Project Report

CARIBBEAN NEWSPAPER IMAGING PROJECT

Phase II: OCR Gateway to Indexing

Context and Proposal

Users of electronic images come to digital media with a set of expectations greater than those they have of other media. They anticipate extensive indexing, directly and interactively linked to the indexed information. With this second phase of the Caribbean Newspaper Imaging Project (CNIP2), the University of Florida tested the viability and costs associated with use of optical character recognition (OCR) as an alternative to manually indexing electronic newspapers.

With funding support from the Andrew W. Mellon Foundation, the University of Florida has scanned its microfilmed newspaper holdings of the ***Diario de la Marina*** (Havana, Cuba), 1947-1960, and ***Le Nouvelliste*** (Port-au-Prince, Haiti), 1899-1979. In the process, these newspapers were indexed selectively by reviewers knowledgeable in the languages. Selective indexing was not ideal, given that it is highly labor-intensive and far from comprehensive. CNIP2 was undertaken to assess the value and cost effectiveness of OCR indexing of these same newspapers.

CNIP2 evaluated OCR effectiveness within the following target groups:

- OCR software technologies;
- Digital image resolution;
- Bit depth;
- Language of the source newspaper;
- Publication dates; and
- Filming methods and technologies.

While OCR of page images smaller than a newspaper's folio dimensions has been successfully demonstrated and cost-effectively applied, OCR application to newspaper images had not been addressed when CNIP2 began in 1999.

Background. Phase One: The Feasibility of Image Capture.

Today, there are only three effective means of reproducing newspapers: (1) image conversion from film, (2) capture using a very-high resolution digital camera, or (3) rekeying from either source newspapers or from film.

Newspapers continue to be too large for extant flat-bed scanners. Lenzar, the Florida company that manufactured large format linear-array fiat-bed scanners, went out of business in 1997. It was the only manufacturer of such products. Alternately, newspaper stock, with its short fibers, is often too fragile for rotary plotter-scanners. And, historic newspapers, universally embrittled, require a great deal of care in handling. It would be unthinkable to pass these newspapers through a rotary plotter-scanner if not also to place them on a flat-bed scanner if one were available.

Rekeying, another alternative, is a labor intensive chore. Though the costs of rekeying can be minimized by sending this work off-shore to nations with lower costs or standards of living, the costs of reproducing an entire run are enormous. While it *might* be every e-newspaper vendor's dream to make issues available retrospectively, the demand for retrospective issues would never be immediate enough to pay the bills. Not surprisingly, the backfiles of electronic newspapers maintained by vendors of e-newspapers is limited. None is retrospective to before the date on which they began making current newspapers available electronically.

Map digitization projects such as those at the University of Florida, employing very-high resolution cameras, have demonstrated the ability to capture great detail from oversized source documents. Digital camera-backs such as those manufactured by PhaseOne are capable of well exceeding minimum resolution guidelines promulgated by Cornell University. Yet, at resolution sufficient to meet these guidelines, the exposure time would average approximately 30 minutes per page.

Newspaper on microfilm is problematic for a number of reasons. The defacto "standard" for production of film intermediaries for oversized source documents calls for 105 mm rather than library "standard" 35 mm film on which newspapers are currently microfilmed. Formulas for digitization of images on film, in comparison against scanner manufacturer's literature and claims, show that no microfilm scanner currently available, whether it scans from contact or from projection, can adequately scan newspaper from 35mm microfilm.

Regardless, phase one of the Caribbean Newspaper Imaging Project (CNIP1) demonstrated that readable newspaper *images* could be captured from film and displayed on computer monitors. The delivery of oversized images and use of scroll compensated for a scanner's inability to meet the resolution guidelines promulgated by Cornell University and commonly employed by library digitization projects. Today, though navigation of newspaper images that scroll vertically and horizontally beyond the average monitor's limits is still problematic, ever increasingly popular high-compression vector image formats (e.g., SID) make viable delivery of these large images via the Internet.

Need. Phase Two: OCR as a Means of Index Construction.

Though CNIP1 demonstrated the ability to *economically* deliver readable newspaper images, it reported costly, labor intensive indexing effort. At four fifths of the total image delivery cost, indexing also under represented the content of newspaper issues. While CNIP1 indexed only 3 articles per issue -- three more than had been indexed previously, three articles far from met the expectations of researchers using the CNIP product. CNIP1 made obvious the need to explore more cost effective and more representative means of indexing.

If the cost of selective indexing by human readers was expensive, the cost of constructing a comprehensive index through rekeying was out of question. CNIP planners turned to optical character recognition (OCR) as a possible means of index construction. Phase two of the Caribbean Newspaper Imaging Project (CNIP2) would compare the utility of indices created through OCR with that of indices created by human readers. Additionally, CNIP2 would assess various off-the-shelf OCR products, their application with the several languages of the Caribbean Newspaper Collection at the University of Florida, and the extent to which "dirty" text could be cleaned cost effectively.

Target Newspapers

Targeted titles included *Diario de la Marina, Le Nouvelliste* and *Trinidad Guardian.* Published in one of the three predominant Caribbean languages and extensive in holdings, each targeted newspaper would afford analysis of OCR application with a variety of language and printing variables. Microfilmed over time to changing standards, comparison of OCR accuracy from images generated from these microfilms also would quantify probabilities of successful OCR.

The *Diario* and *Nouvelliste* had been digitized in CNIP1. For this project, select page images were rescanned for test of additional digital methodologies. Select page images of the *Trinidad Guardian* were digitized and indexed, for the first time, for purposes of this project.

The *Trinidad Guardian* was selected from among the University of Florida's English language newspaper microfilm holdings for its documentation of the colonial British West Indies and of the various independence and republican movements of the English speaking Caribbean nations. Trinidad and Tobago, persuaded by the rhetoric of Dr. Eric Williams, compelled the Caribbean toward a Caribbean identity and nationhood.

Selection Procedure

For each of three newspaper titles, target issues were selected as follows:

- For any given test group, 400 page images were selected in order to maintain statistical validity consistent with +5% accuracy.
- For any given sub-sample, 200 page images were selected in order to maintain statistical validity consistent with +10% accuracy.
- To establish data resolution as to afford comparison across titles, issues were selected from comparable dates, e.g., the first issue every fourth month.

Target Summary

Quantity	Selection
1,200	*Diario de la Marina* images Selected from the CNIP1 project
1,200	*Le Nouvelliste* images Selected from the CNIP1 project
1,200	*Trinidad Guardian* images New images converted from newspaper microfilm
600	Quarter-page scans New images: 200 each of the three targeted newspaper titles
4,200	Total images OCR processed

The target represents two categories of images:

- 3,600 whole-page 400 dpi scans, and
- 600 quarter-page 400 dpi scans.

Targeted page images were selected to represent date and language groups evenly within the bounds specified below. Whole and quarter-page images were made of the same page. All images were scans of projected pages using the same Minolta MS1000 and MS3000 microfilm scanners used in CNIP1. A 400 dpi whole-page newspaper image generated using Minolta projection scanning equipment is the equivalent of an image generated at 50% reduction, relative to the original size of the source newspapers. A 400 dpi quarter-page image generated using this equipment afforded an image which, if partial, approximated the resolution recommended by Cornell University.

Demographic

Languages

%	LANGUAGE
33%	English *Trinidad Guardian* (Port-of-Spain, Trinidad)
33%	French (Française) *Le Nouvelliste* (Port-au-Prince, Haiti)
33%	Spanish (Español) *Diario de la Marina* (Habana, Cuba)

Fonts by Language

FONT NAME	ENG	FRE	ESP
Times New Roman & other serif typefaces	95%	95%	95%
Arial, Helvetica & other sans-serif typefaces	<5%	<5%	<5%
Engravers, Rockwell & other misc. typefaces	<1%	<1%	<1%

Fonts by Size (calculated for source newspaper)

FONT NAME	Smallest "e" on average	Mean "e" on average
Times New Roman & other serif typefaces	1.0 mm	1.0 mm
Arial, Helvetica & other sans-serif typefaces	1.0 mm	3.0 mm
Engravers, Rockwell & other misc. typefaces	3.0 mm	3.0 mm

OCR Accuracy (Summary Findings)

%	CHARACTERIZATION OF TEXT
33%	Article Text Serif text at 1.0 mm
65%	Article Titles Sans-serif text at 3.0 mm
27%	Surnames & Place Names in Article Text Serif text at 1.0 mm
58%	Surnames & Place Names in Article Titles Sans-serif text at 3.0 mm

OCR Software

Images were processed using each of four major off-the-shelf software packages: TextBridge (v.9), OmniPage Pro (v.9), TypeReader (v.5), and Adobe Capture/Exchange. Because of its cost, Prime Recognition software used by the University of Michigan and JSTOR was not tested in this Phase. Adobe Capture is the software engine used by some electronic newspaper distributors (e.g., NewsExpress) of current newspapers issues.

Digital Image Resolution

OCR software is optimized for measures of digital resolution (dpi) associated with the linear CCD arrays found in commonly available scanner hardware. Cornell states that images with dpi not consistent with these measures *may* not be as accurate as those that are consistent with the capacity of these arrays. Evaluation of the resulting text files found no meaningful statistical variation from one OCR package to the other within either of the two categories: whole and partial-page images. Comparing results of the two categories however, accuracy was greater, regardless of the OCR package used, for whole-pages than for quarter-pages, a finding contra-indicative of the Cornell guidelines. The digital resolution of quarter-page scans using Minolta microfilm projection scanners should have approximated the dpi suggested by Cornell for the source newspaper.

Where bigger-is-better in setting digital resolution measured as *dots-per-inch (dpi),* microfilm scanners currently manufactured are not capable of meeting an adequate dpi per the Cornell formulas. Metering projected newspapers into segments for optimal capture was a creative solution but, in terms of workflow had this test produced the anticipated results, the cost of human intervention would likely have been prohibitive.

Bit Depth

Research at Cornell University suggests that scanning at increased bit-depth may enhance the legibility fine detail from the source document. It should be noted, however, the legibility, here, is relative to the human eye's ability rather than to OCR's ability to read a given document. While Adobe Capture and TypeReader are optimized only for bitonal image conversion, OmniPage Pro and TextBridge process both gray-scale (8-bit) and color (24-bit) images. Regardless, the suggestion has *little* utility when scanning from high contrast microfilm rather than from the newspapers themselves. While microfilm is high contrast, microfilmed images do capture tone between black and white. The Minolta equipment available to this project, however, was capable of bitonal capture only.

The adaptive use of a Microtek 9600 XL transparency scanner failed predictably as interpolated dpi was unable to resolve newspaper print at 21:1 reduction with sufficient clarity. Using the scanner's interpolation software, 8,400 dpi resolution was theoretically required for a moderately good (Quality Index 5.5) scan using the Cornell formulas. An 8,400 dpi scan from film with 21:1 reduction should have been the equivalent of a 400 dpi scan from the newspaper itself. Interpolation was unable to compensate for the limitation of the native 600 dpi resolution of the CCD.

With the failure of the Microtek, CNIP2 used a sample of 15 grayscale (8-bit) newspaper images procured from a vendor of microfilm conversion services using Sunrise high-speed microfilm scanners. Images were 200 dpi, the equivalent of those produced by the Minolta microfilm projection scanners. They were produced, however, to current library "standard", with good image quality and lighting balance. The source newspaper, though North American, used type faces and font sizes comparable to those of the CNIP newspapers. Though the sample was small and statistically inadequate, the results were worth note. OCR resulting from the grayscale images was less

than 10% accurate. OCR resulting from bitonal images of the same pages was 82% accurate.

While the depth of grayscale images made it easier on the human eye to read a given page than were their bitonal duplicates, increased bit-depth was a disadvantage to those OCR packages capable of reading it.

Language

OCR is software. One method of programming that software may be more or less effective than other methods in approach to given image characteristics, including "noise", type face, and language. It is reasonable to suggest that individual software packages are more or less reliable than others. Further, all of the OCR packages studied by CNIP2 are off-the-shelf programs written largely for English-language business and personal applications, working with modern type-faces. While each is enabled with multi-lingual dictionaries, none of those dictionaries are equal. Evaluation of the resulting text files representing the whole-page sub-sample found no meaningful statistical variation from one OCR package to the other for any language tested: English, French, or Spanish. Relative to their dates of publication and a subjective assessments of image quality, no one language was converted any more accurately than the other. Microfilm image quality, particularly lighting issues (e.g., contrast and light balance), was more likely to effect the accuracy of OCR than any particular OCR package.

To assess their spell-check routines and to differentiate among otherwise equal OCR packages, a secondary human pass was made against a sub-sample of 300 text files generated by each OCR package. Human native-language readers, with the aid of Microsoft Word running the appropriate language dictionary, assessed the closeness of spelling mis-matches, counting the number of incorrect letters in a word. Each of the OCR packages tested has the ability to "*learn*" from corrected errors. The OCR package which most often

and most closely approximated the correct spelling of words might have an edge in increasing the accuracy of the resulting text file. This was a tedious chore at best; but, it was complicated by the effects of poor microfilm image quality.

While each OCR package converted areas with good image quality more accurately, within these areas their performance varied. Disabling the spell-check routines, in order to assess character recognition alone, produced "anecdotal" evidence. OCR packages with larger dictionaries, it appeared, were able to correct more text. However, it also appeared that OCR packages with smaller dictionaries (e.g., Adobe Capture) had better noise reduction, line formation, other filters; they did not require larger dictionaries. Ultimately, the sub-set of images with good quality among the sub-sample was so small and so uneven as to language that the data was not meaningful.

Regardless the particular language, OCR accuracy at the word-unit level, not surprisingly, was more accurate the shorter the word-unit. Unfortunately, shorter words -- words including articles (*a*, *the*, *le*, *la*, *les*, *los*, etc.), prepositions (*for*, *from*, *in*, to, *à*, *de*, *dans*, etc.), and pronouns (*he*, *she*, *il*, *elle*, etc.) -- are usually regarded as stop-words. Such words have virtually no meaning in an index created from "dirty" text. Words least often corrected, particularly among smaller fonts, were surnames and place names not commonly found in dictionaries. In a sub-sample of 400 items, these names were correctly converted to text below the accuracy of text overall. Only 27% of such names in 1.0 mm serif fonts were accurately converted. Names, usually place names, found in the dictionaries were more frequently corrected than names not in the dictionaries. Unfortunately, these words are among the more commonly searched by researchers.

Publication Date

The condition and characteristics of the source newspaper set bounds on the quality of the film image. Microfilm is a non-additive technology; the film image is never better than the source newspaper. Printing technology, print defects, paper color and aging effects, type faces, and font sizes, among others, are all factors in image quality.

CNIP2 made assumptions about the characteristics of the target newspapers printed at different times. It established four date groups for purposes of analysis.

Date Group	Date Span	Titles in the Group
Early Modern	1890-1920	*Le Nouvelliste, Trinidad Guardian*
Modern	1920-1950	*Le Nouvelliste, Trinidad Guardian*
Late Modern	1950-1970	*Diario de la Marina, Le Nouvelliste, Trinidad Guardian* The *Diario de la Marina* is available only between 1947 and 1960. *Le Nouvelliste* is available only before 1960.
Contemporary	1970-1997	*Trinidad Guardian*

The Early Modern period was characterized, in part, by moveable type and type faces worn as a result of repeated use. The Modern period was characterized by set type as was the Late Modern period. Distinction of these two periods is somewhat artificial. The latter saw the increased use of sans-serif and stylized type faces, albeit primarily in article titles. The Contemporary period saw the introduction of electronic type setting and other automation, albeit largely within the last decade. Somewhat arbitrary as well, the Contemporary period serves as a control-group, for which filming methods and techniques are known. Because copyright restrictions limited reproduction of newspapers in this group, the group was small and solely represented by the Trinidad Guardian with which the University of Florida had negotiated copyright permissions.
(See also, this discussion as regards Filming Methods & Techniques, below.)

CNIP2 found very little deviation in type faces or font sizes from one period to the next, and little more than what might be characterized as a standard deviation from one OCR package to the next. Article titles

become easier to read and were more accurately converted by OCR to text with the introduction of sans-serif article titles. But, because of their size relative to article text, of the two, article titles are more frequently accurate regardless their age, type face, or OCR package used. Worn type, while more common in the Early Modern period, was in evidence only occasionally, and its detrimental effect on OCR was predicted. But, because article titles and text follow standard formats and sizes, OCR accuracy does not necessarily decrease with the age of the newspaper issue. Again, microfilm image quality is a more accurate predictor of anticipated OCR accuracy than were age and artifacts of printing processes.

Filming Methods & Technologies

Factors in microfilming and film characteristics are fundamental to optimal image capture and subsequent OCR accuracy. In newspaper microfilming, there have been four eras, each defined by a set of standards or the lack there of:

Date Group	Date Span	Microfilming Practices Titles in the Group
Pre-Modern	pre-1977	Microfilming defined by *"best-practices"* Titles: *Diario de la Marina, Le Nouvelliste, Trinidad Guardian*
Modern	1977-1986	Microfilming defined by Library of Congress/ANSI standard Titles: *Trinidad Guardian*
Post-Modern	1987-present	Microfilming defined by Research Libraries Group guidelines and revision of Library of Congress/ANSI standard Titles: *Trinidad Guardian*
Contemporary	present select application	Microfilming defined by so called, *"OCR-optimized"* standard, i.e., RLG guidelines modified for allowable 1% skew, fixed reduction, and *"one-up" filming* Titles: *Trinidad Guardian*

Microfilming in the Pre-Modern era was characterized by a set of *best practices* shared among microfilm technicians. Insofar as imaging practices were documented, they were found in recommendations from Eastman Kodak and the MRD/MRE microfilm camera instruction pamphlets. And, film processing, primarily during the early part of this era and outside

the big cities, relied on locally mixed chemicals and the "shake-and-bake" method of fixing and washing exposed films still used today in home dark-rooms. Microfilms produced by the University of Florida during this period, from the early-1950s through the mid-1960s in particular, when a technician with an MRE microfilm was sent packing across the Caribbean on Rockefeller Foundation funding for the Farmington Plan, were subject to environmental conditions, imbalanced lighting, and extended delays between exposure and processing.

Microfilming in the Modern era was marked by concerted effort, centered at the Library of Congress, to standardize practice for newspaper microfilming. In Florida, the era was still without standard and characterized, also, by the use of acetate-base films that deteriorated for lack of cold, dry storage. Deteriorated films were replaced one from another, sometimes in the nick of time. Image quality suffered threefold: (1) inherently detrimental effects of acetate-base aging, (2) deterioration effects associated with climate, and (3) degradation effects of analog-to-analog copying.

Microfilming in the Post-Modern era was distinguished by a more complete set of standards, optimized for image quality and microfilm longevity. In Florida, it was marked by first use of more durable polyester-base films and the adoption of standards for filming, film processing, and film storage. And, the Contemporary era finds the University of Florida's on-going Caribbean newspaper microfilming in lock-step with the preservation standards set revised-for-digitization.

CNIP2 drew primarily from the Early Modern period of microfilming history. Copyright restrictions necessitated that the CNIP project be drawn from the public domain. An exception was made for the *Trinidad Guardian*. The University of Florida negotiated permissions with the newspaper's parent company, Trinidad Publishing Co. LTD., as part of the University's Dr. Eric Eustace Williams project. *Trinidad Guardian* microfilms were examined through 1981, the year of Dr. Williams' death in office. This

small group of Post-Modern and early Contemporary issues served as a control group.

As stated earlier, microfilm image quality was determined to be the most accurate predictor of anticipated OCR accuracy. Regardless of standards, film image quality is conditioned by focus and depth of field, reduction *level,* exposure levels and light dispersion, and the density of imaged film. Microfilm is a high-contrast technology optimized for capture of text, but unsuitable exposure or uneven lighting, in particular among these conditions, can erode the legibility of text.

In general, a microfilm's background density (i.e., density in areas without text, in the unprinted areas between letters) appeared to have no effect. Variations of background density within standard were not recorded in the electronic image. As microfilm images were captured, white-and-black points balanced, and saved as bitonal images much of this area became uniformly white, while text became uniformly black. Quality Index assessment of the inner area of lower case letter "e"s was within the tolerance of analog-to-analog reproduction for microfilms with good image quality. In a microfilm image of good quality, contrast between text and paper should accurately reflect the condition of the original newspaper; the density of text and the density of areas without text each should be relatively uniform. OCR, predictably, was less accurate for microfilms with moderately good and poor image quality, but further assessment of these conditions is a discussion of lighting at the time of microfilming. CNIP2 found two conditions most frequently resulted in poor OCR accuracy: depth of field and light balance.

Nearly all microfilm cameras resolve a depth of field up to three and, in many cases, six inches. Text in the gutter margin can be microfilmed legibly, albeit frequently with shadow from the up-swelling of pages from the binding. When microfilmed pages with shadow are captured electronically as bitonal images, shadow is often recorded as noise, distorting the shapes of letters and reducing the accuracy of OCR.

In these areas, accuracy of OCR fell to less than 5%. Microfilming practices, inasmuch as possible, should be changed to require disbinding and flattening to facilitate future digitization. With volumes that cannot be disbound, microfilming stations should be equipped with additional near-overhead lighting, transforming microfilming stations into those one might find in use by publication-quality photo-reproduction services. A drawback of this recommendation, however, is need to increase the camera operator's skill set at a time when finding and training microfilm camera operators and supervisors is increasingly difficult. While light meters integrated with the camera station should ensure that an appropriate amount of light reaches the source newspaper, balancing an additional two sets of lights would be more problematic than balancing the sets currently in use. An ideal production workflow requiring microfilm for preservation and an electronic version for access would afford successive or simultaneous analog and digital imaging such as that currently made possible by the Zeutschel 300/301 hybrid microfilm camera.

As has been suggested, the most common image quality issue detrimentally affecting OCR accuracy is light balance. Most microfilming stations are equipped with two sets of lights, one situated on each side of the camera head and source newspaper. Ideally, the lights are directed at areas opposite their position. If the beams of light can be envisioned as straight lines, the would all cross below the camera head, approximately equidistant between the lens and the source newspaper. Current RLG microfilming guidelines require an even illumination target the size of the source document be microfilmed at the start of each document and that this target be evaluated for light balance. Newspapers selected for CNIP2, however, predate this requirement and no studies have been published to independently assess either compliance with the requirement or light balance in the target area of microfilms created since this requirement was established. Again, drawing on a small, statistically inadequate sample of newspapers reportedly microfilmed to current library "standard", CNIP2 derived text that was 82% accurate.

In any case, while Caribbean Newspaper Collection microfilms are legible -- light imbalance is frequently noticeable but does not prevent reading, electronic text in raster images (e.g., TIFF files) and text files resulting from their OCR was degraded. Lighting imbalances on the source microfilm produced a spot-light effect of uneven, sometimes starkly contrasting areas on the electronic images. Images were subjectively classed by the size of spot-light into poor, moderate, and good balance. And, within images, areas were subjectively classed into regions of poor, moderate, and good digital background density. Regions of the raster images with poor digital background density were predominantly illegible. In these areas, OCR was wholly inaccurate. Regions of the raster images with moderate digital background density were comparable to that produced by the up-swelling of pages from the binding. In these regions near the outer corners of the page image, accuracy of OCR was less than 5%. Regions of the raster images with good digital background density were legible, though lights often appear to have been directed toward the center of the microfilm frame. In these regions, the accuracy of OCR was 38.5%. OCR of the subset of *Trinidad Guardian* microfilms, representing compliance with Library of Congress/ANSI standard and evidencing more control of light balance, produced much higher OCR accuracy: approximately 79% -- a value close to the more anecdotal 82% accuracy reported from the small test of newspapers microfilmed to current library "standard".

Conclusion: Summary Findings

The accuracy of OCR on the retrospective newspaper collections targeted by the Caribbean Newspaper Imaging Project was disappointingly low. Overall, 33% of article text was accurately converted without any human intervention. The ills of past microfilming practice and the poor image quality of the target films is largely responsible for this poor rate. Anecdotal evidence drawn from contemporary microfilms created to current library "standard" appears to suggest that higher accuracy results from improved microfilming practice.

Human indexing as employed in CNIP1 indexed merely three articles per issue. Relative to the number of articles published on average in each issue, the percentage of indexed articles is also low.

Title	Publication Format	Average Minimum Articles	Percentage Indexed
Diario de la Marina	1 section: 12 pages	72	4.2%
Diario de la Marina	3 sections: 36-48 pages	200	1.5%
Le Nouvelliste	1 section: 4 pages	50	6%

CNIP2 postulated that keyword searching of the "dirty" text resulting from OCR could provide access to newspaper content greater and at less cost than that provided through human indexing. The comparison may be apples and oranges. Results of tests using a sub-sample of articles with both human and machine indices provided no meaningful comparisons. Searching against a word-base constructed from dirty text/OCR product requires different strategies from those used to search against an analytical index constructed from human interface. Nonetheless, 33% accurate text appears to afford broader, if not more meaningful, access the published newspaper content than did CNIP1's human indexing.

CNIP's networked data entry systems will eventually support both human and "machine" indexing. Currently, CNIP is attempting to build automated systems to remove nearly all human intervention from the process of generating dirty text from the extant image files. It is anticipated that this software will eventually remove unrecognized words lacking capitalized initial letters and stop words

(articles, prepositions, and pronouns) in English, French and Spanish. Adding "dirty" text as a search resource should immediately provide the layer of access needed to support additional newspapers and quickly build the content needed to make CNIP economically viable. With the time it buys, we will be able to build the more analytical index entries produced by CNIP1. OCR becomes another tool for indexing but does not necessarily remove the human component at this time.

Currently, the CNIP product is migrating from CD-ROMs to the Internet as a base for delivery of images. The new search resource will be integrated during this migration. As it does so, we will be able to test further the viability of this new resource. CNIP2 still leaves many questions unanswered. There is still no good, cost effective means of providing the researcher with full text or connecting story lines broken by column and page breaks.

CARIBBEAN NEWSPAPER IMAGING PROJECT

Phase II : OCR Gateway to Indexing

Budget (All Funds)

PERSONNEL

Kesse, Erich ..... $ 2,750.00
Director, Digital Library Center; Project Manager (@ 0.05 FTE)

Bressette, Eve ..... $ 1,045.00
Supervisor of Technicians (Student Assistants) (@ 0.05 FTE)

Winston, Harris ..... $ 1,082.83
Senior Programmer (@ 0.025 FTE)

Phillips, Richard ..... $ 1,142.83
Chief Librarian, Latin American Center (@ 0.025 FTE)

Allerton, David ..... $ 1,150.85
Archivist, Supervisor of Indexing Technicians (Student Assistants) (@ 0.05 FTE)

Indexing Technicians ..... $ 385.00
Student Assistants assigned to index selected issues of the *Trinidad Guardian*
70 hours at $5.50/hr

Imaging Technicians ..... $ 517.00
Student Assistants assigned to scan additional images of the *Diario de la Marina* and *Le Nouvelliste*, as well as new images from the *Trinidad Guardian*.
94 hours at $5.50/hr

OCR Technicians & Readers ..... $ 5,203.00
Student Assistants assigned to read OCR out-put against electronic raster (TIFF) images of newspaper pages.
Readers, either native speakers of a target language or a language major, specialized in the Spanish of the *Diario de la Marina*, the French of *Le Nouvelliste*, or the English of the *Trinidad Guardian*.
946 hours at $5.50/hr

Total Personnel Expenditure ..... $ 13,276.51

HARDWARE & SOFTWARE

Optical Character Recognition Software ..... $ 1,014.49
OmniPage Pro ..... $ 87.99
TextBridge ..... $ 86.99
TypeReader ..... $ 262.51
Adobe Acrobat Capture ..... $ 577.00

Workstations (2 at $3968 each) ..... $ 7,936.00
Configured for processing (creating, manipulating and OCR of) newspaper images
Workstations and scanners.

Total Hardware & Software Expenditure ..... $ 8,950.49

BUDGET SUMMARY

Total Expenditure ..... $ 22,227.00

CARIBBEAN NEWSPAPER IMAGING PROJECT

Phase II : OCR Gateway to Indexing

Budget (Andrew W. Mellon Foundation Funds)

PERSONNEL

Indexing Technicians ..... $ 385.00
Student Assistants assigned to index selected issues of the *Trinidad Guardian*
70 hours at $5.50/hr

Imaging Technicians ..... $ 517.00
Student Assistants assigned to scan additional images of the *Diario de la Marina* and *Le Nouvelliste*, as well as new images from the *Trinidad Guardian*.
94 hours at $5.50/hr

OCR Technicians & Readers ..... $ 5,203.00
Student Assistants assigned to read OCR out-put against electronic raster (TIFF) images of newspaper pages.
Readers, either native speakers of a target language or a language major, specialized in the Spanish of the *Diario de la Marina*, the French of *Le Nouvelliste*, or the English of the *Trinidad Guardian*.
946 hours at $5.50/hr

Total Personnel Expenditure ..... $ 6,105.00

HARDWARE & SOFTWARE

Optical Character Recognition Software ..... $ 262.51
TypeReader ..... $ 262.51

Workstations (2) ..... $ 5,732.49
Configured for processing (manipulating and OCR of) newspaper images
Computer workstations ONLY.

Total Hardware & Software Expenditure ..... $ 5,995.00

BUDGET SUMMARY

Total Expenditure ..... $ 12,100.00