Asymptotic cost in document conversion

Dorothea Blostein[*a], George Nagy[†b]

[a]School of Computing, Queen's University, Kingston, Ontario, Canada
[b]DocLab, Electrical, Computer, and Systems Engineering, RPI, Troy, New York

ABSTRACT

In spite of a hundredfold decrease in the cost of relevant technologies, the role of document image processing systems is gradually declining due to the transition to an on-line world. Nevertheless, in some high-volume applications, document image processing software still saves millions of dollars by accelerating workflow, and similarly large savings could be realized by more effective automation of the multitude of low-volume personal document conversions. While potential cost savings, based on estimates of costs and values, are a driving force for new developments, quantifying such savings is difficult. The most important trend is that the cost of computing resources for DIA is becoming insignificant compared to the associated labor costs. An econometric treatment of document processing complements traditional performance evaluation, which focuses on assessing the correctness of the results produced by document conversion software. Researchers should look beyond the error rate for advancing both production and personal document conversion.

Keywords: document processing cost, document recognition, document transformation, performance evaluation, productivity

1. INTRODUCTION

The economics of document image processing have undergone a sea change since the first Document Recognition and Retrieval Conference in 1994. However, there is little evidence that this change is reflected in published research on document processing software. Here we address cost and value issues in *document conversion*, a narrow domain within document processing that is centered on the conversion of hardcopy to computer-readable media, and vice-versa. Thus, we deliberately avoid issues pertinent only to automated content interpretation, information retrieval, translation, summarization, and categorization. These processes generally take place entirely in the digital domain even if source material originates on paper, and results are printed or displayed for human inspection.

The thesis of this report is that the cost of computing resources in document conversion is asymptotically approaching zero and that these costs are already negligible compared to human costs. Research should therefore focus on effective human intervention in addition to more accurate heuristics and algorithms for automation. The effectiveness of human intervention can be measured as the incurred human time converted to monetary units at the rate appropriate for the training and skill required to complete a given document conversion task.

Economic value is associated with documents and the information they contain because documents are instrumental in completing industrial, commercial and personal transactions [1,2]. Documents also represent a significant fraction of our cultural heritage, whose monetary value is impossible to asses (as widely noted during the recent skirmish over copyright-related revenues from mass-digitized library holdings [3]). Costs and values are difficult to quantify due to the complex social and professional environment of document processing.

Document conversion software converts a paper document into an electronic format that supports one or more transactions by providing effective access to the information contained in the document. Conversion in the other direction, from digital media to print or display, is necessary whenever humans must assimilate the content of an electronic document. In practice, "automated" conversion does not mean zero labor costs. The success of document conversion can be quantified

[*] blostein@cs.queensu.ca
[†] nagy@ecse.rpi.edu

by comparing the value of the resulting document files (in the context of particular transactions) to the pro-rated cost of producing them. If labor costs dominate, then we must strive to reduce operator time, and possibly the level of operator training and skill, required to complete the task.

As discussed in Section III, greater algorithm accuracy does not always translate into more effective document processing. In some cases, the cost of human intervention does not decrease (e.g. proofreading time remains high even if the OCR error rate is lowered), and in other cases, the final accuracy is not improved (e.g. if downstream software automatically corrects the errors).

In a research environment, the development of new algorithms or systems for document processing is often justified by listing potentially applicable transactions and document collections. In an operational setting, however, software development must address existing or anticipated transactions on a specified set of documents. The developer must therefore characterize the transactions and document collections (e.g. [4]), and then show that the proposed system adds *value* by reducing the *cost* of the transactions over available methods. Publicly available test data sets are only marginally helpful because they seldom consist of randomly drawn samples from any relevant population.

The ability to retarget conversion software to different types of documents is especially important for small tasks in personal computing. Here, no single family of transactions is large enough to effectively amortize development costs [5]. Retargeting can be carried out manually (e.g. by adjusting parameters of the conversion software) or it can be done automatically by learning algorithms built into the software. Manual retargeting of production software can be carried out only by an expert who knows how to tune the parameters. In contrast, expert intervention is less available in a personal computing environment. In either case, the need for expert intervention is reduced by applying machine learning algorithms to every task execution (e.g. [6,7]).

Section 2 reviews recent trends in the cost and performance of document conversion hardware. Section 3 examines human costs and productivity in document conversion, with an eye to the highly non-linear relationship between machine accuracy and human labor. In Section 4 we note research opportunities that are particularly prominent in high-volume document conversion applications. In Section 5 we discuss what it will take to fully integrate document conversion software into the individual (personal or professional) computer environment.

2. DOCUMENT CONVERSION HARDWARE – COST AND PERFORMANCE

The hardware most relevant to document conversion consists of (1) equipment for document capture (scanners and cameras), (2) computing hardware (microprocessors in personal computers, workstations and servers), (3) storage media (disk and solid state), (4) printers and displays, and (5) transmission facilities (networks, modems and routers). We calculate representative past and current costs and performance with respect to Doc_A, a printed hundred-page mostly-text document. For estimation purposes, we model the text of Doc_A as consisting of 300 five-letter words per page (Table 1). The ASCII representation of Doc_A requires about 1500 bytes per page, 150KB total, without compression, and perhaps three times less with current text compression such as gzip. Assuming an 8.5" x 11" format scanned into a 300 dpi 8-bit gray-scale image, the uncompressed image of Doc_A occupies 100 x 11.5 x 8 x 300^2 bytes, or about one gigabyte. State-of-the-art lossless compression (JBIG2 or DjVu) can reduce this to ~10 MB.

Table 1. Characteristics of the sample text document Doc_A

100 pages		
Symbolic (ASCII) 300 words per page 10 words per line 5 letters per word	Image 8.5" x 11" 300 dpi	
	Gray-level scan	Bi-level scan
150 KB	1 GB	125 MB
Losslessly compressed		
50 KB	10 MB	1.25 MB

2.1 Document Capture.

Forty years ago, a CRT or a drum scanner cost $300,000 and took several minutes to scan one page. Archival documents were often converted to microfilm or microfiche before scanning them with a small-format transparency scanner with a film transport. These scanners required frequent calibration. In the 1990s, high-speed page-feed scanners developed for facsimile cost about $15,000 and could handle 50 pages per minute, or 2 minutes for Doc_A. Flatbed scanners ran to over $1000, and their speed was often limited by that of the upload link. In recent years, 600 dpi flatbed scanners have been available for under $100. Doc_A could be digitized on a flatbed scanner in about five minutes at home. The cameras in high-end cellphones can also capture a page at a resolution adequate for OCR. Recent research explores compensating the resulting images for perspective distortion [8].

Today's production scanners cost under $40,000 and can scan 200 pages per minute, hour after hour and day after day. Page-feed scanners can handle books only with their spine cut off (called *destructive scanning*). The big advance is the advent of robotic (page-turner) book scanners (usually with camera optics) capable of up to 50 pages per minute, priced at ~$200K. Large collections of both ancient and modern books have already been digitized by various organizations, such as Google Books, the Internet Archive [9], and UDL [10]. IMPACT, a European initiative set up to advance and coordinate research for document conversion, digitizes books from the major National Libraries and provides test data and software to researchers [11].

Small-scale document capture can now be carried out via direct text entry using pocket devices. With word-completion software (and lots of practice) thumb-typing on a miniature keyboard is nearly as fast as touch-typing on a regular keyboard. Graphic stylus input allows the conversion of hand-print, including mathematical symbols, to digital symbols.

2.2 Microprocessors.

Before ~1975, all OCR systems ran on special-purpose hardware, cost tens of thousands of dollars, and recognized only a few special fonts. The software eventually migrated to minicomputers, and then to PCs with add-on boards (e.g., Palantir, which morphed into Calera, Caere, ScanSoft,…, Nuance). By 1989 the Intel486 had a peak instruction execution speed of 50 MIPS with a 66 MHz clock, fast enough to run native OCR software. Today's 3.5 GHz multicore processors execute instructions 3000 times faster. Since the cost of PCs has remained flat, the cost of processing has dropped by a factor of 3000 since 1989. Most page-image processing algorithms are essentially linear, with a multiplicative constant based on some window size (downstream algorithms usually work on line, word or character blocks rather than pixels). A typical algorithm with a 5x5 window (binarization, skew estimation, page segmentation) may require 100 instructions per pixel. Therefore in 1989 processing a scanned image of Doc_A on a PC would have taken over half an hour. (That is why many researchers at the time were experimenting with 512 x 512 image arrays). Today it would take less than a second. As a check, we note that conversion from TIF to PDF of a 300 dpi scanned page takes only a few seconds on a modern-day personal computer.

2.3 Storage Media.

From 1990 to 2011 single-platter magnetic disk capacity increased from 100MB to 500GB, with a concomitant increase in transfer rates. Doc_A could not have been stored without compression, in image form, on a 1990 PC drive. Now there is enough room to hold 500 uncompressed copies. The Petabox 4 of the Internet Archive has 650 TeraBytes/rack – the approximate equivalent of 1000 current PC disk drives. As we are nearing the end of motion-based magnetic storage technology (including high-capacity tape), solid state memory technology is poised for a seamless takeover.

2.3 Printers and displays

Bitmapped printers and displays that obviated the need for image output via overprinting have been available for several decades. The price of 20 ppm laser printers has decreased by a factor of 40 over twenty years (from about $20,000 in 1990 to $500 today). None, however, are nearly as fast as the fixed-font chain printers of the mainframe era! Displays large enough to display Doc_A in symbolic form debuted with word-processors in the 1980s. Bit-mapped full-page displays did not come into civilian use until the development of relatively inexpensive flat-screen monitors a few years ago. A major development has been the advent of mobile, small-format, high-resolution and high-contrast displays, including tablet computers, electronic books, iPod-like devices, and even cell-phone displays. Such displays afford both reasonable reading speed and text editing capabilities.

2.5 Networks

Twenty years ago there was no sense in building PC scanners and printers faster than the available buses and connections. Throughout the 1990s, PCs could connect to the internet only through 56Kbits/sec modems. Uploading an uncompressed color scan of Doc_A would have taken more than a day. Cable modems and DSL capable of several megabytes/second transmission have become ubiquitous in the last several years,. The textual content of books can be downloaded in a flash: they are tiny compared to audio and video.

In summary, the cost of processing, storing, transmitting or displaying a pixel has decreased by a hundredfold over two decades. The cost of digitization has decreased less, but the DIA community has not been closely involved in the design and development of document capture machinery. Let us now look at the human side of the equation.

3. HUMAN COST, PERFORMANCE, AND PRODUCTIVITY

Measured in constant dollars, labor costs have remained essentially flat over several decades. We can, however, make a case for a decrease of a factor of two or three based on off-shoring. What about human performance in document conversion?

The average adult reading speed is 300 words per minute. Professional key entry for copying prose remains about 60 words per minute in spite of a spate of new keyboards and word-completion. Speaking and speech comprehension rates are about 150 wpm, intermediate between reading and typing [12].

DIA research has succeeded in partial automation of almost every common document conversion task. Most of these tasks fall into the general categories of converting hardcopy to symbolic computer code, or converting encoded documents to either hardcopy or a temporary visual display. We have not, however, succeeded automating most of these tasks to the level of accuracy expected from human workers. We still depend on human intervention to improve the results to some acceptable level.

Even tasks that are in a sense completely automated, like digitizing, require an enormous amount of human labor in preparing the documents for scanning, disposing them after scanning, and checking that the resulting computer files are complete and properly indexed. The labor cost of operating a book scanner three shifts per week (say 120 hours) at $25,000 per year per operator (including supervision and benefits) is almost twice as much as the hardware cost of the most expensive book scanner if its purchase is written off over five years. The labor cost of document preparation (selection, tagging, transport) is even higher.

In July 2011, advertised prices for automated conversion of a 300-page book ranged from $1.- (destructive scanning and no OCR) to $4 (destructive scanning and OCR without correction) to $45 (non-destructive OCR without correction) to $220 (corrected and edited OCR). The cost of mass digitization is estimated at $6-$10 per book. Vendors advertise 99% per word OCR accuracy on clean printed text. For Doc_A, this translates to an average of three errors per page.

The cost of proofreading and correcting OCR errors is much higher than scanning costs. A lower bound can be based on a reading speed of 300 words per minute and a $25/hour labor cost with overhead. For Doc_A, this amounts to $41.67, or 42 cents per page; the cost would be $125 for a 300-page book of similar print density. An upper bound can be based on the cost of copy editing by language experts; this service is typically available for a few dollars per page. Note that the time necessary for proofreading does not decrease much even if OCR accuracy rates are improved. In contrast, subsequent conversion of OCR output (from PDF to HTML or to one of the Kindle, Sony or iPad e-book formats) is becoming a completely automatic and inexpensive process [13].

There is plenty of hardcopy that remains to be converted [14,15]. Current estimates of the number of existing books (titles rather than copies) run to 130 million [16], almost twice the best estimate ten years ago [17]. A little over 10% of these have been digitized and OCR'd [18]. Only a tiny fraction of the OCR'd books have been proofread and corrected. The OCR version of many older books, technical books, and books in commercially less important languages is almost unreadable, but may still be searchable. Mathematical material also resists conversion to symbolic form (cf. Google Book

version of *Principia Mathematica*). The Mormon Missionary Diaries provide an interesting case study of the costs involved in medium-scale document conversion and related tasks [19].

Another torrent of OCR fodder is litigation. Some law suits require accumulating digital files of tens or even hundreds of thousands of pages, including correspondence, manufacturing manuals, invoices and remittances. As time goes on, the original versions of such materials will be increasingly computer-generated, but their volume and heterogeneity has been a major incentive for the improvement of commercial OCR.

It is not obvious how DIA research can reduce labor costs in existing applications. Even after fifty years of research, key entry remains competitive with OCR in highly context-sensitive and error-intolerant applications like medical form entry. We shall return to this topic in the next section.

4. OPEN RESEARCH PROBLEMS FOR PRODUCTION SYSTEMS IN THE LIGHT OF CHEAP COMPUTING

This section notes research opportunities that are particularly prominent in high-volume document conversion applications such as medical claims processing [20], postal automation [21], legal documents [22] and bank checks [23,24]. Research opportunities for low-volume applications in personal document conversion are discussed in Section 5. This is a fluid distinction since most research advances offer potential improvement for both low- and high-volume document conversions.

4.1 Error/reject ratios

Classifiers – for printed text, for hand-printed digits, for cursive writing – are often run in research settings at zero reject rate, or at a reject rate barely higher than the reported error rates. However, one way to reduce the cost of human proofreading and correction is to guarantee that parts of the output are essentially error-free. This requires a high ratio of rejects to errors [25]. Commercial check amount readers, for instance, typically run at reject-error ratios higher than 1000:1. It is better to have operators enter 30% of the data, using double keying or other forms of verification, than to have them enter only 3%, and leave an error or two in every batch of checks. The desirability of reporting complete error-reject (or ROC) curves has been known since the sixties, but there has been little published on choosing training sets or training regimes for very low-error, high-reject applications. DIA researchers must reconcile themselves sooner or later to the idea that in many applications human interaction is here to stay. Even unmanned space vehicles, among the most automated systems to date, require an alert, 24/7 crew on the earth end of their virtual tether.

4.2 Sampling

The accepted method of estimating the performance of a system on a new, unpredictable set of documents is measuring its performance on a random sample drawn from the same population. This requires the definition of the population of interest, and the development of a sound sampling strategy. While the medical and biological communities seem fully aware of these principles and routinely apply them in research, the DIA community continues to rely on convenience samples from undefined populations. (Some examples of populations and samples appropriate for DIA experiments are listed in [26]). Applying a hypothesis test to confirm the significance of an improvement in segmentation or classification rate is useful only if the sample is drawn from a relevant population. Sound sampling allows setting the optimal human/automatic trade-offs and is the unavoidable cost of predictable and replicable results in the field. We recognize, of course, that random sampling means drawing a fresh sample after any test of a modification of the methodology under development. Fortunately the cost of repetitive sampling has been alleviated by the amount of data now floating in the clouds and the technological developments that allow capturing large chunks of it at will [27,28,29].

4.3 Green Interaction

We define green interaction as the recycling of operator interventions. In successful DIA systems, the size of the training sets used to develop and tune the system is necessarily far smaller than the volume of data processed over the lifetime of the system. We should apply learning algorithms that use routine feedback from the operator to improve classification [30,31]. Furthermore, reduction of the cost of storage has led to the maintenance of complete records – even of systems as large as the whole web. Subsequent human or automated analysis of such data can not only lead to improving the system,

but also to improving operator training. However, one of the problems of experimenting with human-computer systems that improve with use is that so far humans learn much faster than computers, therefore human learning masks the machine learning [32].

4.4 End-to-end performance measures

Errors arise in symbol segmentation, symbol recognition, structure analysis, and recognition of document layout. Performance assessments typically report only the frequency of each of these types of errors. Research opportunities exist in developing end-to-end performance measures that provide an assessment of overall system performance [33]. Such global measures provide an objective function that can be optimized by machine learning. In addition, global measures provide a means for end users to make well-informed selections between competing document conversion systems.

4.5 Adaptive GUIs

User interfaces can adapt to the history of interactions, for example, by reordering and reorganizing requests for operator confirmation or correction. Output can be delayed to ensure that related items are presented in a coherent fashion that minimizes having the operator chase after ancillary data. Similar items can be grouped (via clustering algorithms) to give the operator the option of correcting them all at once if appropriate. Frequent sequences specific to each operator can be stored for easy reinitiation. The operator should never be required to repeat the same action on the same data.

4.6 New DIA problems

The problems listed above have wanted better solutions for decades. However, new document image research problems have been triggered by the growing availability of several images of the same work in multiple versions (different scans, press runs, editions, languages). The underlying methodology is usually text alignment [34] via dynamic programming (applied earlier to OCR output from multiple classifiers). Text alignment may be conducted at various levels of granularity to discover the common content of two anthologies of poetry, or of two editions of a novel or play, or two newspaper reports, or to spot duplicate passages in two works by different authors. Segment matching may be based on character or word shape, OCR output, spacing of stop words, or on mapping a subset of words with a dictionary. Related techniques are used to construct tables of contents, indices and concordances for significant works, and to evaluate OCR errors [35]. Existing manually created transcripts (e.g., from the Gutenberg Project [36]) of historical works can be used to create training sets for classifiers that align the OCR output with the text image to produce searchable PDF.

5. IMPACT OF TECHNOLOGICAL DEVELOPMENTS ON PERSONAL DOCUMENT CONVERSION

What will it take to fully integrate DIA into the individual, personal and professional computer environment? The major difference between production conversion systems and personal conversion systems is that in the latter the conversion operator and the end-user is the same. In personal conversion even more than in production operations, the look and feel of the interface must suggest that it is there to help the user direct the machine rather than vice-versa.

5.1 Black-box systems

Individual users seldom have the desire or the expertise to adapt the source code driving their equipment. However, the systems must be tunable, because typical end-users require a relatively narrow range of tasks that differ from the tasks required by other end-users. Thus research opportunities exist in developing black-box document conversion systems that are effectively tunable. Inspiration can be drawn from other domains, e.g. the use of Excel macros to adjust spreadsheet behavior.

5.2 Predictability

If errors in document conversion are predictable to the user, then the user is able to adapt to the software effectively. If predictability is lacking, the user becomes frustrated. The predictability of direct-entry systems is currently a key factor in persuading many users to prefer direct entry over automated document conversion. An important research challenge is to devise document conversion systems that are similarly predictable. The effect of predictability on user satisfaction is

illustrated in the math domain. Consider a mismatched parenthesis error: the effect of such an error differs markedly depending on whether it arises during direct entry of math expressions (LaTeX) or during application of tablet-based math recognition software. With the entirely-predictable LaTeX system, the user response to a mismatched parenthesis is "Oops, I made a mistake, I will do better next time"; the user takes responsibility for the error and does not feel disappointed with the LaTeX software. When using tablet-based recognition, the user response is "Darn it, the computer didn't recognize my parenthesis, I will try to draw it more clearly next time". The user becomes frustrated and dissatisfied if subsequent attempts to draw parentheses more clearly are still met with recognition errors. Reducing this type of unpredictability is key to the success of on-line document conversion. It is imperative that either the system learns the user's writing style, or that it clearly shows the user what it is unable to handle, why, and what the user can do to accommodate the software.

5.3 100% Quality Control

Unlike the operator of a mass conversion system, the user of a personal computer is very likely to look at the entire output of a paper-to-digital conversion task. Thus, the user may be willing to spend time finding and correcting recognition errors. As was discussed above for production systems, green interaction is called for: apply machine learning to the corrective actions taken by users, to avoid having users correct the same mistake repeatedly, either within the same document or within different documents. In contrast to commercial systems that process vast quantities of similar documents and set aside the unprocessable residue for manual keying, the architecture of personal conversion systems should be versatile enough to ensure that, given enough human help, any task within the range of the system can be accomplished completely and correctly. Repeated tasks should require progressively less effort.

5.4 Interaction based on meaning rather than appearance

A long-term research challenge is to construct DIA interfaces that request user feedback and correction in a manner that allows the user to think in terms of document information rather than in terms of document appearance. Such an interface supports a natural human-computer interaction, one that feels similar to the human-human interaction that occurs when a shared document is discussed [37]. In a social setting, when a person does not understand aspects of a document, he or she asks about the intended meaning of the document, and only rarely asks detailed questions about the correct segmentation or identity of selected symbols. Given that many of the mistakes made by document conversion software occur at the level of segmentation and glyph recognition, it is a challenge to phrase feedback requests in a form that minimizes the extent to which the user is forced to think about the marks on the paper.

6. CONCLUSIONS

Large changes in the economics of document conversion software and hardware, especially orders-of-magnitude reductions of costs for storage and computation, have so far resulted in little observable change in published research on document conversion software. The dominance of human costs in high-volume conversion projects suggests focusing research on reducing human time rather than on improving the performance of automated, stand-alone algorithms. Another strand of research is suggested by the expected increase in individual document conversion due to low equipment costs. Research on personal document conversion should take into account the differences from mass conversion in the quantity of processed data, higher quality and customization of the desired output, larger range of tasks, significant variability in individual skill and usage levels, jarring effect of unpredictable system responses, desirability of content-oriented interaction, and lack of access to the source code.

ACKNOWLEDGMENT

D. Blostein gratefully acknowledges the support of the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] Moody, D., and Walsh, P., "Measuring the Value of Information: An Asset Valuation Approach," in Guidelines for Implementing Data Resource Management (4th Edition), B. Morgan and C. Nolan, (Eds.), DAMA International Press, Seattle, USA (2002).

[2] Raban, D., and Rafaeli, S., "The Effect of Source Nature and Status on the Subjective Value of Information," Journal of the American Society for Information Science and Technology, 57 (3): 321-329 (2006).

[3] Open Book Alliance, "How Many More Books Has Google Scanned Today?", Feb. 2010, http://www.openbookalliance.org/2010/02/how-many-more-books-has-google-scanned-today

[4] Nagy, G., and Sarkar, P., "Document Style Census for OCR," Procs. First Int'l Workshop on Document Image Analysis for Libraries (DIAL04), pp. 134-147, Palo Alto, CA (2004).

[5] Baird, H., and Casey, M., "Towards Versatile Document Analysis Systems," Procs. DAS 2006, LNCS 3872, Springer, pp. 280–290 (2006).

[6] Esposito, F., Ferilli, S., Di Mauro, N., and Basile, T., "Incremental Learning of First Order Logic Theories for the Automatic Annotations of Web Documents," Proc. Ninth International Conference on Document Analysis and Recognition (ICDAR 2071), Curitba, Brazil, pp. 1093-1097, (2007).

[7] Esposito, F., Ferilli, S., Basile, T., and Di Mauro, N., "Machine Learning for Digital Document Processing: From Layout Analysis To Metadata Extraction," In S. Marinai, H. Fujisawa (Eds.), *Machine Learning in Document Analysis and Recognition*, Studies in Computational Intelligence, Vol. 90, pp. 79-112, Springer (2008).

[8] CBDAR2011, Fourth Internal Workshop on Camera-Based Document Analysis and Recognition, Beijing (2011).

[9] "Digitizing Print Collections with the Internet Archive," http://www.archive.org/scanning

[10] Universal Digital Library web site: http://www.ulib.org

[11] Balk, H., "IMPACT: centre of competence in text digitization," Procs. 2011 Workshop on Historical Document Image Processing (HIP 2011), ACM, NY (2011).

[12] Miller, G., Language and Speech, W H Freeman & C (1981).

[13] Marinai, S., Marino, E., and Soda, G., "Conversion of PDF books in ePub format," Procs. ICDAR 2009, pp. 251-255 (2009).

[14] Colye, K., "Mass Digitization of Books," Journal of Academic Librarianship, 32 (6), pp. 641-645, November (2006).

[15] Open Content Alliance, "Economics of Book Digitization," March 2009, http://www.opencontentalliance.org/2009/03/22/economics-of-book-digitization/

[16] Skipworth, H., "Google Counts Total Number of Books in the World," The Telegraph, San Francisco, August 6, 2010, http://www.telegraph.co.uk/technology/google/7930273/Google-counts-total-number-of-books-in-the-world.html

[17] Berkeley website, "How Much Information?" 2000 http://www2.sims.berkeley.edu/research/projects/how-much-info/summary.html

[18] Oder, N., "Google Book Search by the Numbers," Library Journal, February 12, 2010. http://www.libraryjournal.com/article/CA6718929.html (2010)

[19] Harold B. Lee Library, "Mormon Missionary Diaries – Creating the Digital Collection," http://lib.byu.edu/digital/mmd/create.php (viewed July 2011)

[20] America's Health Insurance Plans (AHIP), Center for Policy and Research, "An Updated Survey of Health Care Claims Receipt and Processing Times," Washington DC, May 2006.

[21] United States Postal Service, 2010 Annual Report. http://www.usps.com/financials/ar/welcome.htm

[22] Markoff, J., "Armies of Expensive Lawyers Replaced by Cheaper Software," New York Times, Vol. CLX, No. 55,335, p. 1, March 5, 2011. www.nytimes.com/2011/03/05/science/05legal.html?_r=1

[23] Federal Reserve System, "The 2010 Federal Reserve Payments Study," December 2010.

[24] Shy, G., “Person-to-Person Electronic Funds Transfers: Recent Developments and Policy Issues,” Consumer Payments Research Center, No. 10-1, March 2, (2010).

[25] Houle, G. F., Aragon, D. B., Smith, R. W., Shridhar, M., and Kimura, D., "A Multi-Layered Corroboration-Based Check Reader," in Procs. IAPR Workshop on Document Analysis Systems (DAS-96), October 1996. [full paper w/images: http://alumnus.caltech.edu/~dave/dasbook/dasbook.html], (1996).

[26] Lopresti, D. and Nagy, G., “When is a Problem Solved?,” Proc. Eleventh International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China (2011).

[27] Lamiroy, B., and Lopresti, D., “An Open Architecture for End-to-End Document Analysis Benchmarking,” Proc. Eleventh International Conference on Document Analysis and Recognition (ICDAR 2011), September 2011, Beijing, China (2011).

[28] Lopresti, D., and Lamiroy, B., “Document Analysis Research in the Year 2021,” in Modern Approaches in Applied Intelligence: Proceedings of the Twenty-Fourth International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, June 2011, Syracuse, NY, LNCS Vol. 6703, Springer, 264-274 (2011).

[29] Lamiroy, B., Lopresti, D., Korth, H., and Heflin, J., “How Carefully Designed Open Resource Sharing Can Help and Expand Document Analysis Research,” Document Recognition and Retrieval XVIII (IS&T/SPIE International Symposium on Electronic Imaging), January 2011, San Francisco, CA, Vol. 7874, pp. 1–14 (2011).[

[30] Zou, J. and Nagy, G., “Visible models for interactive pattern recognition,” Pattern Recognition Letters, Vol. 28, 2335-2342 (2007).

[31] Marinai, S. and Fujisawa, H. (editors), “Machine Learning in Document Analysis and Recognition,” Springer Studies in Computational Inteligence, Vol. 80 (2008).

[32] Zou, J. and Nagy G., “Human-computer interaction for complex pattern recognition problems,” in Data Complexity in Pattern Recognition, pp. 271-286, M. Basu and T. K. Ho, Eds., Springer, (2006)

[33] Zanibbi, R., Mouchere, H., Viard-Gaudin, C., and Blostein, D., “Stroke-based Performance Metrics for Handwritten Mathematical Expressions,” Proc. Eleventh International Conference on Document Analysis and Recognition (ICDAR 2011), Beijing, China (2011).

[34] Yalniz, I. Z., Manmatha, R., “A Fast Alignment Scheme for Automatic OCR Evaluation of Books,” International Conference on Document Analysis and Recognition (ICDAR’11), Beijing, China (2011).

[35] Yalniz, I. Z., Can, E. F., Manmatha, R., “Partial Duplicate Detection for Large Book Collections,” International Conference on Information and Knowledge Management (CIKM’11), Glasgow, UK (2011).

[36] Gutenberg Project, http://www.gutenberg.org/

[37] Hayakawa, S.I., Language in Thought and Action, 5th edition (1991).