Automated data entry system: performance issues

George R. Thoma, Glenn Ford
National Library of Medicine, Bethesda, Maryland 20894

ABSTRACT

This paper discusses the performance of a system for extracting bibliographic fields from scanned pages in biomedical journals to populate MEDLINE®, the flagship database of the National Library of Medicine (NLM), and heavily used worldwide. This system consists of automated processes to extract the article title, author names, affiliations and abstract, and manual workstations for the entry of other required fields such as pagination, grant support information, databank accession numbers and others needed for a completed bibliographic record in MEDLINE. Labor and time data are given for (1) a wholly manual keyboarding process to create the records, (2) an OCR-based system that requires all fields except the abstract to be manually input, and (3) a more automated system that relies on document image analysis and understanding techniques for the extraction of several fields. It is shown that this last, most automated, approach requires less than 25% of the labor effort in the first, manual, process.

1. INTRODUCTION

There are two principal and obvious reasons why automated data entry is of interest: first, the gradual rise of labor costs; and second, the unrelenting increase in the amount of data that needs to be entered into databases from paper-based information. The vast majority of the hundreds of databases produced in every discipline rely on laborious keyboard entry of bibliographic information from articles in journals, e.g., article title, author names, institutions, abstract, dates, page numbers, etc. Image analysis and understanding techniques provide the basis for the development of automated systems that promise a cheaper alternative to keyboarding, and a more timely availability of bibliographic data for the public.

We have developed a system, *Medical Article Records System* or MARS, which consists of both automated and operator-controlled subsystems. The first generation system, MARS-1, relied on manual entry of all bibliographic fields except for the abstract which was captured by OCR, and so could be considered equivalent to many OCR-based systems used in production. The current second generation system, MARS-2 (shown in Fig. 1), employs image analysis techniques to capture, in addition to the abstract, also the article title, author names, and institutional affiliations[1]. Fig. 1 shows the automated processes as boxes with thin boundaries, and manual workstations with thick boundaries. The workflow is initiated at the CheckIn stage where a supervisor scans the barcode on a journal issue arriving at the production facility. This barcode number, called the “MRI”, is routinely affixed to every journal issue, and therefore serves as a unique key to identify the issue, all the pages scanned in that issue, and indeed the outputs of all processes performed on those page images. The scanning operator captures the first page of every

article in the issue, since this page contains the fields we seek to extract automatically. The resulting TIFF images go into a file server and associated data into the MARS database for which the underlying DBMS is Microsoft's SQL Server. The OCR system accesses the TIFF images and produces the corresponding text as well as other data descriptive of the text characters such as bounding boxes, attributes (bold, italic, underlined), confidence level, font style and size, and others. The automatic zoning (Autozone) module[2] then blocks out the contiguous text using features derived from the OCR output data, followed by the automated labeling (Autolabel) module[3-6] that identifies the zones as the fields of interest (article title, author names, affiliations, abstract). The Autoreformat module[7] then organizes the syntax of the zone contents to adhere to MEDLINE conventions (e.g., author name *John A. Smith* becomes *Smith JA*).

At this point, two lexicon-enabled modules operate on the data to reduce the burden on the operator performing the final checking and verification of the data: ConfidenceEdit[8] modifies the incorrect confidence levels assigned to the characters by the OCR system, and PatternMatch[9,10] corrects institutional affiliations whose text is frequently recognized incorrectly by the OCR system.

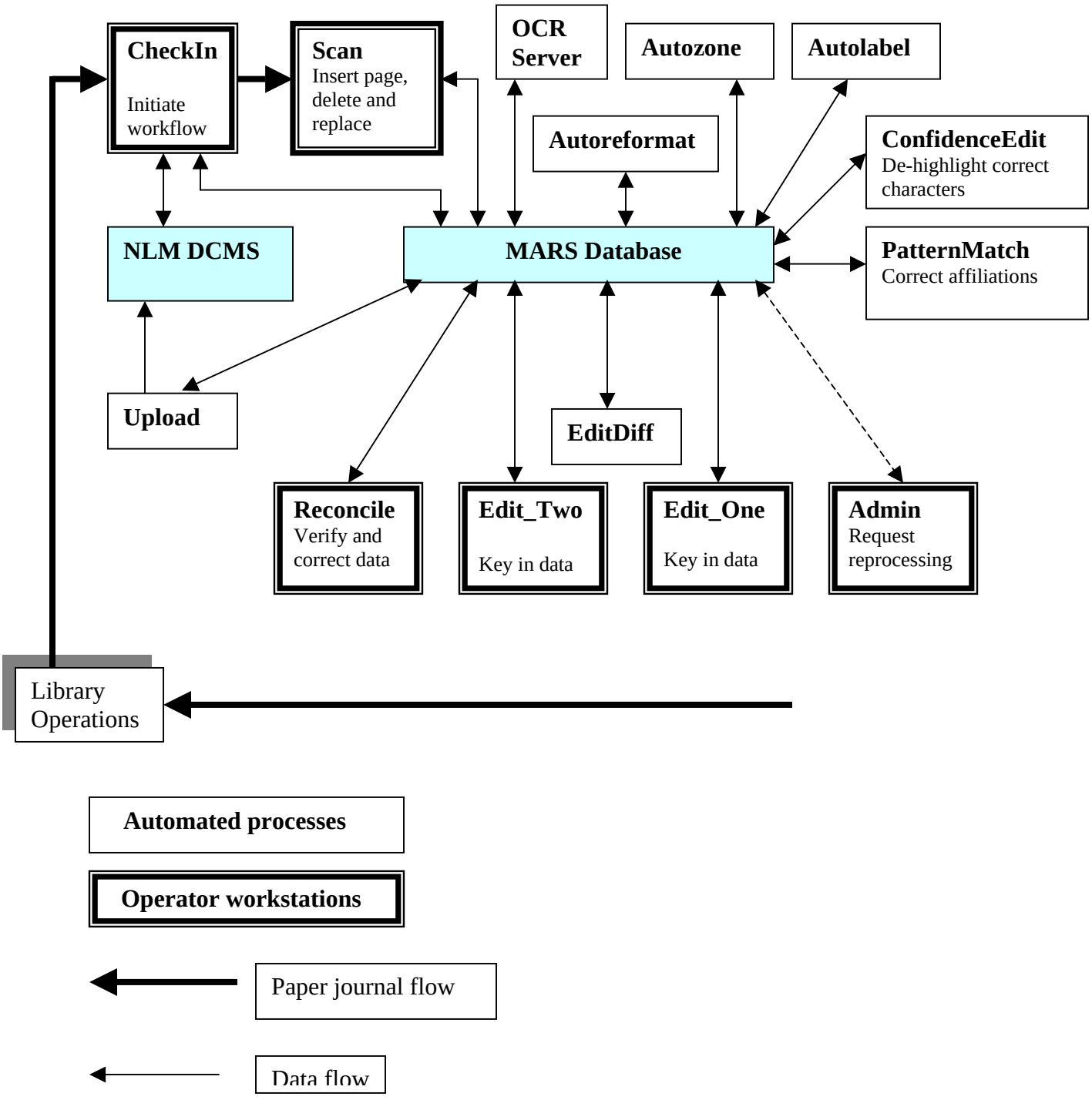
Some data cannot be automatically extracted. The major reason is that they appear in pages other than the scanned first page. Such data is manually entered by a pair of 'Edit' operators, a double-keying process that is commonly used in data entry operations to ensure high accuracy. An EditDiff module then correlates these different entries and highlights differences. The output of the automated processes and the edit operators is then presented to the 'Reconcile' operator who verifies and corrects the text. The Upload module[11] then sends the verified data to the NLM's DCMS (Data Creation Maintenance System) which is accessed by NLM indexers to add MeSH terms and keywords, thereby completing the MEDLINE record.

The Admin workstation shown is used by the production supervisor to send a journal issue back to an earlier processing stage in case of errors.

2. EVALUATION APPROACH

Assessing the performance of the MARS system is an important goal, not only to evaluate the efficiency of its constituent modules, but also to locate potential bottlenecks. In addition, since we seek the best way to create MEDLINE bibliographic records, it is important to compare the productivity (e.g., labor hours per unit record) of the MARS systems, both versions 1 and 2, against each other, as well as with that of the manual keyboarding operation done under contract. Key questions posed as a starting point for performance evaluation are listed as follows:

Figure 1 MARS-2 general schematic



1. What is the time taken by each manual and automatic process for one record?
2. What is the workload distribution of each process?
3. What is the time distribution of manual entry of each field by the Edit operator?
4. What is the overall level of effort (in labor-hours) in the MARS-2 operation as compared to the less automated MARS-1 and the entirely manual keyboarding operation?

These questions are addressed quantitatively by instrumenting the system and analyzing the data recorded, these mainly being event counts and time data. Instrumentation is implemented by two C++ classes written to record such data, one to record times and the other to record statistics generated in a MARS process.

3. PROCESS PERFORMANCE ANALYSIS

The instrumentation data yields information on the processes, both automatic and manual, at different levels of granularity. Figure 2 shows the average time taken by each process to complete its task for one bibliographic record (citation) in July 2001. The processes appear in the figure in the order that they occur in the system. Predictably, the manual processes of scanning, editing and reconciling take much longer than the automated ones. An explanation of the terminology: Edit_First and Edit_Second stand for the first and second Edit operator; Prod is the inhouse-developed daemon that controls the OCR system, hence equivalent to the OCR action; ZoneCzar combines the actions of automated zoning and labeling.

Instrumentation data for a breakdown of some of these processes into their constituents appear in Figures 3 and 4. In Figure 3 we find the actual process of scanning a page (“append”) to take about 20 seconds, including the time taken for the operator to find the page to scan. However, inserting a missing page (“insert”) after the fact and the entry of a new journal issue (“New MRI”) takes much longer. This latter task, found time consuming, motivated the development of the new CheckIn module, which eliminates this function in the scanning operation.

The breakdown of the constituent actions in the Scan process has to be viewed in the context of the workload on the operator, as shown in Fig. 4. The actual workload created by some of these time consuming actions is not high, because they do not occur frequently. For example, as shown in the workload distribution, the burden of inserting pages is very low, since this operation is performed rarely, while “appending” is a major effort as expected since it is done for every page captured.

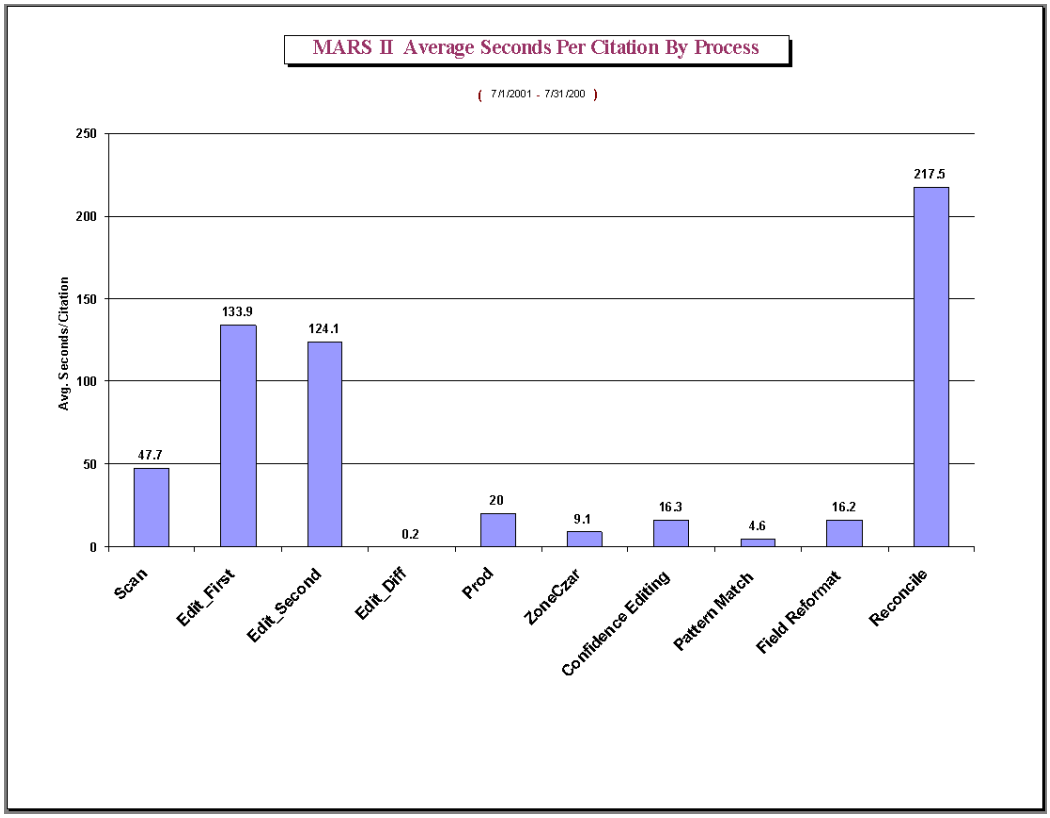


Figure 2

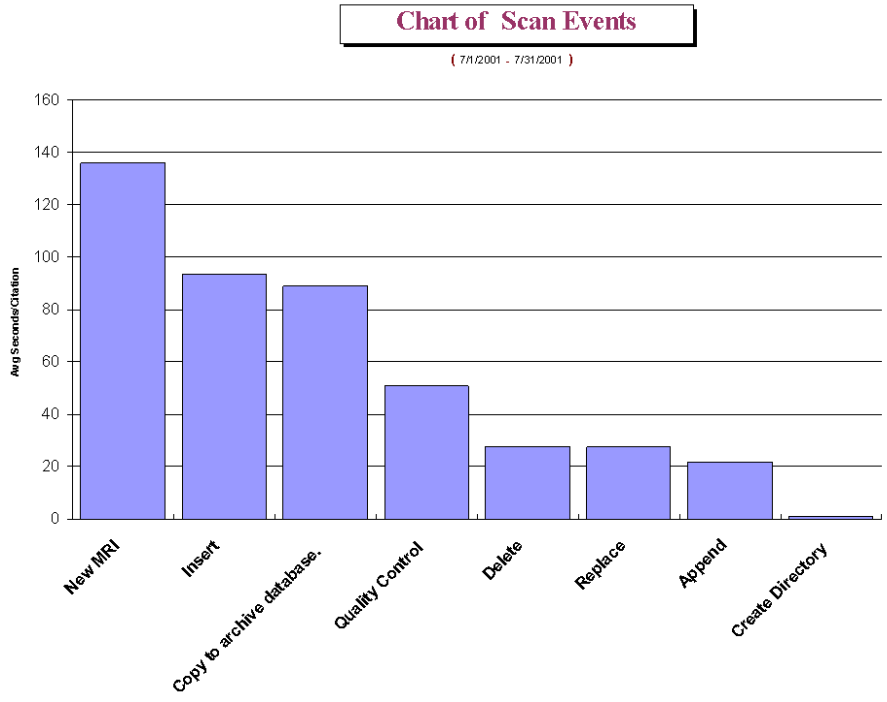


Figure 3

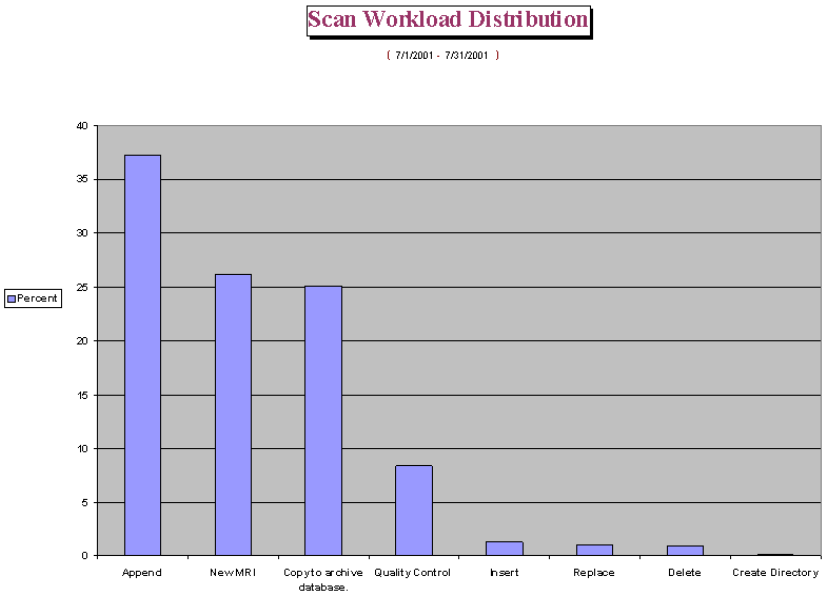


Figure 4

Figure 5 shows the average time taken for the Edit operator to enter the fields not automatically extracted. Only the data for the first Edit operator is shown, since the data for the second operator is approximately the same. In this figure, we show entries for those fields that are automatically extracted in 'compliant' journals (for which we have generated rules from layout analysis for our automated modules), because we are accommodating non-compliant journals also. Furthermore, even for compliant journals, there are pages that are not processed by the automatic modules (e.g., letters to the editor, editorials) requiring the Edit operator to key in the relevant data. The figure, however, indicates opportunities for further automation, especially in the entry of email addresses, corporate authors, pagination and others. The times recorded reflect not just the entry time, but the effort taken in finding the information, often by leafing through the article. The times shown in this figure identify the fields whose automated extraction would contribute most toward productivity.

Since we keep track of operator names in the database, we also offer the supervisor the option of comparing their relative effectiveness, as shown in Figure 6 for scanning and Figure 7 for editing.

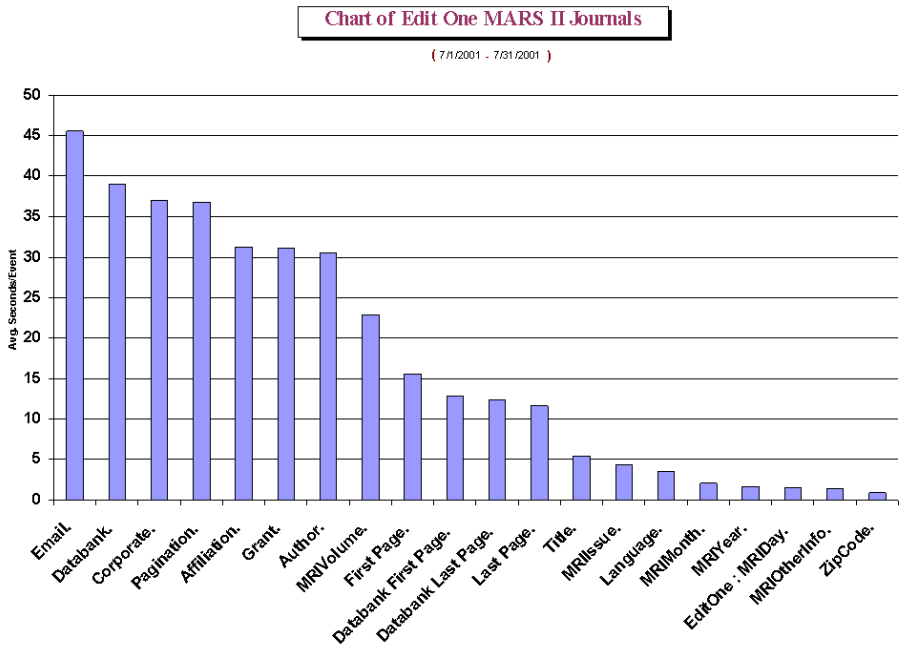


Figure 5

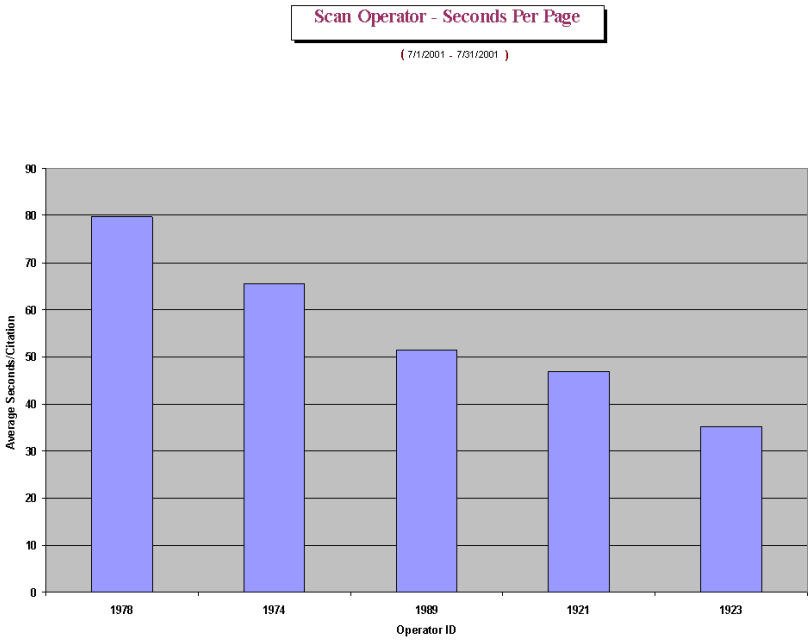


Figure 6

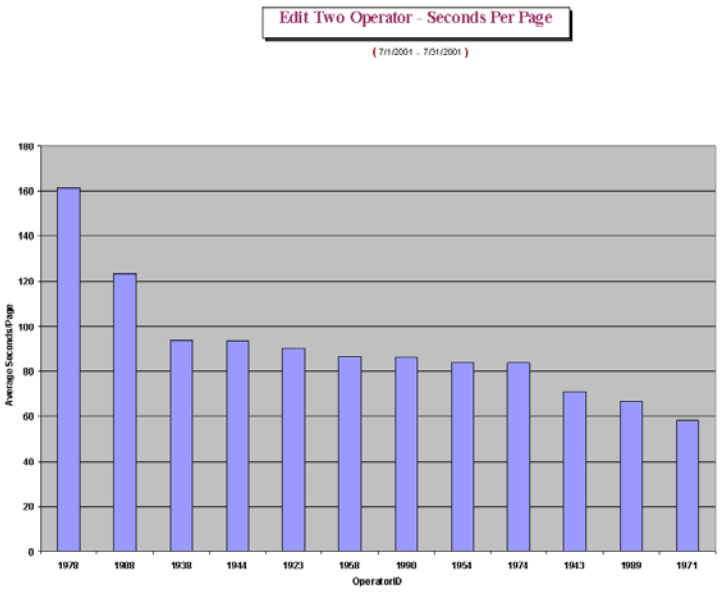


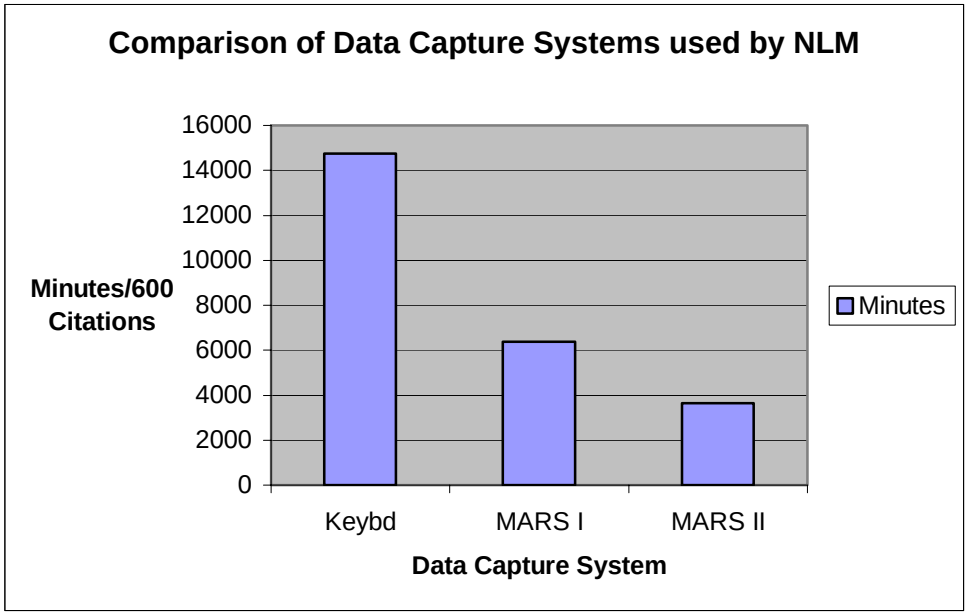
Figure 7

4. COMPARISON OF THE THREE DATA ENTRY SYSTEMS

Here we compare the two systems, MARS-1 and MARS-2, and the manual keyboarding operation on the basis of a workload of 600 completed bibliographic records per day, the average workday load for all of these approaches. The table lists the average number of seconds to scan a page, enter data manually by the Edit operator, and verify the text by the Reconcile operator for each bibliographic record. This is shown for each of the three approaches to generate these records. Data is shown on a per-page (i.e., record) basis in seconds, and the number of minutes per 600 records, and shown in a chart in Figure 8. It can be seen that MARS-2, by eliminating many of the manual functions in MARS-1, is a considerable improvement, and that both are far more efficient than the manual keyboarding operation. To produce 600 records, MARS-2 requires 61 hours of labor per day, while the keyboarding requires 246 hours. In comparison with the keyboarding operation, MARS-2 therefore saves 185 direct labor-hours per day or 51,800 labor-hours per year (based on a year of 280 work days).

Category	**KeyBd** ***Sec/Page***	**MARS I** ***Sec/Page***	**MARS II** ***Sec/Page***	**Keybd** **Min/600**	**MARS I** **Min/600**	**MARS II** **Min/600**
Scan	NA	71	30	NA	706	300
Edit	NA	178	133	NA	1784	1330
Reconcile	NA	388	202	NA	3885	2020
Total	1475	637	365	14750	6374	3650

Figure 8



5. SUMMARY

The comparative performance of three methods to transfer bibliographic information from biomedical journals to the MEDLINE database has been shown. The methods are: a totally manual keyboarding operation, a partially automated system, MARS-1, and a more automated image analysis-based system, MARS-2. The MARS-2 system requires less than 25% of the effort in the manual keyboarding operation. In addition, a breakdown of the times taken in manual and automated processes in the MARS-2 system, and comparisons of different operators performing the same actions are given.

REFERENCES

1. Thoma GR. Automating data entry for an online biomedical database: a document image analysis application. Proc. 5th International Conference on Document Analysis and Recognition (ICDAR'99). Bangalore, India; Sept. 1999; 370-3.

2. Hauser SE, Le DX, Thoma GR. Automated zone correction in bitmapped document images. Proc. SPIE: Document Recognition and Retrieval VII, Vol. 3967, San Jose CA, January 2000, 248-58.

3. Le DX, Kim J, Pearson GF, Thoma GR. Automated labeling of zones from scanned documents. Proc. 1999 Symposium on Document Image Understanding Technology, College Park, MD: University of Maryland Institute for Advances in Computer Studies; 219-26.

4. Le DX, Thoma GR. Page layout classification technique for biomedical documents. Proc. World Multiconference on Systems, Cybernetics and Informatics (SCI 2000); Orlando, FL; Vol. 10, July 2000; 348-52.

5. Le DX, Thoma GR. Automated document labeling using integrated image and neural processing. Proc. World Multiconference on Systems, Cybernetics and Informatics, Orlando, FL; Vol. 6, 1999; 105-8.

6. Kim J, Le DX, Thoma GR. Automated Labeling in Document Images. Proc. SPIE: Document Recognition and Retrieval VIII, Vol. 4307, San Jose CA, January 2001, 111-22.

7. Ford GM, Hauser SE, Thoma GR. Automatic reformatting of OCR text from biomedical journal articles. Proc.1999 Symposium on Document Image Understanding Technology, College Park, MD: University of Maryland Institute for Advances in Computer Studies; 321-25.

8. Hauser SE, Browne AC, Thoma GR, McCray AT. Lexicon assistance reduces manual verification of OCR output. Proc. 11th IEEE Symposium on Computer-based Medical Systems. Los Alamitos, CA: IEEE Computer Society. June 1998; 90-5.

9. Ford G, Hauser SE, Le DX, Thoma GR. Pattern matching techniques for correcting low confidence OCR words in a known context. Proceedings of SPIE, Vol. 4307, Document Recognition and Retrieval VIII, January 24-25, 2001, pp. 241-9.

10. Lasko TA, Hauser SE. Approximate string matching algorithms for limited-vocabulary OCR output correction. Proceedings of SPIE, Vol. 4307, Document Recognition and Retrieval VIII, January 24-25, 2001, pp. 241-9.

11. Pearson G, Moon CW. Bridging two biomedical journal databases with XML: A case study. Proc. 14th IEEE Symposium on Computer-Based Medical Systems. Los Alamitos, CA: IEEE Computer Society. July 2001, 309-14.