Cost analysis of a project to digitize classic articles in neurosurgery*

By Kathleen Bauer, M.S., M.L.S.
kathleen.bauer@yale.edu
Reference Librarian and Librarian for the Yale School of Nursing

Cushing/Whitney Medical Library
333 Cedar Street
P.O. Box 208014
New Haven, Connecticut 06520-8014

In summer 2000, the Cushing/Whitney Medical Library at Yale University began a demonstration project to digitize classic articles in neurosurgery from the late 1800s and early 1900s. The objective of the first phase of the project was to measure the time and costs involved in digitization, and those results are reported here. In the second phase, metadata will be added to the digitized articles, and the project will be publicized. Thirteen articles were scanned using optical character recognition (OCR) software, and the resulting text files were carefully proofread. Time for photocopying, scanning, and proofreading were recorded. This project achieved an average cost per item (total pages plus images) of $4.12, a figure at the high end of average costs found in other studies. This project experienced high costs for two reasons. First, the articles contained many images, which required extra processing. Second, the older fonts and the poor condition of many of these articles complicated the OCR process. The average article cost $84.46 to digitize. Although costs were high, the selection of historically important articles maximized the benefit gained from the investment in digitization.

INTRODUCTION

Most medical researchers are interested in the most current biomedical literature, but there are older classic articles of particular importance to medical historians and to some students, researchers, and clinicians. Barriers exist to the use of these articles, one being the lack of easy access to older material. Many libraries do not have extensive collections going back to the 1800s or early 1900s, when some of these articles were written. Even libraries that have these articles tend to keep them under lock because of their age. Indeed, the fragile condition of some of these articles argues against their widespread use. A second obstacle is the lack of indexing of these articles in searchable, electronic databases such as MEDLINE. Although some articles may be indexed in the print *Index Medicus*, the majority of researchers and clinicians search only electronically available databases [1].

Wider access to these classic print articles is now possible due to the availability of scanning hardware and software, which allow the production of digital copies of print material. Although many individual libraries would not have the resources to undertake major digitization projects, they could do small-scale projects, concentrating on the most important older articles. In this model, an individual library would digitize classic articles in one specialty, and, by adding them to classic articles from other libraries, a useful digital collection of historically significant material would be created. With the eventual goal of encouraging such a cooperative digitization project, the Cushing/Whitney Medical Library at Yale University secured funding to begin a demonstration. Due to the library's long relationship with Harvey Cushing, M.D., a pioneering neurosurgeon, the decision was made to focus exclusively on neurosurgery. Work on selected classic articles in neurosurgery began in summer 2000.

This project was envisioned in two parts: In the first part, articles would be scanned and digital formats would be produced. Scanning technology is fairly straightforward, and standards have begun to emerge for library digitization [2]. The components of the pro-

* This project was supported by a grant from the Yale University Library Standing Committee on Professional Awareness.

cess investigated here are the average time and costs involved in producing a digital format from a printed page. By producing this analysis, library staff wanted to help other libraries in planning their own digitization projects. This article reports on those findings. In the second part of the project, metadata will be added to the digital files, and records for each of the articles will be added to Yale's online catalog.

METHODOLOGY

In this project, the library sought to study producing hypertext markup language (HTML) files from original print articles. HTML, instead of an image format file, was selected so that the articles would be completely searchable. A decision had to be made early in the process whether to use an external scanning service or internal student help for scanning and proofreading articles. Because a goal of this project was to study the process and provide averages for time and costs, we decided to do the work internally with student help. A student was hired for a total of 120 hours, which worked out to be approximately ten twelve-hour weeks. Thirteen articles important in the historical development of neurosurgery were chosen with ten being completed. Articles were selected based on their inclusion in *Morton's Medical Bibliography: An Annotated Checklist of Texts Illustrating the History of Medicine (Garrison and Morton),* a recognized authoritative guide to historically important biomedical literature [3]. Once documents were chosen, they were requested from the locked stacks of the historical collection at the Cushing/Whitney Medical Library. Articles ranged in size from a minimum of two pages to a maximum of thirty-five pages (13.2 average), most with at least one image (7.3 average). In total, the articles represented 172 pages and ninety-five images. No attempt was made to exclude documents based on their condition.

Photocopies were made of each article before scanning. Because some of these articles were more than 100 years old, producing high-quality photocopies was often difficult. Considerable time had to be spent on this, as the quality of the final scanned document depended on it. For some of the articles, directly scanning the article from the original bound journal ended up being the superior method. Either scanning directly or photocopying first required that the scanner or photocopier glass be cleaned often, because of dust and paper particles from the old bound journals. The brittle nature of the originals also necessitated handling them very carefully, a time-consuming process.

Articles were scanned using Caere's OmniPage 9 and Adobe Photoshop. OmniPage, a type of optical character recognition (OCR) software, was used to scan original documents into text formats that could later be converted to HTML. In addition to using OCR to reproduce the text, photographs and drawings in each article were scanned separately using Adobe Photoshop to gain greater flexibility in manipulating images.

Using OCR software on older print material entails many challenges. OCR requires that each scanned letter be recognized and translated, so that the scanned image can be converted into a text file. In some of these older articles, OmniPage did not recognize the fonts and run-on sentences, large spaces between words or sentences, merged columns, merged pages, and merged image captions resulted from the OCR process. In better-quality copies, OmniPage did a very good job. In either case, careful proofreading was important to fix errors that occurred during the conversion to a text file. The easiest and most time-efficient method discovered was to clean up the major problems in OmniPage, such as run-on sentences and merged columns; save it as a Word document; and then do the more detailed proofreading in Microsoft Word. Word documents were finally saved as HTML. Links to images that had been scanned and refined in Photoshop were then inserted into the HTML file.

The ten completed articles are available on the Cushing/Whitney Medical Library Web [4].

RESULTS

Time and costs

In this project, we aimed to scan and format thirteen documents within our allotted 120 student hours. Of those hours, the first ten were spent training the student in getting the articles, scanning the articles, and using the OCR software. In the remaining 110 hours, ten documents were completed and some work was done on all thirteen. The average time spent on each document was 8.5 hours. There was wide variability in the time needed to complete an article. Some of this variability was due to differences in the fonts and poor condition of the older journals. Because the student sometimes worked on more than one article at a time, he found it too difficult to account for time for each specific article. We therefore opted to look at the cumulative time needed to complete all pages (172) and images (95), for a total 267 scanned items. In this report a per-item figure is quoted, not a per-page figure. The number of images in an article directly impacted the cost, and so the per-item figure was considered more truly reflective of the cost libraries would encounter in digitizing articles, in that the cost will depend on the number of pages of text and the number of images. Project time was broken down according to training, photocopying and scanning, and finally proofreading and formatting (Table 1).

The cost estimates provided here excluded the costs of hardware and software, which were already present in the library. The only cost directly incurred was the pay for 120 hours of student work time at $10 per hour, or $1,200. To look at a true per-item cost, training

Table 1
Breakdown of time and associated costs for digitizing print material into final hypertext markup language (HTML) format

	Hours	Hours per item	Total cost ($)	Cost per item ($)
Train personnel	10.00	0.037	100.00	0.37
Photocopy and scan	29.50	0.110	295.00	1.10
Proof and format	80.50	0.301	805.00	3.01
Total	120.00	0.449	1,200.00	4.49
Total excluding training	110.00	0.412	1,100.00	4.12

should be excluded (because training time does not vary according to the number of pages produced). Dividing the $1,100 spent on student work hours by the total 267 items produced a per-item cost of $4.12. Because the average article consisted of 13.2 pages and 7.3 images or 20.5 items, the average per-article cost was $84.46.

The overall cost in this study of $4.12 per item (pages plus images) had to be converted to a cost of $6.40 per page (images not counted) for comparisons with other published cost studies. The $6.40 per-page figure was calculated simply by dividing the $1,100 spent on student work by the total 172 article pages. The $6.40 per-page cost differed significantly from the $2.60 per page cost for OCR projects reported by Puglia [5], representing 146.0% higher cost. The Puglia estimate however was for simple black-and-white text. The articles in this project were quite complex for two reasons. One was the number of embedded images, 7.3 images on average. Images add a level of complexity to an OCR project. The second complication was that the OCR software had difficulty recognizing the older fonts and formats of these articles. This added greatly to the time needed for proofreading and file formatting. The projects reported by Puglia also showed tremendous range in costs for OCR, from $0.65 per page to $5.70.

Puglia reports the cost of rekeying a document as $7.40 per page converted, which would not be too far out of line with the cost of $6.40 per page obtained in this demonstration [6]. If an article is particularly old, is in poor shape, or has very odd fonts, rekeying should be investigated as a possible cost-efficient alternative to scanning. If rekeying were used for the oldest articles in the poorest shape, and OCR were used only for articles in the best shape with relatively easy-to-translate fonts, the cost per page could be brought down considerably.

Alternate formats

The large amount of time necessary for proofreading also indicates that libraries wishing to digitize articles should consider creating images of the articles instead of using OCR software to create text files. The standard, archival-quality, bitmapped image file is called tagged image file format (TIFF). TIFF has the advantages of being a nonproprietary file format and the highest quality image. TIFF files are also the largest image files, and so are not appropriate for distribution via the Internet [7]. Along with the archival TIFF copy, most digitization projects take the saved TIFF, convert it to some smaller image file, and then make the smaller file available. A commonly used image file type for text pages is portable document format (PDF), a proprietary file type owned by Adobe. The decision to produce a PDF or a text file may have major implications for the cost of digitization.

Table 2
Average time in seconds required to convert one printed page to one portable document format (PDF) page

Pages	Seconds	Seconds/ Page
3	174	58.0
12	711	59.3
9	465	51.7
7	294	42.0
7	890	127.1
9	489	54.3
10	664	66.4
10	634	63.4
8	500	62.5
3	190	63.3
16	1,045	65.3
17	980	57.6
4	216	54.0
Average	557.8	63.5

Scanning a document and producing TIFF and PDF files is by far the faster alternative to producing text files, because it does not involve proofreading. Data were gathered at the Cushing/Whitney Medical Library in the process of producing PDF articles for course reserves (Table 2). In producing PDF, time was recorded from the beginning of the scan of a photocopied article to the production of a complete PDF document. These were easy, straightforward items to be scanned: already photocopied clean copies of fairly current articles. Although the quality of one article caused problems and had double the usual scanning time, all other articles were produced in close to one minute per page, for an average of 63.5 seconds per page, or 0.018 hours per page. This translates to a cost of only $0.18 per page when using workers paid $10.00 per hour, sharply less than the $6.40 per page cost found in this study.

The estimate of 0.018 hour per page (or $0.18 per page) is low for two reasons. It does not include time needed to locate and photocopy material, and the PDF files were created from more recent material that was fairly easy and quick to scan. A project at Pennsylvania State University to digitize out-of-print books, which probably included more varied material, showed a cost of $0.28 per page for scanning [8], and Puglia reported

averages of $0.31 to $0.34 for images of simple black-and-white print [9]. The average of these reported figures, $0.31, will be used in calculations of cost estimates in this article. Although scanning to produce an image is quicker and cheaper than using OCR to create a text file, it may not be the appropriate choice in all cases. Some articles are not in good enough condition to produce an adequate image or PDF.

An important consideration for TIFF and PDF is that an image file takes far more storage space than a text file. For a large-scale digitization project, libraries should consider the cost implications of long-term storage, and significantly larger files may mean significantly larger storage costs.

ESTIMATING TIME FOR A DIGITIZATION PROJECT

Based on other published costs for OCR projects of $2.60 per page, this project's cost of $6.40 per page (images not counted) or $4.12 per item (pages plus images) would seem to be quite high. This is probably due in large part to the older nature of the material digitized, which requires extra time for proofreading and correction of errors introduced in the scanning process. If libraries undertake projects with printed articles in excellent condition with easily recognized fonts, the average time found here may be too high an estimate to use for planning. It would seem more likely, however, that projects would involve mixes of articles and fonts. As more libraries carry out digitization initiatives, these averages should be refined.

Any library seeking to estimate time for a digitization project using OCR software to create text files can use the average reported here in the following way. The time necessary for training is fixed: that is, it does not vary with the number of articles to be produced. Time for training in this case is ten hours. Time for photocopying, scanning, creating files, and proofreading are variable and depend on the number of pages and images to be digitized. This estimate counts all pages and all images and then totals them to produce the number X. To digitize X pages and images, libraries would estimate (using the 0.412 hours average found in this project):

$$\text{Hours (text file)} = 10 + X \times 0.412$$

The hours needed to produce an image (TIFF and PDF) would be different, both in the time needed for training and the time needed per page. Y is the number of pages. In this case, images would not be handled separately, so images do not need to be counted separately and added to the number of pages. Training time may be reduced, and the time per page must certainly be lower. Using the per-page time of 0.031 (the average of times reported by other investigators) and five hours for training:

$$\text{Hours (TIFF)} = 5 + Y \times 0.031$$

Cost estimates

If libraries do not already have computers, scanners, and associated software, there will be additional start up costs in digitization projects for purchasing these necessary components. The basic hardware requirements are a personal computer powerful enough to handle several software programs needed for processing documents and a reasonably high-quality scanner. Care should be taken in selecting scanners, as poor-quality, slow scanners will add considerable time to any project. Many articles specify requirements for buying software and hardware for scanning projects. [10]. In general for these types of projects, it is best to purchase the fastest processor and the largest amount of memory possible. A high-quality, large monitor will also help in scanning projects.

Estimating costs for a digitization project is then a matter of adding the costs for hardware and software (if necessary) to the product of the hourly wage paid to workers and the estimate of hours needed as calculated above.

$$\begin{aligned}\text{Cost (text file)} &= (\text{cost of hardware}) + (\text{cost of software}) \\ &\quad + (\text{hourly wage}) \times \text{hours (text file)}\end{aligned}$$

Similarly, to cost out a project to produce image files (TIFF):

$$\begin{aligned}\text{Cost (TIFF)} &= (\text{cost of hardware}) + (\text{cost of software}) \\ &\quad + (\text{hourly wage}) \times \text{hours (TIFF)}\end{aligned}$$

An important factor to be weighed is the hourly wage to be paid to people doing the scanning and proofreading. Although scanning is not much more difficult than photocopying, workers doing this type of work must be familiar with basic computer functions and must be careful and detail oriented. Because fixed costs are the same for very large jobs or small jobs, the cost per item for very small jobs will be quite large. For small jobs, it will perhaps be more cost effective to send the work to companies that specialize in digitization.

As this paper has shown, different projects will experience different per-item costs, depending on the age, condition, and font styles of articles to be digitized. An appropriate step to take before costing out a digitization project would be to perform the entire digitization process on a small sample of one or two articles. Doing a sample should guide the library in selecting the appropriate time estimates.

Finally, these estimates are limited to the digitization process and have ignored the costs involved with storage and file maintenance. Other researchers have listed the costs of this at three [11] to five times [12]

the original digitization cost. This may not be true for the small-scale projects envisioned here, because storage of a small number of files should not put a large burden on a Web server. Whether a project is small or large, however, libraries must consider the many implications of maintaining digital files over time. The Research Library Group has provided a worksheet that can aid in considering all the costs involved in digitization [13]. After weighing all costs, libraries may decide that digitization is not desirable. Access to an article can be improved by simply cataloging a classic article, and some researchers have concluded that maintaining easy access to the print is more desirable for researchers [14]. Certainly, costs and benefits in any project must be carefully weighed before deciding to digitize print resources.

CONCLUSION

The Cushing/Whitney Library has large amounts of historical material of potential usefulness to researchers. This material can be made more accessible by producing digital copies of important older articles from the collection and then making them available on the Internet. The library demonstrated in this small project that older articles could be digitized for average per-item (pages plus images) costs of $4.12 and average per-page costs of $6.40. Other libraries can estimate costs for their own digitization projects by using the formulas described above, which compute cost estimates using the number of items to be digitized, per-item time estimates, and hourly wages.

To use the formulas, libraries must first consider whether they want to produce text files (HTML) or image files (TIFF). A far greater amount of time is needed to convert one page of an article to HTML than to TIFF, mainly because of the need to proofread and format a file produced by OCR software. After a file format is selected, libraries must then consider whether the per-item time average reported here will be appropriate for their projects. The times described in this article were high compared to other studies. Factors that increased the time were the poor condition of some of the articles, the fonts that were poorly handled by OCR software, and the mixture of text with many images, which had to be handled separately. If their projects involve articles in good condition, with easily recognizable fonts, or with few images, they may wish to adopt a lower per-item time estimate. Libraries may test the time estimate by digitizing small samples of articles before starting projects.

Any money spent digitizing materials of doubtful value will represent waste of resources. Libraries must weigh the possible benefit of digitization against the costs they will likely incur. Resources will be used most effectively when a standard is established to concentrate on material of the highest value, and recognized historically important journal articles meet such a standard. If the material is very important, an average cost of $84.46 per digitized article (as was found in this study) represents good value to this library and its patrons. This project demonstrates that, by concentrating on articles already judged to be important, a small investment ($1,200) can be leveraged to produce the greatest possible gain. The benefit of repeating this project in other libraries would be that with relatively small investment historically important articles would be made accessible to more medical historians and researchers.

REFERENCES

1. CURTIS KL, WELLER AC, HURD JM. Information-seeking behavior of health sciences faculty: the impact of new information technologies. Bull Med Libr Assoc 1997 Oct;85(4): 402–10.
2. COLORADO DIGITIZATION PROJECT. General guidelines for scanning. [Web document]. Colorado Digitization Project, 1999. [rev. 4 Jan 2001; cited 12 Mar 2001]. <http://coloradodigital.coalliance.org/scanning.html>.
3. MORTON LT, NORMAN JM. Morton's medical bibliography: an annotated checklist of texts illustrating the history of medicine (Garrison and Morton). 5th ed. Aldershot, Hants, U.K., and Brookfield, VT: Gower, 1991.
4. CUSHING/WHITNEY MEDICAL LIBRARY. Classic articles in neurosurgery. [Web document]. New Haven, CT: The Library, 2001. [rev. 2 Mar 2001; cited 16 Apr 2001]. <http://info.med.yale.edu/library/neurosurgery/>.
5. PUGLIA S. The cost of digital imaging projects. Mountain View, CA: RLG DigiNews [Internet], 1999 Oct 15;3(5). [rev. 19 Oct 1999; cited 13 Apr 2001]. <http://www.rlg.org/preserv/diginews/diginews3-5.html#feature>.
6. IBID.
7. BEARDEN C. Basic scanning for the World Wide Web. Tex Libr J 1999 Fall;75(3):112–6. See also: BEARDEN C. Basic scanning for the World Wide Web. Tex Libr J [Internet], 1999 Fall; 75(3). <http://www.txla.org/pubs/tlj_3/scanning.html>.
8. KRIEGER LA. OP scanning: an acquisitions and preservation solution. Lib Coll Acquis Tech Serv 2000 Autumn;24(3):424–6.
9. PUGLIA, op cit.
10. CHAPMAN S, COMSTOCK W. Digital imaging production services at the Harvard College library. Mountain View, CA: RLG DigiNews [Internet], 2000;4(6). [rev.15 Dec 2000; cited 18 Apr 2001]. <http://www.rlg.org/preserv/diginews/diginews4-6.html#feature1>.
11. LEE S. Digitization: is it worth it? Comput Libr 2001 May; 21(5):28–31.
12. PUGLIA, op cit.
13. RESEARCH LIBRARY GROUP. RLG worksheet for estimating digital reformatting costs. [Web document]. Mountain View, CA: The Group, 1997. [rev. May 1998; cited 16 Apr 2001]. <http://www.rlg.org/preserv/RLGWorksheet.pdf>.
14. ELDREDGE JD, GUENTHER H. Historically significant journal articles: their identification in older bound journal volumes designated for weeding and the creation of new access to these articles. Bull Med Libr Assoc 2001 Jan;89(1):71–5.

Received July 2001; accepted November 2001