

Interactive Document Processing and Digital Libraries

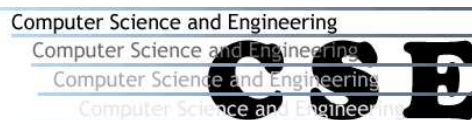
George Nagy *Daniel Lopresti*

February 2006

Technical Report LU-CSE-06-019

Department of Computer Science and Engineering
Lehigh University
Bethlehem, PA 18015 USA

<http://www.cse.lehigh.edu/>



Interactive Document Processing and Digital Libraries*

*George Nagy*¹ *Daniel Lopresti*²

¹ Department of Electrical, Computer, and Systems Engineering,
Rensselaer Polytechnic Institute, Troy, NY 12180
`nagy@ecse.rpi.edu`

² Department of Computer Science and Engineering,
Lehigh University, Bethlehem, PA 18015
`lopresti@cse.lehigh.edu`

February 2006

Abstract

We explore connections between digital libraries and interactive document image analysis. Digital libraries can provide useful data and metadata for research in automated document image analysis, and allow unbiased testing of DIA algorithms. With these goals in mind, we suggest criteria for constructing and evaluating interactive DIA tools.

Discussion

We consider some involuted relationships between digital libraries and document image and content analysis.¹ Exploiting these relationships may accelerate all-around progress. Although we have worked on pieces of these puzzles for many years, we are acutely aware of the need for input by researchers with different backgrounds to channel further research. The questions we propose to explore are the following.

1. How can researchers benefit from the large variety of *documents* (in coded or image form) available in digital libraries to construct data sets for experimenting on specific aspects of document processing?
2. How can researchers make use of the *metadata* available in digital libraries to target problems whose solution would automate, or partially automate, tasks currently performed quasi-manually by librarians and curators?

*Presented at the *Second International Conference on Document Image Analysis for Libraries*, Lyon, France, April 2006.

¹Warning: this paper contains no theory or experiments, and cites references only in the case of the online resources used to illustrate our discussion. Rather, this work is largely an opinion piece reflecting our speculation regarding potential synergies between document analysis research and digital libraries.

3. Can digital library (DL) holdings be used to establish benchmarks for measuring the success of competing approaches without allowing researchers to “train on the test set,” i.e., over-fitting their algorithms, either by design or by training, to a specific collection?
4. What functions should interactive tools perform to facilitate either of the above tasks?
5. What functions performed by an interactive document tool should first be partially automated, and what functions should be reserved for human judgment until significant progress in the state of the art?
6. How can interactive tools for document processing or annotation be tested and validated on statistically significant sample data, without the possibly vicious loop of testing against data sets annotated by similar interactive tools?

Q1. Use of existing document collections to further DIA research.

The title of this section is narrower than Question 1 posed above because the IR community is already making good use of existing document collections to evaluate diverse approaches. An example is the subsets of documents (or abstracts) in the contests of the *Text Retrieval Conferences (TREC) Genomics Track* extracted from the *Mouse Genomics Informatics (MGI)* database of curated publications. Other microbiology contests extract documents from *PubMed* and *NIH MedLine*.

There is a difference of about two orders of magnitude in the size of, and the processing time required for, text in coded and image formats. Consequently IR researchers have no alternative to using existing collections.

Most DIA publications, on the other hand, are based on ad hoc collections of documents assembled by the researchers themselves because they are rich in features of interest to their particular research task. Although the collection, annotation and documentation of such test databases is not a trivial task, we seldom see much reuse by different groups of researchers, except possibly in optical character recognition (OCR), especially in China and Japan.

Until recently, research teams often collected a set of hardcopy documents, and then scanned them at whatever resolution seemed appropriate to the specific task. Within the constraints imposed by their task, researchers strove for diversity. Although this approach tested the range of applicability of the algorithms, it precluded experimentation on adaptation to a large, relatively homogeneous set of documents. Many applications must contend with highly repetitive material: for example, some firms do nothing but convert telephone books to computable readable form.

Whether partial processing of documents is valuable by itself may be open to question, but end-to-end document processing requires a large team with varied resources, and is beyond the capabilities of most university researchers. It would therefore be valuable to have tools that allow the extraction of document collections with specific characteristics - including degree of homogeneity or heterogeneity - from digital libraries.

Collection tools for DIA research require some database of digital libraries with downloadable page images, and a search engine capable of searching the database (or the whole web). The first step is the location of one or more collections with images of the desired type. (Scanned document image formats are often subjected to significant manipulation when added to a digital library). The collection routine must (1) determine on the basis of the metadata of the DL whether the image should be included, or (2) categorize the image with whatever image processing tools it has available in order to decide whether it is acceptable, or (3) present each candidate page to a human screener, who accepts or rejects it.

The three alternatives can of course be combined, normally in the given order. Alternative (1) depends on the availability of the right kind of metadata. Most metadata in digital libraries is still modeled on library catalog cards: source or publisher, edition or copy, author, date, summary, number of components of various kinds (e.g., number of volumes, sections, chapters, pages, illustrations, tables, references). They do not report page skew, contrast, noise, or the coordinates of tables, figures, and text lines. Alternative (2) requires dependable image processing of exactly the kind that we are trying to develop. However, even relatively low accuracy may be acceptable if (2) is combined with (3).

Alternative (3) is perfectly reasonable even by itself for collections of to tens of thousands of images, *provided* that the target DLs are *dense* in the target types. Otherwise the screener must look at far too many “rejects.” The major shortcoming of (1) is that the screener may be part of the research team and, consciously or subconsciously, reject images that would make the proposed algorithm fail.

The paradigm we are proposing was employed in an earlier small-scale study we reported at the 2002 Workshop on Document Analysis Systems involving the Making of America collection [Mak06] (part of Cornell University’s Digital Library), which at the time comprised 267 monographs (books) and 22 journals (equaling 955 serial volumes) for a total of 907,750 pages, making it three orders of magnitude larger than the datasets traditionally used in document analysis research (e.g., the UW1 CD-ROM). Two tasks were evaluated: optical character recognition and table detection. In the case of the former, the textual transcriptions provided by the digital library (primarily for retrieval purposes) were used as the ground-truth, while for the latter, a manual perusal of the pages purported to contain tables was conducted, enabling precision (but not recall) measurements.

Lacking an authoritative index to the digital library, a novel approach for generating random samples was used. This began by issuing a query to the library’s web interface by choosing a random term from the Unix *spell* dictionary (which contains 24,259 words including a number of proper names). From the results of this search, one of the works (book or journal) returned was randomly chosen, and from that work was selected a specific page that contained a match to the query. The implementation of the web interface was automated by programming it using the popular Tcl/Tk scripting language.

Performing such evaluations with only general *a priori* knowledge of the test images and their ground-truths proved to be effective at identifying unforeseen limitations in the document analysis algorithms under study, which is one of the major benefits we are advocating.

Q2. How can metadata in digital libraries help to focus DIA research?

Although digital media have freed the metadata from the size limitation of 3 by 5 cards, further evolution is likely to be slow because of the need for compatibility with existing archives, databases, catalogs, and other library tools. It therefore seems that the DIA community could help the DL community by focusing not only on tools that allow the conversion of hardcopy to both image and coded form, but also on those that allow more rapid generation of metadata.

As an example suggestive of many of the points we are attempting to make, we cite a handwritten manuscript found in the Lehigh *I remain* digital library [I r06] entitled *Synglosson: Fifth Book of Vocabularies of the Languages of Asia, Africa, Europe, & Polynesia* by Constantine Rafinesque, dating from 1832 [Raf32]. Rafinesque was a botanist and linguist who cataloged not only the flora and fauna he encountered in his travels, but the languages as well, including Chinese, Japanese, Arabic, Polynesian, Australian, Indian, Burmese, “Corean,” and Malayan, among others. A screen snapshot of a page from the Synglosson showing excerpts of his word lists for Chinese is shown in Figure 1 and presents interesting problems in off-line handwriting recognition as well as in page layout analysis. The handling of the anglicized transcriptions of terms from other languages could likewise prove a challenge. The entire manuscript comprises 89 pages (178 page images, recto and verso), all of which are freely accessible on the web, offering attractive opportunities for intra-document analysis.

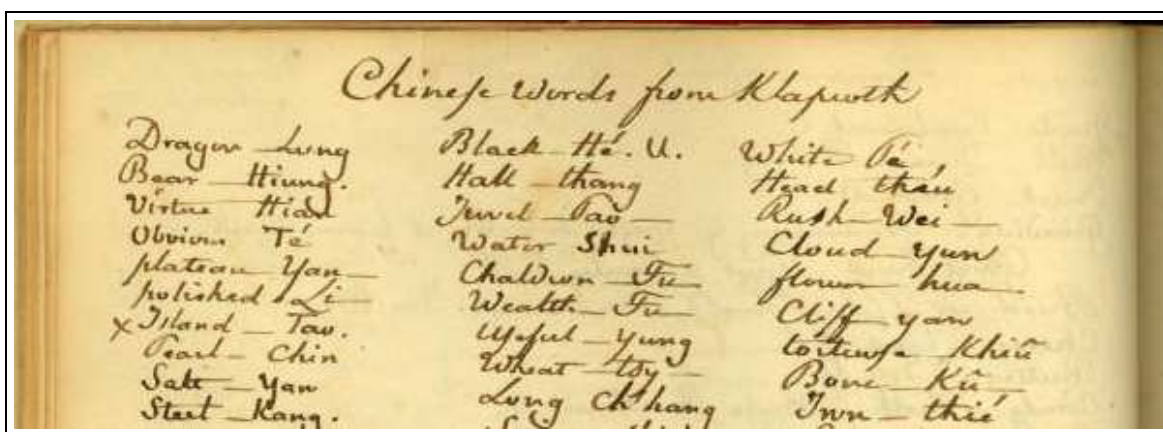


Figure 1: Portion of a page image (015 verso) from Rafinesque’s Synglosson [Raf32].

A screen snapshot of the online metadata accompanying the Synglosson is shown in Figure 2. Here we note standard fields such as “Title,” “Personal Author,” “Date,” “Extent” (length), “Accompanying Material,” “General Note,” “Abstract,” “Personal Subject,” “Subject,” and “Geographical Subject.” It is evident that most of these categorizations go beyond what is contained in the document itself and/or require human intelligence and interpretation to generate. Hence, it is important to recognize that full automation may be an impossible goal, but developing tools to facilitate the human in his/her task of creating metadata is entirely within the realm of feasibility. We should also point out that, as is

currently the case with many digital libraries, a full textual transcription of the Synglosson does not exist online. The development of robust off-line handwriting recognition up to this task would be a boon to both the builders and the users of digital libraries.

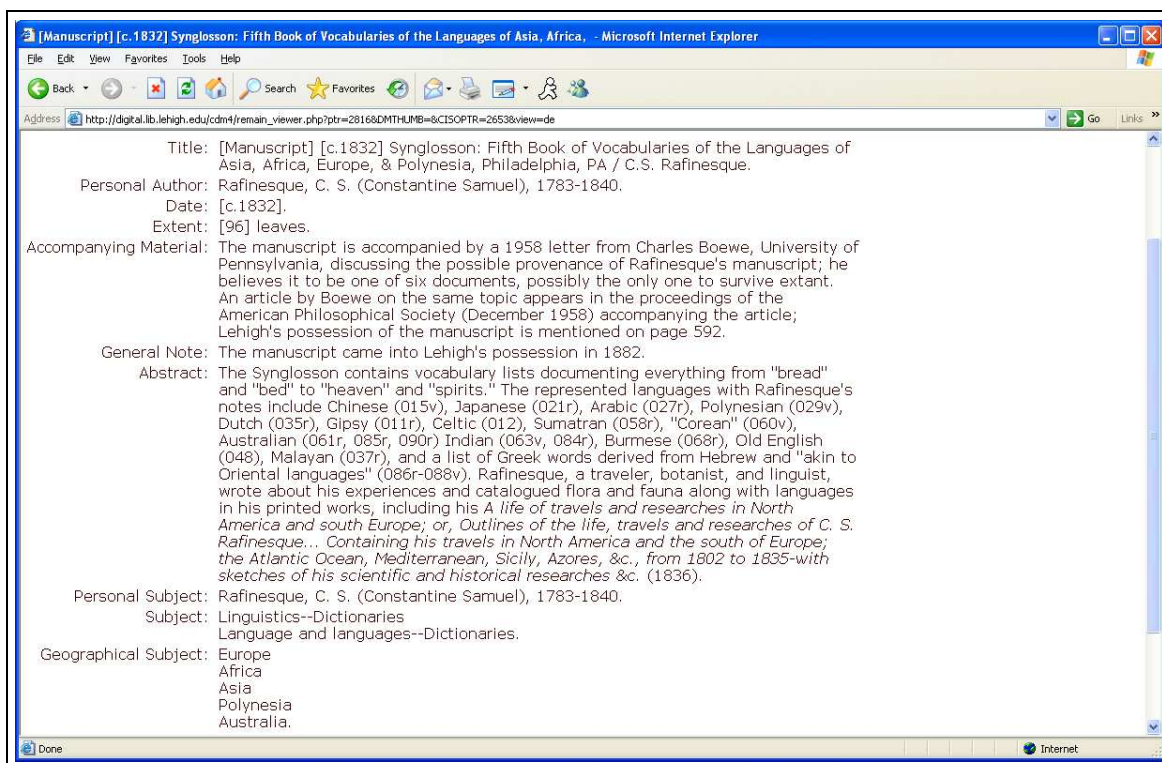


Figure 2: Metadata for Rafinesque's "Synglosson" [Raf32].

The requirements for research on adequate, "loss-free" (re)production of documents in digital form, and more particularly in coded form, were already considered extensively at the previous DIAL and at many DIA conferences. We therefore briefly examine here the production of metadata. This need seems particularly urgent in scientific disciplines that produce an enormous number of publications (millions per year), such as nano-materials, genomics, and molecular biology. It is, of course, also important in business, arts, and news repositories. Although digital libraries almost universally annotate the documents by adding metadata, it is possible that the same research that leads to the automation of annotation will lead to search engines that can access documents without any metadata. We also note that in specialized collections, such as the MGI, the metadata itself is highly specialized, and is often added by subject specialists (curators with a PhD in biology) rather than traditional librarians.

One way to link DIA with the high-level information required by digital library users is to facilitate the conversion of documents into coded digital form for greater ease of human access. The most time consuming part of annotation is the extraction of substantial information from the document. This is a challenging intellectual task requiring specialized

training and broad context, which is not likely to be automated soon.

It is possible that the encoding of documents for librarians and curators, although superficially the same task that we have been working on for decades, is vastly different from that required for other users. One DIA tool that could be useful for all is *text-image reformatting*. After text- and word-level segmentation, such a tool breaks long text lines and reformats the text image to the size of the available window, much as MS-Word or Netscape reformat text coded in .doc or .html format.

Librarians should also be able to rapidly deconstruct documents and to look at and reorder simultaneously all the authors, all the abstracts, all the references, or all the figures in a set of related documents. This needs DIA layout tools. Nevertheless, as long as humans write primarily for humans, we believe that DIA can have only a limited impact here. This may soon change. Already we see web pages constructed to attract the attention of search engines rather than of human readers, and abstracts and keywords in scientific articles that seem to aim at high frequency of machine retrieval rather than reflecting the actual content.

Q3. Can digital libraries help establish benchmarks for DIA?

There have been relatively few successful benchmarks in DIA. Even when researchers use the various publicly available collections of scanned documents, they typically extract different subsets for their training, validation, and test sets, rendering meaningful comparison impossible. Furthermore, any algorithm can be “improved” after observing the errors it makes. This accounts at least partially for the increasing accuracies obtained on a few popular data sets without strict specifications for partitioning.

The best tests are those where the programs are made available to the organizers of the contest, who run them on a set of documents which has not been released to the competitors. Such competitions require an enormous amount of preparation on the part both of the organizing committee and the competitors, and strict standards for interoperability must be met. As a compromise, the test data is sometimes released to the researchers for only a limited period, say during a conference or workshop. (There have been only a few reported instances of contestants modifying their programs after a quick look at the data.) The advent of powerful laptops has made it much easier for researchers to run the necessary experiments on their own platform. Nevertheless, the competitions organized in conjunction with various DIA conferences have attracted relatively few participants.

Instead of sample-by-sample identical test sets, it may be possible to specify *statistically equivalent* test data. This requires independent sampling from a larger set of data. The size of the sample necessary to avoid significant fluctuations in the results of the tests due to sampling alone depends on the variance of the significant features of the test sample - for example typeface, type size, printer and scanner resolution and quality, layout, illustrations. For now, document samples large enough to avoid sampling error, yet small enough to be processed in a reasonable amount of time, may be possible only in narrow domains. Examples of such domains may be hundreds or thousands of pages from a given newspaper (*Le Figaro*), journal (*PAMI*), magazine (*Paris Match*), map series (USGS 15 quads), or from a particular series of books from a publisher (e.g., Random House’s *The Modern Library* or

Gallimard's *Le Livre de Poche*), selected from a period without a change in format. With a large enough source database, such a statistical dataset would preclude most egregious training on the test data. Improvements in results would be more credible than repeated tests on the same data, but they still could not be generalized to other sources.

Independent statistical sampling from a finite number of objects is quite straightforward. It is, however, very difficult to collect a "random sample" from the web. There is no uniform probability distribution on the web any more than on the real line. This is also true for subsets of the web: we do not know how to collect a random sample of technical journal pages, logs, or advertisements, because the web and its subsets, and many DLs, are neither static nor fixed in size.

An often neglected issue is the granularity of sampling. Small-grain independent sampling does not reflect the characteristics of document streams encountered in practical applications. If the last page was cleanly copied hand-printed Turkish, then this page is more likely Turkish than Telugu. If the last page showed a confusion table for one algorithm, then the next page is more likely to show a confusion table for another algorithm than a weather map. Sampling must be performed at large enough grain to preserve such correlations, because good algorithms should exploit them.

The above training and testing scenario was based on the notion of a representative training set, meaning that the training set is obtained from the same distribution as the test set. Mathematical and statistical analysis of the performance of machine learning algorithms is well-nigh impossible without assuming this. But automated systems must keep operating even if the data changes. Typesetting, layout, and even handwriting keep evolving. The formats of newspapers, magazines and technical journals are quite different from what they were only twenty years ago. The change is, of course, most marked in HTML pages where format is determined jointly by the designer and the viewer. The notion of a representative training set is actually an anachronism: we must develop robust trainable algorithms that work well on data statistically different from what they were trained on. This requires further research on quantification of the statistical differences between datasets, which in turn requires lots of new datasets. Digital libraries are the only possible practicable source.

Q4. What functions should an interactive tool have?

From the above discussion it appears that interactive DIA may be useful for (1) assisting research on automated DIA, (2) collecting metadata required for digital libraries, and (3) improving the actual conversion of hardcopy documents to digital form for accession to digital libraries. We focus on (1) and (3).

For producing data and groundtruth for research, document screening, categorization, segmentation, and markup are important. *Screening and categorization* require only the rapid display of pages from selected DLs and an interface with two or more buttons (already provided by most browsers). The selection of DL needs a list of URLs, or an appropriate search engine that checks image formats as well.

Segmentation is more complex, but automated systems have not yet reached human

accuracy. This operation requires a display of all or part of the document image, a point-and-drag mechanism to select regions (in most applications, *rectangular* regions), and a set of buttons to associate region coordinates with component types. This mechanism also suffices for *markup*. First characteristics of the entire document are marked: *language, script, text only or text-and-graphics, machine print or handwriting, technical article, conference paper, etc.* Then the regions are marked up hierarchically. For instance, if graphics is selected then, depending on the document category, buttons for *organization chart, circuit diagram, or chart* may appear.

The entire operation consists only of a succession of *click-and-drag-on-page-image* and various *button-selection* steps. Some buttons may be the mouse buttons themselves rather than screen icons. Fields for free-text notation may be useful to describe anomalies unforeseen in the tool design, or for selection from many alternatives. The tool should not be hard-wired for any particular application, but provide a separate, easy to modify interface for tuning it to any particular collection. (The first version of our *Docutool*, built along these lines 20 years ago, suffered most from the mismatch between display and image resolution. Even today, some applications may require coordinated multiple screens.)

In Figure 3, we show a screen snapshot of *Clicker*, a prototype tool we are developing to assist in the creation of document metadata. *Clicker* is written in Tcl/Tk and is designed for the rapid collection of a variety of high-level attributes of interest to digital librarians and document analysis researchers. The tool features user-configurable hot keys, and it logs every action taken by the user for possible later analysis, e.g., for training recognition algorithms, in addition to studying user behavior as well as categorizing the documents themselves. *Clicker* is web-enabled so that the documents do not need to “live” on a local filesystem - they can be be images anywhere on the web (i.e., pages from a digital library).

As noted previously, digital libraries often include multiple pages from the same source. *Clicker* facilitates the annotation of such images by allowing a new page to be assigned the metadata created for a previous page by default. Preliminary benchmark results have shown that for the high-level annotations that *Clicker* collects, a trained user requires 20 seconds on average to process each page. With the default option set, two mouse clicks are generally sufficient to update the new mark-up.

Moving beyond the simple sort of metadata collected by *Clicker*, manually processing the complete description for a page in a digital library could be slow and hideously boring. For instance a tool we built a few years ago required upwards of 16 hours locating and orienting street names on a map of the District of Columbia. (After manual demarcation, the street names were automatically extracted and rotated to horizontal for ease of OCR or key entry.) In the next section, we discuss some possibilities for speeding up the interaction without giving up final operator control of all accepted entries.

Q5. What is the most effective transition from interaction to automation in DIA?

The suggestion here is that interaction interspersed with algorithmic document processing is more efficient than correction of fully automated processing after proofreading. And

with a well designed architecture, it need not slow down the overall workflow.

Because of the correlation between consecutive documents, and between components in the same document, a *default entry* based on the previous entry should be set for all selections. This is, in fact, what *Clicker* does. More often than not, the operator then needs only to point to all the various fields of type x with identical characteristics y .

Many current DIA algorithms already achieve acceptable accuracy at tolerable reject rates. They can therefore be incorporated in an interactive tool, subject to override by the operator. In some cases, the results can be accepted even without human approval: at DIAL04 we reported on *triage* for OCR, which exempts documents processed with a high confidence level from proofreading. We present some examples of combining algorithmic and interactive processing.

The first step after scanning a document is often either binarization or contrast enhancement. Automated algorithms do not work perfectly on pages with a wide distribution of spatial reflectance. However, when they fail, an operator can easily set the appropriate window size and foreground density either for the whole document, or for selected areas. This still allows local algorithmic processing, and therefore requires much less interaction than setting the parameters manually everywhere, and is much more robust than fully automated local thresholding.

Line finding is another instance where augmenting interaction algorithmically may be effective. In older printed matter, as in handwriting, the orientation of individual lines may change, margins may be ragged, there may be only a few words on a page, and there may be several columns of words or phrases at different angles. Humans can, however, convey this information to the computer by a few well chosen mouse taps or by rotating a superimposed grid. After the computer-proposed skew correction and line finding is corrected, merged pairs of lines can be likewise rapidly separated.

Word segmentation is relatively easy for printed text, except for extremely tightly-set print. In handwriting, however, large spaces often appear within words and, towards the end of a line, words are often squeezed together. In Arabic and other scripts, some inter-letter spaces are mandatory. Underlines can further complicate the task. A simple interface can be designed to correct linked and broken words after line segmentation. Higher-level segmentation algorithms tend to be more error prone, and therefore require higher reject thresholds. Even a 30% reduction in overall human time will be significant in an operational application.

There are opportunities for effective interaction also at the character recognition level. Humans can often tell where perfect accuracy is important, as in proper nouns and dates. If necessary, they can be either entered manually, or selected from the top recognition candidates or from similar items encountered earlier in the document.

The human can also provide global assistance to the character recognition algorithms by indicating the language and script, number of columns, average slant, and in Western scripts, the prevalent type case of a document. The operator may also decide which of the available lexicons would provide the best language model. (The lexicons can be automatically updated with entries from the processed documents that have been deemed correct.)

Most importantly, processing interactively only part of a document may provide enough

training data - to a recognition system designed with this in mind - for fine-tuning the classification algorithms. The underlying assumption is that if the remainder of the document (and perhaps also additional documents) is from the same source, the adjusted parameters will yield more accurate recognition. If that is not the case, the operator can easily separate the portions of the document from different sources.

The order in which DIA algorithms are developed or modified for interactive application can be determined by analyzing their error/reject curves in relation to the time required to perform the same task purely interactively.

Q6. How can interactive tools be tested and validated?

Interactive tools can often achieve accuracy comparable to that of any available groundtruth. The most significant variable, which is responsible for most of the expense of human annotation, is human time. Because of the importance of this variable, precise control of experimentation is necessary. Sound experimental design should account for factors such as education, language and computer skills, training, and experience with the tool. Training and experience are not easily separated. It is perhaps best to provide only *impersonal* training (a manual or computer tutorial, and some practice session sessions), and then monitor the change in the operators speed as a function of experience with the actual data.

The factors that are not generally part of the page samples themselves, but represent the contextual knowledge of the operator, are difficult to quantify. Their absence in DIA routines accounts for the difference in accuracy between human and machine in most tasks.

Any interactive tool must be able to log every keystroke and every mouse click. Most current computing environments render this relatively easy. Again, *Clicker* already incorporates this functionality. Computer response time need to be monitored only to ensure that it does not slow down the operator. In many current systems, the permissible half-second or so is exceeded only when a new page image is loaded. The more difficult part is the transformation of this mass of data into a form that can be readily analyzed with standard spreadsheets or statistical tools. The time logs must be integrated with the information that represents the characteristics of each task and document. For example, the number of mouse clicks necessary to segment a document into text, drawing and photo categories may indicate the complexity of its layout. (The development of the logging and analysis routines of our CAVIAR systems for flower and face recognition took about six person-months.)

What do we want to find out from analyzing interactive document image analysis? A good log should allow not only statistical summarization (e.g., mean and standard variation of human time to locate street names on a map or ads in a magazine, difference between operators in assessing skew angle or "OCR quality"), but also allow complete replay of all or part of every interactive session. Such a digital replay is likely to prove superior for improving the tool than the traditional observation through a one-way window or video recording.

Aside from estimating the time necessary for different types of annotation and different types of document, perhaps the single most important question is how human time can be reduced by the accuracy and reliability of automated components. As machines take over

the simpler tasks, more and more expertise will be required of the operator. The cost of human labor, already the dominant factor in most information processing tasks, will become even more significant.

Conclusion

Traditionally the DIA, IR, and HCI (Human-Computer Interaction) communities have been distinct, although several conferences (DR&R, SDAIR) have targeted the first two. The encoding of printed, typewritten and even handwritten text has been in the realm of DIA, while the processing of coded text was IR. Human factors and interaction are in HCI. We are all aware, however, of the importance of layout, figures and tables in scientific publications. Gestalt vision and text comprehension remain beyond the present ability of computers. Perhaps in time the three communities will unite to make use of *all* of the available information to speed the flow of human knowledge into digital libraries.

Acknowledgments

Daniel Lopresti gratefully acknowledges the support of NSF grant #0430178 and a DARPA IPTO seedling grant administered by BBN Technologies. George Nagy gratefully acknowledges the support of NSF grants #0414644 and #0414854.

References

- [I r06] *I remain: A Digital Archive of Letters, Manuscripts, and Ephemera*, Lehigh University Library, February 2006. <http://digital.lib.lehigh.edu/remain/>.
- [Mak06] *Making of America*, Cornell University Library, February 2006. <http://cdl.library.cornell.edu/moa/index.html>.
- [Raf32] Constantine S. Rafinesque. *Synglosson: Fifth Book of Vocabularies of the Languages of Asia, Africa, Europe, & Polynesia*. 1832. http://digital.lib.lehigh.edu/cdm4/remain_viewer.php?ptr=2816.

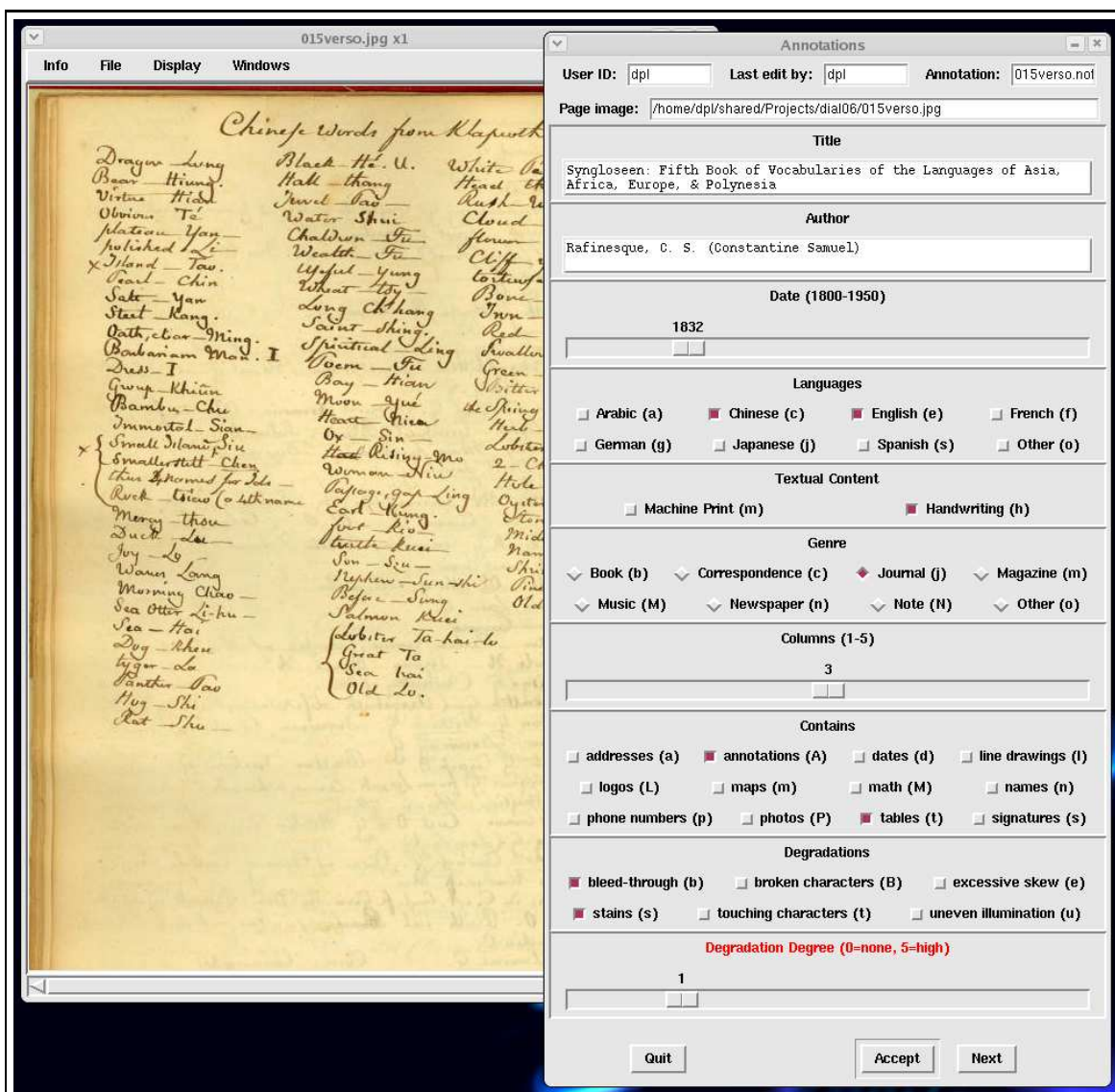


Figure 3: Clicker, a prototype tool for collecting document metadata.