Automatically Creating Biomedical Bibliographic Records from Printed Volumes of Old Indexes

Daniel X. Le and George R. Thoma
National Library of Medicine
Bethesda, MD 20894

ABSTRACT

To provide online access to citations from old hardcopy indexes published from 1879 through 1965, an R&D division of the National Library of Medicine (NLM) is developing an automated system to convert bibliographic information in volumes of the printed Quarterly Cumulative Index Medicus (QCIM) to machine-readable form for inclusion in the OLDMEDLINE® database. The system processes images scanned from a QCIM volume, segments and labels the image records, identifies multiple occurrences of the same record in the volume, and creates unique citation records. The record segmentation and labeling technology is based on a smearing bottom-up approach for text block segmentation, the document page layout formats, and a set of rules for record labeling that is derived from the QCIM document format guideline. Since bibliographic information can be arranged as both "author entries" and "subject entries" in a QCIM document, the duplicate records have to be detected and combined to create a single unique citation. The duplicate records are identified based on matching "cross-reference" information such as author names, journal title abbreviation, volume, pagination, month, and year among different entries of the same citation. The "cross-reference" information can also be used to correct OCR errors resulting in improving the quality of citations created.

The performance of the system has been evaluated using a QCIM volume published in 1929 that consists of 95,717 citation records. Evaluation shows the technical and cost feasibility of building the proposed data conversion system.

Keywords: Quarterly Cumulative Index Medicus, Document image analysis, Document scanning, OLDMEDLINE database, National Library of Medicine.

1. INTRODUCTION AND BACKGROUND

As the world's largest medical library, the NLM's mission is to collect, organize, preserve, and disseminate medical information. The Library is an important source of information for biomedical scientists, health professionals, and the lay public around the world. Advances in computer and communications technologies and the rapid growth of the Internet and World Wide Web technologies help NLM offer quick and cost-effective dissemination of medical information to consumers.

In 1971, NLM introduced MEDLINE, an online searchable database access to citations for journal articles from 1966 forward [1]. Due to format differences and technical difficulties in accurately converting old paper-based document into electronic format, citations to earlier articles were not included in MEDLINE. Later, in response to the increasing demand to search earlier journal articles and in efforts to collect and maintain a comprehensive bibliographic collection from the past, NLM developed OLDMEDLINE in 1996 for the medical literature published from 1879 through 1965. Currently, the OLDMEDLINE database consists of over 1.5 millions article citations converted from hardcopy indexes published from 1953 to 1965 [2].

NLM will continue to convert older printed medical indexes to electronic format and the goal is to cover all citations going back to 1879. However, the current conversion method is completely manual and labor-intensive, requiring the keyboarded entry of citations. Furthermore, since the same citations can appear under different entries in the indexes, there are a lot of duplicate records. As a result, the conversion is very slow and the cost is high because citation entry operators have to spend time keying in, as well as resolving, duplicate records.

To speed up the conversion process, to reduce manual data entry costs, and to prevent duplications, we propose an automated system to convert bibliographic information from 60 volumes of the printed QCIM from 1927 through 1956 to machine-readable form for inclusion in the OLDMEDLINE database. The system processes images scanned from a QCIM volume, segments and labels the image records, identifies multiple occurrences of the same record, and creates unique citation records. The record segmentation and labeling technology is

based on a smearing bottom-up approach for text block segmentation [3, 4], the document page layout formats, and a set of rules for record labeling that is derived from the QCIM document format guideline.

The rest of this paper is organized as follows. Section 2 provides a brief system overview. Section 3 presents the QCIM page layouts, document format guidelines including "author entries" and "subject entries" citations, and cross-reference information. Section 4 describes the process of creating biomedical bibliographic records. Section 5 gives experimental results, and Section 6 contains a summary.

2. SYSTEM OVERVIEW

The system proposed in this paper consists of multiple workstations of two types: scanning and reconciling (text verification). In addition, the system requires five servers: a network file server, an OCR server, a record segmentation and labeling server, a record duplication detection server, and a unique citation record creation server. All workstations and servers are networked via a LAN.

Briefly, the system works as follows. An operator scans all pages of a hardcopy QCIM document, and the bitmapped image files are sent to the network file server. The OCR server performs text conversion, and produces a text file for each scanned page. The record segmentation and labeling server segments the text lines into records, and labels the records as "author entries", "subject entries", "heading entries", "sub-heading entries", "reference entries", or "other entries". The record duplication detection server performs word matching among author and subject records to identify duplicate ones. The unique citation record creation server analyzes text data in the records, and corrects OCR errors using "cross-reference" information and eliminates duplicate information between similar records to generate unique citations. At this point, the OCR output text of the unique citation record and its corresponding bitmapped image file are available for validation and reconciling by a human operator.

3. PAGE LAYOUTS, FORMAT GUIDELINE, AND CROSS-REFERENCE INFORMATION

Each QCIM volume consists of two sections: one for books and the other for periodical literature. The periodical literature section is divided into three subsections: list of publishers, list of journals indexed, and a periodical index. Here, we only capture and process the pages of the last two subsections of the periodical literature section.

In the subsection containing the list of journals indexed, each journal is presented alphabetically with its abbreviation followed by its full title. However, only the journal abbreviations are used later in the periodical index subsection. The paragraph format of each journal entry record is left-aligned with a hanging indentation of about 0.25". The journal abbreviation is separated from its full title by delimiter(s) and its font style is bold. An example of journal title entries is shown in Figure 1.

The periodical index is arranged alphabetically with author and subject entries. For an author entry, its citation starts with the author name(s) and the title of the article in the original language. It is then followed by the journal title abbreviation and ends with the volume, pagination, month, and year. The paragraph format of each author entry record is also left-aligned with a hanging indentation of about 0.25". The author name is capitalized and its font style is bold. Multiple entries of the same author name(s) are arranged using the same paragraph format except that their first line is indented about 0.07". An example of an author entry is shown in Figure 2.

For a subject entry, the citation is in English and grouped under the subject heading. The citation usually starts with the title of the article that is often summarized or expanded to emphasize important points. The paragraph format of each subject entry record is left-aligned with a hanging indentation of about 0.25". However, if the citation is further sub-classified then the title of the article is preceded with a subheading. The subject entry citation is then followed by the names of the authors embedded within square brackets and followed by the journal title abbreviation. Similar to the above author entry, it ends with the volume, pagination, month, and year. An example of subject entries is shown in Figure 3.

Since citations can appear as both "author entries" and "subject entries" in a QCIM document, the duplicate records have to be detected and combined to create a single unique citation. Furthermore, most QCIM documents are old and printed on low quality paper. As a result, many OCR errors occur, which could require labor-intensive manual correction during the reconciling step. However, the multiple occurrences of each citation under different entries help to reduce the OCR errors.

As described in the QCIM document format guideline, the "cross-reference" information between "author entries" and "subject entries" includes author names, journal title abbreviation, volume, month, and year. Additionally, the journal title abbreviation which is found in the list of journals indexed subsection can be used as another cross-reference source to increase the confidence of the journal title abbreviation OCR result. Based on this "cross-reference" information, the proposed automated data conversion system can automatically (1) resolve duplicate records, (2) correct OCR errors, and (3) create unique citations for the OLDMEDLINE database. As a result, the system is able to speed up the conversion process by eliminating duplicate records entries and reducing manual data entry costs.

4. PROCESS OF CREATING BIOMEDICAL BIBLIOGRAPHIC RECORDS

The process of creating biomedical bibliographic records from printed volumes of old indexes described here consists of nine steps: (1) collect the document information, (2) determine the best brightness and contrast setting for scanning the document, (3) scan all pages in the list of journals indexed and the periodical index subsections, (4) collect the page layout-specific information, (5) update the list of journals indexed (6) apply document analysis and labeling processing for pages in the periodical index, (7) conduct quality control on the pages in the periodical index, (8) detect and resolve duplicate records using the "cross-reference" information, and (9) finally, create and reconcile unique citation records. Each step is discussed in detail below.

4.1 Collect the document information

In this step, the document information is collected including volume, published months and year. Since pages containing in the list of journals indexed and the periodical index subsections are scanned, the pagination of the first and last pages of each subsection are recorded for verification and identification purposes. The directories for storing database files, page image files, zoned files, and OCR files are also defined here.

4.2 Determine the best brightness and contrast setting for scanning the document

Since many QCIM documents are old and printed on low quality paper, the selection of the best scanner setting for the brightness and the contrast helps to reduce the number of OCR errors and to improve the quality of the results. The scanner setting selection procedure is as follows:

1. Select a page in the periodical index subsection as a test page.
2. Scan the test page using the normal brightness (0) and the normal contrast (0).
3. OCR the entire image test page and automatically select 25% of text zones to be manually reconciled.
4. Automatically scan and OCR the same test page with different settings of the brightness and contrast values.
5. Analyze the text zones for the different settings and select the best setting giving the minimum number of OCR errors.

4.3 Scan all pages in the list of journals indexed and the periodical index subsections

Using the scanner setting selected in the above step, all pages in the list of journals indexed and in the periodical index are scanned and deskewed. At the end of this step, scanning of the current QCIM document is completed.

4.4 Collect the page layout-specific information

The page layout format for the list of journals indexed subsection is different from that of the periodical index subsection. Therefore, there are two procedures, one for each subsection.

The procedure for the list of journals indexed subsection is as follows:

1. Automatically select a scanned page (excluding the first and last pages) to perform OCR and layout analysis, to create headers, two column zones, text lines and segmented records.
2. Display header and column zones, text line blocks, and record blocks for operator's confirmation and validation.
3. Derive the page layout-specific information based on information in the above two steps:
   a. The locations and sizes of the left/middle/right headers.
   b. The left/right column width and height, and the gap between two columns.
   c. The horizontal and vertical distances between the headers and the left/right columns.
   d. The average text line height.
   e. The hanging indentation distance.
   f. The relative zone locations of journal title abbreviation.

The procedure for the periodical index subsection is as follows:

1. Automatically select three scanned pages (excluding the first and last pages) to perform OCR and page layout analysis, to create header zones and two column zones, to segment text lines and records, and to label records.

2. Display header and column zones, text line blocks, record blocks and labels for operator's validation.
3. Derive the page layout-specific information based on information in the above two steps:
   a. The locations and sizes of the left/middle/right headers.
   b. The left/right column width and height, and the gap between two columns.
   c. The horizontal and vertical distances between the headers and the left/right columns.
   d. The average text line height.
   e. The hanging indentation distances.

4.5 Update the list of journals indexed

In this step, all pages in the list of journals indexed subsection are OCRed. Using the page layout-specific information obtained in the above (step 4.4), each page is segmented, zoned, and labeled. The journal title abbreviations are extracted and compared against a predefined list of journals indexed. If there are new journal title abbreviations, they are presented for the operator to confirm. At the end of this step, the predefined list of journals indexed is updated with new journal title abbreviations, and any obsolete titles are removed from the list.

4.6 Apply document analysis and labeling processing for pages in the periodical index

During this step, all pages in the periodical index subsection are OCRed and followed with a page layout analysis including segmentation, zoning, and labeling operations based on the page layout-specific information collected in step 4.4 above. The results are labeled records that are ready for quality control and for the matching operations. The record segmentation and labeling technology is based on a smearing bottom-up approach starting from characters, words, text lines, and records. The smearing distances among these components are derived from the page layout-specific information. The QCIM document page layout format and guideline are used for decisions on creating and labeling records. The records are labeled as "author-citations", "subject-citations", "headings", "sub-headings", "references", or "others".

4.7 Conduct quality control on the pages in the periodical index

In order to improve the page and record segmentation, each page in this subsection is presented to the operator to confirm and correct any obvious mistakes made during the automated document analysis and labeling process. The system displays headers, columns, indentation vertical lines, record separation lines, and record labels. The operator can confirm or correct the results and if there are any corrections then the entire page is marked for re-processed (repeating steps 4.6 and 4.7).

4.8 Detect and resolve duplicate records using the "cross-reference" information

At this point, all records labeled as "author-citations" and "subject-citations" in the periodical index subsection are matched using their "cross-reference" information to identify duplicate records. As described in Section 3, the "cross-reference" information among these records consists of author names, journal title abbreviation, volume, month, and year. Since these citation records were originally created through scanning and OCR processes, there are OCR errors to be corrected. The detection of duplicate records helps to correct OCR errors and to improve system performance.

4.9 Create and reconcile unique citation records

Finally, unique citation records are created by resolving duplicate records and combining their contents. These unique records are reconciled by the operators to remove any remaining OCR errors. After being confirmed and validated by the operators, these unique records are uploaded to be included in the OLDMEDLINE database.

Figures 4, 5, 6, and 7 summarize the processes of matching duplicate records, correcting OCR errors, and combining records to create a unique citation record. The bold characters signify low confidence OCR output requiring confirmation or correction.

5. EXPERIMENTAL RESULTS

A prototype of the automated system proposed in this paper has been implemented and an experiment has been conducted with 8-bit grayscale document images scanned from QCIM volume 5 published in 1929 [5]. All pages used in this experiment are 8.5 x 11 inches in size and scanned at 300 dpi resolution.

The preliminary experiment result shows that there are 95,717 records in which 74,186 records have at least one match, 13,222 records have no match, and 8,309 reference records. The details of the 74,186 matched records are presented in Table 1. The table shows that there are 16,832 records having one duplication and one record that has up to 23 duplications. The large number of records and their matches demonstrate that our proposed automated

system is capable of detecting duplicate records and thereby saving labor costs.

6. SUMMARY

This paper describes an automated system designed to create OLDMEDLINE citations from 60 volumes of the printed QCIM published from 1927 through 1956. The advantages of the proposed system over manual keyboard entry approach are the automatic elimination of duplicate records, reducing labor costs, and improving accuracy and speed performance. The experimental results on QCIM volume 5 published in 1929 consisting of 95,717 citation records are very encouraging and they show that the system is capable of labeling records and detecting duplicate records with a very high accuracy.

Moreover, in this prototype work, the cross references used for labeling records and matching duplicate records do not include the journal title abbreviations because the predefined list of journals indexed was not available at the time of the experiment. As presented in Section 3, the journal title abbreviations are listed in the list of journals indexed subsection and they appear in every citation entry. Therefore, they are considered a very reliable source of information that can be used for labeling and matching purposes. Next, we plan to modify our program to use the journal title abbreviations as an additional cross reference feature in order to refine our current automated system.

7. REFERENCES

[1] "OLDMEDLINE", www.nlm.nih.gov/databases/databases_oldmedline.html

[2] "OLDMEDLINE Citations Join PubMed®", www.nlm.nih.gov/pubs/techbull/so03/so03_oldmedline.html

[3] L. A. Fletcher and R. Kasturi, "A robust algorithm for text string separation from mixed text/graphics images," IEEE Trans. on PAMI 10: 910-918 (1988).

[4] L. O'Gorman, "The Document Spectrum for Page Layout Analysis," IEEE Trans. on PAMI 5(11):1162-1173 (1993).

[5] Quarterly Cumulative Index Medicus, American Medical Association, Chicago, volume 5, January-June, 1929.

Records	Matches	Records	Matches	Records	Matches
16,832	1	24	7	1	13
8,947	2	12	8	1	15
2,328	3	9	9	2	17
549	4	5	10	1	23
117	5	2	11		
50	6	1	12		

Table 1

Rev. path. comp.--Revue de pathologie comparée et d'hygiène générale. 7, rue Gustave-Nadaud, Paris 16e.
Rev.paulista med.--Revista paulista de medicina. Caixa Postal 2103, São Paulo.

Figure 1: An example of journal title entries.

BLONDIN, M., Sur le choix du greffon dans les greffes de nerfs périphériques, Médecine **9**: 999-1000, Oct. '28

Figure 2: An example of an author entry.

BLOOD, calcium--Continued
—tetany and blood calcium after thyro-parathyroidectomy in goat, [E. Larson & L. A. Elkourie] Proc.Soc. Exper.Biol.& Med. **26**: 210-213, Dec. '28
carbon dioxide, combining properties; capacity of CO_2 for combining with reduced hemoglobin and oxy-hemoglobin; preliminary report, [O. M. Henriques] Biochem.Ztschr. **200**: 18-21, '28
—CO_2 absorption curve and buffer value of blood in physical hyperthermia, [M. Garcia Banus] Am.J. Physiol. **88**: 709-723, May '29

Figure 3: An example of three subject entries.
[Heading: BLOOD, calcium]
[Subheading: carbon dioxide]

Page 69

AALSMEER, W. C., and WENCKEBACH, K. F., Herz und Kreislauf bei der Beri-beri-Krankheit, Wien.Arch.f. inn.Med. **16:** 193-272, Jan. '29

Page 170

BERIBERI

—*heart and circulation in beriberi,* [W. C. Aalsmeer & K. F. Wenckebach] Wien.Arch.f.inn.Med. **16:** 193-272. Jan. '29

Page 265

CARDIOVASCULAR DISEASES

—heart and circulation in beriberi, [W. C. Aalsmeer & K. F. Wenckebach] Wien.Arch.f.inn.Med. **16:** 193-272, Jan. '29

Figure 4: Duplicate records extracted from pages 69, 170, and 265 of QCIM volume 5, 1929

AALSMEER, W. C., and WENCKEBACH, K. F., **Herz** und **K**reislauf **bei** der **B**eri-beri-Krankheit, Wien.**Arch**.f. **inn.Med.** 16: 193-272, Jan. '29

-he**art** and circulation in **beriberi, [W. C. Aalsllieer & K.** F. Wen**ckebachl] Wien.Arvh.f.inn.Med.** 16: **193-**272. **.an. '29**

-heart and circulation in beriberi, **[**W. C. **Aals**meer & K. F. W**e**nckebach] **Wi**en.**A**rch.f.**i**nn.**M**ed. 16: 193-**2**72, Jan. '29

Figure 5: OCR outputs

AALSMEER, W. C., and WENCKEBACH, K. F., **Herz** und **K**reislauf **bei** der **B**eri-beri-Krankheit, Wien.**A**rch.f. **i**nn.**M**ed. 16: 193-272, Jan. '29

-heart and circulation in beriberi, **[**W. C. **Aalsllieer** & K. F. Wenckebachl] Wien.**Arvh**.f.**i**nn.**M**ed. 16: 193-272. **.an.** '29

-heart and circulation in beriberi, **[**W. C. Aalsmeer & K. F. Wenckebach] Wien.**A**rch.f.**i**nn.**M**ed. 16: 193-272, Jan. '29

Figure 6: Correct OCR errors

Authors: AALSMEER, W. C., and WENCKEBACH, K. F.
Original Title: **Herz** und **K**reislauf **bei** der **B**eri-beri Krankhe**i**t
Translated Title: heart and circulation in beriberi
Journal Title Abbreviation: Wien.**A**rch.f.**i**nn.**M**ed.
Volume and Issue: 16
Pagination: 193-272
Published Date: Jan. '29
KeywordList: Beriberi, Cardiovascular diseases

Figure 7: Create a unique OLDMEDLINE citation