Digital Libraries and Document Image Analysis

Henry S. Baird

Palo Alto Research Center
Palo Alto, CA 94304 USA
`baird@parc.com`

Abstract

The rapid growth of digital libraries (DLs) worldwide poses many new challenges for document image analysis (DIA) research and development. DLs promise to offer more people access to larger document collections, and at far greater speed, than physical libraries can. But DLs also tend, for many reasons, to serve poorly, or even to omit entirely, many types of non-digital human–legible media*, such as originally printed and handwritten documents. These media, in their original physical (undigitized) form, are readily — if not always quickly — legible, searchable, and browseable, whereas in the form of document images accessed through DLs they often lose many of their original advantages while of course lacking many advantages of symbolically encoded information. The author explores these issues and illustrates them with brief case studies arising from his experience as a DIA researcher in collaboration with several DL projects in the US. Difficult open DIA technical problems in DL applications are identified in the contrasting advantages of paper and digital displays, at every stage of capture, early processing, recognition, analysis, presentation, & retrieval, and in personal and interactive applications. These support the conclusion that the international DIA R&D community is urgently needed (because uniquely qualified) to provide new technology to help rescue from neglect — even, in many cases, eventual oblivion — the world's vast culturally irreplaceable legacy paper document collections.*

1. Introduction

As more and more information is captured electronically in symbolically *encoded* digital form (*e.g.* ASCII, Unicode, XML), and as more of these resulting data are made avail-

---

able on–line as digital libraries and other web resources, any information that is *not yet encoded* — although still readily human-legible — risks being trapped in a second–class status where it is comparatively difficult to find, access, read, understand, and otherwise reuse. Electronic "digital libraries" (DLs) are designed (in part) to mitigate these difficulties, and to this end they routinely embrace scanned digital images of originally printed and handwritten materials. In fact, much of the excitement surrounding DLs comes from the hope that many 'lost' texts — out of print, deteriorated, mutilated, locked in archives, etc — will find their way back into circulation. But serious obstacles prevent merely *imaged* documents within DLs from playing all the useful roles that encoded documents do, or even many of the roles that their original physical embodiments did. This paper explores these obstacles, discusses the state of the art of document image analysis (DIA) methods relevant to them, and lists open technical problems which the DIA community is uniquely qualified to attack. Solving these problems is, I will argue, an urgent priority: there is reason to expect that much of the knowledge stored in the world's vast legacy collections of paper documents is in danger of being abandoned in an unconsidered rush to a hegemony of 'purely digital' collections of information.

I will focus largely on fresh technical, especially algorithmic and methodological, DIA challenges triggered by DLs. For reasons of space — and due to my limited experience — I will not attempt a thorough review of the state of the art of DLs; in particular I confess and regret my narrowly US-industry-centric perspective. Neither will I discuss every DIA technique relevant to DLs: this Conference's Proceedings alone makes that redundant and futile. I reluctantly but thoroughly ignore financial implications (such as DL business models and impact on publishing revenue), legal obstacles (*e.g.* copyright), and maintenance problems (migration to new storage media, etc). Many of these vital considerations are dealt with in [15] [23] [27] and their references.

The relative advantages of physical (non-digital) doc-

ument media compared to encoded digital media are discussed in Section 2. Section 3 considers document–image capture and its consequences for human and machine legibility, completeness of collections, support for scholarly study, and archival conservation. Early image processing in support of quality control and compression is addressed in Section 4. Section 5 summarizes the large number of open obstacles to the fully automatic, high-accuracy analysis of the content of document images. Presentation, display, printing, and reflowing of document images are discussed in Section 6. Retrieval, indexing, and summarization of document–images is the subject of Section 7. Section 8 is devoted to 'personal' and interactive digital libraries. The urgency of these problems lying at the intersection of DLs and DIA is assessed in Section 9.

To illustrate concretely how the problems arise in practice, I have inserted brief case studies chosen from collaborations between PARC's DIA research teams and DL R&D projects at U.C. Berkeley, Carnegie-Mellon Univ., and Xerox Corporation.

2. Ink-on-Paper *versus* Digital Displays

The physical properties of high-quality paper makes certain helpful functions ('affordances') possible or easier for users [35], including:

- lightweight, and so usually easy to carry, hold, and position;
- thin, and so easy to grasp;
- flexible, and thus convenient to position, bend, and fold;
- reflective, able to be illuminated for a wide range of brightnesses and contrasts;
- markable by a variety of means in a simple and uniform manner;
- allowing detailed high-resolution markings;
- opaque and two-sided, and so efficiently legible on both sides;
- unpowered, and so portable and 'always-on';
- stable, and so self-conserving and maintenance-free for many years;
- cheap and movable, so many can be used, *e.g.* spread out side by side; and
- simple, easily learned, and widely understood methods of use.

Digital display technologies used by today's DLs to deliver document images — a rapidly evolving ecology of desktop, laptop, and handheld computers, plus eBook readers, tablet PCs, etc — offer often contrasting affordances, *e.g.*:

- automatically and rapidly rewritable;
- interactive;
- connected (*e.g.* wirelessly) to a network and so can deliver potentially unlimited data;
- radiant/back-lit, and so legible in the dark, but often limited in range of brightness and contrast; and
- sensitive (to, *e.g.*, touch, capacitance), and so markable.

This catalogue is incomplete but long enough to suggest the multiplicity of ways in which information conveyed originally as ink-on-paper may, and may not, be better delivered by electronic means favored by DLs (for an extended discussion, see [21]). One result is that, as Sellen and Harper [35] report, "paper [remains at present] the medium of choice for reading, even when the most high-tech technologies are to hand." They point to four principal reasons for this:

1. paper allows "flexible [navigation] through documents;"
2. paper assists "cross-referencing" of several documents at one time;
3. paper invites annotation; and
4. paper allows the "interweaving of reading and writing."

It is illuminating to bear these considerations in mind when identifying obstacles to the delivery of document images via DLs.

Of course, efforts are underway [5][7] to commercialize electronic document displays offering even more of the affordances of paper including flexibility, low weight, low power, and low cost.

3. Capture

The capture of document images for use in DLs is often carried out in large-scale batch operations. The operations are almost always designed narrowly to meet the immediate purpose. For reasons of cost, only rarely will the documents ever be rescanned. In fact, documents can be damaged or destroyed in the process, sometimes deliberately: *e.g.* spines of books cut off to allow sheet-fed scanning. Even more drastically, many libraries have discarded large collections of books and documents after they have been scanned, triggering anguished charges [16] of violations of the public trust. The wretched image quality of many microfilm archives of documents scanned in the 1950's and 1960's is a particularly egregious precedent. The Society of American Archivists has defended professionally designed scanning operations followed by destruction or deaccessioning, as reported in [20]. This debate will certainly not halt, and may not even slow, the replacement of hardcopy documents

with scanned images, but it highlights the urgency of controlling scanning so that the resulting images will serve a wide variety of uses for many years.

3.1 Quality Control

Image quality is most often quantified through the technical specifications of the scanning equipment, *e.g.* depth/color, color gamut and calibration, lighting conditions, digitizing resolution, compression method, and image file format.

3.1.1 Scanner Specifications

In the recent past, most large-scale document scanning projects, constrained by the desirability of high throughput and low storage costs, produced only bilevel[1] images; this is now yielding rapidly to multilevel and color scanning. Digitizing resolutions (spatial sampling frequency) for textual documents typically range today between 300 and 400 pixels/inch (ppi); 600 ppi is less common but is rapidly taking hold as scanner speed and disk storage capacity increase.

For what downstream uses are these rough guidelines sufficient? Tests of commercial OCR machines in 1995 [34] showed that accuracy on bilevel images fell for documents digitized at less than 300 ppi but did not appreciably rise at 400 ppi. They also showed that some systems were able to exploit the extra information in greylevel images to cut error rates by 10%–40%. Researchers have suggested that greylevel processing will allow OCR to read documents whose image quality is now well below par [32][41]. Of course, many document images are printed in color, and the costs of color scanning and of file storage and transmission are falling rapidly: the DIA research has, within the last five years, begun to take this challenge seriously – but, in my view, not as fast as it should.

Some attempt has been made to issue refined scanning standards. The Association for Information and Image Management (AIIM) [3] publishes standards for the storage of textual images, including ANSI/AIIM MS-44-1988 "Recommended Practice for Quality Control of Image Scanners" which defines "procedures for the ongoing control of quality within a digital document image management system." It is designed predominantly for use in bilevel imaging. MS-44 test targets include:

- IEEE Std 167A-1987, a facsimile machine test target that is produced by continuous-tone photography, with patterns and marks for a large range of measurements of moderate accuracy;
- AIIM Scanner Target, an ink-on-paper, halftone-printed target; and

---

[1] black and white only, 'bi-tonal,' 'binary'

- RIT Process Ink Gamut Chart, a four-color (cyan, magenta, yellow, and black), halftone-printed chart for low accuracy color sensitivity determinations.

DIA professionals are well equipped to investigate the utility of existing targets for the manual or automatic monitoring of image quality. It may be necessary to design new targets to assess aspects of image quality directly relevant to the success of later DIA processing.

Far higher standards are required for certain archival and scholarly uses[6][14]. To pick one example: the US Commission on Preservation and Access [4], in a 1995 report on the "Digital Imaging of Papyri," recommended:

- full continuous-scale, calibrated color;
- 600 ppi for primary archival images, 300 ppi for larger pieces, > 600 ppi for pieces with unusually high information density;
- 24–bit TIFF (Rev. 6.0) file format;
- do not perform image compression for archival files; use JPEG for Internet transmission;
- include textual identifying information (ruler, color scale, watermark);
- keep metadata ('management data') in separate data base files; and
- system designs minimizing heat and light damage to originals.

Recent AIIM proposals relax these, to allow compression and metadata storage in the file. A joint activity between AIIM and the Association for Suppliers of Printing, Publishing and Converting Technologies (NPES) is discussing an international standard (PDF-Archive) [10] to define the use of PDF for archiving and preserving documents.

The Octavo Digital Imaging Laboratory [1] scans rare books for digital preservation and scholarly study, using 24-bit color up to 800 ppi. Paul Needham, the Scheide Librarian at Princeton University, along with physicist Blaise Agüera y Arcas, studied such images of Princeton's copy of the Gutenberg Bible and discovered [42] that the received wisdom about the invention of 'type cast with a reusable matrix" must be revised.

The DIA community has begun to involve itself in recommending standards for document-image scanning, especially to allow scholarly investigations sensitive to details of paper, ink, bleedthrough, evidence of age, marks of history of use, indicators of changing cultural context of the document, etc. The DAS'02 Working Group on Digital Libraries [36] floated the idea of a tiered set of standards, *e.g.*:

- any document printed before the year 1600 should be scanned at 2000 ppi;
- documents from 1601-1800 at 600 ppi; and
- documents 1801 and later at 400 ppi.

3.1.2 Measurement & Monitoring of Quality

Since specifications of printing and scanning conditions are rarely preserved and attached to the images, tools for the automatic estimation of scanner parameters from images of text could be an important contribution to the success of DLs. Work along these lines is well underway in the DIA community (*e.g.* [37]). At their current state of development it is not yet clear when they will work robustly while running fast run enough to be applied to every page image in high–throughput scanning operations. Thus an important open problem is how to monitor image quality in real-time to ensure that all downstream DIA stages will succeed.

There are already a few DIA studies attempting to predict OCR performance and to choose helpful image restoration methods, guided by automatic analysis of images (cf. [40] and its references). It is not yet clear whether improving image quality, by itself, will ever improve OCR results enough to obviate the need for post-OCR correction.

3.2 Handling Rare & Fragile Documents

The process of imaging can damage documents in various ways. Some documents are so fragile than even careful manual handling is dangerous. Some books cannot safely be opened fully — else their spines and bindings may crack, and delicate pages tear — and so cannot be pressed flat without special hardware design. The extra care required can slow scanning down so much that it is unaffordable. These constraints, in turn, limit choices of illumination, optics, and sensors and thus affect image quality and demand special DIA methods to enhance quality.

3.2.1 Case Study: the PARC BookScanner

In response to a US Library of Congress draft specification, PARC designed a bookscanner [33] suitable for rare and fragile books, which is shown in Figure 1.

The wedge platen moves up and down by hand and is lowered onto the book for imaging. The book cradle holds the book and has adjustments for the angle and the separation to allow for different book sizes. The camera assembly contains a 4000x4000 pixel CCD camera, along with shutter, optics and a counterweight. The assembly is connected to the wedge and is lowered when the wedge is brought down onto the book. All book and page handling is manual. The scanable page area is a few mm larger than 8-1/2" x 11"; otherwise, there are no limitations on the size of the book. Throughput can be as high as 360 pages/hour greyscale and 120 pages/hour color.

The software user interface is Java based with specially developed image capture and processing routines written in C and C++. The operator is initially presented with an image viewing sub-window and a book structure selection
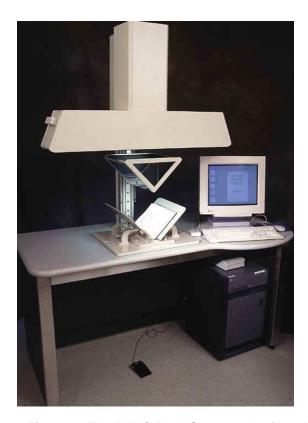

Figure 1. The PARC BookScanner, designed to scan rare and fragile books at 300 ppi 8-bit greyscale or 24-bit color. The book is opened to only $90°$ and placed, facing up, in the cradle; the two-paned glass 'wedge platen' descends to press both left & right pages flat simultaneously, while also pulling the camera assembly down to ensure precise geometry. Special DIA processing is required to enhance image quality due to many constraints on illumination & optics.

portion of the application window. Scanning is initiated by a foot switch (both pages), freeing the operator's hands to concentrate on the book/platen interaction to minimize impact on the book and ensure the quality of the image during capture by inspecting each image immediately after it is scanned and processed.

The operation of the bookscanner is designed for optimization of throughput and image quality. A scanning session begins with the capture of calibration images. Dark images and images of white sheets for each channel are used to calculate per-pixel gain/offset to compensate for fixed system noise and nonuniform illumination. An IT8 color target [2] is then imaged and analyzed to produce a color correction lookup table. The image of the target and lookup ta-

ble is saved to accompany the collections to be scanned to verify scanning accuracy at a later date. The subsequent scanning operations produce gain, offset, color and bad pixel corrected, full field images intended as archival quality images. Image processing enhancements were needed to sharpen the contrast in order to highlight text and suppress slowly varying background illumination — this also reduced some "bleed-through" artifacts. The images are not compressed and are stored as 8-bit-deep TIFF files.

The profile template describes a set of image processing routines geared for rendering the image in the best possible way. Initially, this may be set for viewing the images on the computer screen. At present, this includes crop area (right and left), orientation (auto-orient, flip, rotate), deskew, bitonal threshold, invert, sharpen, and the application of a tone reproduction curve (TRC) for background removal. A profile is setup as the default initially, but individual pages may have specific profiles as needed.

4. Initial Processing

The PARC BookScanner illustrates the wide range of early-stage image processing tools needed to support high-quality image capture. Note the importance of image calibration and restoration specialized to the scanner. Image processing should, ideally, occur quickly enough for the operator to check each page image visually for consistent quality. Tools are needed for orienting the page so text is rightside-up, deskewing the page, removing some of the pepper noise, and removing dark artifacts on or near the image edges. Software support for clerical functions such as page numbering and ordering, and the collection of metadata, are also crucial to maintaining high throughput.

In addition to these, it would be helpful to be able to check each page image for completeness and consistency. Has any text been unintentionally cropped? Are basic measures of image consistency — *e.g.* brightness, contrast, intensity histograms — stable from page to page, hour after hour? Are image properties consistent across the full page area for each image? Are the page numbers — located and read by OCR on the fly — in an unbroken ascending sequences, and do they correspond to the automatically generated metadata? Techniques likely to assist in these ways may require imaging models that are tuned to shapes or statistical properties of printed characters. Perhaps it will someday be possible to assess both human and machine legibility on the fly (today, this may seem a remote possibility; but cf. [19]).

4.1 Case Study: Jepson's "Flora of California" (UC Berkeley & PARC)

The complete page images of a rare scholarly publication, "Flora of California," by W. L. Jepson (1943) [25], have been captured and made available on the web in a collaboration between the UC Berkeley Digital Library Project, PARC, The Jepson Herbarium, and The California Academy of Sciences Library. The publication consists of five volumes containing a total of 1901 pages.

The *Flora* was chosen for several reasons. It has been out of print for many years (a few are still available from the Jepson Herbarium), but it is still in demand: systematic botany differs from other sciences in its persistent reliance on its early literature, going back as far as the eighteenth century; each new study of a species must be based on a thorough analysis of all previous published studies. Jepson's *Flora* contains information with specific references to individual specimens, flowering times, original descriptions, and other detailed information. Several fonts and text sizes are used to indicate the complex structure of entries. Line drawings, keyed to the text, illustrate specimens (Figure 2). Attempts to apply commercial OCR systems to the *Flora* have failed, for reasons that will be familiar to DIA professionals and to their more demanding users: variable (often low) image quality, unusual typefaces, and a highly specialized domain of discourse (uncommon names, terms, and abbreviations). Perhaps the most serious obstacle is the book's reliance on an elaborate convention (special to the book) employing typeface, typesize, and textual layout to indicate the logical structure of the botanical entries: the state of the art of commercial OCR is not reliably sensitive to such differences and cannot exploit them to improve accuracy even if it detects them; recent DIA research has attacked this (*e.g.* [26], [29]).

The *Flora* was scanned using the PARC BookScanner yielding contrast-enhanced greylevel images. Despite all possible care by the scanning operator, a second manual 'sanity-check' pass revealed that four pages had been omitted and 6.5% of the pages failed a visual check due to motion blur, too light or low contrast, and cropped-off text (due to narrow margins and thin gutter space). Also, page numbers assigned automatically did not track the printed page numbers and had to be remapped.

The high-resolution TIFF files were batch-processed at UC Berkeley to produce screen-resolution GIF files. Metadata was entered about each volume. Although these final images are plainly legible over the web (they have been rated 'excellent' by botanists), experiments have since revealed that binarization of the TIFF files to support consistent OCR remains a significant technical challenge due to page-to-page variations in intensity and contrast, and to a lesser degree variations across each page image.
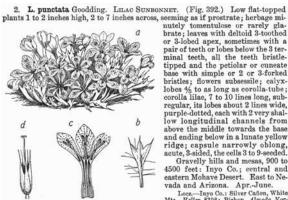
2. **L. punctata** Goodding. LILAC SUNBONNET. (Fig. 392.) Low flat-topped plants 1 to 2 inches high, 2 to 7 inches across, seeming as if prostrate; herbage minutely tomentulose or rarely glabrate; leaves with deltoid 3-toothed or 3-lobed apex, sometimes with a pair of teeth or lobes below the 3 terminal teeth, all the teeth bristle-tipped and the petiolar or cuneate base with simple or 2 or 3-forked bristles; flowers subsessile; calyx-lobes ⅘ to as long as corolla-tube; corolla lilac, 7 to 10 lines long, sub-regular, its lobes about 2 lines wide, purple-dotted, each with 2 very shallow longitudinal channels from above the middle towards the base and ending below in a lunate yellow ridge; capsule narrowly oblong, acute, 3-sided, the cells 3 to 9-seeded.

Gravelly hills and mesas, 900 to 4500 feet: Inyo Co.; central and eastern Mohave Desert. East to Nevada and Arizona. Apr.–June.

Locs.—Inyo Co.: Silver Cañon, White Mts., *Heller* 8308; Bishop, *Almeda Nordyke*; Argus Range (n. end), *C. N. Smith* 139; Hanaupah Cañon, Panamint Range, *Jepson* 7010. Mohave Desert: betw. Kelso and Baker, *Jepson* 20,582; Amboy, *Newlon*; Ludlow, *Jepson* 5494; Lavic (5 miles s.), *J. Grinnell*; Newberry (lava fields sw.), *Newlon* 498; Ord Mt., *Jepson* 15,496; Barstow, *Jepson* 5361.

The anthers, borne on very short filaments inserted unequally in the throat, approximate about the style at mouth of throat, leaving pin-hole entrances to the tube beween the filaments.

Refs.—LANGLOISIA PUNCTATA Goodding, Bot. Gaz. 37:58 (Jan., 1904); Hel., Muhl. 1:57 (Feb., 1904); Jepson, Man. 808, fig. 780 (1925). *Gilia setosissima* var. *punctata* Cov., Proc. Biol. Soc. Wash. 7:72 (1892), type loc. Surprise Cañon, Panamint Range, *Coville* 716. *Navarretia setosissima* var. *punctata* Cov., Contrib. U. S. Nat. Herb. 4:154 (1893). *Gilia punctata* Munz, Man. 402 (1935). *L. lanata* Brand; Engler, Pflzr. 4²⁵⁰:169 (1907), type loc. Candelaria, Nev., *Jones* 3965.

Fig. 392. LANGLOISIA PUNCTATA Goodding. *a*, habit, × 1; *b*, leaf, × 1; *c*, long sect. corolla, × 1½; *d*, calyx and pistil, × 1½.

Figure 2. Part of a page of Jepson's *A Flora of California***, a richly illustrated botanical masterpiece published in several volumes spanning decades. This image was scanned by the PARC BookScanner and the images were processed in realtime by PARC document image analysis tools. The entire multi-volume** *Flora* **is available on-line, as images, via the U.C. Berkeley Digital Library website [13].**

4.2   Restoration

The principal purposes of document image restoration are to assist:

- fast & painless reading;

- OCR for textual content;

- DIA for improved human reading (*e.g.* format preservation); and

- characterization of the document (age, source, etc).

To these ends, methods have been developed for contrast and sharpness enhancement, rectification (including skew and shear correction), superresolution, and shape reconstruction (for a survey, see [28]).

4.2.1   Rectification

The DIA community has developed many algorithms for accurately correcting skew, shear, and other geometric deformations in document images. It is interesting how inconsistently these have been applied to document images provided by DLs; although uncorrected they are easily detectable by eye, and cause some users to complain, they do not affect legibility and reading comfort except in extreme cases (for example more than 3 degrees of skew). However, not all DIA toolkits that may later be run on these images will perform equally well, so it could be a significant contribution to rectify all document images before posting them on DLs. It is also possible — although it is seldom discussed in the DIA literature — to "recenter" text blocks automatically within a standard page area in a consistent manner; again, it is not clear that this, although a clear improvement in aesthetics, matters much to either human or machine reading.

4.2.2   Degradation a Desirable Quality?

As part of a "free-to-read-over-the-web" policy, some publishers provide images of all pages of their books on-line, of an image quality carefully chosen to be sufficient for browsing for short intervals but irritating when read for very long. This policy mimics the unconstrained – but eventually uncomfortable — browsing that is commonly permitted, even encouraged, in bookstores. Some publishers experimenting with 'browsing over the web' policies have enjoyed hardcopy sales increases in spite of the obvious potential for theft. However this delicate balance could potentially be upset if image restoration algorithms were readily available to enhance the legibility of freely downloaded image. This raises a technical question which I believe has never been addressed in the DIA literature: can document images be deliberately degraded so that they are easy to browse, painful to read for long, and yet unrestorable automatically?

5. Analysis of Content

The analysis and recognition of the content of document images requires, of course, the full range of DIA R&D achievements: page layout analysis, text/non-text separation, printed/handwritten separation, text recognition, labeling of text blocks by function, automatic indexing and linking, table and graphics recognition, etc. Most of the DIA literature is devoted to these topics so I will not attempt a thorough survey in this short space.

However, it should be noted that images found in DLs, since they represent many nations, cultures, and historical periods, tend to pose particularly severe challenges to today's DIA methods, which are not robust in the face of multilingual text and non-Western scripts, obsolete typefaces, old-fashioned page layouts, and low or variable image quality. To pick only one example, the 'Making of America' DL [9], containing scanned images of 8,500 books and 50,000 journal articles with 19th C. imprints, offers a daunting array of DIA challenges.

5.1 Accurate Transcriptions of Text

The central classical task of DIA research has been, for decades, to extract a full and perfect transcription of the textual content of document images. Although perfect transcriptions have been known to result, no existing OCR technology, whether experimental or commercially available, can guarantee high accuracy across the full range of document images of interest to users. Even worse, it is rarely possible to predict how badly an OCR system will fail on a given document.

The open problems here are clearly legion, but they are also thoroughly discussed in the DIA literature. I will thus content myself with two references: (a) to the most recent (and most thorough) published test comparing commercial OCR machines [34]; and (b) a valuable survey of the state-of-the-art of DIA methods published in IEEE Trans. on PAMI in the last twenty years [30].

5.2 Case Study: Historical Newsprint (CMU & PARC)

Especially challenging problems are illustrated by the Historical New York Times project, which is aimed at providing, free to read on the Internet, the day-to-day history of civilization as reported by the New York Times from September 1851 to September 1923. This was a cooperative venture by the Universal Library Project in the School of Computer Science at Carnegie Mellon University, PARC, Seagate Industries, and the New York Times.

Figure 3. Image of part of the first page of The New York Daily Times, November 2, 1852, scanned from microfilm, in the CMU Historical New York Times Project. This image is rotated, sheared, and warped in a non-linear fashion; and the column gutters are narrow and noisy.

The images of the New York Times, scanned from microfilm (*e.g.* Figure 3), need to be segmented into columns and lines of text so that they can be displayed more legibly as images over the web (the full page image is far too large to be conveniently panned-and-zoomed). The challenges include poor image quality, widespread heavy noise,

tight column spacing, and non-linear spatial warping (worse than affine deformations such as skew and shear). Existing commercial OCR products can't cope with the unusual and obsolete typefaces, poor original print quality, crowded and irregular layouts, and low image quality typical of these – and many other – historical documents. The challenges are illustrated in Figure 4, in which the text is small, noisy, and in an unusual font, yielding passages which are difficult for even a highly motivated human reader to interpret, and even when displayed at an optimal size. The conversions from paper to microfilm and then to digital image have seriously compromised future legibility and reuse.

Figure 4. Image of part of the first page of The New York Daily Times, from the 1850s, scanned from microfilm, in the CMU Historical New York Times Project. These images of text are barely legible by a motivated human reader.

5.3 Determining Reading Order of Sections

Determining the reading order among blocks of text is of course a DIA capability critically important for DLs since it would allow more fully automatic navigation through images of text. This however remains an open problem in general, in that a significant residue of cases cannot be disambiguated through physical layout analysis alone, but seem to require linguistic or even semantic analysis. However the number of ambiguous cases on one page is often small and might be made manageable in practice by a judiciously designed interactive GUI presenting ambiguities in a way that invites easy selection (or even correction): such capabilities exist in specialized high-throughput scan-and-conversion service bureaus, but are not now available, to my knowledge, to the users of any DL, allowing them to correct reading order and so improve their own navigation.

5.4 Tabular Textual Data

Detecting and analyzing tabular data is a problem which has received sustained attention by the DIA community. It is of course harder in general than the analysis of images of bodytext; it appears however to be far easier than detecting

and analyzing images of mathematics. It represents perhaps the easiest of an important family of layout styles that express two-dimensional data: thus it is a natural target of *queries* (not just searches) which could significantly enlarge the utility of DLs by allowing data mining and aggregation.

An interesting evolution of technology is revealed in the DIA literature on table recognition: methods originally developed for use on document images are now finding application on Web pages, where tables are often miscoded in HTML but remain unambiguously well-structured to the eye. It is the *visual appearance* of a Web page, rather than its HTML labeling, which determines its intended meaning: this holds true for a wide range of partially or even fully encoded document representations such as PDF, LaTeX, and MSWord.

5.5   Labeling of Structure

In the most general case, DLs would benefit from DIA facilities that label every part of document structure within images to a degree of refinement possible using markup languages such as XML. Of course the general case remains a resistant class of DIA problems. Solutions to this problem might be especially useful in DLs since they would aid in navigation within and among documents, capturing some of the flexibility that keeps paper competitive with DLs. Navigation can be assisted by a wide range of partial DIA labelings, *e.g.* automatic indices, overviews (at various levels of detail), jumping for one section to the next, following references to Figures, and such.

5.6   Representations

Just as there exists a bewildering variety of raster image file formats, — JPEG, PNG, GIF, TIFF, BMP, DjVu, SilX, etc — there are many competing document file formats: 'pure raster' images, mixed image-and-text formats such as PDF, HTML, MSWord, and PostScript, and 'pure text' formats such as ASCII/Unicode. Of particular interest to DIA researchers is the ability of a representation to express the result of every stage of document-image analysis, from pure image all the way to completely encoded and tagged data, and especially the many varieties of partially interpreted (or alternative) versions in between. Progress has been made towards this goal in academia and industry, but consensus has been unsurprisingly elusive.

5.6.1   Multivalent Documents

Thomas Phelps and Robert Wilensky have richly developed the notion of *multivalent* document representation [31], in which each document (or each page of each document) is represented in multiple ways ('views,' or 'layers'), often including: raster image, OCRed text, PDF, DVI, annotations,

highlights, and hyperlinks. A versatile software environment — a multivalent web browser — is able to infer a common physical document hierarchy from many (not all) of these views, to the extent permitted by its particular representation, and so enables uniform methods for annotation of the appearance of the document and sharing of annotated versions with other users. Clearly, it would be helpful if all raster document images could be reliably treated in the same manner: this would of course require fully automatic extraction of layout structure which is, as we have seen, still a challenge to DIA R&D.

5.6.2   Adobe's PDF

Adobe's Portable Document Format (PDF) [11] is an open file format specification design to assist the distribution and exchange of electronic documents and forms. It preserves fonts, images, graphics, layout, and annotations and can be used to develop tools to create, view, or manipulate documents. It appears to be evolving towards a broadly inclusive multivalent representation.

6. Presentation, Printing, & Reflowing

Paper is cheap and movable enough to invite the spreading out of many pages over a large surface. The relative awkwardness of digital displays is felt particularly acutely here. When attempting to read images of scanned pages on electronic displays it is often difficult to avoid panning and zooming, which quickly becomes irritating and insupportable.

6.1   Highly Legible Display

This problem has been carefully and systematically addressed by several generations of eBook design, and progress is being made towards high-resolution, greyscale and color, bright, high contrast, lightweight, and conveniently sized readers for page images. But even when eBooks approach paper closely enough to support our most comfortable habits of reading, there will still be significant needs for very large displays so that large documents (*e.g.* maps, music, engineering drawings) and/or several-at-once smaller documents can be taken in at one glance. Perhaps desktop multi-screen 'tiled' displays will come first; but eventually it may be necessary to display documents on wall-sized surfaces. The DIA community should help the design of these displays and should investigate versatile document-image tiling algorithms.

6.2   Preserving Original Appearance

In many printed materials the author's and editor's choice of typeface, typesize, and layout are not merely

aesthetic: they are meaningful and critical to understanding. Preserving all of these stylistic details through the DIA pipeline remains a difficult problem. Even if DIA could provide "perfect" transcription of textual content (as ASCII/Unicode/XML), many critical features of its original appearance may have been discarded. One solution to this problem is, of course, multivalent representations where the original image is always available as one of several views.

6.2.1 Mobile Access

Recently, DIA researchers have investigated systems for the automatic analysis of document images into image fragments (*e.g.* word images) that can be reconstructed or "reflowed" onto a display device of arbitrary size, depth, and aspect ratio (*e.g.* [17]). An example is shown in Figure 5. The intent is to allow imaged documents to be read on a limited-resolution, perhaps even hand-held, computing device, without any errors and losses due to OCR and retypesetting. Thus it mimics one of the most useful features of encoded documents in DLs. It also holds out the promise of customizable print-on-demand services and special editions, *e.g.* large-type editions for the visually impaired.

This is a promising start but, to date, document image reflowing systems work automatically only on body text (and still have some problems with reading order, hyphenation, etc). Automation of link-creation (to, *e.g.*, Figures, footnotes, references) and of indexes (*e.g.* tables of contents) would greatly assist navigation on small devices. It would be highly useful to extend reflowing to other parts of document images, such as tables and graphics, difficult as it may be to imagine, at the present state of the art, how this could be accomplished.

7. Indexing, Retrieval, & Summarization

The indexing and retrieval of document images are critical for the success of DLs. To pick a single example, the JSTOR DL [8] includes over 12 million imaged pages from over 300 scholarly journals and allows searching on (OCRed) full text as well as on selected metadata (author, title, or abstract field). This is of course a well-studied problem: I content myself with a single reference to an excellent survey [22]. Almost without exception these attempt recognition and transcription followed by indexing and search operating on the resulting (often errorfull) encoded text. Although for some retrieval tasks, error rates typical of commercial OCR machines do not seriously degrade recall or precision of statistical 'bag-of-words' methods, some textual analysis tasks (*e.g.* depending on syntactic analysis), whether modeled statistically or symbolically, can be derailed by even low OCR error rates.

Figure 5. On this PDA a passage of text-image from Jepson's *Flora* has been reflowed onto the small display, *i.e.* the images of its words have been moved about and broken into shorter text-lines in order to fit legibly. No OCR was attempted; thus no recognition errors are visible. The original appearance of the specialized and meaningful typography, reading order, and the proximity of line-art Figures to their first mention in the text, are preserved.

7.1 Summarization & Condensation

There has been, to my knowledge, only one DIA attack on the problem of summarization of documents by operating on images, not on OCRed text. In this work [18], word-images were isolated and compared by shape (without recognition), clustered, and the cluster occurrences and word sizes used to distinguish between stop words and non-stop words, which were then used to rank (images of) sentences in the usual way.

This successful extension of standard information retrieval methods into a purely image domain should spur investigation of similar extensions, for example, methods for condensing document images by abstracting them into a set of section headers.

8. Personal & Interactive Digital Libraries

Research has recently gotten underway in 'personal digital libraries,' with the aim of offering tools to individuals willing to try to scan their own documents and, mingling imaged and encoded files, assemble and manage their own

DLs. All the issues we have mentioned earlier are applicable here, but perhaps there is special urgency in ensuring that all the images are legible, searchable, and browseable. Thus there is a need for deskilled, integrated tools for scanning, quality control and restoration, ensuring completeness, adding metadata, indexing, redisplay, and annotation. An early example of this, using surprisingly simple component DIA technologies informally integrated, is described in [38]. In addition, this might spur more development and wider use of simple–to–use, small–footprint personal scanners and handheld digital cameras to capture document images, with a concomitant need for DIA tools (perhaps built into the scanners and cameras) for image dewarping, restoration, binarization, etc.

In addition one may wish to detect (near) duplicates, either to prune them out or to collect slightly differing versions of a document; the DIA literature offers several attacks on this problem, but it appears not to have been effectively solved. Even when the document content starts out in encoded form, document image analysis can still be important. For instance, how might duplicate detection be performed when one of the versions is in PDF and the other is in DjVu? The common denominator must be the visual representation of the document, and from the point of view of individual (especially non-professional) users, the visual representation will be, more than in other contexts, normative and canonical.

Often, users may wish to be able to perform annotation using pen-based input (on paper or with a digital tablet/stylus). A role for document image analysis here could be annotation segmentation/lifting or word-spotting in annotations.

8.1   Interactive/Shared Digital Libraries

As publicly available DLs gather large collections of document images, opportunities will arise for collective improvement of the DL services. For example, one user may volunteer to correct an errorfull OCR transcription; another may be willing to indicate correct reading order or add XML tags to indicate sections. In this way a multitude of users gradually improve the usefulness of the DL collection without reliance on perfect DIA technology. Within such a community of volunteers, assuming it could establish a culture of trust, review, and acceptance, DIA tools could be critically enabling.

An example of such a cooperative volunteer effort, which is closely allied intellectually to the DIA field, is The Open Mind Initiative [39], a collaborative framework for developing "intelligent" software using the internet. Based on the traditional open source method, it supports domain experts, tool developers, and non-specialist "netizens" who contribute raw data.

Another example, from the mainstream of the DL field, is Project Gutenberg [12], the Internet's oldest producer of free electronic books (eBooks or eTexts). As of November 2002, a total of 6267 'electronic texts' of books have been made available on-line. All the books are in the public domain. Most of them have been typed in, and corrected (sometimes imperfectly), by volunteers working over the Web. Such databases are potentially useful to the DIA community as source of high quality ground-truth associated with known editions of books, some of which are available also as images. These collections have great potential to drive DIA R&D relevant to DLs, as well as to benefit from it.

8.2   Offering DIA tools

To assist such interactive projects, the DIA field should consider developing DIA tool sets freely downloadable from the web, or perhaps run on DL servers on demand from users. These could allow, for example, an arbitrary TIFF file (whether in a DL or privately scanned) to be processed, via a simple HTML link, into an improved TIFF (*e.g.* deskewed). Each such user would be responsible for ensuring that their attempted operation succeeded — or, less naively, there could be an independent review; the result would then be uploaded into the DL, annotated to indicate the operation and the user's assurance (and review). In this way even very large collections of document images could be improved beyond the level possible today through exclusively automatic DIA processing.

9. Urgency of the Need for Solutions

Cultural trends point to growing irrelevance and even the invisibility of any information that remains symbolically unencoded or if difficult to access digitally: "If it's not in Google, I don't need it." Compounding this is the technical infeasibility of automatically extracting highly accurate transcriptions of the content a wide range of document images, together with the awkwardness of digital displays and of most DL delivery and navigation user interfaces. The result is that most paper documents, even when represented as images in DLs, in a dangerously inconvenient state: compared with encoded data, they are relatively illegible, unsearchable, and unbrowseable. Thus a large fraction of our vast culturally irreplaceable legacy of paper documents seems to be threatened by a growing hegemony of 'purely digital' information. The DIA R&D community is uniquely qualified in many ways to assist in the rescue of this heritage from oblivion.

10. Acknowledgement

References

[1] Octavo, 134 Linden Street, Oakland, CA 94607-2538 USA; www.octavo.com.

[2] ANSI Standard IT8, American National Standards Institute. 1819 L Street, NW, Suite 600 Washington, DC 20036; www.ansi.org.

[3] Association for Information and Image Management, International. 1100 Wayne Avenue, Suite 1100, Silver Spring, Maryland 20910; www.aiim.org.

[4] Commission on preservation and access. 1400 16th Street, NW, Suite 740, Washington, DC.

[5] E Ink, 733 Concord Avenue, Cambridge, MA 02138. www.eink.com.

[6] European Commission on Preservation and Access. Royal Netherlands Academy of Arts and Sciences, Kloveniersburgwal 29, P.O. Box 19121, NL-1000 GC Amsterdam The Netherlands; www.knaw.nl/ecpa.

[7] Gyricon Media, Inc., 6190 Jackson Road, Ann Arbor, MI 48103. www.gyriconmedia.com.

[8] JSTOR Digital Library. Univ. of Michigan and Princeton University, www.jstor.org.

[9] Making of America Digital Library. Univ. of Michigan and Cornell University, moa.umdl.umich.edu.

[10] NPES/AIIM PDF-Archive Project. www.aiim.org/standards.asp?ID=25013 .

[11] Portable Document Format. Adobe Systems Incorporated, 345 Park Avenue, San Jose, California 95110-2704 USA, www.adobe.com/products/acrobat/adobepdf.html.

[12] Project Gutenberg. promo.net/pg.

[13] The UC Berkeley Digital Library Initiative – II Project. www.elib.berkeley.edu.

[14] *The State of Digital Preservation: An International Perspective.* Council on Library and Information Resources, Washington, DC, July 2002.

[15] Nabil R. Adam, Bharat K. Bhargava, and Yelena Yesha, editors. *Digital Libraries: Current Issues.* Springer-Verlag, December 1995. Lecture Notes in Computer Science vol. 916, ISBN: 3540592822; selected papers from Digital Libraries Workshop (Dl '94), Newark, NJ, May 19-20, 1994.

[16] Nicholson Baker. *Double Fold: Libraries and the Assault on Paper.* Vintage Books, April 2002. ISBN 0375726217.

[17] Thomas M. Breuel, William C. Janssen, Kris Popat, and Henry S. Baird. Paper to PDA. In *Proc., IAPR 16th ICPR*, pages Vol. 4, 476–479, Quebec City, Canada, August 2002.

[18] Francine R. Chen and Dan Bloomberg. Summarization of imaged documents without OCR. *Computer Vision and Image Understanding*, 70(3), June 1998. Special Issue on "Document Image Understanding and Retrieval," J. Kanai and H. S. Baird (Eds.).

[19] M. Chew and H. S. Baird. BaffleText: a human interactive proof. In *Proc., 10th IS&T/SPIE Document Recognition & Retrieval Conf.,*, Santa Clara, CA, January 23–24 2003.

[20] Richard J. Cox. Vandals in the stacks?: A response to Nicholson Baker's assault on libraries, August 2002. ISBN 0313323445; (Contributions in Librarianship and Information Science).

[21] A. Dillon. Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics*, 35(10):1297–1326, 1992.

[22] D. Doermann. The indexing and retrieval of document images: A survey. *Computer Vision and Image Understanding*, 70(3), June 1998. Special Issue on "Document Image Understanding and Retrieval," J. Kanai and H. S. Baird (Eds.).

[23] Daniel Greenstein and Suzanne E. Thorin. *The Digital Library: A Biography*. Council on Library and Information Resources, Washington, DC, December 2002. ISBN 1-887334-95-5.

[24] M. Hearst, G. Kopec, and D. Brotsky. Research in support of Digital Libraries at Xerox PARC. *D-Lib Magazine*, June 1996. ISBN 1082-9873.

[25] Willis Linn Jepson. *A Flora of California*. Associated Student Store, Berkeley, CA, 1943. All its page images can be viewed at elib.cs.berkeley.edu/docs/rare.html.

[26] Gary E. Kopec. Document image decoding in the UC Berkeley Digital Library. In *Proc., SPIE Document Recognition III Conference*, volume 2660, pages 2–13. SPIE–the International Society for Optical Engineering, SPIE, January 1996.

[27] M. Lesk. *Practical Digital Libraries: Books, Bytes, & Bucks*. Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1997.

[28] Robert P. Loce and Edward R. Dougherty. *Enhancement and Restoration of Digital Documents: Statistical Design of Nonlinear Algorithms*. Society of Photo-optical Instrumentation Engineers, January 1997. ISBN 081942109X.

[29] Huanfeng Ma and David Doermann. Bootstrapping structured page segmentation. In *Proc., SPIE/IS&T Document Recognition and Retrieval X*, pages 179–188, Santa Clara, CA, January 2003.

[30] George Nagy. Twenty years of Document Image Analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[31] T. A. Phelps and R. Wilensky. Multivalent documents. *Communications of the ACM*, 43(6):83–90, June 2000.

[32] Kris Popat. Decoding of text lines in grayscale document images. In *Proc., 2001 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, Salt Lake City, Utah, May 2001. IEEE.

[33] Steve Ready and Robert Street. The PARC Bookscanner. Technical report, Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA. www.parc.com/eml/members/ready.

[34] S. V. Rice, F. R. Jenkins, and T. A. Nartker. The fifth annual test of OCR accuracy. Technical report, Information Science Research Institute, Univ. of Nevada at Las Vegas, Las Vegas, Nevada, 1996. ISRI TR-96-01.

[35] A. J. Sellen and R. H. R. Harper. *The Myth of the Paperless Office*. The MIT Press, Cambridge, MA, 2002.

[36] E. H. Barney Smith and David Monn. DAS'02 working group on digital libraries & document image analysis of antique documents report. *Int'l J. on Document Analysis and Recognition*. In press; from the 5th IAPR Workshop on Document Analysis Systems, August 19–21, 2002, Princeton, NJ; www.research.avayalabs.com/DAS02.

[37] E. H. Barney Smith and X. Qiu. Relating statistical image differences and degradation features. In *Proceedings, 5th IAPR International Workshop on Document Analysis Systems*, pages 1–12, Princeton, NJ, August 2002. Springer Verlag. LNCS 2423.

[38] A. Lawrence Spitz. SPAM: A scientific paper access method, 1998.

[39] David G. Stork. The Open Mind initiative. In *Proc., IEEE Expert Systems and Their Applications*, pages 16–20, May/June 1999. www.openmind.org.

[40] Kristen Summers. Document image improvement for OCR as a classification problem. In T. Kanungo,

E. H. Barney Smith, J. Hu, and P. B. Kantor, editors, *Proc., SPIE/IS&T Electronic Imaging Conf. on Document Recognition & Retrieval X*, pages 73–83, Santa Clara, CA, January 2003. SPIE Vol. 5010.

[41] L. Wang and T. Pavlidis. Direct gray scale extraction of features for character recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15:1053–1067, October 1993.

[42] Blaise Agüera y Arcas. Computational analytical bibliography. In *Proc., Bibliopolis Conference 'The future history of the book'*, The Hague (Netherlands), Koninklijke Bibliotheek, November 2002.