS | 1 | 7426 | 7426 | KS | |
| School of Eco. & Soc. | 1 | 7762 | 7762 | | /A |
| Geography | 1 | 435 | 435 | | /G |
| Economics | 1 | 23777 | 23777 | | /E |
| Politics | 1 | 8538 | 8538 | | /P |
| ISBMS | 1 | 507 | 507 | | /ISBMS |
| Sociology | 1 | 13684 !! | 13684 | | /S |
| Statistics | 1 | 6031 # | 6031 | | /St |
| ICSSR(IC) | 1 | 2847 | 2847 | | /IC |
| FL (Fort Library at Churchgate) | 1 | *386922 ! | 386922 | | |
| JNL (Jawaharlal Nehru Library at Santacruz) | 1 | 150000 | 150000 | K | |
| JNL | 300001 | *311254 | 11254 | K | |
| JNL (Presentation) | 150001 | *173979 !!! | 23979 | K | |
| ISAE | 1 | 12709 | 12709 | | /ISAE |
| Applied Psychology | 1 | 3642 | 3642 | | /AP |
| Law | 1 | 3776 | 3776 | | /L |
| LS(FL) | 1 | 13041 | 13041 | | /LS |
| Club House Reading Room Collection | | | | | |
| Total Book Collection | | | 676330 | | |

*as on 08/06/06

! 69384, 129528-129537, 131400-131431, 283637, 283877, 305365-305370, 305682-305688 Missing

!! 12313/S missing

!!! K15344 missing

# not entered as either books were not received by the library or they correspond to journals/question paper/pamphlets/reports

**Annexure 2**

**List of authority files to be updated before Retrospective Conversion**

| Description of Fields considered for Authority files to be updated before retrospective conversion | Total No of values under respective field | Total No. of Unique values under respective field |
|---|---|---|
| Author - Tag 100 $a (Personal Name) + 110 $a (Corporate Name)+ 111 $a (Meeting Name) | 756719 | 277018 |
| Publisher - Tag 260 $b | 1948945 | 171426 |
| Class No. - Tag 082 $a | 769093 | 82747 |

**Annexure 3**

**The alphabetical list of Author surname count and unique surname**

| Column 1 | Column 2 | Column 3 |
|---|---|---|
| Author surname starting with Alphabet | Total no. of authors with surname starting with alphabet in column I | Total no. of authors having unique surname |
| A/a | 36338 | 13781 |
| B/b | 45841 | 23899 |
| C/c | 30233 | 15700 |
| D/d | 29767 | 15219 |
| E/e | 10416 | 4392 |
| F/f | 12452 | 7245 |
| G/g | 29257 | 14645 |
| H/h | 26662 | 14843 |
| I/i | 3756 | 2197 |
| J/j | 18644 | 9664 |
| K/k | 31727 | 15920 |
| L/l | 18322 | 10229 |
| M/m | 45815 | 24476 |
| N/n | 13486 | 7272 |
| O/o | 4453 | 2677 |
| P/p | 30395 | 15670 |
| Q/q | 989 | 600 |

| | | |
|---|---|---|
| R/r | 28344 | 15070 |
| S/s | 60501 | 32106 |
| T/t | 17473 | 8999 |
| U/u | 2254 | 1331 |
| V/v | 12267 | 6305 |
| W/w | 17296 | 9587 |
| X/x | 89 | 59 |
| Y/y | 2605 | 1483 |
| Z/z | 2264 | 1408 |
| Total No. of authors | 531646 | 274777 |

**Reference**


1. RAJA Nalini A. Digitized contents and index pages as alternative subject access fields. (2012). In Neelameghan, A., & Raghavan, K. S. (Eds.). Categories, contexts and relations in knowledge organization: Proceedings of the Twelfth International ISKO Conference (Mysore, India, August 6-9, 2012). 12th International ISKO conference on Categories, Contexts and Relations in Knowledge Organization, ISKO (India Chapter), Mysore, India, 6-9 August 2012, organized by SRIELS & University of Mysore. p.
2. ALAN Danskin. (2004). Mature consideration: developing bibliographic standards and maintaining values. New Library World, Vol. 105 (3/4), 113 – 117
3. ANN Chapman. Owen Massey. (2002). A catalogue quality audit tool. Library Management, Vol. 23 (6/7), 314 – 324
4. BIA Alejandro. Muñoz Rafael, and Gómez Jaime. (2010) DiCoMo: the digitization cost model. International Journal on Digital Libraries, Vol. 11 (2), 141-153
5. BLANKE, Tobias. Bryant Michael and Hedges Mark. (2012). Ocropodium: open source OCR for small-scale historical archives. Journal of Information Science, Vol. 38 (1), 76-86.
6. CHRISTELLE Creff. (2002). Opening interlending services to end users: the Catalogue Collectif de France. Interlending & Document Supply, Vol. 30 (3), 126 – 129
7. KARIC, Miran; KRPIC, Zdravko; MARTINOVIC, Goran. (2013). Optical Character Recognition On Grid And Multi-Core Systems - Performance Analysis Technical Gazette, Vol. 20 (4), 647-653.
8. LARS E. Leon. (2004). Linking four libraries 9,012km apart: steps to global resource sharing. Interlending & Document Supply, Vol. 32 (1) 30-37
9. MANDELL, Laura. (2012). Texas A&M Developing Ocr For Early English Texts, Library Journal, Vol. 137 (20), 18-19.

10. OGHENEVWOGAGA Benson Adogbeji. Esharenana E. Adomi, (2005) Automating Library Operations at the Delta State University Library, Nigeria. Library Hi Tech News, Vol. 22 (5), 13-18

11. PORTIA Bowen-Chang, Yacoob Hosein, (2009) Cataloguing training at the University of the West Indies, St Augustine, Library Review, Vol. 58 (2), 97 - 108

## About Author

**Ms. Nalini A Raja,** University of Mumbai, J N Library, Vidyanagari, Satacruz (W), Mumbai. Email: nalinir99@gmail.com