Microfilm, Paper, and OCR:

Issues in Newspaper Digitization at the Utah Digital Newspapers Program

By

Kenning Arlitsch and John Herbert
J. Willard Marriott Library
University of Utah

Kenning Arlitsch – Title: Head of Information Technology. Address: 295 S. 1500 East, Room 463, Salt Lake City, UT 84112. Email: kenning.arlitsch@library.utah.edu

John Herbert – Title: Program Director – Utah Digital Newspapers. Address: 295 S. 1500 East, Room 418, Salt Lake City, UT 84112. Email: john.herbert@library.utah.edu

Microfilm, Paper, and OCR: Issues in Newspaper Digitization at the Utah Digital Newspaper Program

History of the UDN Program

The Marriott Library at the University of Utah (U of U) has a long history of large-scale newspaper projects beginning with the National Endowment for the Humanities' United States Newspapers Program (USNP) in the 1980's, in which the Library led the effort to catalog and microfilm Utah newspapers. This involvement continues today with the Utah Digital Newspaper (UDN) program, which is digitizing historic Utah newspapers, making them searchable and available on the Internet.

UDN's Grant History: 2002-2004[1]

With the first of three Library Services and Technology Act (LSTA) grants, in 2002 the Marriott Library digitized 30 years of three weekly newspapers. During this first phase of the program, the newspaper digitization process was developed and the UDN website was launched with some 30,000 total pages. (http://digitalnewspapers.org)

A second LSTA grant, which ran from January-September, 2003, digitized 106,000 new pages, effectively quadrupling the collection. The grant also funded a project director to run day-to-day operations and secure ongoing funding, and funded a publicity campaign to insure broad knowledge of the program across the state.

In September, 2003, the program was awarded a $1 million federal grant to continue for another two years. This grant was awarded by the Institute for Museum and Library Services (IMLS), an agency within the Department of Health and Human Services. IMLS is providing $470,000, with the U of U and Brigham Young University (BYU) providing matching funds of $450,000 and

$100,000 respectively. With this grant, the program will digitize 264,000 newspaper pages, with portions distributed to other sites, namely BYU and Utah State University. The metadata, (including searchable full text) from these sites will be harvested and combined with the metadata from the U of U's collection. This will present a combined, or aggregated, collection to readers so they can search on the entire collection at once, regardless of where the data is located. Another major goal of the grant is to administer a training program to other academic and historical institutions in the West, providing information on launching a digital newspapers program, managing the digitization process, and writing compelling grant proposals.

In March, 2004, the program was awarded a third LSTA grant to digitize 10,000 pages each of five specific Utah newspapers in five different counties. In administering this grant, the Utah State Library is providing $74,000, with matching funds of $25,000 raised locally, $5,000 each from public libraries in the five newspaper communities. These matching funds in particular show how the program has substantial grass roots support in local communities throughout the state. By the time the two current grants expire in September, 2005, the program should have 450,000 newspaper pages digitized.

Impact of the Program

As the program has grown during the past three years, it has had an increasing impact on Utahans. Monthly website usage has increased five-fold from June, 2003 to March, 2004[2]. Numerous emails and phone calls have been received from patrons who either want more information about the program or who are willing to support it in some way. What the program has done, at a very high level, is break down the traditional barriers between a major university and the general citizenry. Not only is the program telling the unique story of Utah's history to the

world via the Internet, it is also helping to create a new generation of "citizen historians" who are experiencing Utah history more easily and effectively than ever before.

Digitizing Microfilm

The first newspapers digitized by the UDN were scanned from microfilm. After decades of independent newspaper microfilm creation and USNP participation, the U of U's newspaper microfilm was clearly the most complete and accessible source for scanning. Many newspaper originals were destroyed following filming, so the expectation was that paper would be difficult to locate[3]. But problems with the quality and availability of our microfilm caused us to pursue print archives, and during 2003, 65% of the 106,000 pages were digitized from paper.

Service Bureaus

Libraries have long used service bureaus to convert their documents to microfilm, and the quality of work performed by these bureaus can have repercussions long after contracts are completed. Lockhart and Swartzell[4] conducted extensive tests on five vendors in the late 1980's, determining that while "all vendors met the basic technical standards…each test batch had problems which would require detailed attention in project initiation.[5]" In the UDN these problems would have a significant impact.

The U of U began microfilming newspapers through a service bureau in 1948 and by the time the USNP was launched in 1983, "the Marriott Library had almost complete microfilm holdings for 30 years' worth of Utah Newspapers."[6] Some of that microfilm was digitized in 2002 and its defects had an impact on the digitized images, both visually and for optical character recognition (OCR) processes. Uneven lighting plagued many of the newspapers. An image might go from an acceptable exposure on one side of the frame to a one or two f/stop difference on the other side. Consistent focus across the frame was another challenge; letters were sharp on one side but

sometimes more softly focused on the other. (This can easily occur when a copy-stand-mounted camera is not level.) Black smudges infected many frames, blocking out words or entire columns.

Most of these visual defects appear in the early years of the newspapers, leading us to conclude they were the first to be microfilmed and that service bureaus of the late 1940's had not perfected their techniques. There may also have been little or no quality control efforts on the part of the U of U; recommendations from the American Library Association, RLG, and ANSI/AIIM for inspection of microfilm only became available in the 1990's[7].

Even the ownership of master reels can come into question with a service bureau. The Library's microfilm service bureau had changed ownership several times, and a misunderstanding resulted in the master reels being shipped out of state. The service bureau erroneously believed it had acquired the master reels as a part of the purchase from the previous owner. During the 2002 processing, we discovered, shockingly, that the master reels were in Texas and the service bureau refused to return them. The University reacted by contacting the Utah Attorney General's office, and after several months of correspondence, the reels were returned. Now the Library is using storage and duplication services offered by BYU.

Physical Condition

The physical condition of microfilm can also affect scan and OCR quality. Cellulose acetate film, used widely through the 1970's before being replaced by stronger polyester, is known to tear.[8] Cellulose acetate is also prone to the same kind of "vinegar syndrome" chemical decomposition (though it is not as flammable) as the older cellulose nitrate base.[9] This decomposition leads to "buckling and shrinking, embrittlement, and bubbling,"[10] causing distortions in the image. Separately, chemical "redox blemishes", resulting from oxidative attack

in less-than-ideal storage conditions, have been noted in microfilm throughout the country,[11] and have been seen in a few instances in film used by the UDN. These reddish spots adversely affect the quality of the scanned image.

Advantages of Microfilm

Despite the problems mentioned above, scanning newspapers from microfilm offers several distinct advantages:

- Inexpensive scanning. With the right equipment, microfilm can be scanned in an automated fashion, allowing an operator to load a reel of film and essentially walk away from the scanner. These scanners can cost $100,000, but the UDN experience shows that firms with this equipment can offer pricing at approximately $0.15/page.
- Low conservation costs. Whereas paper may require conservation treatment prior to scanning, microfilm is usually physically stable and requires no such treatment. Barring the physical problems described above, preparation costs are limited to making scanning copies from the master reels. Scanning from microfilm is best done from a clean copy, free of the defects found in service copies.
- Availability. Thanks to the USNP, newspaper microfilm collections are available and fairly complete in each state.

Digitizing Paper

In 2003 sixty-five percent of the 106,000 pages were digitized from paper. When in good condition, paper represents original source material, whereas microfilm represents as much as a third-generation copy (paper-to-master-to-scan copy). Our hypothesis that scanning from paper produces better images and more accurate searching is discussed in the section "OCR Accuracy – Microfilm vs. Paper." For all its promise of cleaner images and better search accuracy, though,

original newsprint has its own set of challenges, not the least of which is finding the collection in the first place. The UDN is constantly canvassing the state trying to locate original collections.

Scanning Equipment

The oversized nature of newspapers makes them difficult to scan on conventional equipment; a book scanner or high-resolution digital scanning camera with copy stand and lighting are requisite. Equipment that scans this size at a minimum of 300 dpi (400 dpi is recommended), and at an economically feasible speed, costs $50,000 - $100,000. The UDN out-sources its scanning at $.20 - $.30/page, depending on whether the newspaper is loose or bound.

Conservation costs

Newsprint from the mid-19th century can still be in very good condition, if it was properly stored and handled minimally. Some collections, though, have been stored in adverse conditions, and have deteriorated over time, requiring conservation work to render them stable enough for scanning. This repair work generally consists of minor mending and cleaning. While the time and effort for this can vary widely from one collection to another, our overall cost average for this minimal conservation is $0.19/page.

Advantages of Paper

- Cleaner digital images, more accurate OCR. Provided the paper is in fair (or better) condition, better digital images are achieved by scanning directly from originals. And cleaner digital images produce more accurate searchable text. These assumptions are reinforced by our initial testing (see "OCR Accuracy – Microfilm vs. Paper"), though more testing is needed. That said, the variable quality of each source medium, microfilm or paper, makes it difficult to state categorically that one is always a better choice than the other.

- Color scanning. Using originals makes it possible to scan in full color, though the desirability of this is questionable. Color scanning offers a more accurate representation of the original, but storage costs for the larger color files are much greater, if not prohibitive. Issues of file presentation and storage are discussed in the next section.

Digital File Storage

As the UDN program grows, we face the problem of storing and managing both low-resolution PDF files presented on the Web and high-resolution archival TIFF files. The bi-tonal PDF files presented on the Web are quite small: articles average 10-50KB, and full pages average 300KB. The entire current collection of images and metadata requires only 130GB of online disk space. But because a separate file is generated for each article, the files are numerous. The 136,000 pages currently in the collection comprise 1.6 million files, or nearly 12 files per page. Providing this average continues, we will have nearly 5.4 million files for our 450,000 newspaper pages by the end of 2005. The good news is that even with the Library's current total of 2 million files from all digital collections, including metadata and full text, we are not experiencing performance issues.[12]

Numerically daunting as the eventual 5.4 million files is, it is not the most serious problem we face. The more serious problem is the long-term storage of the full-page archival files for the newspapers, which are 4-bit grayscale TIFF files averaging 14MB in size. Standard procedure for the other digital collections at the Library (photographs, books, documents, maps, art prints, etc.) is to store the archival files directly on the server.[13] But the sheer number and size of the newspaper files preclude archiving them online. In 2002 we stored these files on DVD but soon realized that with two copies of each file, we would be quickly overwhelmed with creating and maintaining discs. So, we are now storing them on magnetic tape, realizing that while tapes do

not have the longevity of optical discs, they offer more flexibility and reusability. We are using LTO Ultrium tapes, which have an uncompressed capacity of 100GB and cost $50 each.

Storage of Color Files

Our partners at BYU have decided to present the 40,000 pages of the *Deseret News* in full color on the Web. At 80MB each, the full-page archival images are considerably larger than the 14MB grayscale images at the U of U, and present an extreme case of the storage and management concerns described above. BYU has decided to archive these files on DVD, with a second copy stored off-site.

To present the color files on the Web, BYU is using CVista compression software from CVision Technologies Inc.[14] Achieving a 64:1 compression ratio, BYU is able to reduce the 80MB full-page TIFF files to 1.25MB PDF files. While this is still large for some dial-up Internet users, the remarkable compression rate does allow BYU to present full-color images on the Web.

Distributing the Collection

In 2002 the Utah Academic Library Consortium (UALC)[15] established the Mountain West Digital Library (MWDL). Four digitization centers in Utah and two in Nevada support educational and cultural heritage partners by providing digitization infrastructure, training, and standards, as well as creating their own digital collections. An aggregating server[16] at the U of U harvests metadata from each center and provides a single searchable index at http://mwdl.org. Images are called from the CONTENTdm server where they reside in real time.

As the MWDL matures, it has become clear that the UDN would benefit from its distributed network. The four centers in Utah plan to host newspaper titles in their region, thereby reducing the burden and cost of collection storage at a single location. As the aggregating server at the U of U harvests metadata, we believe we will create the first distributed digital newspaper

collection in the country. We anticipate the MWDL will begin harvesting newspaper metadata from BYU by early summer 2004.

Optical Character Recognition

OCR is the centerpiece of creating full-text from digital images. Today there are literally hundreds of commercially available OCR packages with wide ranges in sophistication and price. Digitizing historic newspapers is much more complex and difficult than generating text from most other documents. Some common problems are deteriorated originals, unusual fonts, faded printing, shaded backgrounds, fragmented letters, touching/overlapping letters, skewed text, curved lines (which is very common in bound volumes), and bleed through.[17] In fact, running OCR against a gray-scale image (rather than bi-tonal) can actually reduce OCR accuracy where bleed through has occurred.[18] Consequently, nothing but the most robust OCR software should be used for historic newspapers.

One method of measuring OCR effectiveness is how accurately it determines which words are on the printed page. This is normally expressed as the percentage of words on the page that are accurately "read" by the software. Of course, "reading" a word involves piecing it together letter-by-letter (see the next section), so sometimes OCR accuracy is measured as the percentage of letters accurately "read". It's important to realize, however, that these two levels of accuracy are fundamentally different. Word accuracy is by definition significantly lower than letter accuracy because it is effectively the joint accuracies (or joint probabilities) of the letters in the word. For example, OCR accuracy at the letter-level for a document may be 98%. But computing the accuracy of a five-letter word in that same document is done by taking 0.98 to the fifth power (the joint probability of five letters), which is 90.4%. This distinction between letter and word accuracy is critical to note when analyzing OCR accuracy.

One other consideration with newspaper OCR is that articles often have a lot of redundancy. The same important search keyword, such as a last name or a city, often appears more than once in the same article. So in these cases, word accuracy need not be 100% for a successful search.[19]

General Description of Software

Included in the processing services offered by our service provider, iArchives Inc. of Lindon, Utah, is OCR. Their OCR software is not only very sophisticated and state-of-the-art, it is proprietary and patented. Consequently, we are limited in the details we can present here. What follows is a fairly generic description of how OCR software operates.

The OCR process begins with a clean-up of the raw TIFF images created by the scanning process. This clean-up involves:

- Cropping each image, which determines where the edges of each page are and removes everything outside them;
- De-skewing each image to present the printed lines on the horizontal;
- De-speckling each image to remove extraneous spots/speckles from the original.

Once the clean-up is complete, each page is "zoned" into individual articles. This creates a separate image of each article, which is important because the OCR runs on the individual articles, not the full page.

The next step is the OCR software itself. iArchives' OCR framework uses multiple engines, each of which runs with a different orientation. For instance, one engine may work better with lighter images and thinner fonts, while another may work better with darker images and bolder fonts. These orientations are needed because images span a wide range of quality from article to article and page to page.

Each engine inspects each image pixel by pixel, looking carefully at contiguous dark pixels, determining their overall shape, and comparing the shape to known letters in many different fonts. It also closely examines the adjacent white areas. One of the most important decisions the software makes is, given the size, shape, and location of a white space, is it: 1) part of a letter that's fragmented; 2) the space between letters; or 3) the space between words? Various algorithms are run to help answer these questions, including: a) growing and shrinking potential letter fragments to see if dark areas can be connected to form letters; b) whitening the background and darkening the text to improve the contrast; and c) changing near-white to white and near-black to black. In the end, the software assembles the contiguous/connected dark pixels into a letter, sometimes noting that more than one letter is a possibility. Letters within a certain (very small) distance of each other are assembled into a "node", which is what the engine believes is a word.

Finally, because the engine may have multiple possibilities for a particular letter, it may as a result have multiple possibilities for the associated word that the letter is in. For example, if the engine can't decide about the third letter in the node "be*t", possible word options include beat, beet, bent, and best. The engine will rank order the possible words from best to worst option, and then, depending on the setup parameters, provide the requisite number of words for the text. These words are accumulated into the text file for the article.[20]

The Dilemma of Multiple Word Options

The UDN program presents full-text searching for 136,000 newspaper pages. When searches are performed on the entire collection, the search-hit limit of 10,000 can easily be reached, especially if a single and somewhat common word is used in the search. One method of limiting a search, of course, is to search on more than one word. As is quite popular, a user will search on

a first and last name together, instead of the last name only. For example, searching the collection for "cassidy" results in 392 hits, but when that's limited by searching for "butch cassidy", 62 hits result. Additionally, if an exact-phrase search is used to further limit the search, only 40 hits result. This is where things get interesting, as the example below illustrates.

On page 2 of the *Eastern Utah Advocate*, November 14, 1901, there is an article titled "Eastern and Southern Utah". On the printed page is the phrase "George Parker alias Butch Cassidy", and the OCR has generated this corresponding text: "george parker alia alfas[21] butch dutch cassidy". The OCR had some difficulty with the words "alias" and "Butch", putting two options in for each. In particular, the two options for "Butch" (butch and dutch) present an intriguing problem. The three words "butch dutch cassidy" as they appear in the text preclude a successful search on the exact phrase "butch cassidy" because "dutch", as the second word option for "butch", is in between.

So the dilemma is this: when the OCR generates more than one word for a particular node, we enhance our ability to successfully search on that single word because there is more than one possibility to search on. In the example, both a search for "butch" and for "dutch" will generate a hit on the article. But at the same time, almost paradoxically, a multiple word option **reduces** the chances of a successful search if that word is used in an exact phrase. As noted above, the article is not included in the hits for "butch cassidy" because the word "dutch" is in between "butch" and "cassidy". In fact, an exact-phrase search for "butch dutch cassidy" reveals five hits, so the identical OCR result occurs four other times in the collection.

Testing to Find a Solution

Working closely with iArchives, in October, 2003 we ran a series of formal tests of their OCR framework. The testing was designed to find the best word-option setting for their software. The

test set included 16 randomly selected full pages, representative of the entire collection. The text for each page was keyed and verified to nearly 100% accuracy, becoming what was called the "ground truth", which was used to compare against actual OCR results. Then, the OCR framework was run four times: with one word generated for each node, with two words generated, with three words generated, and with no limit on the number of words. The results are in Table 1 below[22].

No. of Words Generated	OCR Accuracy	Accuracy Improvement	Excess Words Generated	Additional Words
1	77.9%	n/a	18,990	n/a
2	80.3%	2.4%	26,473	7,483
3	80.7%	0.4%	28,293	1,820
No limit	80.8%	0.1%	28,676	383

TABLE 1 – OCR ACCURACY FOR MULTIPLE WORD OPTIONS

The data shows that as the limit on the number of words generated increases, OCR accuracy increases, as does the number of excess words. These excess words, like "dutch" in the example, can reduce exact-phrase search accuracy. It should be noted that OCR accuracy and excess words are strongly dependent on the quality of the images and format complexity. Good image quality and simple formats will generally have higher OCR accuracy and fewer excess words, while poor image quality and complex formats will likely have lower OCR accuracy and more excess words.

Our analysis concluded that the best **overall** search accuracy would come from the two-word option. This conclusion tried to strike the right balance between single-word and exact-phrase accuracy. The two-word option had a significant increase (2.4%) in accuracy over the one-word option, while the three-word and unlimited options increased accuracy only slightly at 0.4% and 0.1%, respectively. The 2,203 additional excess words generated by these options, which have

the undesired side-effect of further reducing phrase search accuracy, were not, we concluded, worth the small 0.5% increase in single-word accuracy.

In spite of this well tested solution, we consider it less-than-optimal because it merely strikes a balance between two competing interests (single-word and exact-phrase searches) rather than strongly supporting both. The ultimate resolution to this tricky problem lies with utilizing a "proximity search". Proximity searches, which are growing in popularity among search engines, allow a search for a phrase, like "butch cassidy". But instead of the words literally having to be together in the text, these searches allow them to be within a certain pre-set number of words (say, three) of each other. So, any time "cassidy" is within three words of "butch", a hit would be generated. This longer-term solution will allow us to support multiple words generated by the OCR, providing higher accuracy for single-word searches, and at the same time be able to search accurately on phrases.

Dictionary Filters

After the initial text is generated by the OCR, it is filtered through a number of different dictionaries to insure only valid words are in the final text. In our early days, we used a small English dictionary of only 28,000 words. There are any number of ways to count them, but according to Oxford Dictionaries, "…there are, at the very least, a quarter of a million distinct English words."[23] So our filtering dictionary was too small by an order of magnitude (!), and we found, not surprisingly, that it was leaving out far too many important words. iArchives located and incorporated a two-million item dictionary containing all English words, common foreign language words, surnames, and place names. Additionally, we augmented the place names dictionary with a set of Utah place names provided by BYU. We felt it particularly important, since we are digitizing Utah newspapers, that we be able to accurately search on all Utah place

names, and, given the high genealogical use of the collection, to have a robust surnames dictionary.

Once we incorporated the new dictionaries into the text-generation process, we re-filtered the originally generated text and re-built the text files using the expanded dictionary. This re-filtering merely involved re-running a script at the very back end of iArchives process and was accomplished in a short timeframe with little additional expense. The result is that the entire collection is now filtered properly through the new dictionaries.

OCR Accuracy – Microfilm vs. Paper

One of the important processing issues involved in newspaper digitization is deciding what source material to digitize from: microfilm or original newspapers. The USNP has made available almost every important newspaper title in microfilm. However, this often decades-old microfilm, while available, doesn't necessarily provide the best source material for OCR. We wanted to examine how well OCR operates on originals, too, because intuitively we believed originals would provide higher quality images for the OCR. When paper is scanned, a new digital photograph is taken of the original newspaper page, which is often a tremendous improvement in overall image quality from that paper's microfilm

To test this assumption, during 2003 UDN scanned originals when they could be found in good condition. After completing the 2003 processing, we ran a series of acceptance tests to insure the work was ready for installation onto our servers. The QA testing involved, in part, performing keyword searches and determining overall search accuracy for each newspaper title. The results of the QA testing are in Table 2 below.

	Issues Sampled	Keyword Searches	Hits	Pct
Original Paper	43	294	218	74.2%
Microfilm	30	219	142	64.8%
TOTAL	73	513	360	70.2%

TABLE 2 – QA TESTING RESULTS

As this somewhat small sample shows, original newspapers provide approximately a ten-percentage-point improvement in OCR accuracy over microfilm. While these results certainly support our assumption of originals being better source materials, we consider these results preliminary and in need of further sampling and study to confirm the numbers. The 2004 and 2005 processing for UDN should have an even higher percentage from originals and provide us with more material to test from.

Next Steps

Assessing Microfilm "OCR-Ability"

The economic conditions and technical solutions are in place today for launching and expanding all types of digital collections, complete with full-text searching. Generating accurate full-text, of course, relies upon securing good quality source materials and using effective OCR software. Microfilm has been widely used as the storage medium of choice for newspapers and other media for a generation or more, and the amount of material on microfilm is staggering. Moreover, as noted earlier, many originals were destroyed once they were filmed, reducing the ability to find originals in any condition. Even though the UDN has had some good fortune in acquiring originals, we understand that not all digitization projects will be as fortunate. So it's not difficult to foresee circumstances when microfilm may be the only available source material for some digitization projects. Our UDN experience shows, however, that while some microfilm

is digitized as accurately as any original, others clearly have much poorer results. While microfilm may be prevalent and less expensive to digitize, it's not necessarily the best source material.

How then do we decide whether microfilm should be digitized or whether we should incur the additional time and expense of locating and scanning original paper? An assessment methodology to predict the accuracy of OCR-generated text extracted from a microfilm scan is needed. In other words, what is the "OCR-ability" of a reel of film? We need the ability to estimate the overall OCR accuracy of film without having to go through the entire digitization process and the expense that would necessarily be incurred. This type of evaluation model would have broad applicability well beyond newspapers as many different media have been copied to microfilm.

Distributed Collections and Aggregated Searching

Our current plans at UDN call for aggregating the distributed collections at BYU and USU, and presenting a single searchable index at our website. Developing this technology will greatly enhance our ability to link digital newspaper collections together, enabling powerful search engines to provide users with nearly complete data in very short response times. This is the real promise of digital newspaper collections, indeed of digital collections of all types – providing to practically everyone immediate access to meaningful data.

We would like to gratefully acknowledge the contributions of Scott Christensen and Frederick Zarndt of iArchives Inc., and of Randy Silverman, Preservation Librarian at the Marriott Library in the preparation of this manuscript.

[1] John Herbert and Kenning Arlitsch, "digitalnewspapers.org: The Digital Newspapers Program at the University of Utah", *Serials Librarian*, 47, nos. 1 and 2 (2003)
[2] Website visits averaged 433 per month from April-June, 2003, and 2,347 per month from January-March, 2004.
[3] Nicholson Baker. *Doublefold: Libraries and the Assault on Paper,* New York: Random House (2001).

[4] Vickie Lockhart and Ann Swartzell. "Evaluation of Microform Vendors," *Microform Review*, 19, no. 3 (Summer 1990)
[5] Ibid.
[6] Robert P. Holley, The *Utah Newspaper Project Final Report, Project No. PS-200010-85 National Endowment for the Humanities United States Newspaper Program* (1987).
[7] Walter Cybulski. "You Say You Want a Resolution? Technical Inspection and the Evaluation of Quality in Preservation Microfilming", *Microform & Imaging Review,* 28, no. 2 (1999)
[8] Michael J. Gunn, "Poly or Cell?" *Microform Review,* 16 (Summer 1987)
[9] Thomas A. Bourke, "The Curse of Acetate; or a Base Conundrum Confronted." *Microform Review,* 23 (Winter 1994)
[10] Ibid.
[11] James M. Reilly, Douglas W. Nishimura, Kaspars M. Cupriks, and Peter Z. Adelstein. "Stability of Black and White Images, with Special Reference to Microfilm." *Abbey Newsletter,* 12, no. 5 (1988) http://palimpsest.stanford.edu/byorg/abbey/an/an12/an12-5/an12-507.html
[12] We use CONTENTdm digital asset management software from DiMeMa Inc. to manage and present all our digital collections, including the newspapers. See http://contentdm.com for product information.
[13] We track the online files using CONTENTdm's Full Resolution Manager feature. Regular tape backups of the server files and rotating copies sent off-site ensure long-term viability.
[14] See http://www.cvisiontech.com/ for CVista product information.
[15] See http://www.ualc.net for information about the higher education library consortium.
[16] Each MWDL site runs a CONTENTdm server, and the aggregator at the University of Utah is a CONTENTdm product known as the Multi-Site Server. See http://contentdm.com for details.
[17] Frank R. Jenkins, Thomas A. Nartker, and Stephen V. Rice, "Testing OCR Accuracy", *Inform* (September, 1996)
[18]Thomas A. Nartker, Stephen V. Rice, and Frank R. Jenkins, "OCR Accuracy", *Inform* (July, 1995)
[19] Kazem Taghva, Julie Borsack, Allen Condit, and Srinivas Erva, *Journal of the American Society for Information Science*, 45, no. 1 (January 1994)
[20] Included with the text are the x- and y-coordinates of each node's location within the image. These coordinates are used to highlight the word when a successful search for it is done.
[21] The words "alfas" and "alia", while uncommon, do actually pass a dictionary filter. When searching Google, "alfas" retrieves 80 pages of hits, with the most common reference appearing to be as the shortened plural of "alfa romeos". "Alia" retrieves 87 pages of hits from Google, and is the acronym of both the "Australasian Lighting Industry Association" and the "Australian Library Information Association".
[22] "OCR Accuracy" was defined as the percentage of single words in the ground truth which are also found in the OCR text. "Excess Word" was defined as a word in the OCR text which was not in the ground truth. The total number of ground truth words was 62,228.
[23] AskOxford.com, http://www.askoxford.com/asktheexperts/faq/aboutwords/numberwords