Final Report

CRL/LAMP Brazilian Government Serials Digitization Project

by

Scott Van Jacob

Project Coordinator

December 2001

Table of Contents

1. INTRODUCTION

In 1994, the Andrew W. Mellon Foundation funded the Center for Research Libraries (CRL) and the Latin American Microfilm Project (LAMP) to conduct a joint project to scan and index about 700,000 pages of microfilmed Brazilian Government Documents and then provide access to them over the Internet. The project was completed in December, 2000.

This final report on the Brazilian Government Serial Documents Digitization Project describes the process of scanning and indexing these materials, originally issued between the 1820s and the 1990s.   Standards for reformatting microfilm to digital images are still emerging, and this report may help others to identify “best practices” for digital access projects.[1] Many of the lessons learned here may be more generally applicable to projects focused on Internet access to historical materials.

The source materials for this project were microfilms of inkprint, text based reports, with relatively few illustrations and little use of color. One of our first decisions centered on whether to provide full-text searchable files by rekeying all the texts or deploying Optical Character Recognition software. The quantity of material, variable page formats and microfilm image quality, inherent variability of hand-press originals, and unpredictable orthography of the Portuguese language for much of the period represented by the originals, led us away from this approach.

The documents were instead scanned as page images, which are essentially digital black and white snapshots of each page. The files produced in the project therefore cannot be searched for specific words or phrases. Intellectual access, beyond the general level of the title of each digitized piece, has therefore relied on subject indexing. We have here experimented with searchable volume indexes produced by rekeying original tables of contents and indexes, and providing links to specific page images. We similarly rekeyed and linked a separate subject index to some of the source materials, employed image mapping to link some subject indexes to specific pages without rekeying, and explored other approaches to offset the limitations inherent in a database of page images.

Considerations of cost, user preference, and available technologies all affected the final structure of the database. In the final analysis, though, anyone with Internet access can now effectively access, search, and retrieve the entire contents of nearly 700,000 pages concerning Brazilian federal and state history. The project has successfully enhanced scholarly access to materials that were heretofore scarce, fragile, and widely scattered.

[1] See works by:
-- Paul Conway. *Conversion of microfilm to digital imagery: a demonstration project*;
-- Anne Kenney. *The Cornell/Xerox/Commission on Preservation and Access Joint study in digital preservation: report, phase 1, January 1990-December 1991*;
-- A. Kenney, “Digital-to-microfilm conversion: an interim preservation solution,” *Library Resources & Technical Service*s;
-- A. Kenney, *Moving theory into practice: digital imaging for libraries and archives*;
--*The Digital Library Toolkit*. This Sun Microsystems publication lists several relevant digital library projects.

1.1 Acknowledgments

Financial support from The Andrew W. Mellon Foundation, reflecting its ongoing commitments to Latin American Studies and to scholarly utilization of digital technologies, allowed this project to take shape and develop. The Foundation's support and guidance as new possibilities emerged enabled us to go a good bit beyond our initial expectations.

The Center for Research Libraries provided both staff and financial resources for the project, especially in the areas of general administrative support and in housing and providing access to the database.

The members of the Project Committee invested considerable time into this project. Their consistently high expectations stimulated rigorous responses.

Dickinson College and the University of Notre Dame provided broad-based support for the Project Coordinator in a host of ways, both obvious and indirect.

Finally, numerous scholars and librarians offered comments and criticisms that helped to clarify the project's purpose and scope, and that significantly enhanced the quality of the result. Our thanks to all.

2. PROJECT OVERVIEW

2.1 The Latin American Microfilm Project and the Center for Research Libraries

This Project took shape as a cooperative endeavor between two organizations that have worked together closely over many years. The Latin American Microform Project was formed in the 1970s as a membership organization with the goal of acquiring, preserving, and loaning microfilms of scarce, endangered, rare, or voluminous research materials pertaining to Latin America. Forty-three North American libraries comprise the current membership, whose annual dues fund most ongoing microfilming activity. LAMP is administered through the Chicago-based Center for Research Libraries. Its policies and projects are determined by its entire membership, which meets once a year. An elected Executive Committee, chaired by Dan Hazen during most of the project's early phases and then by David Block, coordinates LAMP activities between the annual meetings.

The Center for Research Libraries, founded in 1949, is the country's oldest cooperative, membership-based repository for research materials. Its membership currently consists of 200 university, college, and research libraries throughout the United States and Canada. CRL's mission is to provide research materials that are rarely held in North American libraries to the broad scholarly community. The Center accomplishes this mission by acquiring, preserving, providing bibliographic access to, and loaning resources from its collections in Chicago. Its holdings encompass archival materials, dissertations, newspapers, monographs, scientific and technical serials, and more than 1,400,000 units of microfilm. The Center also hosts a number of area microfilm projects, including LAMP, and sponsors or participates in other cooperative efforts.

2.2 Project Background and Administration

As its name suggested, the Latin American Microfilm Project was founded well before electronic technologies offered new possibilities for reformatting and access. Moreover, microfilm remains a durable preservation medium, and the capacity to produce and utilize microfilm is widely distributed within North America (albeit much more thinly in many developing countries). However, microfilm is clumsy, unpleasant to use, and only available to remote users through traditional forms of physical shipment. LAMP, and CRL as well, was therefore eager to explore digital technology as a means to promote ease of use and broader accessibility. The proposal submitted by these two organizations to The Andrew W. Mellon Foundation was funded in 1994. Both LAMP and the Center have used the results in refining their strategies for the preservation, description, and distribution of information resources. (See Project Chronology in Appendix 10.1.)

The size and complexity of this project demanded oversight and administrative support. LAMP and the Center began by organizing a Project Committee to provide broad guidance in matters of policy and priorities. The members have included:

David Block, Ibero-American Bibliographer, Cornell University
Ann Hartness, Assistant Head Librarian, Nettie Lee Benson Latin American Collection, University of Texas at Austin
Dan Hazen, Librarian for Latin America, Spain and Portugal, Harvard College Library
Don Simpson participated on the Project Committee during his tenure as CRL President.
Marlys Rudeen, CRL's Program Officer for the Area Studies Projects, contributed until her 1998 departure;
James Simon has filled this role since.
Scott Van Jacob, Iberian and Latin American Studies Subject Librarian, University of Notre Dame, was appointed Project Coordinator soon after funding was approved.

The Project Committee has met annually throughout the project, usually during conferences of the Seminar on the Acquisition of Latin American Library Materials, SALALM. Three separate meetings have taken place at CRL, along with innumerable e-mail exchanges and conference calls.

Scott Van Jacob, as Project Coordinator, has been responsible for the day-to-day work of creating the database. He served as the principal liaison to the companies who contributed to the project, and also provided regular reports to the Project Committee and the Mellon Foundation.

CRL's project responsibilities have included fiscal oversight, a logical extension of the Center's continuing role in managing LAMP's account.   The database itself resides on a mass storage device at the Center and is accessible through CRL servers. Three individuals have managed the Center's systems operation during the project's five-year life; each has played a crucial role (which continues even now) in maintaining and improving the database.

A number of companies and individuals contributed to the Project. PFA Inc., of Sun Valley, California, performed all the scanning and most of the indexing. Hollyer and Schwartz, of Chicago, Illinois, created the database that allows the Provincial Presidential Reports to be searched. The Project Coordinator supervised indexing for the provincial presidential reports; with one exception, Brazilian graduate students from Notre Dame and neighboring colleges did most of the work.

2.3 Project Goals and Summary of Accomplishments

The initial proposal (See Appendix 10.2.)to the Mellon Foundation suggested seven Project goals . These goals were either achieved or, in a few cases, rejected in view of high costs or technical obstacles. This section briefly reviews these goals, which are treated in detail in chapters 3-6.

The Project from the outset concerned itself with delivering images to users, rather than creating images that would fulfill requisites for long-term preservation. Other projects and agencies, among them Project Open Book and the Council on Library and Information Resources, have focused on image specifications that fully represent original source materials and therefore can serve as enduring surrogates. This approach requires minimal—ideally no—loss of information between original materials and reformatted versions. For our purposes, in light of the availability of preservation microfilm at CRL, the focus was instead on quickly providing readable digital images over the Internet. Organizing the database has been a higher priority, since intellectual access was problematic to both the original paper sources and the microfilm. Put another way, this Project's primary concern is access, not preservation.

Goal 1: Facilitate scholarly access to a central and coherent body of high profile research resources for Brazilian Studies

The Project has provided researchers with Internet access to 3,593 official reports that contain 673,449 page images and span 170 years of Brazilian history. Four separate sets of documents have been digitized, with each one offering its own views and perspectives:

Table 1: Collections in the Database

Title	Type	From-To	Reports	Number of Images
Provincial Presidential Reports	State	1823-1930	2,572	216,187
Ministerial Reports	National	1821-1960	93	329,159
Presidential Reports	National	1890-1993	872	18,103
E.H. Laemmert Almanak*	National	1844-1889	56	110,000
		Totals	3,593	673,449

*E.H. Laemmert (publisher) *Almanak Administrativo, Mercantil e Industrial do Rio de Janeiro*

Goal 2: Expand the emerging corpus of digital page images aimed at a scholarly audience

More than 670,000 page images, representing a major assemblage of historical documents from Latin America's largest country, are now freely available over the Internet. (See wwwcrl.uchicago.edu/info/brazil/.) These images capture text, tabulated statistical data, and illustrative material including some maps and diagrams.

Goal 3: Implement mechanisms to ensure traditional bibliographic access to each digitized publication

CRL has cataloged each document set as an electronic file. Bibliographic access is thereby assured to those searching CRL's online catalog, in libraries that have loaded CRL's catalog as a complement to local records, and through national cataloging utilities such as OCLC via its public interface, WorldCat. These records can also be individually loaded into local online catalogs.

Goal 4: Provide structured access to the individual volumes within each serial set

Several levels of intellectual access have been provided to the digitized resources. The broadest measures, described above, provide catalog records for electronic files corresponding to the four categories of reformatted publications. (The total number of separate catalog records is much larger: the "ministerial reports" category, for instance, also includes separate records for each ministry.) The next level centers on providing access to specific volumes within each title. An index to each set of reports, supplemented with hypertext links, permits this level of navigation to and between reports.

Goal 5: Provide, as necessary, electronic indexing to the sections within these documents, single issues of which can be hundreds of pages long

Structured access within each report has been created in a variety of ways. Solutions have primarily reflected the availability of internal or external indexes, as informed by technological and financial constraints. The basic database associated with each report provides access to every page within the corresponding document, facilitating browsing from page to page and also direct entry to specific pages, supplements, or tables. When available, internal organizing aides such as tables of contents and subject indexes have been recreated electronically in order to improve access. Thus, the table of contents for each report was linked to the respective page images. The extensive subject indexes included in the *Laemmert Almanak* were linked directly to the corresponding page images with page mapping software. Ann Hartness's *Guide to Statistics in the Presidential Reports of the Brazilian Provinces, 1830-1889* (Austin, 1977) indexes statistical materials in this set of materials. The Project therefore prepared a digital version of this guide, linking each citation to the associated page image. Users can also search across these reports, combining presidents, provinces, years, and subject terms within a relational database.

Goal 6: Explore relative levels of demand and patterns of use for digitized materials by issuing them both as CD-ROMs and as files available over the Internet

The Project Committee decided at an early point to make the database available only over the Internet. The rapid emergence of the World Wide Web and ready availability of free Web browsers promised broad and easy access to the files—albeit overseas users, particularly those in Brazil, are even today somewhat hindered by difficult and expensive Internet connections. Moreover, short runs of CD-ROMs were at that time costly to produce (they've since become far less expensive), and distribution posed formidable logistical challenges.

Database use is continually monitored with "WebTrends" software, installed in May 1998. This program tracks database use (numbers of hits and requests), and also addresses matters associated with site administration. A user survey form was provided online in 1997, before many of the files could be consulted. Current responses provide a better picture of demand and use, and will help us to improve further the user interface and search options.

Goals 7: Refine the process of creating digital images files from preservation microfilm

Many analysts argue that text rich electronic documents are best provided as searchable files, in ASCII or marked-up formats. Quite apart from theoretical discussions, the costs of this approach proved prohibitive. Rekeying the text, at about $1.50 per thousand characters, would have required more than twice the entire grant amount of $225,000. Variations in typeface, font size, and density and optical resolution in our microfilm "originals," moreover, precluded the use of optical character recognition (OCR) software. OCR would also have proved problematic for charts and tables, as well as maps and illustrations.

The Project Committee therefore elected to prepare scanned page images as an approach that was cost effective and that would capture all the information, including tables and maps, that could be retrieved from the source microfilm. By 1995, when production scanning began, the basic process was quite well established. The Committee opted for two image formats. Each microfilm image was first scanned as a 300 dpi TIFF (Tagged Image File Format) image. GIF (Graphical Interface Format) images were then derived from the TIFF files. Since Web browsers routinely support the GIF format, these images became the default means for Internet delivery. The low resolution (100 dpi) bitonal GIF images, file sizes generally range between 30 and 70 kilobytes, transmitted quickly over the Internet. The higher quality TIFF images can also be downloaded, at a somewhat slower speed. Since current Web browsers do not support TIFF, a separate graphics viewer is needed as well.

While the basics of scanning from microfilm were understood before the Project began, both the Project Committee and the company contracted to perform the scanning faced numerous challenges during the production process. Most had to do with the sometimes problematic quality of the source microfilm, produced several decades ago (when preservation microfilming standards were less thoroughly developed) from originals that were often in poor condition. We all learned by doing, and can now point to microfilming refinements that should simplify similar projects in the future.

2.4 The Source Materials

The Project has focused on Brazilian serial documents. Brazil carries special importance both within the region and for North American scholars. Brazil is far and away Latin America's largest country, enjoying tremendous geopolitical significance and immense potential as an emerging world power. At the same time, Brazil exhibits all the economic, social, political, ethnic and ideological tensions that characterize Latin America as a whole. It is the object of a great deal of research by scholars both within and beyond its borders.

The Project has worked with four distinct yet overlapping sets of materials. The national level reports, ranging from the 1820s to the 1990s, cover Brazil's imperial and republican periods. Annual reports predominate. They, predictably, describe activities of the preceding year and establish goals for the year ahead. Each set of materials offers its own characteristics.

2.4.1 Provincial Presidential Reports (1823--1930) -- 2,572 reports

Provincial presidents, equivalent to state governors, prepared annual reports that were published by the provincial administration. These reports usually contain a narrative that discusses achievements and activities, typically with separate sections for each department. Many also include supplemental reports from the provincial departments. It's not at all unusual to find four to ten supplements, many with detailed statistical information, appended to one of these reports.

2.4.2 Federal Presidential Reports (1889--1993) -- 93 reports

These "State of the Union" reports continue to be issued annually. Each document narrates the previous year's achievements and outlines future goals. Some reports, especially from more recent years, include supplements and statistical information.

2.4.3 National Ministerial Reports (1821--1960) -- 872 reports

Brazil's ministries are operational organs of the executive branch, responsible for carrying out policies and programs. Even during periods of normal rule, and within constitutional structures separating executive from legislative powers, Latin American ministries typically enjoy substantial authority to promulgate rules and regulations. During periods of authoritarian rule, ministries have sometimes played significantly larger roles in making as well as implementing policy. Ministries often enjoy broader and more pervasive powers than might be expected in a North American context.

Thirteen federal ministries are represented in this collection. Their annual reports offer fuller contexts and more detailed analysis than the broad brush characterizations offered by the president.

2.4.4 Almanak Laemmert (1844--1889) -- 56 reports

This annual publication, issued by the Imperial Court in Rio de Janeiro, typically includes three sections:

1) The Imperial Court's report on the year's activities, including a lists of members of the Court and its supporting bureaucracy;
2) A supplement with census information, imperial decrees, advertising, etc.; and
3) A report on the province of Rio de Janeiro, seat of the Imperial Court.

The *Almanak* also provides extensive lists of businesses and property owners. A supplement to each issue describes legislation passed during the proceeding year as well as statistical data. Most reports include extensive subject indexes.

2.5 The Microfilm Collection

The microfilms that served as source materials were created through several separate projects. The Library of Congress, for instance, had microfilmed many of the ministerial reports. Other microfilmed materials were themselves the result of special projects. Scholars and librarians had long been aware of the importance of the provincial reports. Ann Hartness's *Subject Guide to Statistics in the Presidential Reports of the Brazilian Provinces, 1830-1889* galvanized interest in pulling together and preserving these materials. Hartness drew heavily on documents from Brazil's National Library and National Archive. LAMP, in the late 1970s, therefore embarked upon a joint project with the National Library. The Library filmed its own holdings, as supplemented by reports borrowed from local repositories throughout the country. LAMP provided raw preservation quality film stock, then virtually unavailable within Brazil, and received positive copies of all the reels. Creating unified microfilm collections from fragile and widely dispersed original publications was an essential first step in providing the source materials that allowed this digitization Project to take shape.

3. SCANNING FROM MICROFILM TO CREATE DIGITAL IMAGES

The Project relied on scanning to create digital images from analog images stored on microfilm. The process consisted of many steps, the first of which centered on defining our scanning procedures. Choosing image formats that could represent the microfilm images, and identifying the steps involved in scanning the microfilm and processing the images so that they could be indexed were part of this process. Once these procedures were established, the film was scanned by previewing the microfilm and then performing the scanning itself. Finally, the digital images were post-processed for retrieval and viewing over the Internet. Arranging the scanned images into an ordered collection to facilitate user access was the most complex portion of the process.

This section more fully describes all of these steps, and also provides cost figures for the work.

3.1 Scanning Procedures

3.1.1 Sampling Microfilm

We began by sending a sample of the microfilm to be scanned to PFA so that they could estimate their costs, recommend an image format for the project, and devise a hierarchical hypertext access structure based upon the documents' intellectual content. This sample included microfilm rolls selected by the Project Committee from two of the sets identified for scanning.[2]

PFA devised its initial scanning procedures on the basis of these sample reels. The film was viewed over a light table using a digital densitometer to measure the background density. Letter quality was evaluated by using a 15X loupe.   This process was consistent with PFA's normal film inspection procedures.[3]

This initial review of the microfilm identified approximately one percent of the sample images as partially or completely illegible. (See PFA Report in Appendix 10.4.) Other problems were discovered as well, as both film density and contrast were not always suitable for scanning. Further, some of the original images had been skewed during microfilming, creating unexpected post-processing requirements. Some of the paper originals had also deteriorated significantly before they were filmed, for instance with speckling. In some cases the printing was barely legible.

[2] Rio Grande do Sul - Provincial Presidential Reports, 1829-1890
Relatorios Ministeriais. Relaçoes Exteriores 1871-1888
Relatorios Ministeriais. Guerra 1827-1925

[3] Background Density: This is the numerical measurement of the contrast between the image and its background. Density is important because it affects the legibility of the image.

3.1.2 Evaluating the Microfilm

PFA rated the film sample in terms of its scanning difficulty on the basis of two primary criteria: frame separation, and contrast (density and letter brightness). Samples that fell within either of the first two categories listed below were considered good candidates for successful scanning. Sufficient contrast was necessary to produce on-screen readability and legibility. Adequate frame separation allowed the scan aperture to accurately position itself before each scan by means of an edge-detection routine, permitting automatic scanning rather than a more expensive manual process.

Ratings of Frame Separation and Contrast:

1) Acceptable edge-detection for frame positioning.
Overall contrast within reasonable quality limits.

2) Acceptable edge-detection for frame positioning.
Contrast variation approaching the lower limit for acceptable output quality.

3) Problematic edge-detection for frame positioning and/or contrast variation
at or below the limit for acceptable output quality.

4) Pull-down (operator defined) frame positioning likely to be necessary, and/or
contrast below the limit for acceptable output quality. (This situation indicates manual, rather than automatic scanning.)

3.1.3 Selecting Image Formats for Scanning and Serving

The Brazilian documents were scanned as TIFF files. The decision reflected the following criteria:

1) We sought a standard format for displaying images through a Web browser;
2) We wanted a format that adequately presented readable textual and tabular information found within the reports;
3) We wanted a file format that balanced image quality and file size, the latter being the primary determinant of the time required to transmit files to Internet users.

We also sought a stable file format that would not soon be discarded. However, as the Project unfolded, it became clear that we needed two image file formats, one as a format for creating the digital image and the delivery format to enable the page images to be viewed on Web browsers.

Both needs were addressed by scanning and saving each page as a 300-dpi TIFF. A graphic file-conversion utility program then converted each page-image as a 100-dpi GIF that was saved as the primary delivery format.

Choosing TIFF as the format to create and store digital images was easy, since TIFF was already the *de facto* industry standard in 1994. TIFF files can be compressed with no loss of information, guaranteeing fidelity to originals as well as relatively quick transfers over the Internet.

The Project opted for Group-4 compression, also an industry standard. TIFF files can be saved at virtually any resolution, including the 100-dpi which was chosen for the GIFs. However, TIFF is not yet (late 2000) supported for direct viewing within Web browsers. TIFF files must therefore be downloaded to the user's hard drive and then viewed using a separate graphical viewer program. (This technique provides access to the Project images that are illegible as 100-dpi GIFs.).

As of 1994, the most widely used formats for viewing images on the World Wide Web were GIF and JPEG (Joint Photographic Engineering Group). Both provide a "snapshot" of page-images. We opted for GIF because it provides a higher quality image for an equivalent file size. While JPEG performs better for color images, our microfilm source materials were all in black and white.

As of December 2000, both the TIFF and GIF formats continue to be industry standards.

3.1.4 Image Quality

A wide variety of conditions affected the readability of the microfilms from which we scanned, and therefore the legibility of the page images we produced. "Noise" around and within the text was variously caused by environmental degradation of the original paper reports, poor microfilming practices that led to muddied film images, the scanning process itself, and poor contrast between the microfilmed text and background.

3.1.5 Resolution Levels for Digital Images

We were able to select resolution levels for both of our file formats.[4] Higher resolutions mean sharper images and can in some cases enhance legibility. The resolution of an image is commonly described in terms of dots-per-inch, a figure obtained by squaring the number of dots-per-linear-inch. A 300-dpi image, for instance, has nine times as many pixels as a 100-dpi image, regardless of the file format used to store it. The file size is correspondingly larger.

The Project's master file of TIFF images was designed to preserve as much detail as possible from the source microfilm, permit lossless storage and transmission, and allow the eventual cre-

[4] Digital Resolution: The density of the dots within a mapped image, known as the resolution, determines how sharply the image is represented. This is often expressed in dots per inch (dpi) or simply by the number of rows and columns, such as 100 by 100. http://www.pcwebopedia.com/TERM/b/bit_map.html (10/16/01)

ation of additional derivative or “use” files as required by changing technology and new user needs. These files were therefore captured at an “intermediate” resolution of 300 dpi: lower than the levels suggested for the archival quality digital masters, but sufficient to capture essentially all the relevant information in our sometimes problematic source microfilm.

These 300 dpi files also provide an alternative when the low resolution, 100 dpi GIF’s are difficult to read. In this situation, users can access the TIFF file by clicking on the hypertext button titled “300-dpi TIFF image,” which is found at the top and bottom of each page-image file. The TIFF image will then download onto the user’s hard drive where a specific TIFF viewer, once installed on the machine, will open it for viewing. Should a page-image remain illegible in the 300-dpi TIFF format, the reader's only recourse is to refer to the microfilm or, more likely, a hard-copy original.

In 1994 (and still in 2000), computer monitors use video drivers that format on-screen images at various resolutions. High-resolution images can only be fully represented on monitors capable of showing many pixels. Most drivers in 1994 allowed for a maximum screen resolution of 800x600 dpi, so we chose this resolution for our delivery files in order to minimize the need for scrolling.Many drivers in 1994 were fixed or set at 640x480 dpi, which unfortunately required users to scroll from left to right in order to read the 800x600 dpi page-images. Larger monitors are by now more common and affordable.

For the reasons described above, the GIF and TIFF formats were chosen for the Project. These image file formats were used for scanning and storing the Provincial Presidential Reports, Presidential Messages and Ministerial Reports.

TIFF -- 300-dpi resolution 1-bit, bitonal (file size = approx. 30-250 kilobytes)
(The TIFF 6.0 files were compressed with ITU Group 4 method)
GIF -- 100-dpi resolution, bitonal   (file size = approx. 20-70 kilobytes)

In 1998, after the scanning of the previous three sets had been completed, PFA scanned the *Almanak* images, again utilizing the GIF and TIFF formats. However, due to evolving industry format standards, a different version of the GIF was used for the Almanak. Whereas the 1-bit GIF format provides only bitonal, black-and-white values for pixels, the 2-bit GIF gray-scale format represents four possible values of gray -- white, light gray, dark gray or black (values 0-3). This choice of the 2-bit GIF format yielded higher quality images with only a nominal increase in file size and transmission time.

TIFF -- 300-dpi resolution 1-bit, bitonal (file size = approx. 30-250 kilobytes)
(The TIFF 6.0 files were compressed with ITU Group 4 method)
GIF -- 100-dpi resolution 2-bit gray-scale (file size = approx. 70 kilobytes)

3.1.6 Transmission Time over the Internet

Connection or modem speed and available bandwidth being equal, the size of the file will determine the amount of time required for downloading. When a user is moving through documents one page at a time, quick downloads are crucial. Under normal circumstances, the relatively small GIF images open very quickly within a Web browser’s window.

The TIFF images, even though about nine times larger, travel over the Internet more quickly than their size might suggest. They are stored and transmitted as compressed files, that are then expanded on the client’s computer by whatever TIFF-viewer has been installed and linked to the machine’s Web browser.

3.2 Costing Out Scanning and Post-Processing Procedures

PFA initially categorized scanning costs into two levels depending on whether the camera’s automatic frame-detection could be used. Post-processing had its own separate costs. (Section four addresses these processes and costs in greater detail.)

The costs mentioned here do not reflect current prices: technology and scanning procedures have change dramatically since 1995.

LEVEL ONE - $0.195 per frame – This category subsumes scanning from rolls that are conducive to reliable, automatic detection of frame edges, and in which the images contain sufficient contrast to allow the scanner’s enhancement system to produce acceptable digital images without operator intervention. Many microfilm frames include two page-images. Microfilm frames that fall within the first two categories of the microfilm rating table transcribed in section 3.1.2 were ranked at level one.

LEVEL TWO - $0.225 per frame – These images entail scanning from microfilm rolls that contain frames with insufficient separation or density to allow for reliable, automatic frame edge detection. Frames that fell within the third and fourth categories in section 3.1.2 on the microfilm rating table were assigned to this level.

RE-SCANNING - $1.35 per frame – The cost to re-scan non-sequential film frames primarily reflects the labor required for each scan. Reasons for re-scanning include efforts to improve upon poor film images, and image-cropping problems arising out of detection difficulties.

PFA originally estimated that ten percent of the frames would require manual rescanning. The actual figure was much lower, since PFA found that careful inspection of the film during the preview phase identified most problems before the initial scan.

3.2.1 Post-Processing Costs

COST- $0.0403 per delivered image
Post-processing included the following procedures:
1) Cropping the images of frames that contain a single page or document; splitting and cropping images corresponding to microfilm frames containing two pages of text;
2) rotating images to display correctly on a monitor;
3) deleting images that were irrelevant to the report set (microfilming targets, punch-outs, etc.); and,
4) placing images within an HTML hierarchical structure to create an organized representation of the source documents.

3.2.2 Film Duplication Costs

COST - $18.50 per roll
The density levels of some microfilm reels were inadequate for scanning. These reels were duplicated at appropriate levels allowing some (though not all) of these pages to be scanned.

3.2.3 Renegotiation of Post-Processing Costs

The initial microfilm sample submitted to PFA failed to represent the nature and volume of problems that were subsequently encountered. These challenges led PFA to repeatedly review and revise its scanning and post-processing procedures. For instance, the image within a microfilm frame could reflect any of approximately twenty-four page positions. These variations included page orientation, the number of pages per frame (one or two), and shifts in page size due to foldouts, maps, and the like. Continuous encounters with variation within the microfilm led PFA to request that we renegotiate the contract costs for post-processing. The eventual outcome was to increase these charges from $0.0403 to $0.05 per frame image.

When the Project began, neither the Project Committee nor PFA was aware of the large number of problems affecting both the film and the original documents that would complicate both scanning and indexing. Given each group’s limited initial experience with this kind of scanning project, it’s no surprise that the sample provided an inadequate representation of the document set.[5]

Future projects might minimize unexpected cost variations of this type through either of two measures. First, the company or institution scanning the documents should build into its contract a preliminary cost to allow a review of the entire document set. (For the Project’s second phase to scan and index the *Almanak Laemmert*, PFA charged an inspection fee of $7.50 per roll. This

[5] PFA, with a scanning capacity of 500,000 microfilm images per week, found the microfilm for this Project to be as difficult as any they have processed.

allowed the company to anticipate the cost associated with difficult image scans.) Second, an experienced microfilm provider might undertake an initial preview of the film. Even with such careful pre-planning, however, both parties must realize that unforeseen problems may significantly affect costs.

3.3 Microfilm Acquisition and Duplication

Several generations of the source microfilm were usually available for scanning. CRL typically circulates service (positive) microfilm reels to users, while retaining a printing negative in storage for use in making both service copies and sale copies. We determined that these printing negatives would produce the cleanest scans for three reasons: these copies are minimally removed from the original paper documents; service copies will have suffered some loss of legibility during the imperfect process of duplication; and any service copy circulated by CRL will have accumulated additional “noise” (i.e. scratches) from use.

CRL stores its microfilm negatives with Preservation Resources, a firm based in Pennsylvania. As the Project advanced, CRL would notify Preservation Resources to send the microfilm cores to PFA. In most cases the shipping went smoothly, though there were some instances in which shipments were not complete and had to be reordered.

Each document set was scanned in its entirety before the next set was begun. The Provincial Presidential Reports were completed first, followed by the National Presidential Reports, the Ministerial Reports, and the *Almanak*. Compartmentalizing the scanning process minimized confusion and allowed orderly transitions to the indexing phase.

PFA previewed every microfilm reel twice on a light table in order to establish scanner settings in advance. The first inspection determined whether the film was scannable. The second focused on density and resolution levels, image arrangement (cine or comic), frame spacing, and such film characteristics as frame sizes, splices, mixes of different sized documents, shadows, and the like. The tables of contents were also identified during the second review.

The first preview quickly revealed any portions of the film that were unscannable due to inadequate background density. The minimum acceptable level for scanning was 0.5; lower densities can cause the text and background to blend. Optimal densities are in the 0.8 – 1.1 range unless the characters are extremely faint, in which case a lower density may make the letters clearer.[6]

When duplication was necessary, PFA and the Project Committee opted for densities that would allow good electronic images. All the film identified for reduplication had initially been prepared in Brazil. Twenty-one reels of the Ministerial Reports and thirteen of the Provincial Presidential Reports were reduplicated at the National Library of Brazil. Direct duplicating microfilm stock is quite costly in Brazil, so CRL sent film stock to the Library of Congress’ Field

[6] Film could also be unscannable from being out-of-focus or including dark (underexposed) images. These problems could not be corrected by duplicating the microfilm, but would rather require a new microfilm of the original paper documents.

Office in Rio de Janeiro, from whence it was delivered to the National Library. The Library of Congress Rio Office then shipped the completed film back to CRL.

Most microfilm that meets archival standards can be scanned, but adherence to certain criteria regarding density and resolution will lead to better electronic images. PFA created the following list of requirements for film duplication in order to create microfilm images that lend themselves to legible electronic images.

1) Produce a direct duplicate negative off the camera negative (using Kodak Direct Duplicating Microfilm or equivalent).
2) Choose the exposure for each roll of microfilm based on its own characteristics of density and contrast.
3) Produce rolls with good appropriate contrast, taking into consideration the density, contrast, and content of frames on that roll.
4) Recommended density range:
The background density should be no lighter than the original film.
The background density should be no darker than 0.3 above the original film.
5) The Dmin value, or density of "clear" film, should not be below 0.25.

Despite these reduplication efforts, the following reels of Ministerial Reports could not be scanned:

1) Agricultura
1946-1950 The film is too dark. The text around the edges of the page is faint.
2) Guerra
1846-1854 The film is too light.
1865-1869 The density is acceptable but the tops of the pages are out of focus.

The complicated logistics of reduplicating film in Brazil precluded further efforts, even though PFA thought that the problems with the "Guerra" film could be overcome. All these films were also legible, even though they were not scannable according to PFA's criteria.

3.4 Scanning Equipment and Processing

The scanner used for the Brazil Project was a SunRise Imaging SRI-50 Microfilm Scanner. The software used to run the scanner was Scscan^{TM}, a DOS-based program that is now obsolete.

Scanning was accomplished by loading a roll of microfilm on the scanner, and adjusting the settings in accord with the condition of the microfilm images as determined during the preview process. These settings included an adjustable pull-down feature for the aperture to accommodate the varying frame sizes found in many rolls of film.

The scanner automatically lined up the film so that the camera could capture a 300-dpi TIFF image. The scanner could not be preset to manage automatically all of the problems that would lead to unacceptable page-images, so a PFA staff member had to view each image after it was loaded and make manual adjustments as necessary.[7]

Once the page was scanned, the TIFF page-image was saved to a server to await post-processing.

3.5 Post-Processing Scanned Images

The post-processing prepared the TIFF images for viewing and indexing. These steps were conducted by the PFA lab in California and also a subcontracted programmer in Hawaii.

The TIFF images for each report were first placed in a multi-image "pipe file," so that a technician could view each image for screen readability and legibility using a high-resolution monitor. Most of these images consisted of frames containing two pages of text. PFA employed the Scindex program to split the two pages within the frame. Images deemed irrelevant (targets, punch-outs, blank pages found at the beginning of reports, etc.) were deleted. Each page was then given a unique locator code consisting of the report number and page number (e.g. 40/000128.tiff).[8]

At this point, the images were saved onto 4-mm DAT Tape and sent to a programmer in Hawaii to create the GIF files from the TIFF images and then organize these files using a data structure based on HTML. Converting the 300-dpi TIFF to 100-dpi GIF was accomplished by taking 3x3 blocks of pixels from each TIFF image and representing them as one black or white pixel. The images were also rotated so that they displayed in the correct orientation on a monitor. The two microfilm image modes, "cine" (pages filmed top-to-bottom) and "comic" (pages filmed side-by-side), were sometimes intermixed within a single document, requiring manual rotation.

The processed GIF and TIFF files were saved back onto 4-mm DAT tapes and sent directly to CRL. Data from each tape were then transferred to the jukebox at CRL in a process that required about 7.5 hours per tape, reflecting a transfer rate of about one hundred and fifty megabytes per hour.

[7] While only 30 frames per minute could be scanned in this manner, more recent (and straightforward) PFA projects have achieved production rates of 90 frames per minute without manual intervention.

[8] The Scindex software, a DOS product by Image Retrieval Inc., was created especially for this project. It is now being used in a few other projects at PFA, and will probably be migrated to a Windows environment.

4. INDEXING THE DOCUMENT COLLECTION[9]

While the creation of images from these microfilm sets was relatively straightforward, the development of “finding aids” required a more creative approach. The image database was like an opaque glass that needed an indexing overlay to become transparent. Full-text databases, by contrast, can be searched with information retrieval systems that offer relatively high levels of both precision and recall.[10]

An image database demands extensive indexing work. We invested far more time and effort in creating database indexing structures than in scanning and processing the images. PFA’s rule of thumb is that two-thirds of total project time will be spent on indexing.

The Brazil Project ultimately utilized five different approaches to index the four image collections. The traditional organization according to page numbers and chapters was recreated electronically for all of the materials. Subject indexes, tables of contents, and a freestanding 1997 guide to statistics found in the Provincial Presidential Reports provided supplemental opportunities for access.

4.1 Indexing Phases

4.1.1 Phase One Indexing

All of our indexing has been based on hypertext technology through which each index citation can be linked to an HTML file which includes embedded pointers to each and every page image. During Phase One we established a website hierarchy and devised three indexing approaches:

1) A hypertext hierarchy was established to permit navigation between the four collections;
2) A report-level access structure was designed and applied globally to all documents;
3) The few tables of contents found in the collections were keyed and hyperlinked to their respective page-image GIF documents; and,
4) Ann Hartness, *Subject Guide to Statistics in the Presidential Reports of the Brazilian Provinces, 1830-1889*, which identifies selected quantitative information in the Provincial Presidential Reports, was keyed as a digital file and its citations were hyperlinked to the corresponding page-images.

[9] Portions of “Section 4” can be found in Scott Van Jacob. “Six Ways from Sunday: Approaches to Indexing Digital Text Images.” *Computers in the humanities*.

[10] The definitions of precision and recall found in the “International Encyclopedia of Information and Library Science,” p. 211, will be used here. “The recall ratio measures the proportion of those relevant documents in a database which are retrieved, whilst the precision ratio measures the proportion of the retrieved items which are relevant.”

4.1.2 Phase Two Indexing

During Phase Two we addressed the difficulties of finding particular pieces of information within files that could be several hundred pages long. Many of the scholars and librarians who reviewed the database recommended better access to specific information. Phase Two's approach to indexing was greatly improved by the thoughtful responses received from all of these users and reviewers.

Two indexes were created during Phase Two. The first, building on the subject headings used in the Hartness *Guide*, created both a controlled vocabulary to each of the 2,572 Provincial Presidential Reports and an interface for searching this index using an Access^{TM} relational database system and SQL Server. The second index was constructed by image-mapping the *Almanak's* detailed subject indexes to link the citations directly to their respective page-images. (See Appendix 10.4 for HTML coding of pagination files.)

4.2. Indexing Approaches to the Image Files

All four collections were organized by a hypertext structure to provide basic access to every report. Table Two charts the application of each of the five indexing methods to provide more detailed access within each document set. It also reveals that not all indexing methods were used on all sets.

The Provincial Presidential Reports have been indexed by four of the five methods listed in Table Two. These materials describe activities for nineteen provinces over more than a one hundred years including the Imperial and Republican eras. They also comprise about seventy-two percent of all the reports represented within the Project.

Table 2: Access to Document Sets

	Provincial Pres. Reports	Presidential Reports	Ministerial Reports	Almanak Laemmert
Hypertext structure providing chronological and/ or regional access to reports	x	x	x	x
Report pagination files	x	x	x	x
Report table of contents	x	x	x	
Hartness' Guide to Statistical Information...	x			
Search interface by Subject Thesaurus	x			
Image-Mapping of Subject Index				x

The following sections (4.2.1 -- 4.2.8) describe the hierarchical organization of the Brazil Project's website and each of the five indexing approaches noted above. Each section includes a description, and often a table showing an example of the indexing.

4.2.1. The Homepage and Beyond

The Brazil Project's homepage (See Example A.) provides overall access to the database, serving as a finding aid rather than an index. The home page offers direct hypertext links to all of the Reports, as well as providing descriptions of the collections, information on the arrangement of the Reports, and access to the search interface for the Provincial Presidential Reports. These pages and data are available in both English and Portuguese.

Access to specific reports within the four document sets varies according to their arrangement. The *Almanak* and National Presidential Reports, issued by a single body, are organized chronologically. The Provincial Presidential Reports are organized by province and then chronologically within each province. Ministerial Reports are organized by ministry and then chronologically within each ministry. For users who already know which report they are searching for, the Homepage hypertext structure provides a means of moving directly to the document. For example, the following pages would have to be negotiated to reach a specific page image within a Ministerial Report.

a. Project Homepage

b. Ministerial Reports Collection

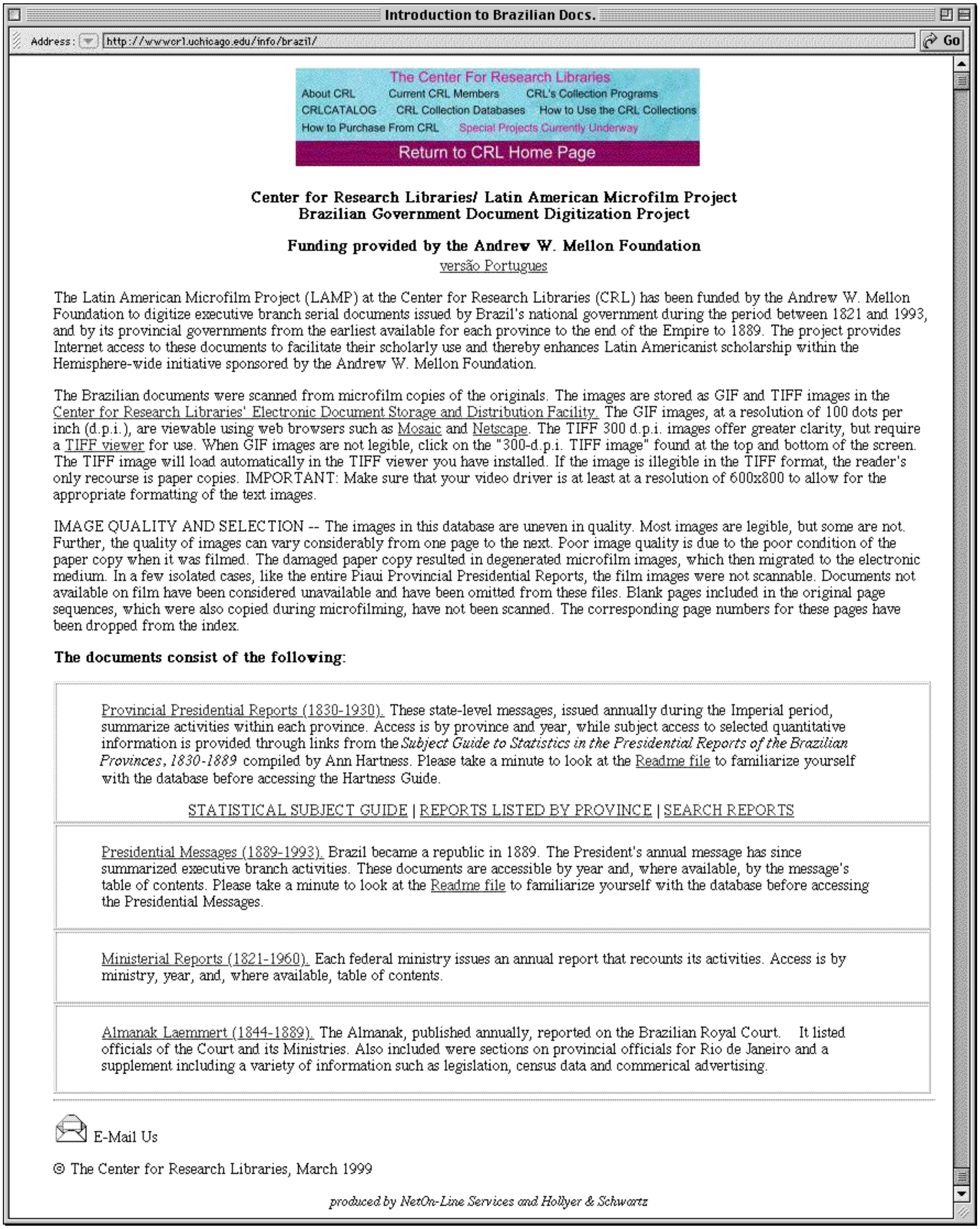
c. Ministry

d. Report Year

e. Pagination File

f. Image

Example A: Project Website[http://www.crl.uchicago.edu/info/brazil]


Introduction to Brazilian Docs.
Address: http://wwwcrl.uchicago.edu/info/brazil/
Go
The Center For Research Libraries
About CRL
Current CRL Members
CRL's Collection Programs
CRLCATALOG
CRL Collection Databases
How to Use the CRL Collections
How to Purchase From CRL
Special Projects Currently Underway
Return to CRL Home Page

Center for Research Libraries/ Latin American Microfilm Project
Brazilian Government Document Digitization Project

Funding provided by the Andrew W. Mellon Foundation

versão Portugues

The Latin American Microfilm Project (LAMP) at the Center for Research Libraries (CRL) has been funded by the Andrew W. Mellon Foundation to digitize executive branch serial documents issued by Brazil's national government during the period between 1821 and 1993, and by its provincial governments from the earliest available for each province to the end of the Empire to 1889. The project provides Internet access to these documents to facilitate their scholarly use and thereby enhances Latin Americanist scholarship within the Hemisphere-wide initiative sponsored by the Andrew W. Mellon Foundation.

The Brazilian documents were scanned from microfilm copies of the originals. The images are stored as GIF and TIFF images in the Center for Research Libraries' Electronic Document Storage and Distribution Facility. The GIF images, at a resolution of 100 dots per inch (d.p.i.), are viewable using web browsers such as Mosaic and Netscape. The TIFF 300 d.p.i. images offer greater clarity, but require a TIFF viewer for use. When GIF images are not legible, click on the "300-d.p.i. TIFF image" found at the top and bottom of the screen. The TIFF image will load automatically in the TIFF viewer you have installed. If the image is illegible in the TIFF format, the reader's only recourse is paper copies. IMPORTANT: Make sure that your video driver is at least at a resolution of 600x800 to allow for the appropriate formatting of the text images.

IMAGE QUALITY AND SELECTION -- The images in this database are uneven in quality. Most images are legible, but some are not. Further, the quality of images can vary considerably from one page to the next. Poor image quality is due to the poor condition of the paper copy when it was filmed. The damaged paper copy resulted in degenerated microfilm images, which then migrated to the electronic medium. In a few isolated cases, like the entire Piaui Provincial Presidential Reports, the film images were not scannable. Documents not available on film have been considered unavailable and have been omitted from these files. Blank pages included in the original page sequences, which were also copied during microfilming, have not been scanned. The corresponding page numbers for these pages have been dropped from the index.

The documents consist of the following:

Provincial Presidential Reports (1830-1930). These state-level messages, issued annually during the Imperial period, summarize activities within each province. Access is by province and year, while subject access to selected quantitative information is provided through links from the *Subject Guide to Statistics in the Presidential Reports of the Brazilian Provinces, 1830-1889* compiled by Ann Hartness. Please take a minute to look at the Readme file to familiarize yourself with the database before accessing the Hartness Guide.

STATISTICAL SUBJECT GUIDE | REPORTS LISTED BY PROVINCE | SEARCH REPORTS

Presidential Messages (1889-1993). Brazil became a republic in 1889. The President's annual message has since summarized executive branch activities. These documents are accessible by year and, where available, by the message's table of contents. Please take a minute to look at the Readme file to familiarize yourself with the database before accessing the Presidential Messages.

Ministerial Reports (1821-1960). Each federal ministry issues an annual report that recounts its activities. Access is by ministry, year, and, where available, table of contents.

Almanak Laemmert (1844-1889). The Almanak, published annually, reported on the Brazilian Royal Court. It listed officials of the Court and its Ministries. Also included were sections on provincial officials for Rio de Janeiro and a supplement including a variety of information such as legislation, census data and commerical advertising.

E-Mail Us

© The Center for Research Libraries, March 1999

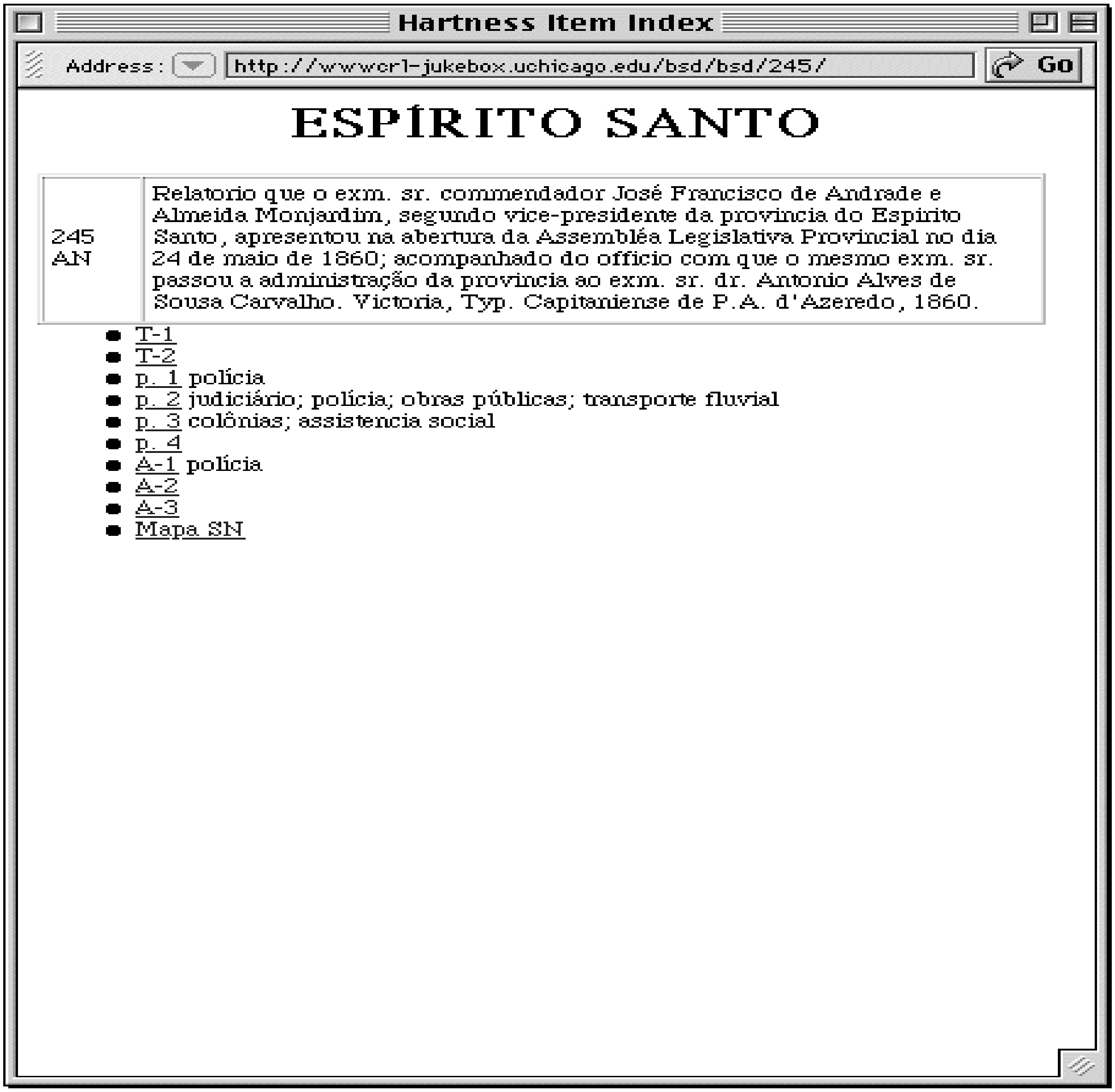
produced by NetOn-Line Services and Hollyer & Schwartz

4.2.2 Pagination Files

PFA created a page index for each and every report. Each of these indexes details the length of the document, the number of supplements included therein, and the number of tables.

Each of the 3,593 reports has been assigned a unique number ranging from 001-U2461. In the case of the Provincial Presidential Report in Example B, the number “245” reflects the report item number from the Hartness *Guide*.

Example B - Pagination File 245 [http://wwwcrl-jukebox.uchicago.edu/bsd/bsd/245/]



ESPÍRITO SANTO

245 AN	Relatorio que o exm. sr. commendador José Francisco de Andrade e Almeida Monjardim, segundo vice-presidente da provincia do Espirito Santo, apresentou na abertura da Assembléa Legislativa Provincial no dia 24 de maio de 1860; acompanhado do officio com que o mesmo exm. sr. passou a administração da provincia ao exm. sr. dr. Antonio Alves de Sousa Carvalho. Victoria, Typ. Capitaniense de P.A. d'Azeredo, 1860.

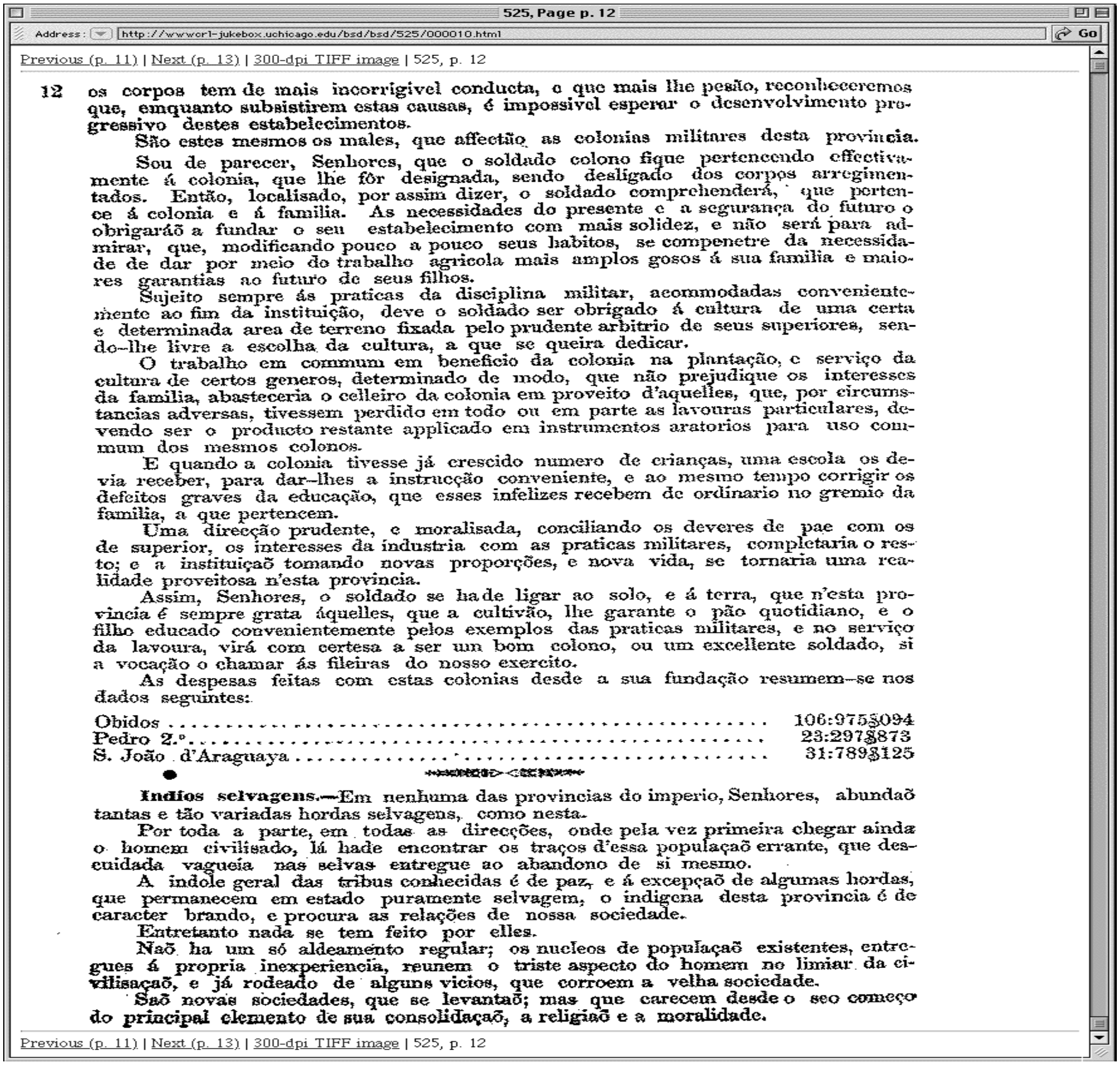
- T-1
- T-2
- p. 1 polícia
- p. 2 judiciário; polícia; obras públicas; transporte fluvial
- p. 3 colônias; assistencia social
- p. 4
- A-1 polícia
- A-2
- A-3
- Mapa SN

Page-images are coded sequentially within each pagination file, beginning with "000000" and continuing to the final page. Since none of the reports exceeded 1,000 pages, the coding could have been limited to four digits.

Example C illustrates page 12 of report number "525". The GIF and TIFF images were assigned the same numerical name in order to match them during the pagination process. They differ in their format suffix, e.g. 00001.gif and 00001.tif.

Example C -- Pagination File "Report 245: Espirito Santo," Page 12

[http://wwwcrl-jukebox.uchicago.edu/bsd/bsd/525/000010.html]



http://wwwcrl-jukebox.uchicago.edu/bsd/bsd/525/000010.html

Previous (p. 11) | Next (p. 13) | 300-dpi TIFF image | 525, p. 12

12 os corpos tem de mais incorrigivel conducta, o que mais lhe pesão, reconheceremos que, emquanto subsistirem estas causas, é impossivel esperar o desenvolvimento progressivo destes estabelecimentos.

São estes mesmos os males, que affectão as colonias militares desta provincia.

Sou de parecer, Senhores, que o soldado colono fique pertencendo effectivamente á colonia, que lhe fôr designada, sendo desligado dos corpos arregimentados. Então, localisado, por assim dizer, o soldado comprehenderá, que pertence á colonia e á familia. As necessidades do presente e a segurança do futuro o obrigaráõ a fundar o seu estabelecimento com mais solidez, e não será para admirar, que, modificando pouco a pouco seus habitos, se compenetre da necessidade de dar por meio do trabalho agricola mais amplos gosos á sua familia e maiores garantias ao futuro de seus filhos.

Sujeito sempre ás praticas da disciplina militar, acommodadas convenientemente ao fim da instituição, deve o soldado ser obrigado á cultura de uma certa e determinada area de terreno fixada pelo prudente arbitrio de seus superiores, sendo-lhe livre a escolha da cultura, a que se queira dedicar.

O trabalho em commum em beneficio da colonia na plantação, e serviço da cultura de certos generos, determinado de modo, que não prejudique os interesses da familia, abasteceria o celleiro da colonia em proveito d'aquelles, que, por circumstancias adversas, tivessem perdido em todo ou em parte as lavouras particulares, devendo ser o producto restante applicado em instrumentos aratorios para uso commum dos mesmos colonos.

E quando a colonia tivesse já crescido numero de crianças, uma escola os devia receber, para dar-lhes a instrucção conveniente, e ao mesmo tempo corrigir os defeitos graves da educação, que esses infelizes recebem de ordinario no gremio da familia, a que pertencem.

Uma direcção prudente, e moralisada, conciliando os deveres de pae com os de superior, os interesses da industria com as praticas militares, completaria o resto; e a instituiçaõ tomando novas proporções, e nova vida, se tornaria uma realidade proveitosa n'esta provincia.

Assim, Senhores, o soldado se ha de ligar ao solo, e á terra, que n'esta provincia é sempre grata áquelles, que a cultivão, lhe garante o pão quotidiano, e o filho educado convenientemente pelos exemplos das praticas militares, e no serviço da lavoura, virá com certesa a ser un bom colono, ou um excellente soldado, si a vocação o chamar ás fileiras do nosso exercito.

As despesas feitas com estas colonias desde a sua fundação resumem-se nos dados seguintes:

Obidos	106:975$094
Pedro 2.º	23:297$873
S. João d'Araguaya	31:789$125

Indios selvagens.—Em nenhuma das provincias do imperio, Senhores, abundaõ tantas e tão variadas hordas selvagens, como nesta.

Por toda a parte, em todas as direcções, onde pela vez primeira chegar ainda o homem civilisado, lá hade encontrar os traços d'essa populaçaõ errante, que descuidada vagueia nas selvas entregue ao abandono de si mesmo.

A indole geral das tribus conhecidas é de paz, e á excepçaõ de algumas hordas, que permanecem em estado puramente selvagem, o indigena desta provincia é de caracter brando, e procura as relações de nossa sociedade.

Entretanto nada se tem feito por elles.

Naõ ha um só aldeamento regular; os nucleos de populaçaõ existentes, entregues á propria inexperiencia, reunem o triste aspecto do homem no limiar da civilisaçaõ, e já rodeado de alguns vicios, que corroem a velha sociedade.

Saõ novas sociedades, que se levantaõ; mas que carecem desde o seo começo do principal elemento de sua consolidaçaõ, a religiaõ e a moralidade.

Previous (p. 11) | Next (p. 13) | 300-dpi TIFF image | 525, p. 12

While based on hypertext technology, this access structure essentially recreates the paper-based organization of each report. Users can enter a document at any page, read pages consecutively, or move around in any other order. This approach also makes it easy to view supplemental materials.

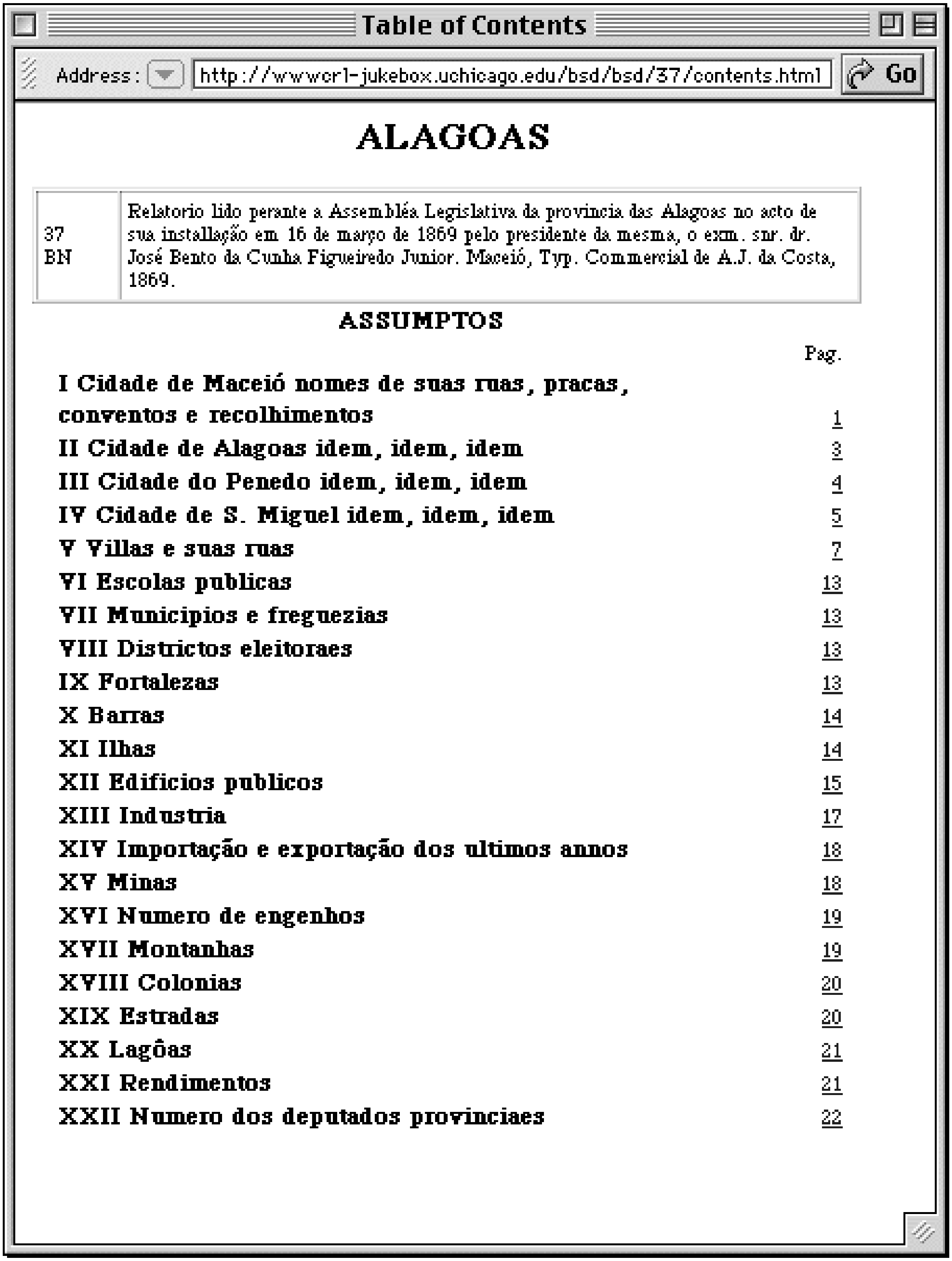
Each page image also contains direct hypertext links that allow users to move sequentially to the next or previous page. The *Almanak* collection, created in Phase Two, adds more options at the page level including a link back to the pagination file and a link to the report's subject index. Every page-image also provides a button that allows access to the corresponding TIFF image. It was relatively easy to paginate the narrative section of the Project documents. However, the many appendices and tables found in some reports proved a greater challenge. We therefore devised an alphanumeric coding system to accommodate multiple appendices. "A-1, A-2" was utilized for annexes (*Anexos*), (Appendix A, pg. 1; Appendix A, pg. 2), while "S-1, S-2" was used for sub-reports, which was the default category when an appendix was not clearly identified as an annex. Tables were called "*Mapas*" in accordance with Portuguese spelling. *Mapas* are numbered ("*Mapa 1*, *Mapa 2*") in accordance to the numbering within the report. Maps or other pages without numbers are referred to as "s.n." (*sem número*), the term for unnumbered pages. The Provincial Presidential Reports, in particular, include many unnumbered pages.

4.2.3 Tables of Contents

Only some of the Project documents contain tables of contents. There seems to be little rhyme or reason to explain which reports have a table of contents and which do not. Some reports were too short to need such a guide. Many documents of the early nineteenth century did without. Tables of contents found within Project reports were rekeyed into the database, after which PFA linked the page numbers to their corresponding page-images. (See Example D.) This process cost $2.70 per table of contents.

Tables of contents were not rekeyed for the *Almanak* since these materials contained extensive subject indices that offered superior access to their contents.

Example D - Table of Contents (excerpted due to length) [http://wwwcrl-jukebox.uchicago.edu/bsd/bsd/37/contents.html]



ALAGOAS

37 BN	Relatorio lido perante a Assembléa Legislativa da provincia das Alagoas no acto de sua installação em 16 de março de 1869 pelo presidente da mesma, o exm. snr. dr. José Bento da Cunha Figueiredo Junior. Maceió, Typ. Commercial de A.J. da Costa, 1869.

ASSUMPTOS

Pag.

I Cidade de Maceió nomes de suas ruas, pracas, conventos e recolhimentos 1
II Cidade de Alagoas idem, idem, idem 3
III Cidade do Penedo idem, idem, idem 4
IV Cidade de S. Miguel idem, idem, idem 5
V Villas e suas ruas 7
VI Escolas publicas 13
VII Municipios e freguezias 13
VIII Districtos eleitoraes 13
IX Fortalezas 13
X Barras 14
XI Ilhas 14
XII Edificios publicos 15
XIII Industria 17
XIV Importação e exportação dos ultimos annos 18
XV Minas 18
XVI Numero de engenhos 19
XVII Montanhas 19
XVIII Colonias 20
XIX Estradas 20
XX Lagôas 21
XXI Rendimentos 21
XXII Numero dos deputados provinciaes 22

4.2.4 *Subject Guide to Statistics in the Presidential Reports of the Brazilian Provinces, 1830-1889*

In the 1970s, Ann Hartness, a librarian at the Nettie Lee Benson Latin American Collection at the University of Texas, Austin, compiled a guide to selected statistics from 1,085 Provincial Presidential Reports issued during the Imperial period, 1830 -- 1889. Reports issued from 1890 -- 1930 were not indexed in this publication. The Hartness *Guide* is arranged by ninety-two subject headings, such as “elections”, “gold”, and “public work expenditures.” Citations within each subject category are arranged by Province, year, report (Referred to as an “item.”), and page or table number.

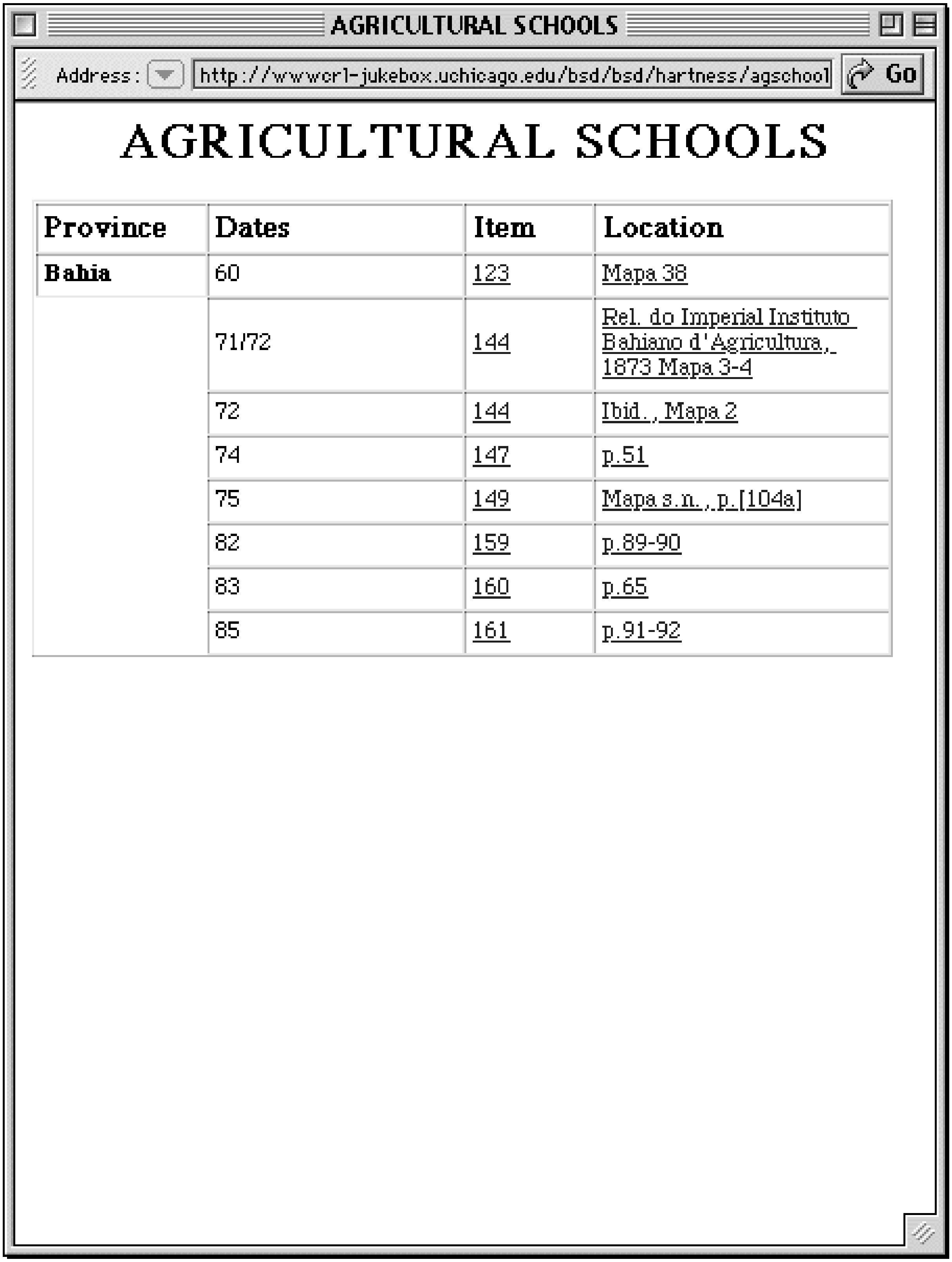
The Project Committee was eager to use the Hartness *Guide* as an access tool to the image files. We therefore created an electronic *Guide*, following the same format as the paper publications. Specific citations are linked under the electronic *Guide’s* “Location” and “Item” (i.e. pagination files) columns. (See Example E.)

PFA rekeyed the *Guide* to create this electronic database. Post-processing of the page-images included matching between the index database created with the Scindex program and the Hartness *Guide* database in order to create hyperlinks to the individual images. About sixty percent of the links were correctly identified during this matching process. However, the complexity of the electronic pagination files’ numeration meant that forty percent of the citations could not be automatically linked to their respective page images.

These remaining hyperlinks had to be created manually, a time-consuming, detail-oriented procedure that was carried out by student workers:

1) The students used a Web browser to scan each report file in order to locate unmatched citations. Most of the “missing” statistics cited in the *Guide* were found in a supplement to the report.
2) The remaining unmatched citations were checked against other report files that were issued during the same year and from the Province.
3) Finally, the microfilmed reports were inspected to locate citations that could not be found in the electronic files. The microfilm reader enabled students to move quickly through the myriad supplements found in some of these reports as they searched for the appropriate statistical table.

Example E - Hartness *Guide* [http://wwwcrl-jukebox.uchicago.edu/bsd/bsd/hartness/agschool.html]



AGRICULTURAL SCHOOLS
Address:
http://wwwcrl-jukebox.uchicago.edu/bsd/bsd/hartness/agschool
Go

AGRICULTURAL SCHOOLS

Province	Dates	Item	Location
Bahia	60	123	Mapa 38
	71/72	144	Rel. do Imperial Instituto Bahiano d'Agricultura, 1873 Mapa 3-4
	72	144	Ibid., Mapa 2
	74	147	p.51
	75	149	Mapa s.n., p.[104a]
	82	159	p.89-90
	83	160	p.65
	85	161	p.91-92

4.2.5 Indexing the Provincial Presidential Reports and Creating the Search Interface

Each report's page index ("pagination file") provided direct links to the individual page images. Nearly fifteen percent of the Provincial Presidential Reports also had an internal table of contents. However, users seeking more specific information on, for instance, "elections in Paraná in 1870," were faced with the daunting tasks of browsing numerous pages and /or supplements. Scholars who beta-tested the database strongly recommended that a finding aid be added to each report to describe its contents. Enhanced indexing for the Provincial Presidential Reports was addressed during the Project's second phase.

Enhanced indexing was accomplished by using controlled subject terms to index each chapter within these reports. Almost every report included chapter headings that clearly identified significant issues and were easy to spot. (The alternative approach of analyzing each page was dismissed as too labor intensive and as unlikely to significantly improve access.)

Our controlled subject terms were carefully selected to adequately represent the chapter headings found in the reports themselves. A sampling of chapter headings from fifty of the reports revealed 320 terms. Simply repeating these natural language headings proved unfeasible due to the variant terms for a single concept. For example, we found eight headings for public health, twenty-five for the judicial system, and the following fourteen variants for "police":

companhia de policia
corpo policial
divisão policial
departamento de segurança pública
força policia
força pública
ordem pública
policia
policia do porto
segurança individual
segurança individual e de propiedade
segurança pública
tranquillidade pública, e segurança individual
tranquillidade e segurança pública

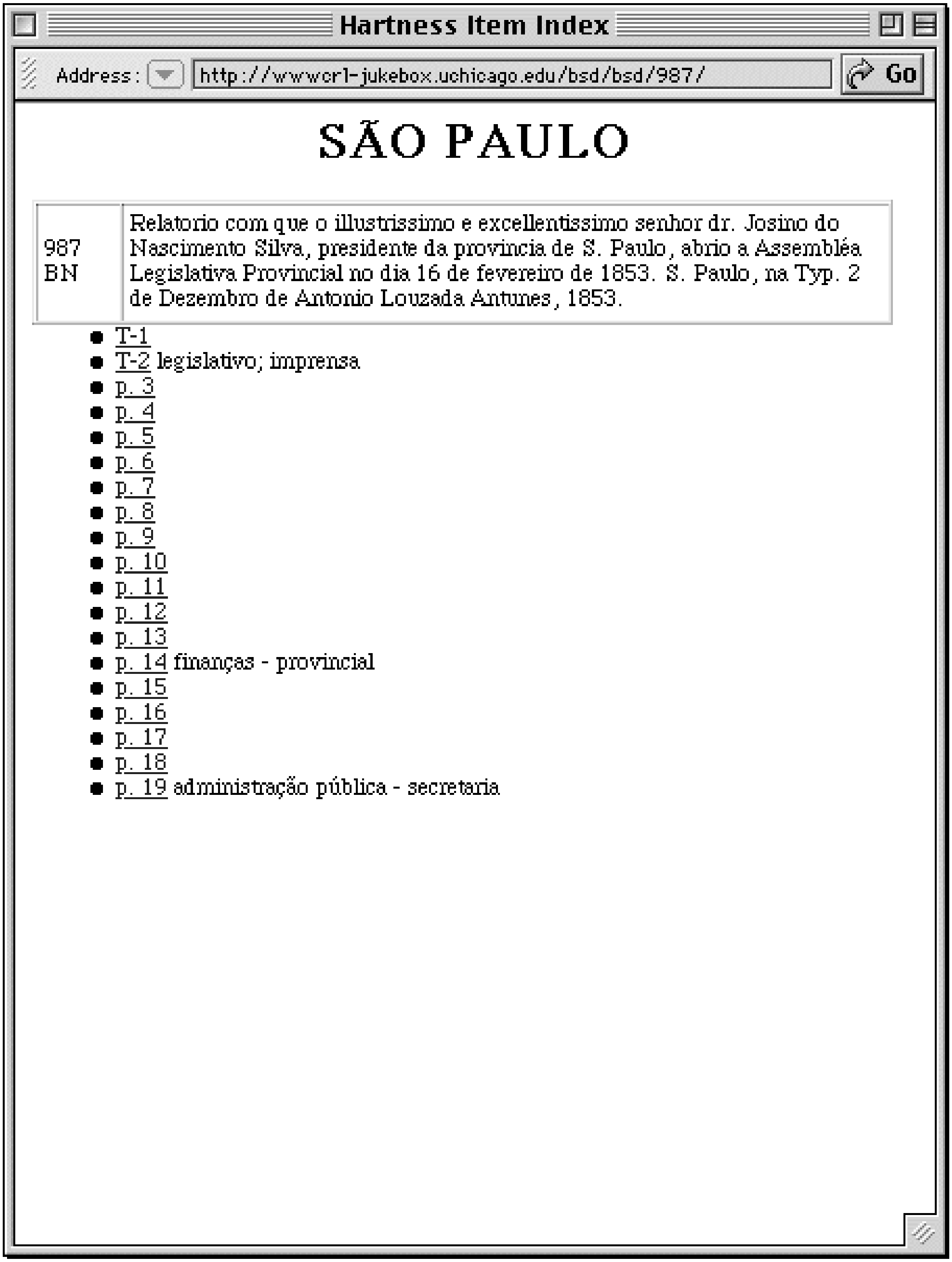
The 320 chapter headings identified in our sample were thus distilled into a controlled vocabulary of 193 subject headings, including the ninety-two employed in the Hartness *Guide*. Since the Hartness *Guide* describes only statistical information, the other 101 headings refer to other topics and activities. For example, the Hartness *Guide* includes no direct references to the Imperial family or legislation, and only indirect references to wars and social movements. An online thesaurus (wwwcrl.uchicago.edu/lamp/thesaurus.html) directs users to the subject terms employed for our indexing.

Once our controlled vocabulary was in place indexers at ten dollars an hour were hired to review all the Provincial Presidential Reports and to add the appropriate subject terms to the pagination files. All these indexers, Brazilian graduate students at the University of Notre Dame or Indiana University, South Bend, were comfortable with Portuguese and familiar with Brazilian history.

The indexing began before a relational database for indexing and searching had been put into place. Initially, each indexer had to identify a chapter heading in a report, find and copy the appropriate subject heading from a spreadsheet program, and then paste the term into the report's HTML coding via Notepad, a basic word processor. This cut-and-paste approach proved time consuming, requiring an average of 20:00 minutes per report. By contrast, the relational database method reduced this time to an average of 13:48 minutes per report for 228 reports. This enormous improvement primarily reflected the ease of selecting subject headings from a database menu.

Example F provides a sample of an indexed report. Page T-2 includes two subjects: "legislativo" and "imprensa." The semicolon indicates that the indexed page includes more than one subject. Page 14 and 19 have subjects with qualifiers: finanças – provincial" and "adminstração pública – secretarias." The hyphens indicate that the first term is qualified by the second. The appearance of a subject term indicates that the following pages fall under that same subject until another subject appears. "Legislativo; imprensa" thus runs from pages 2 to 14, at which point "finanças – provincial" denotes the beginning of a new range of five pages.

Example F - Subject Terms[http://wwwcrl-jukebox.uchicago.edu/bsd/bsd/987/]



SÃO PAULO

987 BN	Relatorio com que o illustrissimo e excellentissimo senhor dr. Josino do Nascimento Silva, presidente da provincia de S. Paulo, abrio a Assembléa Legislativa Provincial no dia 16 de fevereiro de 1853. S. Paulo, na Typ. 2 de Dezembro de Antonio Louzada Antunes, 1853.

- T-1
- T-2 legislativo; imprensa
- p. 3
- p. 4
- p. 5
- p. 6
- p. 7
- p. 8
- p. 9
- p. 10
- p. 11
- p. 12
- p. 13
- p. 14 finanças - provincial
- p. 15
- p. 16
- p. 17
- p. 18
- p. 19 administração pública - secretaria

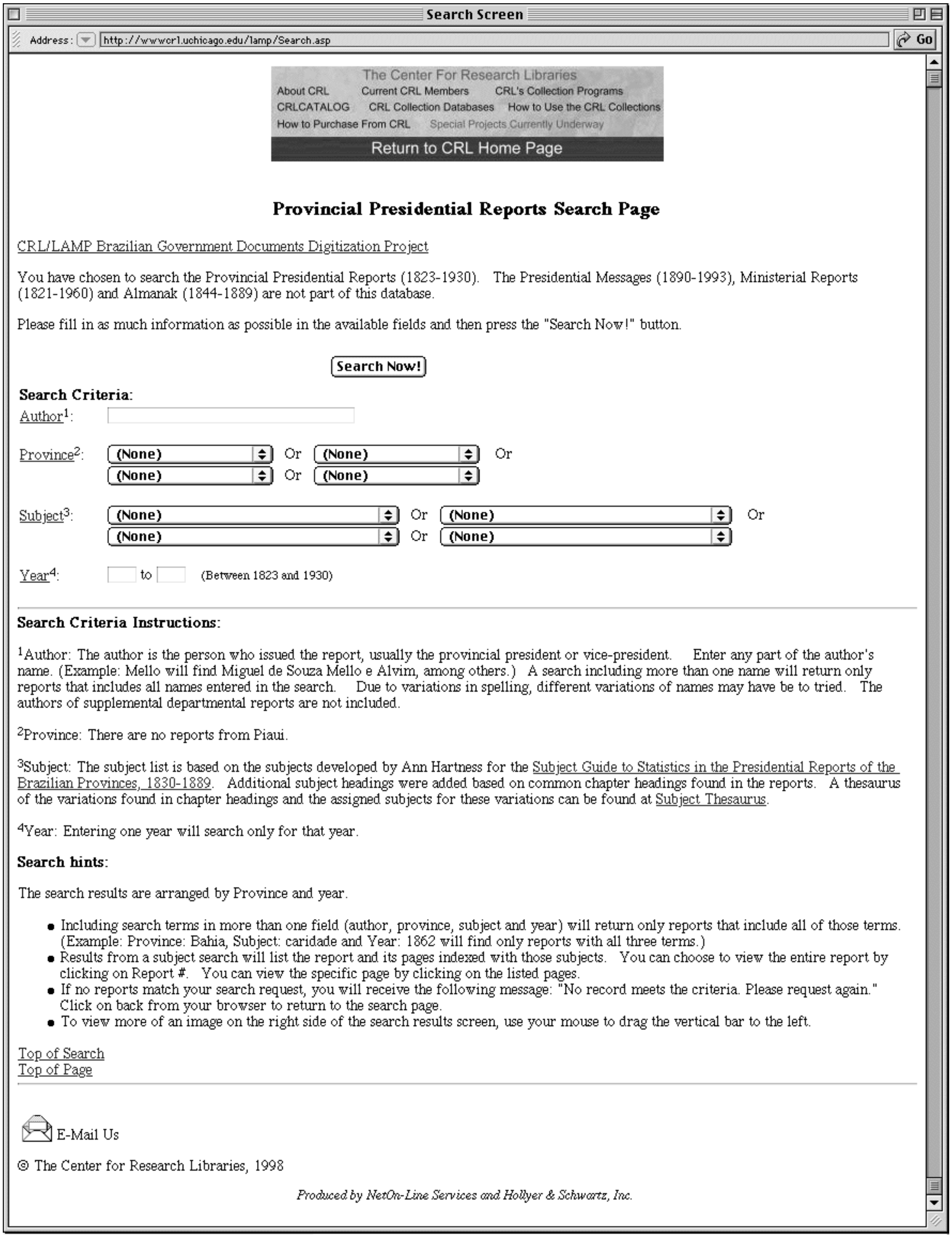
4.2.6 Search Interface Design

A special interface was designed to ensure access to the controlled vocabulary indexing. An information technology company, Hollyer and Schwartz (acquired by XOR in 1999, (http://www.xor.com), was hired to create this SQL (Structured Query Language) relational database. Users can thereby conduct Boolean searches by author, year, province, and subject, or by any combination of these fields.

Additional programming allowed us to migrate about 1,000 of the 2,572 previously indexed pagination files into the SQL database, and also to add subject terms to the remaining 1,500 files. Microsoft Access was used for the data fields (author, province, subject, and year) that were captured through an interface that imported the pagination file; allowed subjects to be entered, changed and/or deleted; and then sent the data to the SQL relational database. The “Province” and “year” values were extracted directly from the pagination file, while indexers entered data for the “Author” and “Subject” fields.

Subject searches may be carried out by entering terms alone or in any combination, as seen in Example G.

Example G - Search Page [http://wwwcrl.uchicago.edu/lamp/Search.asp]


Search Screen
Address: http://wwwcrl.uchicago.edu/lamp/Search.asp
Go
The Center For Research Libraries
About CRL
Current CRL Members
CRL's Collection Programs
CRLCATALOG
CRL Collection Databases
How to Use the CRL Collections
How to Purchase From CRL
Special Projects Currently Underway
Return to CRL Home Page

Provincial Presidential Reports Search Page

CRL/LAMP Brazilian Government Documents Digitization Project

You have chosen to search the Provincial Presidential Reports (1823-1930). The Presidential Messages (1890-1993), Ministerial Reports (1821-1960) and Almanak (1844-1889) are not part of this database.

Please fill in as much information as possible in the available fields and then press the "Search Now!" button.

Search Criteria:

Author[1]:

Province[2]: (None) Or (None) Or
(None) Or (None)

Subject[3]: (None) Or (None) Or
(None) Or (None)

Year[4]: to (Between 1823 and 1930)

Search Criteria Instructions:

[1]Author: The author is the person who issued the report, usually the provincial president or vice-president. Enter any part of the author's name. (Example: Mello will find Miguel de Souza Mello e Alvim, among others.) A search including more than one name will return only reports that includes all names entered in the search. Due to variations in spelling, different variations of names may have be to tried. The authors of supplemental departmental reports are not included.

[2]Province: There are no reports from Piaui.

[3]Subject: The subject list is based on the subjects developed by Ann Hartness for the Subject Guide to Statistics in the Presidential Reports of the Brazilian Provinces, 1830-1889. Additional subject headings were added based on common chapter headings found in the reports. A thesaurus of the variations found in chapter headings and the assigned subjects for these variations can be found at Subject Thesaurus.

[4]Year: Entering one year will search only for that year.

Search hints:

The search results are arranged by Province and year.

- Including search terms in more than one field (author, province, subject and year) will return only reports that include all of those terms. (Example: Province: Bahia, Subject: caridade and Year: 1862 will find only reports with all three terms.)
- Results from a subject search will list the report and its pages indexed with those subjects. You can choose to view the entire report by clicking on Report #. You can view the specific page by clicking on the listed pages.
- If no reports match your search request, you will receive the following message: "No record meets the criteria. Please request again." Click on back from your browser to return to the search page.
- To view more of an image on the right side of the search results screen, use your mouse to drag the vertical bar to the left.

Top of Search
Top of Page

E-Mail Us

© The Center for Research Libraries, 1998

Produced by NetOn-Line Services and Hollyer & Schwartz, Inc.

4.2.7 Image-Mapping for the *Almanak Laemmert* Report

Advances in technology over the Project's five-year life span provided new and exciting ways to utilize the extensive indices found in the *Almanak*. As with the previous Project collections, PFA created pagination files for each *Almanak* volume. But the *Almanak* also contained detailed subject indices of a sort not found in any of the other Project sets. These were ideal finding aids for scholars seeking specific information, and provided access to the three sections that comprised each *Almanak* volume: the *Almanak* per se (covering the Imperial Court), Province (Rio de Janeiro), and Supplement.

Fortunately, there were few structural variations within the *Almanak*'s subject indices over its fifty-year life. In 1846, two years after its inception, the *Almanak* included separate indices for the Imperial and Supplement sections, only. By 1848, a single subject index covered all three sections.

4.2.8 Paginating the *Almanak Laemmert* Subject Indexes

We employed two approaches to these subject indices. The first identified the subject index pages within the *Almanak* pagination files by labeling the index pages "*Indice Alphabetico,*" as shown below. Since single volumes sometimes included many index pages, the first three letters of the first and last terms on each index page were added to the pagination file entry:

INDICE ALPHABTICO ABR-AGE p. 229
INDICE ALPHABETICO AGE-ARA p. 230
INDICE ALPHABETICO ARA-ARM p. 231
INDICE ALPHABETICO ARM-AUS p. 232

As with all the pagination files, the "Indice Alphabetico" listings were linked directly to the corresponding page images.

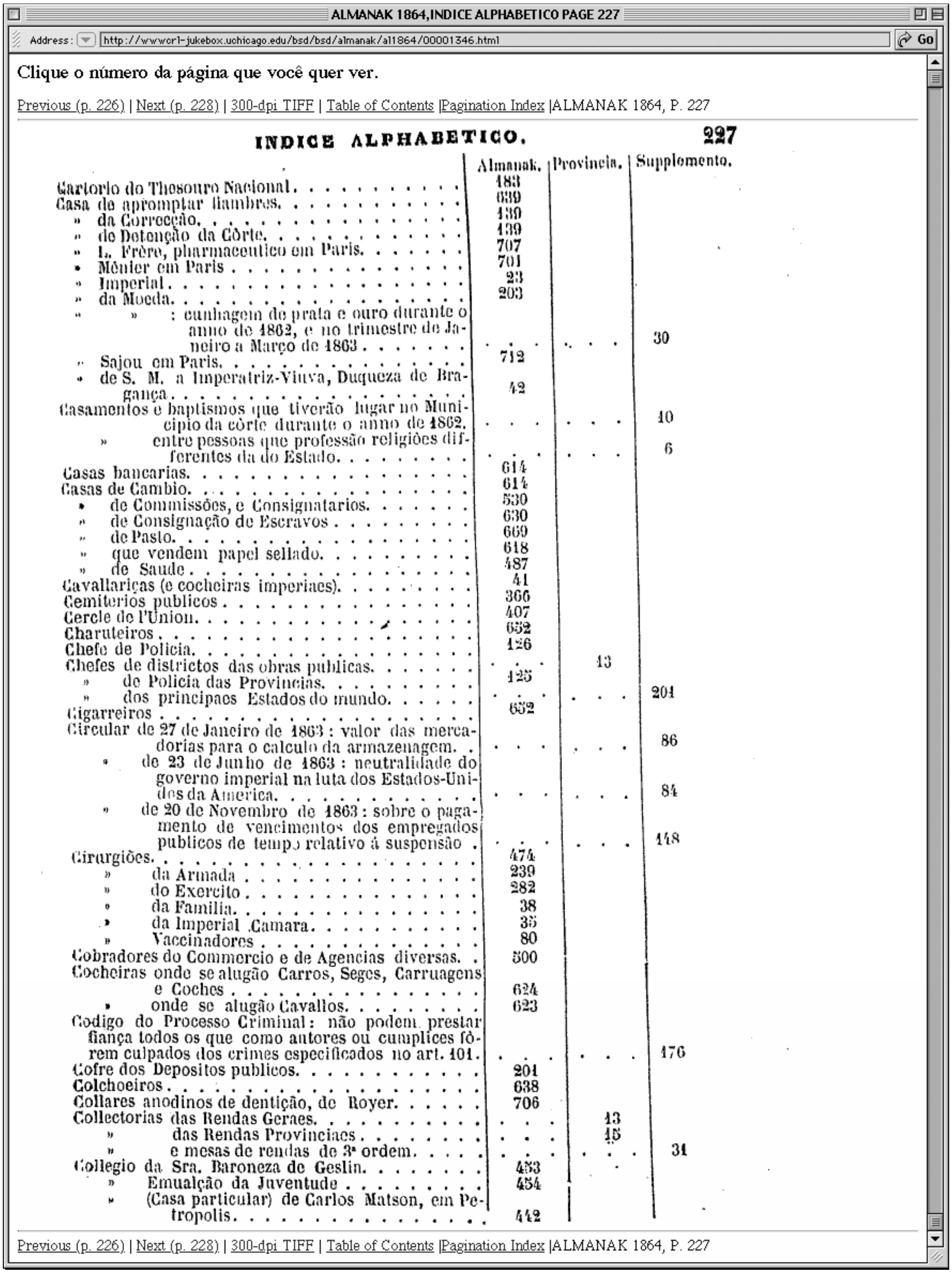
4.2.9 Character-Mapping the Subject Index to the Page-Images

Our second approach, by contrast, linked the page citations within each "Indice Alphabetico" to the corresponding page-images. Live Image software [http:www.liveimage.com/] was employed to link each citation to the appropriate page-image within the report. Users can therefore click on the page number of the citation within the "*Indice Alphabetico*" and go directly to the page. (The program builds on the spatial coordinates associated with some element of a page-formatted file, which then serve as the reference point for external links.) This approach was systematized by PFA and sent offshore to be completed.

In the following Example H: *Indice Alphabetico*, the three columns headed "Almanak," "Provinçia," and "Supplemento" provide access to the three different sections. (The left-hand column corresponds to the heading "article." The basic units within the report are numbered consecutively and referred to as articles.)

Example H - *Indice Alphabetico*

[http://wwwcrl-jukebox.uchicago.edu/bsd/bsd/almanak/al1864/00001346.html]


ALMANAK 1864,INDICE ALPHABETICO PAGE 227
Address:
http://wwwcrl-jukebox.uchicago.edu/bsd/bsd/almanak/al1864/00001346.html
Go

Clique o número da página que você quer ver.

Previous (p. 226) | Next (p. 228) | 300-dpi TIFF | Table of Contents |Pagination Index |ALMANAK 1864, P. 227

INDICE ALPHABETICO. 227

	Almanak.	Provincia.	Supplemento.
Cartorio do Thesouro Nacional	183		
Casa de apromptar fiambres.	639		
» da Correcção.	139		
» de Detenção da Côrte.	139		
» L. Frère, pharmaceutico em Paris.	707		
» Ménier em Paris	701		
» Imperial	23		
» da Moeda.	203		
» » : cunhagem de prata e ouro durante o anno de 1862, e no trimestre de Janeiro a Março de 1863	. . .	. . .	30
» Sajou em Paris.	712		
» de S. M. a Imperatriz-Viuva, Duqueza de Bragança	42		
Casamentos e baptismos que tiverão lugar no Municipio da côrte durante o anno de 1862.	. . .	. . .	10
» entre pessoas que professão religiões differentes da do Estado	. . .	. . .	6
Casas bancarias.	614		
Casas de Cambio.	614		
» de Commissões, e Consignatarios.	530		
» de Consignação de Escravos	630		
» de Pasto.	669		
» que vendem papel sellado.	618		
» de Saude	487		
Cavallariças (e cocheiras imperiaes).	41		
Cemiterios publicos	366		
Cercle de l'Union.	407		
Charuteiros	652		
Chefe de Policia.	126		
Chefes de districtos das obras publicas.	. . .	13	
» de Policia das Provincias.	125		
» dos principaes Estados do mundo.	. . .	. . .	201
Cigarreiros	652		
Circular de 27 de Janeiro de 1863 : valor das mercadorias para o calculo da armazenagem.	. . .	. . .	86
» de 23 de Junho de 1863 : neutralidade do governo imperial na luta dos Estados-Unidos da America.	. . .	. . .	84
» de 20 de Novembro de 1863 : sobre o pagamento de vencimentos dos empregados publicos de tempo relativo á suspensão	. . .	. . .	148
Cirurgiões.	474		
» da Armada	239		
» do Exercito	282		
» da Familia.	38		
» da Imperial Camara.	35		
» Vaccinadores	80		
Cobradores do Commercio e de Agencias diversas.	500		
Cocheiras onde se alugão Carros, Seges, Carruagens e Coches	624		
» onde se alugão Cavallos.	623		
Codigo do Processo Criminal: não podem prestar fiança todos os que como autores ou cumplices fôrem culpados dos crimes especificados no art. 101.	. . .	. . .	176
Cofre dos Depositos publicos.	201		
Colchoeiros	638		
Collares anodinos de dentição, de Royer.	706		
Collectorias das Rendas Geraes.	. . .	13	
» das Rendas Provinciaes	. . .	15	
» e mesas de rendas de 3ª ordem.	. . .	. . .	31
Collegio da Sra. Baroneza de Geslin.	453		
» Emulação da Juventude	454		
» (Casa particular) de Carlos Matson, em Petropolis.	442		

Previous (p. 226) | Next (p. 228) | 300-dpi TIFF | Table of Contents |Pagination Index |ALMANAK 1864, P. 227

4.3 Discussion of the Indexing Approaches

The various indexing applications employed in this project allow scholars to access the reports in a number of ways. At this point, there has been no testing to see whether any approach works better than the others. Indexing decisions have been made on the basis of technical limitations, advice from the Project Committee, and suggestions from librarians and scholars familiar with the source materials. Utilizing the finding aids available in the documents themselves – pagination, table of content, subject indexes - saved considerable time and money since they reflected a completed intellectual process. Project resources were instead deployed to resolve technical issues such as how to format the Hartness *Guide* in a usable form, and then link the citations to the proper locations within the documents.

This section discusses the advantages and disadvantages of each indexing approach.

4.3.1 Indexing: Pagination Files

The pagination file allows users to see a screen document whose structure mimics the paper original. Each page is numbered consecutively, and HTML code links the pagination file to the corresponding images. Users also can move page-by-page through the documents, reading the narrative as it was originally organized.

The *Almanak* set includes links between page-image and both the subject index and the pagination file. Adding these features to each page in the other three sets would allow those images, too, to be linked to software applications and to supplementary information.

While creating the pagination file was straightforward for most of the Project materials, including the Presidential Reports narratives, it proved problematic for the Presidential Provincial Reports and their supplements. These were often numbered separately from the base report, and sometimes they were not numbered at all. In the paper format, they were bound together with the President’s report. Our digital format required us to devise a new numbering system. A standard pagination structure to represent complex combinations of reports and sub-reports would be very useful.

4.3.2 Table of Contents

Tables of contents provide access to these reports by identifying chapters and other major elements within the works. These guides are also familiar to users. Most were easy to rekey and link to the appropriate pages, and costs were usually low. In this case, character mapping was not used with the table of contents since the process was not employed until they were already created using HTML.

Relying on tables of contents for access to extensive documents will often identify only the most general information, forcing users to review many pages to locate specific data. Enhanced subject access, for instance by character mapping for existing subject indices or through the creation of new indices, offers greater precision.

4.3.3 *Subject Guide to Statistics in the Presidential Reports of the Brazilian Provinces, 1830-1889*

Locating and electronically recreating guides and indices to digitized materials can significantly enhance access. The Hartness *Guide* to the Provincial Presidential Reports was a case in point. The *Guide's* tabular format was easily recreated using HTML coding. Many of the citations were linked to specific page images with special software, though some links could only be created manually.

The statistics indexed in the Hartness *Guide* are selective, and omit the report years 1890-1930. Indexing all of the statistics from these reports, however, would have been both costly and time-consuming. Some citations from the *Guide* could not be matched to our image files, since some reports were never filmed and some page-images were illegible. Additional time and money would, of course, allow ever more comprehensive results.

4.3.4 Indexing with a Controlled Vocabulary for the Provincial Presidential Reports

Creating a useful thesaurus of controlled subject terms requires considerable time and an understanding of users' needs and their approaches to the texts. Applying this sort of thesaurus to the Provincial Presidential Reports was particularly valuable since these documents rarely contained tables of contents or extensive internal indexes. A controlled subject term was applied to every chapter in each report and also to each supplement.

Researchers can utilize the indexing in two ways. First, they can view the pagination file for any report to see the subject headings. (See Example F.)

They can also begin from the headings themselves. (See Example G.) Search results will show every report, and page, to which the search term(s) have been applied.

Subject indexing was a time-consuming and costly process that was affected by the availability of qualified labor. Not only was there a limited pool of Portuguese speakers in South Bend, Indiana, where the Project Coordinator resided, but the indexers were university students who could only commit limited amounts of time to the Project. Indexing projects such as this must consider the skills required and the availability of potential employees.

The creation of a searchable index for the Provincial Presidential Reports will benefit users undertaking comparative investigations across geography (nineteen states) and time (108 years). The ability to combine 193 subject terms with provinces, presidents, and years offers a powerful research tool.

Creating this indexing database was by no means cheap. The approximate costs included $18,500 for programming, $1,300 for the SQL Server software, and $20,000 for indexing services. Not all institutions will have both the supervisory and technical staff to create large relational databases. Outsourcing the work would add an additional layer of expense.

4.3.5 Character Mapping the *Almanak Laemmert*

The character mapping process utilized for the *Almanak* affords an attractive alternative to rekeying extensive indexes. The *Almanak* was produced with detailed subject indexing, in sharp contrast to its very general tables of contents. Moreover, character mapping retains the format of the original document. The process may also produce fewer errors than creating HTML files for subject indexes, since there are fewer steps and keystrokes. Finally, PFA, charged the same fee, $2.70, to rekey a table of contents or to map the same page.

On the other hand, character mapping limits the index to information that is available in the original document itself. The approach would not have worked with, say, the Provincial Presidential Reports, which contain so few tables of contents and virtually no subject indices. Many other documents that are good candidates for image scanning may also lack the internal indexing that lends itself to character mapping. Finally, users who are not familiar with character mapping may initially be confused since the hypertext links are typically not highlighted: the browser's "hand icon," indicating an active link, only appears when the cursor rolls over a page number from the index.

4.4 Alternative Indexing Strategies

One more indexing alternative merits some discussion. While OCR software cannot provide accurate text files from the collections discussed here, it might capture enough of the text for searching purposes. Page-images could be presented for viewing, while searchable texts accommodated user queries.

We considered this approach for the Provincial Presidential Reports, but did not pursue it because of the variety of terms used to describe a particular topic, for instance, the previously noted fourteen variations for police. Software that would launch a text-file search for all variants of a term would solve this problem, but we were unaware of this kind of product.

When we devised this Project, in 1994, several markup languages were under development. Standard Generalized Mark-up Language (SGML) and Extensible Mark-up Language (XML), and such applications as "Encoded Archival Description (EAD), combined HTML's linking capabilities with additional tagging to improve recall and precision.[11] These approaches, however, were judged too new to employ.

5. DATABASE DELIVERY AND FILE MANAGEMENT

Systems support at the Center for Research Libraries has played a very important role in the storage and delivery of the Project files.

5.1 Website

The Project website [http://wwwcrl.uchicago.edu/info/brazil/] was created in 1996 to direct users to the database and inform them about its peculiarities. The site was initially hosted on a server at Notre Dame so that the Project Coordinator could update it. During this phase, users were seamlessly redirected from CRL's website to the server at Notre Dame.

Portions of the site were translated into Portuguese in 1998, and in 1999 the site was moved permanently to CRL. The Project Coordinator has since edited the site via FTP.

The website provides access to the page-images by allowing users to select a document set and then navigate by links to a specific desired page-image. To view a page within a Presidential Provincial Report, for instance, a researcher chooses the province of interest and the year, and then selects the report title before clicking on the desired page from the pagination index. The searchable subject database for the Provincial Presidential Reports also allows users to search by the president's name, province, selected subjects, and year.

During its first few years, low screen resolutions (at that time ranging between 640x480 and 800x600) hindered viewing. Both required horizontal scrolling to see the complete pages, and the site instructed users to set their screen resolutions at a minimum of 800x600. Screen resolutions of 1024x768 had become commonplace by the late 1990s, negating the need for horizontal scrolling.

One particularly important page on the website lists options for finding a TIFF viewer, since neither Netscape Navigator nor Microsoft Explorer can display TIFF images. Fortunately, both browsers simplify access to TIFF viewer applications through their internal configuration settings. Some viewers, like Docuview, are free of charge. Others, for example Vueprint, can be purchased for a reasonable price. Many academic institutions will have site licenses for TIFF viewers. At some point, Netscape and Explorer may themselves support TIFF images, so that the Project's low-resolution GIF files will no longer be needed.

5.2 Image Storage and Access

CRL was very interested in using the Project to explore the issues of providing onsite maintenance and Internet access for a very large image database. This responsibility has required significant staff and financial resources. CRL's commitment has been repeatedly tested by hardware

[11] "SGML is a set of rules for defining and expressing the logical structure of documents thereby enabling software products to control the searching, retrieval, and structured display of those documents. The rules are applied in the form of markup (tags) that can be embedded in an electronic document to identify and establish relationships among structural parts." Development of the Encoded Archival Description Document Type Definition. [http://www.loc.gov/ead/eadback.html] (10/18/01)

and software failures, difficulties in loading the image and indexing files into the storage device, and a turnover of system managers.

5.2.1 Choosing a Storage Medium: WORM vs. RAID

When this Project began, in 1995, CRL examined a number of mass storage devices before narrowing the field to a WORM jukebox or a RAID unit.

WORM refers to a “Write Once, Read Many” magnetic-optical storage medium. WORM devices are referred to as “jukeboxes” because their internal mechanisms include an armature which accesses many disks. The WORM stores data on a number of disks that are physically placed in separate columns. The drive arms move from column to column, then disk to disk, in order to access the data. The storage system is slower than other systems and also less durable.

RAID, “Redundant Array of Independent Disks,” stores the same data on multiple hard disks (therefore, redundantly). Reading performance is improved since different disks can be read at the same time.[12]

The WORM device, utilizing magnetic optical disks of 1.3GB each, was eventually purchased for the Project because it cost substantially less than RAID devices and also reflected an established product for large database projects.

5.2.2 Moving the Page-Images from DAT to WORM

PFA supplied page-images on Digital Audio-Tapes (DAT). The data then had to be extracted and recorded onto the WORM device. Loading page-images into the server was a labor intensive process due to differences in file-names and formats between the two systems.

The image files were first extracted from the tapes and run through simple programs, or “scripts” to convert three-letter filename extensions to four letters: “HTM” to “HTML” and “TIF” to “TIFF”. This proved necessary because the original scanning and file-naming was performed on DOS systems using the “8.3” convention for file names.

Some corrupted data also meant that the jukebox storage capacity was lower than planned. “Write Once” technology permanently alters the medium so that “bad” data render the corresponding disk space unusable. At one point, more than 50% of the space was thus unavailable on six of the twelve optical disks. The lost space did not affect the WORM drive’s performance, but CRL’s systems administrator recommended their replacement. This entailed purchasing and installing the new disks, and then transferring the data from the disks being replaced.

[12] Http://searchstorage.techtarget.com/. (10/18/01).

5.2.3 Downtime Leads to WORM Upgrade

Unfortunately, the original jukebox was afflicted by recurring hardware problems. The first two years, for instance, saw three drive-arm failures. The original repair personnel hired to maintain the storage device were likewise problematic, and a different company was eventually hired. Further, disk failures required data to be reloaded from the vendor's original DAT tapes onto new WORM disks, a relatively slow process insofar as individual files had to be identified within a linear medium. Files had to be relocated in order to be reloaded, and the tapes had to be fed through a relatively slow reading mechanism. The sequence was not unlike searching for a particular song in a group of audiocassette tapes.

By 1998, as mechanical problems continued, the jukebox manufacturer offered CRL a favorable price on a new and more efficient model that used a single column of disks. Reduced movement in the armatures was expected to minimize the misalignment problems that caused the initial failures.

5.2.4 Upgrading to Dual Servers to Balance the Load

The server that provided user access to the jukebox storage device also failed from time to time. This was sometimes due to a "hung" process, such as when the tape drive did not respond to commands for data access when the jukebox was taken off-line for repairs. The server had to be re initialized or "rebooted" before service could be resumed.

In 1998, CRL moved the Project website from its UNIX server to a Windows NT platform. This required updates to a number of the hyperlinked pages related to the database and its hierarchy. CRL had two main reasons for this change. First, this measure separated two major systems so that a problem with one would not affect the other. Second, the shift made it easier to manage the site. The NT server in particular allowed CRL's staff to modify webpages directly from their workstations, thereby reducing the burden on the system administrator.

While the UNIX server continued to serve page-images from the jukebox, the NT system hosted the website itself. This migration of tasks went smoothly, and dividing the load between the servers enabled more efficient operations as well as greater access for the staff.

Also in 1998, as large amounts of data arrived from PFA, an additional external hard drive was purchased in order to hold these data until they could be transferred to the jukebox.

The tremendous growth of the UNIX file system, which contained the index to the page-images, became an additional concern. A file system which approaches 90% capacity is considered overloaded, and ours had reached 82%. The solution was to "soft-link" the two largest index files, with nearly 180MB of information, to another file-system on the server.

5.2.5 Upgrading of Storage Devices To Improve Reliability

The issue of dead space was resolved by installing a new jukebox in October, 1999. By this time, some of the 1.3GB disks were entirely unusable. By contrast, the second jukebox had only four percent of the of unusable space on each single disk. This new jukebox could hold thirty-two optical disks of 5.2GB each, or more than 160GB altogether. It also had two hard disk drives that can be used to read and write data to the disks. As with the first jukebox, there were mechanical failures, albeit fewer of them.

The external hard drive did crash in July, 1999, mandating a delicate procedure to extract the massive 180 MB index of database files from the old hard drive and rewrite it to the replacement. Since then, the system has only been rebooted once.

Maintaining and upgrading our servers and storage devices has been expensive and has also required knowledgeable staff who can respond quickly to emergencies. Several system administrators supported the Project during their respective tenures, and each one provided exemplary support. Funding will remain an issue since it will still be necessary to update the server and storage devices and also to migrate the data. Housing this kind of database requires an ongoing, day-to-day commitment by systems staff, as well as continuing administrative oversight and support.

5.2.6 Migrating the Data from the Optical Jukebox to Hard Drives

When the new jukebox proved unreliable and slow, the data were moved again - this time to an external SCSI hard drive.

The jukebox, a mechanical device with many moving parts, often broke down. A common problem was the misalignment of the optical disks when inserted into the disk drives. The process looked like this: a user would request data that was on a disk; and a lift mechanism would retrieve the disk and move it to a disk drive much like a CD ROM drive. However, the lift mechanism used a carefully precisioned shaft and pulley system that was prone to misalignment. When this happened the drive would take itself offline as a defense mechanism against future requests. When the misalignments affected both drives all the data became unavailable to web users.

Another problem concerned the elapsed time between the initiation of a user request and on-screen delivery of the document. This was due to the process of retrieving one optical disk at a time, leading to a queue of requests.

These drawbacks were overcome when CRL installed an external hard drive on the Brazilian Documents web server in October, 2001. The data were restored from tape to the new hard drive in a process that took just over a week to complete. Since the restoration, the hard drive has performed perfectly and users have noticed a significant improvement in speed. An additional, identical drive was also bought as a backup. The drive will be installed and the data from the original drive will be copied to it. This backup will be pulled off the live web server and kept in safe storage as “hot,” pluggable backup.

6. EVALUATING THE PROJECT

The Brazil Digitizing Project was one of the first of its kind. We therefore sought to gather as much information as possible about the database's creation and use. Earlier chapters of this report have described how the database was created. This chapter focuses on evaluation, including the Project's stated goal to "Explore relative levels of demand and patterns of use for the digitized materials byissuing them both as CD-ROMs and as files available over the Internet."[13]

This chapter discusses four different aspects of evaluation. The first entails the Project Committee's effort to monitor vendor performance in creating the image files. Three other activities gathered user feedback in order to improve the database.

6.1 Image and Indexing Evaluation

It was essential to evaluate the image database in an ongoing manner in order to identify and resolve production problems as they occurred. Every member of the Project Committee helped to evaluate the database in order to ensure that image quality, database construction, and indexing were all correct.

PFA's developmental work was monitored with particular care, especially during the initial design of the database structure, and the selection of image file formats and attributes. These crucial decisions defined the quality and usability of the entire effort. PFA routinely provided beta versions for review following extensive conversations with the Project Committee. These versions were then improved as needed and put into production.

PFA began to deliver the first digital tapes containing scanned page-images and their accompanying database structure in September, 1995. The Project Committee then devised an ongoing evaluation schedule to review page-images and also the HTML files that made them accessible.

An initial preview of the page-images served to establish a range of image quality and to test the database design. A system to rank image quality was then prepared to ensure consistency among the evaluators. The Project Coordinator assigned images and image links to each evaluator. The large number of images, and the contractual need to report problems to PFA within thirty days, meant that evaluations were based on samples drawn from the document files.

In April, 1998, the Project Committee agreed that image quality and the database structure were in general terms acceptable. Sampling additional materials was therefore delegated to graduate students. The students, like the Project Committee members before them, reported image and database problems to the Project Coordinator, who in turn notified PFA. Once a set of files had been accepted, PFA issued the corresponding invoice to CRL for payment.

Each image was assigned one of four rankings:

[13] The decision to forego the production of CD-ROMs and rely exclusively on Internet access meant that database usage and nature could be monitored by the frequency of Internet "hits" as well as responses to an online user survey. Evaluating use of CDs, beyond sales tallies, would have been very difficult.

Excellent -The image is clearly legible and easily readable, with no internal variation in quality.
Good - The image is legible and readable, though its internal quality may vary.
Fair - The image is legible, but may not be readable due to significant noise, i.e. bleedthrough, density problems, etc.
Poor - The image or significant portions of the image are not readable and are largely illegible. Contact PFA for possible rescanning.

Many of the images rated "poor" could not be improved due to substandard images on the original microfilm. It was deemed too expensive and time-consuming to seek paper originals, so these images were accepted in their illegible form. A notice on the Project's Home Page alerts users to these cases.

Example K, below, transcribes one evaluator's report, along with the Project Coordinator's comments. The Project Coordinator inspected problematic images with a TIFF viewer, so his annotations reflect images with higher resolutions than those examined by the evaluators. Most of the TIFF page-images were legible, even when their GIF counterparts were not.

Example K: Sample Evaluation

File #	Year	Evaluator and Project Coordinator's Comments about Image Quality
u1843	1851	Evaluator: Some minor skewing.0046-0047 illegible as GIFs Project Coordinator: 0046-0047 LEGIBLE AS TIF
u1844	1852	Evaluator: Bleed through 0004. Some minor skewing.0023-0024 illegible as GIFs. Project Coordinator: 0023-0024 LEGIBLE AS TIF
u1845	1853	Evaluator: Skew 0060, 0062. All illegible as GIFs0069-0075, 0082-0083, 0091, 0094-0111, 0116-0122 Project Coordinator: 69-70 ILLEGIBLE AS TIF, ALL OTHERS LEGIBLE AS TIF
u1846	1854	Evaluator: Small images include0035-0040, 0300-0301,0305,0307 (illegible as GIFs) Project Coordinator: 301 AND 307 LEGIBLE AS TIF, ALL OTHERS LEGIBLE AS TIF

Example L, below, transcribes the Project Coordinator's summary report on the Provincial Presidential Reports for the state of Rio Grande do Sul. About 10,267 of 100,644 images were sampled, or ten percent. This summary also assesses links between the Hartness *Guide* and the page-images.

Example L: Subject: Rio Grande do Sul Evaluation

Report on Province: Rio Grande do Sul
Total number of images evaluated:10,267 of 100,644

1) Items incomplete or missing: Files 872-908: Not scannable.
2) Links to Hartness Guide: Fine
3) Image quality: Wide variation; many pages included different density levels.
4) Scanning quality: Fine
5) Category Breakdown:
   a. Excellent: 0
   b. Good: 22
   c. Fair: 26
   d. Poor: 2
   /914/000010 image broken up and low density
   /921/000028 “ “

6.2 User Feedback

Specialist librarians and Brazilianist scholars gather regularly at conferences. We therefore arranged several database demonstrations in order to publicize the Project and elicit feedback. Users at the University of Notre Dame, home base for the Project Coordinator, also worked with the database.

The database was demonstrated before audiences from the following core groups of Latin American Studies organizations:

1) Seminar on the Acquisition of Latin American Library Materials (SALALM)
   -- audience of about 40 Latin American Studies bibliographers;
2) Latin American Studies Association (LASA) 1997 and 1998 meetings
   -- audience of about 60 university teachers and researchers;
3) Brazilian Studies Association (BRASA)
   -- audience of about 20 university teachers and researchers, some from Brazil.

Hands-on demonstrations and beta-testing were carried out with two scholars and three graduate students, together representing the fields of anthropology, history, sociology, and political science, at the University of Notre Dame.

Phase Two’s indexing efforts were greatly improved by the thoughtful responses of scholars and librarians who were familiar with the Project’s first phase. Both the controlled vocabulary index to the Provincial Presidential Reports and the image mapping for the *Almanak Lammert’s* detailed subject index reflected these concerns.

6.3 Website Usage

We have used WebTrends software to track website use since late in 1998. Fourteen activity reports have followed, the first from December, 1998 and the most recent from October, 2000. Each report tallies a month's worth of database activity. Twenty-eight months have elapsed between the WebTrends installation and the end of 2000. Jukebox crashes and other problems account for the modest tally of results, which includes one report from 1998, eight from 1999, and five from 2000.

All four database sets were online and available when WebTrends was installed in 1998. Indexing of the Provincial Presidential Reports began shortly before WebTrends was installed and continued through 2000. This ongoing indexing inflated our usage statistics. On the other hand, jukebox downtime intermittently halted all use of the database.

The following five tables provide information on database use, most frequently requested pages, visitor statistics, regional visitors, and downloaded file types. The WebTrends reports include much more information, but the data in these tables are particularly useful in understanding the level and nature of use. (Appendix 10.4 provides the complete table of contents for the October, 2000 WebTrends report.)

Table 3: Database Use [a]

Categories (Number of reports used)	High Month	Low Month	Mean	Median
Homepage Hits (14 reports)	257 Jun-99	62 Dec-99	115	103
Successful Hits for Entire Site (14 reports)	126,498 Jun-99	10,280 Dec-99	55,340	42,398
Ave. Successful Hits per Day (14 reports)	8,426 Jun-99	331 Dec-99	1,882	1,235
Page Views - Impressions (14 reports)	59,854 Sep-99	439 Dec-98	36,042	43,669

a. Homepage Hits -- Number of times the home page was visited.
Successful Hits for Entire Site -- A count of all successful hits (a successful hit occurs when the database responds to a requests) including HTML pages, pictures, forms, scripts, and downloaded files.
Average Hits per Day -- Number of Successful Hits divided by the total number of days in the log.
Page Views (Impressions): Total -- A count of the number of pages viewed, not including the supporting graphic files within the page.

The above data are difficult to assess without a baseline or comparative data. It's clear that the database is being used on a regular basis, sometimes quite heavily. Ongoing use is perhaps best represented by the median data. The “successful hits for the entire site”; including all files and images; and “page views and impressions,” representing only the image file, reveal thousands of hits each month.

The homepage is a consistent entry point, but by no means the only one. Since any image or finding aide can be easily bookmarked, experienced users may go directly to specific pages of particular interest.

Table 4: Most Requested Pages (October, 2000)[a]

Pages -- Each URL began with: http://wwwcrl-jukebox.uchicago.edu/	Views	% of Total Views	Visitor Sessions	Avg. Time Viewed
1. /robots.txt	2,302	10.44%	379	00:00:25
2. /bsd/bsd/hartness/minopen.html	166	0.75%	145	00:01:22
3. /bsd/bsd/hartness/crl.html	155	0.7%	110	00:01:16
4. /bsd/bsd/hartness/relacoes.html	101	0.45%	91	00:01:36
5. /bsd/bsd/hartness/prestoc.htm	96	0.43%	82	00:02:05
6. http://wwwcrl-jukebox.uchicago.edu/	128	0.58%	58	00:01:35
7. /bsd/bsd/hartness/imperio.html	47	0.21%	38	00:00:35
8. /bsd/bsd/almanak/al1856/	62	0.28%	31	00:00:53
8. /bsd/bsd/almanak/al1844/	68	0.3%	31	00:00:25
9. http://wwwcrl-jukebox.uchicago.edu/bsd/	47	0.21%	26	00:01:04
10. /bsd/bsd/	50	0.22%	26	00:01:39
11. /bsd/bsd/hartness/fazend.html	28	0.12%	25	00:00:33

[a] This section identifies the most popular pages on the site. The number of views includes only the successful hits for the page itself. The percentage of total views is the percentage of hits for that page compared to all other page types.

Table 4 tallies the most visited sections of the site during October, 2000. Column one shows that three of the four collections were represented among the month’s most requested pages. Several specific subjects from the Hartness *Guide* were requested repeatedly. Entry number 5, “/bsd/bsd/hartness/prestoc.htm”, is the access page for the Presidential Messages. The *Almanak*, the most recent addition, shows up twice in this report. The most requested page, “/robots.txt”, was a script to limit access for software programs that check links for a particular user. “Robots” are very common, particularly for such major search engines as Yahoo, Excite, etc. Entry number 10, “/bsd/bsd/”, is the jukebox location for all the reports. The Project Coordinator accessed this location on a regular basis to transfer files for indexing.

Table 5: Visitor Statistics[a]				
	High Month	Low Month	Mean	Median
Visitors (12 reports)	2,347 May-99	732 Mar-99	917	1,293
Ave. Visitor Session (hrs.:mins.:secs.) (13 reports)	1:13:28 Jun-99	30:23 Oct-00	49:56	NA
Visitors Who Visited More Than Once (6 reports)	168 Oct-00	121 Dec-99	132	141
International Visitor Sessions (5 reps.) – On average, about 30% of all visitors.				

a. Average Visitor Session Length -Average of length of visitor sessions in the log.
Visitors - Unique visitors are counted using their IP address, domain name, or cookie.
Visitors Who Visited More Than Once - The count of visitor sessions that appeared more than once in the log file. By default a visitor session is 30 minutes.
International Visitor Sessions determined by the User Domain field in the log.

The most interesting piece of information in Table 5 is the number of visitors who visited the database more than once, i.e. a mean of 132 a month. While these statistics were affected by the project indexers who used the database on a daily basis, there were never more than five indexers accessing the database during any reporting month. The average visitor session length, almost fifty minutes, is striking. Users seem to be reading these reports, or viewing the statistical data for extended periods of time. The final statistic, concerning international visitors, must be framed within the context that 30-40 percent of all visitors cannot be identified. Moreover, international visitors using Internet Service Providers that are international as well (Hotmail.com, etc.) cannot be identified. International visitors account for about half of the 70 percent of visitors that we can track. This figure corroborates the data in Table 6, which reports a high number of Latin American visitors.

Table 6: Regional Visitors				
Region	High Month	Low Month	Mean	Median
North America	535 Jun-00	249 Apr-00	346	417
South America	413 Jun-00	245 Oct-00	283	341
Ratio of North American visitors to South American visitors – 1.22 to 1				

South Americans are some of our most frequent visitors. The Internet Protocol addresses further indicate that most of these South American visitors come from Brazil. Visitors from areas outside of North America and Brazil account for about five percent of total use.

Table 7: Downloaded File Types (13 Months Counted)				
File Types	High Month	Low Month	Mean	Median
GIF	31,810 Dec-99	596 Feb-99	9302	4422
TIFF	3,474 Sep-99	118 Feb-99	1732	1773
Ratio of Total GIFs to TIFFs Downloaded – 5.4 to 1				

This table identifies the file types that were accessed by users. These statistics are particularly revealing in documenting use of low-resolution GIFs. The TIFF files would normally be opened only when the GIF is not legible. The ratio of opened GIFs to opened TIFFs, 5.4 to 1, indicates that TIFFs were viewed about 18.5 percent of the time. Overall, 8 out of 10 pages were viewed without opening the TIFF. Further analysis is needed to confirm whether the gray-scale GIFs produced for the *Almanak* reduce usage of the TIFFs.

6.4 Online User Survey

We designed an online survey in 1997 to elicit feedback from our users. While the survey does not ask why users have visited the database, their responses do suggest their reactions to what they've found. (The User Survey and selected results can be found in Appendix 10.6.)

Most responses are very positive, and we have received only a few negative comments. As of this writing, fifty-two users have responded to the survey, forty-six of whom are based in Brazil. Forty-seven respondents found the database to be very useful. Most users located the materials that they were attempting to find. Twenty-nine found *The Hartness Guide* easy to use. Most scholars using the database were historians, 31 of the 47 recording a profession. The sole negative comment, to question nine, arrived when the database was unavailable due to a jukebox problem. This complaint reflected the user's frustration at being unable to access the database.

The online survey defaulted to the most favorable response to each question rather than a neutral selection or "no response." This may have biased the results. Moreover, the survey was inadvertently made available only on the Portuguese-language homepage, which would account for the preponderance of responses from Brazil.

Nonetheless, the favorable responses suggest that the database will find a useful place among scholars working on 19th and 20th century Brazil.

7. OTHER NOTABLE PROJECT ELEMENTS

This section addresses elements of the project not covered in previous chapters.

7.1 Outsourcing

The Project Committee determined from the first to rely on resources outside of CRL for specific portions of the work. CRL had outsourced its microfilming for many years, and was comfortable sending Project work outside as well. Outsourcing was utilized throughout the project when neither CRL nor the Project Committee could mobilize the staff and/or expertise needed to complete some task. The Project relied on outside firms to scan and index all of the Brazilian microfilm documents, and also to create a relational database for the Provincial Presidential Reports. PFA, in turn, subcontracted portions of the hypertext work and indexing to a company in the Philippines.

Three vendors were initially invited to bid on the Project's scanning and indexing. (See the RFP in Appendix 10.7.) One firm went out of business during the initial discussions. The other two vendors had extensive experience with microfilm, but had never scanned microfilm into a digital format on a contractual basis. Both vendors were attracted to the Project because they anticipated an emerging market for scanning from microfilm. Samples of the source microfilm materials were sent to each vendor, and they both presented their results to the Project Committee in 1995.

The Project Committee selected PFA, Inc., to scan and index the documents. One telling factor was PFA's creation of a preliminary website containing samples of the scanned images, which allowed the committee to visualize the interface and anticipate functionality. PFA and CRL signed a contract in July 1995.

Several features in the PFA contract will apply to other scanning projects, as well. The costs for scanning, post-processing, and indexing were clearly defined. In our case, scanning costs were further divided into two levels dependent upon the qualitative difficulties of handling with each image. Regardless of whether paper or microfilm is the source material, costs must be based on individual segments, i.e. frames, images, and pages. Our scanning was costed out on a per frame basis, though many frames included two images. Post-processing charges were calculated on a per image basis.

In the Spring of 1996, after thousands of images had been scanned, indexed, and delivered, PFA approached the Project Committee to renegotiate the contract. The initial sample had not represented the full scope of problems, which PFA was unable to address within its original budget. After considerable discussion, the post-processing fee was increased from $0.0403 to $0.05 per image. Provisions for preliminary vendor review of the entire source material, careful initial analysis, and flexibility in the face of unexpected problems have all proved essential.

7.2 Publicizing and Providing Access to the Database

7.2.1 Cataloging the Database

One of the project's original goals was to "Implement mechanisms to ensure traditional bibliographic access to each digitized serial." CRL has cataloged all of the digital serials sets. Users of CRL's online catalog can thus link directly to the database. Further, these newly created records will be added to the OCLC (Online Computer Library Center) database, so that any OCLC member can download them and add them to its local catalog. Users can then find these records in their local catalogs and locally link directly to the database.

7.3 Project Phasing

7.3.1 Cost Analysis of Phase I

The Project Committee established very clear goals from the start. A May, 1996 project review indicated that we would meet the initial goals without spending out the grant. In fact, almost forty percent of the original $225,000 was still on hand.

This underspending reflected four main elements. First, the Project Committee initially took a very conservative approach to indexing the images, due to concerns that the funds would run out before the completion of Phase One. As Don Simpson noted, "We will build a Ford, not a Cadillac."

Second, the original project proposal estimated that the document sets included one million pages. In reality, these materials totalled fewer than 600,000 pages:

Provincial Reports	216,187 images
Presidential Reports	18,103 images
Ministerial Reports	329,159 images
Total	673,449 images

The third and fourth factors contributing to lower-than-expected costs centered on scanning fees and mechanisms to deliver content. Scanning costs were reduced because PFA was able to digitize virtually all of the microfilm images at the Level One cost of $.195 per image, rather than the Level Two cost of $.225 per image. The Project Committee also opted to cancel the planned creation of CD-ROMs due to the then high by 1996 costs of "burning" CDs.

In sum, the Project Committee faced the unexpected task of seeking new ways to utilize the balance of the grant.

Table 8: Table of Costs for Phase I

Project Goal	Projected	Actual	Balance
CD-ROMs	$10,000	0	$10,000
Scanning	$107,000	$66,110	$40,890
Indexing/Post Processing	$55,000	$36,961	$18,039
Labor	$20,000	$20,000	0
Miscellaneous	$10,000	$10,186	-$186
Interest Earned on Grant	--	--	$13,576
Total: Grant $225,000	$202,000	$133,257	$82,319

7.3.2 Phase II Planning

The experience gained in constructing a usable database of more than 500,000 images was crucial as we planned for a second phase. The Project Committee, working closely with PFA, considered two main alternatives. The first, would employ established procedures to scan and index additional microfilm document sets. This sequence would bypass the labor intensive start-up efforts that affected both PFA and the Project Committee during Phase I. The second alternative would explore new approaches to recognized problems, assuming acceptable start-up and maintenance costs. The Project Committee and PFA were eager to improve the database, even if this required us to develop new processes. A proposal combining both approaches was eventually submitted to the Mellon Foundation, which needed to approve the changes in workplan and budget. The new proposal was approved.

The following activities were considered for Phase II:

1) Hiring additional labor for the Project -- Part-time employees were hired to take over various ongoing tasks from the Project Committee. These duties included evaluating the images and links prepared by PFA and cleaning up various portions of the database. Employees with Portuguese language experience were hired when possible.

2) Scanning and indexing additional materials -- The *Almanak Laemmert* -1844-1889 (*Almanak Administrativo, Mercantil e Industrial do Rio de Janeiro*), held in microfilm at CRL, was seen as an ideal candidate for scanning and indexing. The *Almanak* was published annually between 1844 and 1889 and included legislation, census data, and commercial advertising, as well as listings of Court and Ministry officials. The entire set included about 57,000 pages.

Each annual report also included a detailed index that the project reworked into a searchable online index. PFA carried out the *Almanak* scanning and indexing.

3) Indexing the Provincial Presidential Reports -- The Project Committee decided to index more than 2,500 reports, as described in Chapter Four.

4) Preparing a monograph to publicize Project results -- Phase I of this project created a usable database of Brazilian historical materials that is now available to scholars throughout the world. Its findings complement those of earlier studies on digital reformatting of microfilm source materials, for instance Paul Conway's Project Open Book at Yale University.[14] The proposed publication would assess the conversion process itself, vendor relations and provide a detailed economic analysis.

Several scholars suggested that Brazilian census reports would comprise a useful addition to the project database, given their importance and heavy use. The Project Committee was already familiar with the challenges of producing legible digital images of numerical tables. Nonetheless, we wanted to revisit this problem and consider any new approaches.

In the end, census materials were not included in Phase II. PFA attempted to scan microfilm samples of the 1872, 1920, and 1970 censuses, and then create quantitative data that could be opened or imported directly into a spreadsheet program. Small fonts made much of this information very difficult to read even from the microfilm, and most of the table format was lost. Substantial clean-up would have been necessary for each image.

The Almanak Laemmert was completely scanned and indexed at the following costs:

Table 9: Almanak Laemmert Costs (Phase II)

Function	Cost / Unit	Total (Est.)	Final Costs
Inspection of 78 reels	$7.50 / reel	$575	$575
Scanning* - level 1 for 56,789 frames	$.195 / frame	$4,195	$6,992
Rescanning	$1.35 / frame		$38
Image split & indexing for 113,578 images	$.0863 / split image	$3,625	$6,014
Table of Contents - 1,575 estimated pages	$2.70 / TOC page	$4,252	$0
Image-mapping**	$7.85 / hour	$21,625	$7,792
TOTAL		$34,262	$21,411

*Estimate based on 420 frames per roll of film. **Presumes an estimated 1,100 mapped index pages.

[14] Paul Conway. *Conversion of microfilm to digital imagery: a demonstration project*. Performance report on the production conversion phase of Project Open Book.

7.4 Project Delays

The Project took much longer to complete than expected – five years rather than the projected two or three. In retrospect, we can identify several significant reasons for the delays.

Manual Indexing - Indexing the Provincial Presidential Reports took longer than expected, mostly because it was difficult to find indexers who were both qualified and available to work. The average time to index a file was about seventeen minutes, so that the 2,500 reports could have been completed in about 23.5 weeks of thirty hours per week. This work actually required more than two years.

The project's indexers, all students, were only available to work for three to ten hours per week. Outsourcing was not considered, given our conviction that indexing had to be closely monitored and controlled by the Project Committee.

Project Coordinator's Availability - The Project Coordinator averaged between one and four hours a day on this assignment, five days a week, while maintaining a full-time library position. The work moved ahead, but progress would have been faster had the coordinator arranged for a three-quarters time position.

A leave of several months would also have freed up time for the project, but in our experience there were unexpected obstacles and delays during the entire process. Any Project Coordinator must make sure that there is adequate time to complete tasks on schedule. Such judgments are best based on previous experience, which in our case did not exist.

Failures with the Mass Storage Device – Some months were lost due to problems with the software and hardware purchased to store the images. The WORM jukebox suffered numerous technical problems requiring repeated visits from technicians. The systems administrator at CRL spent an inordinate amount of time addressing these problems, which were only solved by purchasing a single hard-drive to hold all the data.

Delivery of Scanned Images and Indexes - Scanning the microfilm was at first quite slow due to the uneven quality of the filmed images, the unpredictable rotation of images from comic to cine mode and back again, and unevenness in the borders. PFA found that manual intervention was needed more often than predicted, adding several weeks to the time allotted for scanning in Phase I.

Project Phasing – The Project Committee devised the Project's second phase when it became clear that we had a surplus. The new initiatives took time to plan and implement. The activities associated with Phase II largely account for the difference between the original three-year project term and the five years actually required for completion.

8. CONCLUSIONS

This section of the report recapitulates the lessons that we've learned during the Project. Reading these recommendations without first reading the body of the report may be confusing, as many of these comments are based upon details described in Chapters 1-7.

8.1 Scanning from Microfilm in Order to Provide Internet Access

One of the Project's primary objectives was to explore scanning from microfilm for the purpose of Internet access. Yale University's Project Open Book, an earlier microfilm scanning project, documented the process for 2,000 books. The Brazil Project created images in a similar fashion, but then made them accessible on the Web, exploiting a technology that was not available during Project Open Book. Projects based on scanning from microfilm remain uncommon. Many purchased microfilm collections are under copyright. Moreover, many libraries prefer to focus on scanning original documents. Some microfilm collections nonetheless merit digital reformatting, particularly when demand is high and widely dispersed.

The digital images produced in this Project are two generations removed from the original paper documents. While image legibility was a concern, ninety-eight percent of the 100 dpi GIFs are in fact readable. The Project has demonstrated a time- and cost-effective means of migrating a collection to an electronic medium while providing both preservation via microfilm and digital access via scanning. Fortunately, many of these reports are still available in their paper format, and users can examine the original texts if and when the digital version do not provide the needed information.

8.2 Scanning from Microfilm: Image Formats for Internet Access

One of the Project decisions was to select an image format that would be viewable in the 1994 versions of popular Web browsers. Mosaic was already available and Netscape, Microsoft Explorer and America Online's browser all appeared shortly thereafter. Mosaic supported two common image file formats that appeared to be emerging industry standards: JPEG and GIF. The GIF format was chosen because it provided better image quality, particularly for the Project's black and white images, in relatively compact files that minimized transfer times.

The TIFF format was selected to provide high quality master images. These also served as backups when the GIFs were illegible.

As of December 2000, both GIF and JPEG continue to be supported by all major programs and utilities. TIFF remains the standard for digital images created from analog media like microfilm.

8.3 Scanning from Microfilm: Image Resolution

The choice of image formats was followed by decisions concerning the resolution that would best represent information contained in the source microfilm. Given the need for small images that would quickly transport through what was then a rather slow Internet, bitonal 100-dpi GIF images were created for screen display. Higher quality, bitonal 300-dpi TIFF files provided masters. The Project moved from bitonal images to much-improved gray-scale GIFs in Phase II. By this time, too, Internet connections had become far less problematic.

An emerging "best practices" calls for archival quality images to be scanned at 600-dpi. While Project images would certainly be enhanced if they were all available at 600-dpi, most are legible without such high resolution. Those that could be better include manuscript pages, documents with small typeface, and crowded tables of numeric data that are difficult to read even in the 300-dpi TIFF format. Cost, scanning procedures, file sizes, and transfer times all argued against this approach, however.

Paul Conway in *Conversion of microfilm to digital imagery: a demonstration project* recommends at least 400-dpi as "essential for preservation quality digital conversion of text." We found that scanning from microfilm at 100-dpi provided reasonably good images that transmitted quickly and were generally legible over the Internet. Again, this Project has focused on access, not preservation.

8.4 Scanning From Paper Versus Scanning from Film

Scanning from paper originals will as a rule produce better images than scanning from microfilm surrogates. For this project, the microfilm sets had been prepared from widely dispersed originals: scanning from film was the only real option.

Nonetheless, other projects will need to explore both possibilities. Costs, in these cases, will be one consideration. PFA's Jim Harper notes:

> "Where film formatting and image quality are reasonably good, the cost to scan from microfilm should be much lower than scanning from paper (particularly when the paper is difficult to handle). Labor is the primary issue here. Less labor should be needed for film scanning than for paper scanning.
>
> 1) If there is existing microfilm and it is of reasonably good quality, film scanning should be a cost effective solution. This will be particularly true if the source material is difficult to handle.
> 2) If the material needs to be filmed for preservation purposes or because it is brittle, microfilm scanning should be cost effective if reasonable film quality and formatting can be produced.
>
> If the source material has a legible image, it will certainly produce a legible digital image. If the film that exists is of poor quality, then scanning the paper is the best solution."

8.5 Indexing

Providing subject access to image-based collections of text-rich documents poses many challenges. Until OCR technology improves or rekeying costs drop significantly, most image-based collections will have to rely on manual indexing schemes. Such traditional indexing guides as tables of contents offer effective access to textual materials. Enhanced indexing will work as well, provided that index terms are linked directly to the appropriate page-images.

Indexing is slow and expensive. Finding indexers with appropriate skills – in our case Portuguese – can be difficult. Indexing our project documents took more than two years due to a limited pool of indexers who were available for limited amounts of time. Improved technology may ultimately allow word recognition and synonym-matching software to substitute for indexer expertise. Rapid developments suggest that this time may be drawing near.

8.6 Knowledge of the Collection

One perhaps obvious finding is that one must know the content and intellectual organization of the collection to be reformatted. Even then, some assumptions may turn out to be wrong. For example, initially we expected to scan about 1,000,000 page-images, but Phase I only produced half that amount. We had anticipated two page-images for every microfilm frame, and we had a rough count of 500,000 frames. While our frame count was reasonably close, at 483,545, the actual ratio of images to frames was only 1.16 to 1.

We also failed to anticipate the complex pagination structure that was needed to accommodate the numerous supplements and annexes that we encountered. Likewise, many of the microfilm images could only be scanned with custom programming and manual intervention. In each case, greater initial awareness would have simplified the work.

8.7 Technological Change: Running to Stay in Place

Creating a digital collection during a period of great technological change almost guarantees that some elements of a five-year project will have become obsolete by its completion.

For instance, the World Wide Web has emerged as a very efficient mechanism to transfer visual data desktop PCs. HTML coding is ceding ground before its successor, XML. Sun's Java has become a powerful extension to Web browsers, as well as a virtual operating system in its own right. Stand-alone personal computers have been linked in Local Area Networks (LAN), and more broadly to wide area networks, intranets, and the World Wide Web itself. Desktop PC clients may soon also function as personal Web servers. And the changes continue: by 2005, a typical workstation might be able to manage the Brazil Project's entire 50GB database, currently housed at CRL on a 70GB hard drive, with nothing more than a multi-disc DVD drive.

The Brazil Project database was created using image file formats and a hypertext language that remain standards today. The GIF format continues to be viewable on the Web, and this will not change in the near future. The TIFF format, while still not viewable with current Web browsers, is accessible by means of plug-ins. HTML remains commonplace, even though it lacks the power of new programming languages and codes. Our database continues to meet user needs.

The delays associated with Phase II allowed the Project to exploit new knowledge and software, and thereby improve both the quality of the images and the indexing interface. The first three sets of documents were made available as bitonal (black and white) files. The *Almanak* images in Phase II were prepared using gray-scale, significantly improving their legibility. Phase II also allowed us to devise a relational database to provide subject access to the Provincial Presidential Reports.   Finally, we used image-mapping software to link the *Almanak* indexes to the respective page-images.

Turning to hardware, our original choice of a WORM jukebox for mass-storage delivery now appears mistaken. Technological improvements have resulted in much better alternatives. The transfer of the data to a single hard drive in October, 2001 has improved both delivery and performance.

9. CONCLUSION

Some digitizing projects have created archival quality images for preservation. The Brazil Project, by contrast, has focused instead on providing access to page-images. Text images delivered at 100-dpi, instead of the emerging preservation standard of 600-dpi, are easily transferred and generally legible. Just as important, carefully constructed finding aids, indexes, and guides allow users to find specific reports and information. These collections are now freely available throughout the world to anyone with Internet access.

The Project's initial choice of TIFF and GIF image formats, in 1994, was based on emerging industry standards. These formats remain dominant in 2000. HTML, which was used to construct the webpages, likewise persists – though XML seems likely to take its place. The Access^{TM} relational database and the Microsoft Sequel Server software used to create the index and search interface to the Provincial Presidential Reports, are both fairly standard.

Other challenges required homegrown solutions. The Scindex software employed by PFA was created especially for this project in order to automate processing for the scanned images. Live Image software allowed us to link the extensive *Almanak* subject indexes to the appropriate pages within each report. These new approaches may help others who face similar data processing and indexing challenges. If nothing else, the Brazil Project has shown that indexing need not be a one-way street: numerous access points can be created by utilizing standard Internet technology combined with traditional indexing techniques.

Indexing for close to 700,000 images lies at the core of this project. Users can quickly and efficiently find what they seek. The digital environment has allowed creative approaches to indexing that include traditional finding aides constructed to take advantage of hypertext technology (page numeration, tables of contents, bibliographies, and subject indices) and also relational databases designed to link subjects terms to particular report pages. What we have learned should add to the body of literature on creating and indexing text-rich images.

Our Project was one of the first to demonstrate that historical text collections can be successfully migrated to the Internet. The Brazil Project has received very positive responses from its users, especially those unable to consult the original documents or their microfilm surrogates. Brazilians in the farthest reaches of their immense country are now able to view these documents. So can scholars in other parts of the world. This digital collection will continue to serve users ranging from students seeking to understand the fabric of Brazilian culture to researchers in fields as diverse as epidemiology, economics, and criminology.

10. APPENDICES

Appendix 10.1 -- Project Chronology

Proposals -- 1994-1995

May 1994:	Proposal submitted to The Andrew W. Mellon Foundation.
June 1994:	Grant awarded for $225,000.
October 1994:	Microfilm sent to PFA for review.
July 1995:	Contract signed by PFA Inc. and The Center for Research Libraries.

Phase I -- 1995-1997

August 1995:	Test images mounted at CRL.
September 1995:	Ongoing evaluation of image files begins.
September 1995:	Draft "Home Page" for the Project mounted for comment.
January 1996:	Hartness Guide available on the Homepage.
March 1996:	Four Provinces (Bahia, Alagoas, Amazonas, & Espirito Santo) available for use.
April 1996:	Pricing structure of contract with PFA is amended.

Original Contract		Amended Contract
Level 1 Scanning	$0.195 per frame	$0.195 per frame
Level 2 Scanning	$0.225 per frame	$0.225 per frame
Post Processing	$0.403 per image	$.50 per image
Pre-Scanning	No cost	$1.35 per roll
Table of Contents	$2.70 per page	$2.70 per page

October 1996:	All scannable Provinces are available for use, with the exception of a handful of images that need further processing.
October 1996:	Presidential Messages (1890-1993) available for use.
January 1997:	Online user survey implemented.
March 1997:	Project Homepage completed.
January 1998:	Expended $120,000 of $225,000 project funds.

Phase II -- 1998-2000

December 1997: Phase II Goals

1. Hire students to evaluate images, $5,000 to $10,000.
2. Scan Almanak (57,000 pages), $25,000 to $30,000.
3. Scan Census information, $15,000 to $25,000.
4. Hire students to index the Provincial Reports, $10,000 to $25,000.
5. Produce a monograph study of the project, $5,000 to $7,000.

March 1998: Mellon Foundation approves Phase II.
March 1998: Alamanak images mounted.
May 1998: Provincial Reports indexing begins.
Aug.-Oct. 1998: Jukebox down at CRL. Hardware replaced.
October 1998: Hollyer and Schwartz hired to create search interface and data maintenance interface for Provincial Records.
February 1999: Search Interface and Data Maintenance Interface completed.
June 1999: Pages linking all Provincial Presidential Reports, arranged by Provinces, made available on CRL’s website.
June 1999: Hartness Guide links updated; subject links available from Project website.
March 2000: Indexing of Provincial Presidential Reports completed.
December 2000: Database cleanup completed.
October 2001: Jukebox storage system replaced by single harddrive.
December 2001: Final report completed.

Appendix 10.2 -- RFP to Mellon Foundation

Creating a Digital Core Collection of Brazilian Serial Documents:

A Proposal to
The Andrew W. Mellon Foundation
by
The Center for Research Libraries
on behalf of the
Latin American Microform Project

May 5, 1994

Project Summary

The Latin American Microform Project at this time proposes to digitize a coherent body of high-use Brazilian documents that are both substantial and important for Latin Americanist scholarship. This project will address several specific aspects of electronic access, bibliographic description, and indexing. It will assess user demand for different formats of electronic information. It will increase current understanding of and capabilities to create digital image files of textual documents. As it explores these issues and produces image files, the project will benefit both Latin Americanists and the research library community. The project budget is $222,000.

Contact Persons

Dan C. Hazen
Selector for Latin America, Spain, & Portugal
Harvard College Library
197 Widener Library
Cambridge, MA 02138
Tel: 617-405-1749 or 617-495-2425
Fax: 617-495-0403
Internet: hazen@widener1.mhs.harvard.edu

Donald B. Simpson
President
The Center for Research Libraries
6050 South Kenwood Avenue
Chicago, IL 60637-2804
Tel: 312-955-4545x335
Fax: 312-955-4339
Internet: simpson@crlmail.uchicago.edu

The Center for Research Libraries
6050 South Kenwood Avenue
Chicago, IL 60637-2804
Creating a Digital Core Collection of Brazilian Serial Documents:
A Proposal to The Andrew W. Mellon Foundation by
The Center for Research Libraries on behalf of the
Latin American Microform Project

The Latin American Microform Project (LAMP) through The Center for Research Libraries proposes to digitize a core set of executive branch serial documents issued by Brazil's national and provincial governments during the period between independence and 1990. This endeavor will complement other efforts to exploit electronic technologies for Latin Americanist scholarship within the Hemisphere-wide initiative sponsored by The Andrew W. Mellon Foundation.

The project will accomplish the following:

- Facilitate scholarly access to a central and coherent body of high-profile research resources for Brazilian studies.
- Expand the still-tiny corpus of digital image files aimed at a scholarly audience.
- Implement mechanisms to ensure traditional bibliographic access to each digitized serial.
- Provide structured access to individual volumes within each serial set.
- Provide, as necessary, electronic indexing to the sections within these documents, single issues of which can be hundreds and even thousands of pages long.
- Explore relative levels of demand and patterns of use for the digitized materials by issuing them both as CD-ROMs and as files available over the Internet.
- Refine the process of creating digital image files from preservation microfilm.

This proposal briefly describes the Latin American Microform Project and The Center for Research Libraries. It addresses at somewhat greater length the materials intended to be digitized, the processes expected to be employed, and some of the technical complexities sought to be addressed. A specific work plan and a (necessarily tentative) budget are also outlined.

1. The Latin American Microform Project and The Center for Research Libraries.

The Latin American Microform Project was formed in the 1960s in order "to acquire, preserve, and maintain for its subscribers microform collections of unique, scarce, rare, and/or bulky and voluminous research materials pertaining to Latin America." About thirty-five North American libraries with Latin American collections comprise the current membership. Most of LAMP's preservation work is funded by member dues.

LAMP is administered through the Chicago-based Center for Research Libraries. LAMP's policies and projects are determined by its entire membership, which meets once a year. An elected Executive Committee, currently chaired by Dan Hazen of Harvard University, coordinates activities between the annual meetings.

LAMP has heretofore focused on preservation via microfilm. The approach is proven, the medium durable, and filming capacity fairly widely distributed--at least in North America. Nonetheless, microfilm is clumsy, unpleasant to use, and only available through traditional (and slow) delivery mechanisms. LAMP now seeks to explore digital technology in order to promote easy access and use. The proposed project will create, describe, manage, and monitor demand for image files of some of LAMP's most important holdings. This will enable LAMP to refine its strategies for preservation, access, and distribution for the future.

The Center for Research Libraries, founded in 1949, is the nation's oldest cooperative, membership-based research library. The Center's mission is to make available to the scholarly community research materials that are rarely-held in North American libraries. In working toward this goal through a program of cooperative collection development, the Center acquires, preserves, provides bibliographic access to and delivers from its collections. These collections, comprised of more than 3.7 million volumes and 1.4 million units of microform, are housed in Chicago and include newspapers, dissertations, archival materials, scientific and technical serials and monographs, area studies microforms and special collections. The Center administers five area studies microform projects. Center membership consists of 130 university, college and research libraries throughout the United States and Canada.

2. The Materials.

The project will concentrate on two broad categories of executive branch serial documents from Brazil and will include all available materials issued between independence and 1990. Brazil was selected because it carries particular importance both within the region and for North American scholarship. Brazil is far and away Latin America's largest country, with tremendous geopolitical significance and immense potential as an emerging world power. At the same time, Brazil exhibits all the economic, social, political, ethnic, and ideological tensions that characterize Latin America as a whole. The nation is the object of a great deal of research by scholars both within and beyond its borders.

The documentary focus will in the first instance address materials produced by Brazil's national government. The annual messages of the president often include declarations of goals and purpose akin to those found in the United States Presidents' "State of the Union" messages. They also report on the government's activities and accomplishments during the year immediately preceding. They thus provide an indispensable overview of executive branch plans, priorities, and achievements.

Presidential messages offer a general picture of official policies and practice. Fuller contexts and more detailed analyses are provided by the annual reports of each ministry. The ministries are the operational organs of the executive branch. Even during periods of normal rule, and within constitutional structures separating executive from legislative powers, Latin American ministries enjoy substantial authority to promulgate rules and regulations. During periods of authoritarian rule, ministries have sometimes played significantly larger roles in making as well as implementing policy. Models of the state based on strong central government, which have recurrently predominated through Brazil's history, have often given the ministries broader and more pervasive powers than would be expected in a North American context. The annual reports of Brazil's ministries--the second and quite voluminous group of national documents to be digitized- -are thus essential for understanding the goals and role of official policy.

Presidential messages and ministerial reports reflect national plans and priorities. Also to be digitized are the annual reports of Brazil's provincial "presidents" for the period between independence and 1930. A centripetal perspective is important for several reasons. Brazil is a land of sharply differentiated regions. The modernizing Center-South contrasts with the less dynamic, drought-ridden Northeast, which in turn plays against the Amazon basin, and on. Moreover, Brazil is large--so large that many states are bigger than entire nations in the rest of the Continent. Regionalism, often expressed in protests against centralist policies and proclamations, has lain at the heart of much provincial politicking. Reports from provincial executives add counterpoint and balance to central visions of Brazil and its destiny. Finally, contemporary scholarship more and more focuses on regions and localities. The stories told by these provincial reports are increasingly important in their own right.

Original research on virtually any aspect of Brazil's national period requires the use of these materials. Their importance for scholarship in and of itself makes them good candidates for digitization. These publications also comprise a coherent whole. They reinforce one another, thereby magnifying the utility of individual volumes as well as the joint corpus. More practical considerations further enhance their appeal. All these official documents are in the public domain, so they can be digitized and distributed without concern for copyright. LAMP already owns microfilm of almost all the materials, so the project can focus on scanning from existing film with an only occasional need to create preservation microfilm as well.

The potential audience for these materials is substantial. The "Brazilian Studies" regional committee of the predominantly North American "Conference on Latin American History" (CLAH) lists 118 members. Historians not associated with this subcommittee, the many academics from other fields whose research centers on Brazil, and Brazilian scholars themselves, comprise a far larger group of potential users: a 1985 National Directory of Latin Americanists thus identified 635 researchers interested in Brazil. The "Brazilian Studies Association" (BRASA) was formed, with an initial membership of more than four hundred, during the 1994 meeting of the Latin American Studies Association. Easily accessible materials may further attract undergraduate and graduate student use. The demand for these materials is likely to be high.

Alternate sources of access, on the other hand, are quite limited. The original volumes are either widely scattered or altogether unavailable. The Library of Congress (LC) had to borrow single volumes from numerous institutions to assemble runs of the ministerial reports that it filmed for LAMP; LC subsequently discarded its own paper holdings. The fullest available set is on LAMP film, some of which is not yet cataloged on-line. The provincial materials were filmed within Brazil through a LAMP- financed project at the National Library. This effort had to rely on generally dispersed holdings within the states of origin. The original materials remain scattered in Brazil and are very sparsely represented in the United States. Rosa Quintero Mesa's Latin American Serial Documents: A Holdings List, volume 2 of which (1968) concerns Brazil, shows no holdings for these reports. Quick searches on RLIN for presidential reports from four states (Alagoas, Bahia, Ceara, and Minais Gerais) indicate holdings of but 12 hard copy volumes in the entire database. At Harvard, finally, pre-1890 holdings from all provinces are limited to about forty reports. (Ann Hartness's Subject Guide to Statistics in the Presidential Reports of the Brazilian Provinces, 1830-1889, by contrast, enumerates 1,085 of these documents.) For all their importance, these are scarce resources that have been severely under utilized. Digitizing LAMP's film will allow the assessment of the use of an electronic resource for which hard copy equivalents simply do not exist. It will also, for the first time, enable scholars to take fuller advantage of these extraordinary materials.

LAMP now holds about 635 microfilm reels of the serial documents that are proposed for digitizing. Perhaps thirty additional reels of film will be needed to fill gaps and complete runs. The 665+/- reels to digitize will represent about 650,000-700,000 frames of film. Some materials were filmed with one page per frame, others with two. About one million pages of text are estimated.

3. The Process.

Pilot projects with digital imaging suggest that it can be combined with preservation microfilm to ensure both longevity and fluid access. The preservationist community is in general agreement that, at the present time, one should not rely on digital files for long-term preservation. Since LAMP already possesses preservation microfilm for almost all the materials under consideration, the project can particularly focus on issues of electronic access without compromising LAMP's mandate for long-term preservation.

Pilot projects at Cornell University (among other institutions) have established that binary scanning at 600 dots per inch (dpi) results in preservation-quality images of most printed materials. Such high-resolution scanning may be more than what is required for typical scholarly uses of LAMP's holdings. Initially, the project will experiment with a resolution of 300 dpi, which has proven acceptable in other digitizing projects. Higher resolutions will be employed if and as the materials so require--for instance if very small typefaces or unusual fonts must be captured. The quality of LAMP's preservation microfilm (which in turn reflects the quality of the hard copy originals) will further affect both scanning resolutions and scanning costs.

Even 300 dpi images take up relatively large amounts of raw storage space. The files will thus be compressed and stored according to the Group 4 TIFF standard, the emerging norm for straight text materials. The proposed documents for digitization are overwhelmingly textual, with occasional charts and tables and relatively rare illustrations. The proposed scanning procedure should capture their content.

Producing digital image files of microfilmed materials is but the first element in the project. Internet tools such as Gopher and Mosaic will provide minimal access to files available over the network. However, these tools only operate within the Internet environment. In order to serve researchers seeking these materials via traditional bibliographic tools, the project will thus augment existing bibliographic records in the Center for Research Libraries' online catalog (CRL-CATALOG) with data for the Internet address ("Uniform Resource Locator" or URL and Uniform Resource Name or URN) of online files and with information on CD-ROM versions. CRL's catalog records are broadly available through the Internet, the major bibliographic utilities and also in a growing number of universities' local online catalogs. Scholars seeking Brazilian serial documents will thus be directed to electronic as well as microform versions.

The project will focus on serial documents, where a single bibliographic record may represent a great many volumes. Ministerial reports can be particularly voluminous: a single issue may fill several volumes, and the complete run of a title can extend to tens of thousands of pages. Users must be able to navigate through these very large files as they conduct their research. This will become possible through two layers of control. The project will provide both year and volume access for each title. Where appropriate, it will also index major sections within each report. Users will thus gain access to elements of each report, as defined by the structure of the documents themselves. The files produced will become effectively available to the Latin Americanist research community, rather than remaining an undigested mass of raw images.

The project will also explore the feasibility of more specialized access to parts of the state presidential messages. Ann Hartness's 500-page Subject Guide to Statistics in the Presidential Reports of the Brazilian Provinces, 1830-1889, published in 1977, provides detailed indexing to this subset of LAMP's holdings. Copyright to this publication is held by the Institute of Latin American Studies at the University of Texas, though the publication itself is no longer in distribution. The author, Ann Hartness, works in the Benson Latin American Collection at the University of Texas, and she anticipates no difficulty in securing copyright release. Once the copyright issue is settled, the project will digitize the Guide and create electronic links between the index and the original texts.

A central objective of the project is to evaluate different means of providing electronic files to the scholarly community. Online access via a client/server will ensure instantaneous access over the Internet. This approach assumes a certain level of user equipment and sophistication, and is also most cost-effective for materials enjoying fairly high use. CD-ROMs work well for files generating less intensive use and in situations where telecommunications links and Internet access are not assured. CDs will be particularly important in Latin America. The Brazilian market for these digitized materials should be significant, since complete runs even of the presidential messages are by no means commonplace. Researchers in other Latin American countries will use these materials, despite the language difference and a common scholarly tendency to focus on one's own nation. There are several reasons for cross-national interest. Recent regional developments, including Mercosul/Mercosur (a free trade zone encompassing Brazil and the Southern Cone) and Brazil's increasing presence throughout the Amazon Basin, are provoking interest and attention. Historical studies of the Continent more specifically require the sources to be digitized. Brazil's nineteenth-century expansion, for instance, particularly affected Paraguay, Bolivia, Uruguay, and Argentina. Brazilian documents include material essential in analyzing these events. Finally, the original microfilm will also remain available for loan.

The project initially intends to offer Internet access without charge to all comers. The CD-ROMs will be loaned free of charge to members of the Center for Research Libraries and LAMP and will be sold at cost to anyone at all. Requests for the LAMP microfilms have been tracked by the Center for Research Libraries, providing a baseline against which the project can assess changes in demand as well as patterns of use. Use of the existing microfilm, some of which has not yet been cataloged, has been light. (RLIN searching for presidential reports from four states revealed no catalog record at all for reports from Ceara, and a record created only in March, 1994 for Minais Gerais.) Eighty-three reels of microfilm have been requested by eight users from but five institutions during the first three months of 1994. One can expect substantially greater use once the materials are digitized, publicized, and made readily accessible. The project will deliberately minimize user expense in order to develop an accurate sense of scholars' preferences among alternate digital formats (or for the original microfilm!). What is learned about scholars' needs, possibilities, and preferences will inform CRL and LAMP's access, distribution, and pricing policies for the future.

4. Workplan.

LAMP will appoint two or three representatives as a special Project Committee for this digitizing effort. The Project Committee will be selected by the LAMP membership during its 1994 annual meeting. The Committee will actually convene if and when this proposal is funded. Dan Hazen, current Chair of LAMP's Executive Committee and compiler of this proposal, is a probable Committee member. The other(s) will be LAMP representatives versed in both historical research methodologies and electronic technologies. The Committee will provide broad guidance for the project and will either resolve operational and policy issues or refer them on to the membership as a whole.

The project will also be supported by a .33 FTE Project Manager who will supervise, coordinate, evaluate, and report on the work. The Project Manager will be an individual--not necessarily from a LAMP member institution--attuned to electronic information and also able to add project tasks to his or her continuing job responsibilities. The project will seek someone with Latin Americanist experience who is conversant with electronic information, and particularly with intelligence and energy. A candidate has been identified, and this candidate has responded favorably to very preliminary inquiries. Further negotiations will be contingent upon project funding. The Project Manager will identify the specific titles and volumes that will be digitized. (S)he will help negotiate digitizing contracts, oversee additional preservation microfilming that may be needed, actively contribute to the indexing process, publicize the project, and monitor the results.

Commercial vendors as well as a few library operations are producing digital files from microfilm. LAMP intends to subcontract these operations: the equipment is expensive, staff training a significant consideration, the prospect of idle capacity following this initiative quite real, and LAMP's supervisory capability for production operations virtually non-existent. The companies or service bureaus that will eventually perform project tasks of scanning microfilm, indexing electronic files, and preparing CD-ROMs will be chosen on the basis of written proposals. The Project Manager will submit sample microfilm reels to the three firms that have provided the fullest preliminary cost information for scanning and indexing, plus any additional organizations that come to light before the project begins. The Project Committee will review the ensuing bids and choose the most suitable candidate. We may need to conduct a similar but separate process for CD-ROMs, since none of the most promising scanning companies has yet developed an in- house CD-ROM capability.

Materials already available on microfilm will be scanned forthwith. The Project Manager will identify volumes not yet on film and prepare them for preservation filming according to national standards. The filming itself will be subcontracted to an established filmer either within a library or in the private sector. Only a limited amount of preservation microfilming will be conducted for this project. The annual messages of Brazil's president are not on film; they together comprise about 8,000 pages of text, or four or five film reels. Scattered additional volumes of ministerial reports, for volumes published after 1960 (the closing date for LAMP's original project) and very occasionally to fill in gaps, will account for no more than fifteen or twenty additional reels of film. The perhaps twenty-five reels of film will be prepared in the three generations (master negative, print master, positive use copy) mandated for preservation microfilming.

Bibliographic control and basic internal indexing are essential for image files to be fully useful. LAMP's microform holdings are (or will shortly be) recorded in the Center for Research Libraries' bibliographic files. CRL's records are directly accessible over the Internet. The tapes have also been loaded into local online catalogs at many research libraries, and the records are included within the OCLC and RLIN databases. CRL will annotate these records with an Internet address and CD-ROM information for all materials digitized during the project.

Internal indexing--which in the case of serial documents will distinguish each separate volume as well as some major divisions within each volume--will be performed by the agency preparing the electronic files, with advice from the Project Manager and LAMP's Project Committee. The Project Manager will further explore arrangements to digitize Ann Hartness's Subject Guide to Statistics... and then link its references to the appropriate page images. Achieving this access is contingent on copyright clearance, the resolution of several technical details, and affordability.

The digital files will be mounted on the Center for Research Libraries' client/server and also made available on CD-ROMs. The most likely possibility is for the service agency performing the scanning to prepare the electronic products. The Project Manager will also explore whether an operation like Mexico's Universidad de Colima could prepare CD-ROMs from tapes created somewhere else. Some technical issues, for instance how to bundle image decompression software, need clarification before one can actually produce CD-ROMs. Because CD-ROM mastering costs are high, the project will prepare CDs for only part of the digitized materials. Any savings realized in other aspects of the project will be used to prepare more CD-ROMs.

While the specific server site will not be critical, several considerations favor The Center for Research Libraries. The Center has rapidly developed an impressive electronic capability encompassing internal operations, its online catalog, Telnet, and electronic mail. Gopher, list-serv and world wide web servers software are currently in operational testing, and CRL's Access Services Department will shortly add RLG's ARIEL II to the CCITT Group IV fax system presently available. CRL is certainly the location of choice to distribute and track CD-ROM use. Locating the

Internet files at the Center would facilitate reports comparing demand for CD-ROMs, Internet files, and microfilm originals.

A convincing test of the demand for these materials will require broad publicity so that scholars throughout the Latin Americanist community are aware of the possibilities. Over the long term, standard forms of bibliographic and Internet access should prove sufficient in and of themselves. For a shorter-term assessment, though, special publicity is essential. The project will be announced and its results in the LASA Forum (the quarterly newsletter of the Latin American Studies Association), the CLAH Newsletter, the SALALM Newsletter, CRL's Focus on the Center for Research Libraries, and through postings on Internet discussions lists for librarians and Latin Americanists.

5. Tentative Budget.

The budget for the one-year production phase is necessarily tentative: prices for emerging technologies are volatile as new vendors and service agencies appear almost daily, and some costs will only become apparent by working with LAMP's existing film. (One commercial vendor, for instance, would digitize a microfilm sample without charge in order to test procedures and develop an accurate overall budget.) Budget figures for processing and electronic production costs are based on the following sources:

- Preservation microfilming: average costs for recent LAMP microfilming projects; in-house filming costs from Harvard's Photographic Services unit.
- Scanning and indexing: extended discussions and verbal estimates from PFA, Inc. (Sun Valley, California), Preservation Resources (formerly MAPS) (Bethlehem, Pennsylvania), and International Archives Management (Sterling, Virginia).
- CD-ROM mastering and production: discussions and verbal estimates from PFA, Inc., Preservation Resources, Disc Manufacturing (Wilmington, Delaware), and The One-Off CD Shop, Inc., (White Plains, Maryland office). One-Off also supplied a detailed rate sheet.

The following areas and amounts of project expense are anticipated:

The project will require a part-time coordinator to oversee progress and manage the work. This Project Manager will negotiate arrangements with outside vendors, manage business operations, verify the quality of finished digital files, and assemble and collate necessary project materials in microfilm or hard copy, as required. (S)he will oversee indexing and help specify appropriate levels of internal access. The project requires a one-third FTE position for someone combining some knowledge of electronic technologies with experience in Latin American librarianship, at an estimated cost (including benefits) of $20,000.

Estimates are that the project will produce preservation microfilm of about 25,000 pages of text from volumes of standard size. Current prices for the corresponding 12,500 microfilm frames, for three generations of silver halide film, suggest microfilming costs of about $3,000.

Scanning and related prices remain volatile. Different vendors have, sight unseen, suggested approximate costs ranging from 10-12 cents per page (two contacts) to close to 20 cents a page (one contact). Thus, a tentative budget of $140,000 to digitize the estimated one million pages is estimated.

Title-level cataloging will be performed at The Center for Research Libraries at its standard rate of $20/hour. Online records already exist for many of these materials, and serial documents by their nature encompass many volumes within only a few titles. An estimated $2,000 will cover these costs.

Volume-level indexing will be built into the scanning process. More detailed indexing, including links between Ann Hartness's Subject Guide... and the page images themselves, will cost more. A very tentative guess anticipates $25,000 for indexing costs.

CD-ROMs can currently accommodate about 7,500 compressed page images at 300 dpi. Preliminary estimates place mastering costs as high as $1,500 per disk. These price levels are associated with masters for large, "production" runs of hundreds or thousands of CDs. Techniques based on limited runs and recordable CD-ROMs will probably prove more appropriate to our project. With this approach, three copies of each CD could be prepared for about $500, and a single copy of each disk would cost about $300. These costs are still relatively high, so the project will initially prepare CDs only for the provincial reports. This will enable the provision of a coherent and affordable subset of all project materials on CD-ROM, and also to create portable access to the materials to which the expected most detailed indexing (via the Hartness Subject Guide...) will be available. The 255 microfilm reels of provincial reports will fill about fifty CDs at a cost of about $25,000.

Maintenance of electronic files, in the short term while small, should be fairly cheap. Mounting the files at the Center for Research Libraries means that file maintenance will fall within overall CRL project overhead. Overhead will also include administrative costs, communications charges, supplies, etc. All such charges total $4,000.

The Advisory Committee will meet with the Project Manager and CRL staff, either at the Center for Research Libraries or at the scanning service center, soon after image files have begun to arrive, in order to assess progress and to make any necessary adjustments. Travel and related costs are estimated at $3,000.

Many aspects of cost will only become clear once the work is underway. Accordingly, a phased approach to the project is proposed. The project may, for instance, discover that unexpected type sizes and styles, or challenging microfilm quality, require repeated scanning adjustments to produce usable image files. Labor costs would jump, pushing expenses beyond the budgeted fourteen cents per page. Similarly, costs to master CD-ROMs may be higher than those budgeted, since the estimates currently available are surprisingly imprecise.

The project will be phased in terms of its three sets of materials (presidential messages, ministerial reports, provincial reports) and the major operations associated with each. The project will begin by scanning the ministerial reports and creating preservation microfilm for the presidential messages. Then will follow scanning of the provincial reports from after 1889 (that is, the ones not indexed in the Hartness Subject Guide...), the provincial reports from the Empire, and the presidential messages. CD-ROM production will first focus on the provincial reports, and then follow with presidential messages and the ministerial reports. Should funds run short, the Project Committee will determine the best stopping place.

BUDGET SUMMARY

Project Manager	$20,000
Preservation Microfilming	3,000
Scanning	140,000
Cataloging	2,000
Indexing	25,000
CD-ROMs	25,000
CRL administration and overhead	4,000
Travel	3,000
Total	$222,000

6. Project Follow-up

This project here proposed will provide digital access to a core collection of textual materials. Digital imaging, while a proven means to capture textual data, does not now allow full-text searching or manipulation. These and similar activities become possible with electronic files in ASCII format. Certain categories of publications, for instance statistical reports, will be most usefully digitized as ASCII files that researchers can manipulate into their own tables and graphs. The difficulty lies in scanning technology: Optical Character Recognition remains unacceptably erratic, requiring slow and expensive proofreading or rekeying. The technology is improving. LAMP is very interested in a future initiative that might combine image files of statistical volumes with ASCII files in order to facilitate analysis and manipulation.

Appendix 10.3 -- PFA Microfilm Evaluation

(PFA submitted this evaluation of a microfilm sample for the Project Committee in 1995. From this sample, PFA established the scanning and indexing guidelines, as well as the costs for these, for Phase I.)

PFA INC.
Written by Jim and Chris Harper

EVALUATION OF CRL MICROFILM

As I indicated a week or so ago we have completed our inspection of the microfilm cores that were provided to us by CRL. in addition to this evaluation, I have provided the raw data we compiled for you to refer to.

INSPECTION CRITERIA

Our inspection procedure consisted of viewing each core on a foot-by-foot basis over a light table. We used a digital densitometer to make the background density measurements and a 15X loupe to make the letter quality evaluations. This process is consistent with our normal film inspection procedures. After we compiled and reviewed the data, we assigned a numeric rating to each roll to generally define its scanning difficulty. The headings on the raw data sheet can be interpreted as follows:

<u>HEADING</u> <u>INTERPRETATION</u>

CORE: CRL’S CORE NUMBER

ROLL: ORDER OF THE ROLL ON THE CORE

DENSITY: THE AVERAGE DENSITY RANGE *ON* THE ROLL
(THERE ARE SOME LIGHTER & DARKER FRAMES)

QUALITY: RATING OF LETTER BRIGHTNESS WHERE:
G = GOOD, F=FAIR, P = POOR
(COMBINATIONS SUCH AS FG, MEAN "FAIR TO GOOD").

F/L: FRAME SEPARATION AS IT RELATES TO THE SCANNER'S ABILITY TO DETECT FRAMES.

RATING: AN OVER-ALL RATING FROM 1-4 ESTIMATING THE SCANNING DIFFICULTY.

:

COMMENTS OBSERVATIONS MADE BY THE EXAMINER.

The "RATING" scale breaks down as follows:

RATING CRITERIA

1.ACCEPTABLE EDGE DETECTION FOR FRAME POSITIONING.
OVER ALL CONTRAST WITHIN REASONABLE QUALITY LIMITS.

2. ACCEPTABLE EDGE DETECTION FOR FRAME POSITIONING.
CONTRAST VARIATION APPROACHING THE LIMIT FOR
ACCEPTABLE OUTPUT QUALITY.

3.PROBLEMATIC EDGE DETECTION FOR FRAME POSITIONING
AND/OR CONTRAST VARIATION AT OR BELOW THE LIMIT FOR
ACCEPTABLE OUTPUT QUALITY.

4."PULLDOWN" BASED FRAME POSITIONING LIKELY AND/OR
CONTRAST BELOW THE LIMIT OF ACCEPTABLE OUTPUT QUALITY.

As you can see, we have sorted the raw data sheets in two ways; by core number and by rating. Together they should give you a reasonable overview of contents at the roll level and core level. In case you are not familiar with the labeling of the cores, it is as follows:

CORE # CONTENTS YEAR

428 RIO GRANDE DO SUL. PROVINCIA 1829 - 1930
469 BRAZIL. RELATORIOS. MINISTERIALS RELACOES EXTERIORES 1830- 1870
470 BRAZIL. RELATORIOS. MINISTERIALS RELACOES EXTERIORES 1871 - 1888
484 BRAZIL. RELATORIOS. MINISTERIALS GUERRA1827 - 1900
485 BRAZIL. RELATORIOS. MINISTERIALS GUERRA1901 -1925

In general, there are two basic criteria that a microfilm frame must meet to be successfully scanned. The first is the ability to be accurately positioned in the scan aperture (the issue is reliable edge detection). The second is sufficient contrast to produce acceptable legibility, In assigning a rating number to a roll, each of these factors was taken into consideration.

FRAME POSITIONING

Early on, we were concerned that the frames contained in some rolls would produce edge detection problems. Our review of the sample found that the problems of variable frame density, format, size, spacing and page shadowing were present frequently enough and severely enough to necessitate fundamental changes in our scanning methodology. The good news is that we believe that the two new approaches that we have developed will solve the majority of the frame positioning problems that we have seen (the ratings presume these new methods). The bad news is each will cause a decrease in scanning and/or image extraction speed. The impact this will have on the cost of the project is something we are looking into at this time.

FRAME QUALITY

The other key factor in positioning is the difference in density between the document or copy board and the area between the framers. Sufficient contrast here is important for reliable edge detection. Contrast is also a primary factor in determining image quality. As the data shows, the background densities in this sample cover a very broad range. Variations of this magnitude present potential problems for legibility and, in densities below 0.6, for reliable edge detection.

Again, there is good news and bad news. The good news is, the letter density appears to be reasonably low regardless of the background density (there are some exceptions). This means we are cautiously optimistic about the image quality considering the wide variations present in the film. The bad news is, setting up the scanner to reliably detect frames with densities in the 0.5 to 0.6 range will be difficult and, for frames with densities below 0. 5, it will be increasingly problematic as the density drops.

In our opinion, these density problems are rooted in a combination of poor quality documents and poor microfilming techniques. They were then exaggerated by the two reverse polarity duplication procedures that produced the copy that we have.

We believe that, in most cases, the density variations could be significantly reduced by making a good quality direct duplicate from the camera negative. As I have indicated in prior telephone conversations, reverse polarity film is a high contrast medium and tends to make light areas on the film lighter and dark areas darker. Reverse polarity film works well when the images are of consistently good quality throughout a roll. When however, the densities on the original vary significantly, the high contrast nature on this film exaggerates the variations (this was made even worse by duplicating the film on cores without apparent regard to background density on individual rolls).

CONCLUSION

We believe that we can successfully scan existing rolls that are classified with a rating of “1” or “2." Generally, this means negative film with adequate background density and good contrast if some rolls are only available in positive form, we'll do our best with them). It, also, presumes that we will be using the new scanning methodology developed from this sample. It must be understood, however, that there may be frames on these rolls that fall outside of the criteria for a “1” or “2” rating and may require re-scanning. These re-scans would be subject to the $1.35 charge that we presented in our proposal.

For rolls that fall under the "3" classification, we would encourage CRL to obtain a better silver, duplicate, if possible, preferably duped individually in negative mode. For the most part, the primary problem with these rolls is marginal contrast. This is something that we believe could be sufficiently improved by a more careful duplicating effort from the camera negative. Our hope is that this process will result in rolls that are category "2" or higher.

If better duplicates cannot be acquired, we may be able to scan the existing rolls. This may require a "pull down" based positioning methodology which will affect scanning and extraction speed (and, therefore, the price). Using category "3" rolls will, also, produce a higher percentage of re-scans.

Rolls that have a "4" rating have a problem that will require an individualized solution. In the sample, we reserved this rating for rolls containing: A). 0.4 background densities or below, B) exceedingly wide density variations and C) rolls with non-existent or unacceptably narrow frame separation.

For rolls that are consistent with "A” and "B" above we strongly recommend a better duplicate from the camera negative (hoping, in most cases, for a category "2" result). Although it may be possible to scan these rolls, the result would not be good and the effort required to improve the output would be time very consuming and therefore expensive.

Rolls with characteristics similar to "C," will require the slower "pull down" based scanning methodology previously mentioned. Although we have developed this technique in principle, we have not written the software to operate the scanner as yet.

RE-SCANNING

The accuracy with which we can predict the number of chargeable rescans that will occur in a collection as varied as this one is difficult to estimate. We have examined the data in each rating category in an effort to determine an approximate percentage of re-scans that each could produce, our estimates are as follows:

RATING CATEGORY	ESTIMATED PERCENT OF RE-SCANS
A.	10.5%
B.	21.5%

C. 35.0%
D. 41.0%+

CRL RAW DATA BY CORE --

Table 10: Selections from PFA Report on Film Scanning Quality

CORE	ROLL	DENSITY	QUAL	F/L	RATE	COMMENTS
428	1	.70	FG	OK	1	
428	11	.90	FG	OK	1	Good roll
469	10	1.00-1.40	FG	OK	1	Density varies from frame to frame
484	4	1.50	FG	OK	1	Some blotchiness
484	12	1.50	FG	OK	1	Varies within frames - good
428	10	.70-1.05	FG	OK	1	20% handwritten, 40% typed 40% blotchy
470	3	1.00-1.25	FG	OK	2	Tight pulldown last 5%
469	2	1.40	FP	Tight	3	Bad. Frame separation extremely narrow
485	10	.40-.70	FG	OK	4	Sections of .30-.50
484	15	.30-.80	FG	OK	4	Frame to frame, blotchy & balloony

These percentages include re-scans to improve legibility and to correct any positioning problems that were not adequately handled by the new scanning methodologies. We believe these numbers make a compelling case for obtaining a better copy of film whenever possible.

We have scanned a sampling from some of the cores to give you an idea of just what kind of quality the various densities produce. Laser prints of these sample images are being sent to you and CRL so we can continue to develop a common understanding of acceptable image quality.

PRICING

I'm sure you've noticed my references to decreased productivity and price changes. After analyzing the entire sample, we concluded that modifications in our scanning approach were essential to successfully capture the broad spectrum of "stuff' (no other word fits) that this collection contains.

Believe me when I say that our purpose in making these changes is to avoid as much re-scanning as possible and, therefore, save a considerable amount of money.

The magnitude of these pricing changes is something we are working on now and we should have this information for you shortly.

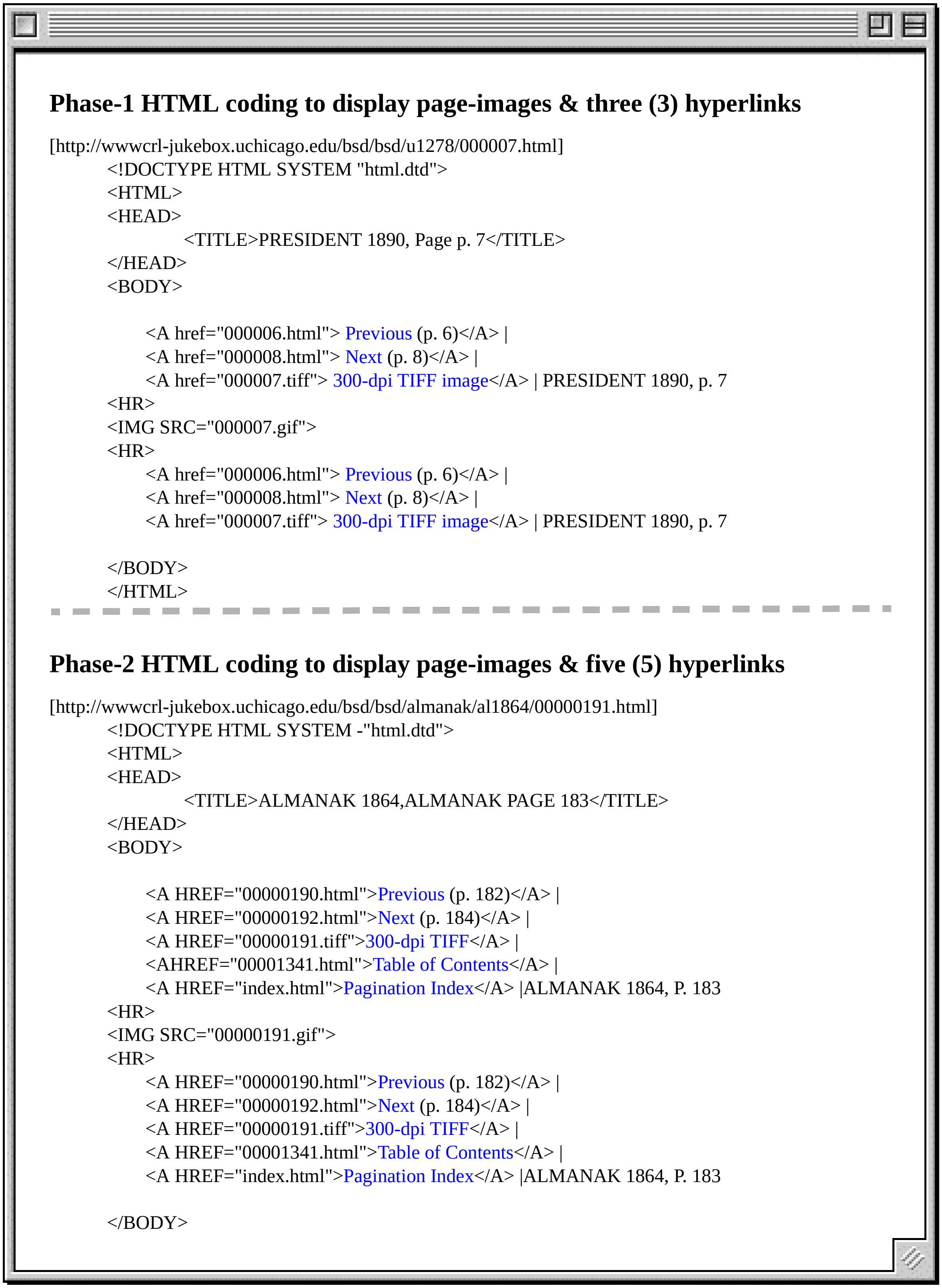
......

Sincerely,

James Harper

Appendix 10.4 -- HTML coding of pagination files

A. Phase-1 vs. Phase-2: Comparative HTML Codes for Hyperlinks. Results on following page.



Phase-1 HTML coding to display page-images & three (3) hyperlinks

[http://wwwcrl-jukebox.uchicago.edu/bsd/bsd/u1278/000007.html]

```
        <!DOCTYPE HTML SYSTEM "html.dtd">
        <HTML>
        <HEAD>
                  <TITLE>PRESIDENT 1890, Page p. 7</TITLE>
        </HEAD>
        <BODY>

             <A href="000006.html"> Previous (p. 6)</A> |
             <A href="000008.html"> Next (p. 8)</A> |
             <A href="000007.tiff"> 300-dpi TIFF image</A> | PRESIDENT 1890, p. 7
        <HR>
        <IMG SRC="000007.gif">
        <HR>
             <A href="000006.html"> Previous (p. 6)</A> |
             <A href="000008.html"> Next (p. 8)</A> |
             <A href="000007.tiff"> 300-dpi TIFF image</A> | PRESIDENT 1890, p. 7

        </BODY>
        </HTML>
```

Phase-2 HTML coding to display page-images & five (5) hyperlinks

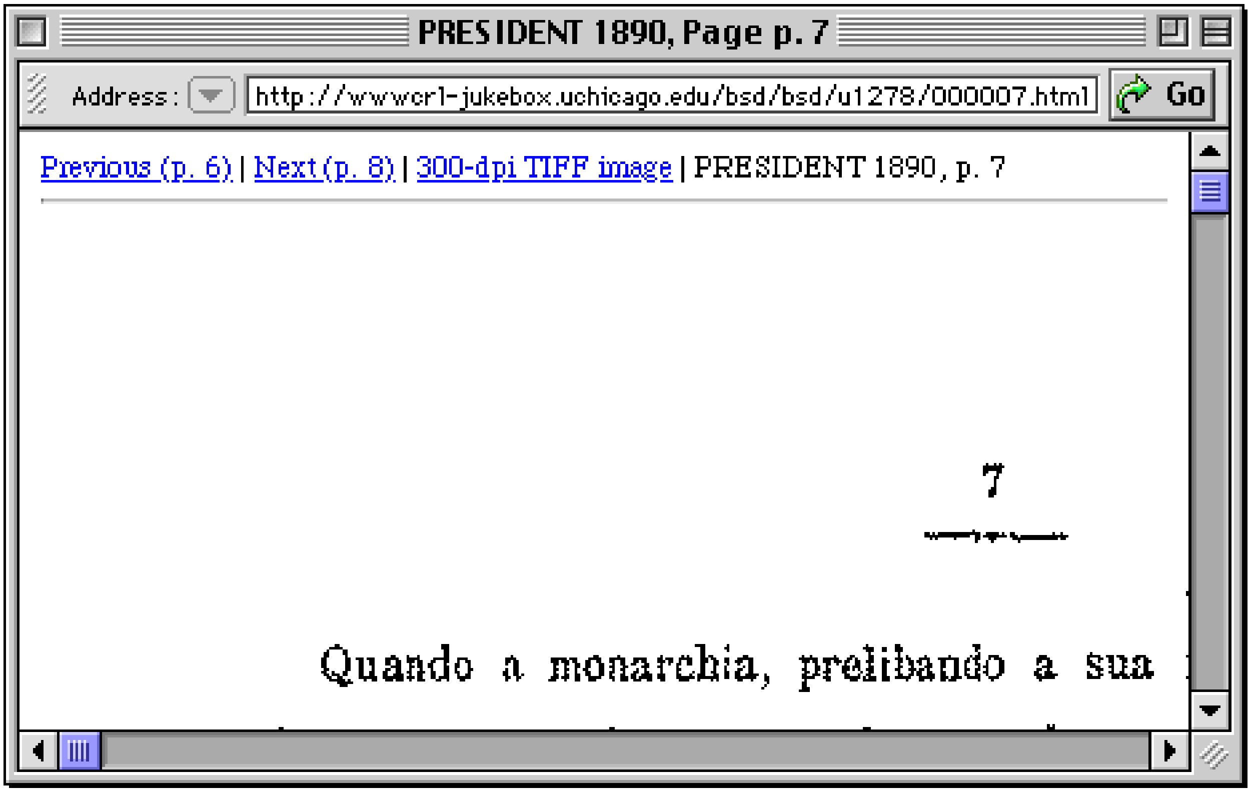
[http://wwwcrl-jukebox.uchicago.edu/bsd/bsd/almanak/al1864/00000191.html]

```
        <!DOCTYPE HTML SYSTEM -"html.dtd">
        <HTML>
        <HEAD>
                  <TITLE>ALMANAK 1864,ALMANAK PAGE 183</TITLE>
        </HEAD>
        <BODY>

             <A HREF="00000190.html">Previous (p. 182)</A> |
             <A HREF="00000192.html">Next (p. 184)</A> |
             <A HREF="00000191.tiff">300-dpi TIFF</A> |
             <AHREF="00001341.html">Table of Contents</A> |
             <A HREF="index.html">Pagination Index</A> |ALMANAK 1864, P. 183
        <HR>
        <IMG SRC="00000191.gif">
        <HR>
             <A HREF="00000190.html">Previous (p. 182)</A> |
             <A HREF="00000192.html">Next (p. 184)</A> |
             <A HREF="00000191.tiff">300-dpi TIFF</A> |
             <A HREF="00001341.html">Table of Contents</A> |
             <A HREF="index.html">Pagination Index</A> |ALMANAK 1864, P. 183

        </BODY>
```

B. Phase-1 vs. Phase-2: Comparative Page-Image Display

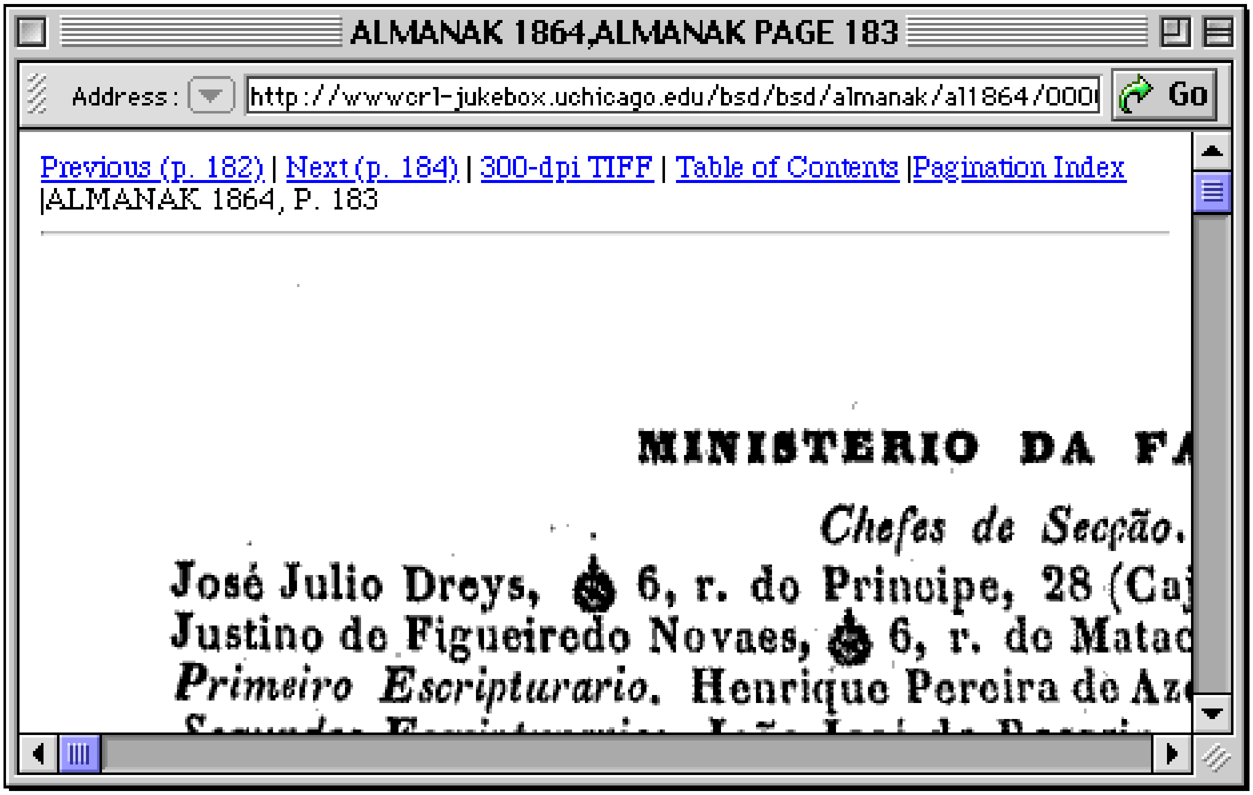


http://wwwcrl-jukebox.uchicago.edu/bsd/bsd/u1278/000007.html

Previous (p. 6) | Next (p. 8) | 300-dpi TIFF image | PRESIDENT 1890, p. 7

7

Quando a monarchia, prelibando a sua



http://wwwcrl-jukebox.uchicago.edu/bsd/bsd/almanak/al1864/000

Previous (p. 182) | Next (p. 184) | 300-dpi TIFF | Table of Contents |Pagination Index |ALMANAK 1864, P. 183

MINISTERIO DA FA

Chefes de Secção.

José Julio Dreys, ✠ 6, r. do Principe, 28 (Ca
Justino de Figueiredo Novaes, ✠ 6, r. do Matac
Primeiro Escripturario. Henrique Pereira de Az

Appendix 10.5 -- WebTrends: Table of Contents for October 2000

Appendix 10.6 -- User Survey Questions

Total responses:52

Brazil	46
U.S.A.	1
U.K.	1
N.A.	4

1. How useful did you find the Database?

Very	47
Somewhat	3
No	1
N.A.	1

2. Did you find the materials you were looking for?

Yes	45
No	7

3. If you used the Hartness Index, did you find the tool easy to use?

Very Easy	29
No	4

4. If you could expand the database, what other resources would you want included? (Optional)

Alamanak Litterario de Sao Paulo 1876-1885 publicado por Jose Maria Lisboa.

Anais do Senado e da Camara Federal.

“The Jornais provinciais.” (Mentioned two times.)

“The reports of the Dom Pedro II railroad after the Republic Proclamation named Estrada de Ferro Central do Brasil. The statistics of union railroads since 1897 Estadistica das estradas de ferro da uni...”

“I would like brasilians families - surnames.”

“Documentos anteriores especialmente dos XVII. Earlier documents, especially those from the 17th Century.”

“I would like very much to see the Brazilian economic and population censuses and the provincial almanaks as well.”

“The Almanak Laemmertz after 1899. That is the republican period. The years reports of the Dom Pedro II railroad after the Republic Proclamation named Estrada de Ferro Central do Brasil. The statistics of union railroads since 1897 Estadisticas das estradas de ferro...”

"Gostaria de ver disponivelos discursos feitos pelos presidentes do regime militar."

"All the Brazilian Archives... I know it is a utopia."

5. Are you a scholar?

Yes	40
No	10
Other	1

6. If you are a scholar, then what type of scholar are you?

Historian	31 (1 not a scholar)
Economist	3
Sociologist	1
Other	7
NA	5

7. What is your name? (Optional)

8. What is your email address? (Optional)

9. Feel free to enter any additional comments here. (Optional)

"Parabens pela iniciativa que cria enorme facilidade de acesso as fontes."

"Great!!!!"

It's very important this site but I just can't find the pages with the Ceara Reports. Is it not ready yet or is it a problem with the system?

"It's impossible to access the Almanak Laemmertz. When you click in Almanak betatest you only access the year 1844. There is no way to access the other years."

"A most useful source. There are problems accessing the Relatorios for Maranhao which seem to have become confused with Para. This should be resolved."

"Sorry by write in Portuguese. fabuloso poder ter acesso a informa es importantes sobre a Historia do Brasil e seus personagens Este site deveria ser mais divulgado e ter seu acesso mais facilitado a todos AParabens pela iniciativa "Fabulous to be able to access a report is important about Brazilian history and its actors. This site ought to be more widely availalbe [sic] and to have its access available to all."

"A database on Brazilian Documents is a very good idea. There is nothing like that in Brazil about Brazil. Congratulations."

"I have to say that I am really impressed with this site. My Great-grandfather was a Brasilian President of the Provincia de Sao Paulo and later of the Republica do Brasil among other political positions. I've been looking online for documents on him during Brasil Imperio for a long time. And found them here together with many other important documents not accessible anywhere on-line. It is really amazing the work you've been doing. Congratulations.”

“I'm writing to express support to the work of the Brazilian Digitalization Project. I had come in contact with the project about two years ago (?) on the web, and went back last week to check on progress. I was really impressed with the results. It is going to be extremely useful for my research to have easy access to ministerial reports, reports from the provincial presidents and to the famous “Almanak Laemmert”. I have two questions: are the ministerial reports - justice going to be digitalized? When do you expect to have the Almanak available on the web?”

“It is going to be very helpful. I was *dying* to read the reports of the ministers of justice, and can hardly believe I'll be able to do it from home! By the way, I work with the nineteenth century, abolition of the slave trade in Brazil - British interference. You can count with my inconditional [sic] support for your project. Please let me know of its developments.”

Gentlemen, I need help. There is some time I am trying to enter in the link http: crlmail2.uchicago.edu almanak 1849 and I don’t get. I would like to know if it is necessary to have some password type to have access or I are having some problem with the link. I am Brazilian. I live in Pernambuco I make masters degree in History and I need to access this documentation. If somebody can me to help. I will be very grateful.”

“This is the most useful and brilliant initiative ever made for Brazilian researchers.”

“It has been very difficult to access the database from Brazil. I don’t think this problem is related just to the state of our telephone lines.”

“You should permit to apply for more than one field in the item n. 6. What about a sociologist-historian or a Political Scientist with strong interest in the historical field. Inter-disciplinary fields are not contemplated by the what items should be filled.”

“Very Good. Congratulations.”

“Well organized. Very useful for me.”

“Achei extremante relevante este trabalho de voce. Pude completar minha dissertacao de mestrado na Universidade Federal de Rio de Janeiro.”

“Your homepage is very useful and I hope to use it more times.”

10. If you entered any additional comments, may they be made public?

Appendix 10.7 -- RFP to Vendors

October 3, 1994

REQUEST FOR PROPOSAL:
Creating a Digital Core Collection of Brazilian Serial Documents:

A Proposal by The Center for Research Libraries
and the Latin American Microform Project.

INTRODUCTION

The Center for Research Libraries, CRL, and Latin American Microform Project, LAMP, has been funded by the Andrew W. Mellon Foundation to digitize a core set of executive branch serial documents issued by Brazil's national and provincial governments. The project will accomplish the following:

1.We will facilitate scholarly access to a central and coherent body of high profile research resources for Brazilian studies.

2.We will expand the still-tiny corpus of digital image files aimed at a scholarly audience by issuing them both as CD-ROMs and as files available over the Internet.

3.We will implement mechanisms to ensure traditional bibliographic access to each digitized serial.
We will provide structured access to individual volumes within each serial set.
We will provide electronic indexing to the sections within these documents, single issues of which can be hundreds and even thousands of pages long.

4.We will explore relative levels of demand and patterns of use for the digitized materials.

PARTICIPATING INSTITUTIONS

The Latin American Microform Project (LAMP) was formed in the 1960s in order "to acquire, preserve, and maintain for its subscribers microform collections of unique, scarce, rare, and/or bulky and voluminous research materials pertaining to Latin America." About thirty-five North American libraries with Latin American collections comprise the current membership. Most of LAMP's preservation work is funded by member dues.

The Center for Research Libraries, an Illinois not-for-profit corporation, is a cooperatively supported library for libraries. The center collects and provides access to research materials that often cannot be afforded by public or university libraries.

The Andrew W. Mellon Foundation provides grants on a selective basis to higher education projects. The Mellon Foundation has made a commitment to improve access to Latin American research materials.

MATERIALS

Our project will concentrate on Brazilian executive branch and provincial serial documents issued between 1830 and 1990. The executive branch materials include the annual messages of the president and annual federal ministerial reports that often offer fuller contexts and more detailed analyses than provided by the presidential messages. The provincial documents we will digitize are the annual reports of Brazil's provincial "presidents" from 1830 to l889.

All these official documents are in the public domain, so they can be digitized and distributed without concern for copyright. LAMP already owns microfilm of almost all the materials, so we can focus on scanning from existing film with an only occasional need to create preservation microfilm as well.

LAMP now holds about 635 microfilm reels of the serial documents that we propose to digitize. We will need to prepare perhaps thirty additional reels of film as we fill gaps and complete runs. The 665+ / - reels that we digitize will represent about 650,000-700,000 frames of microfilm. Some materials were filmed with one page per frame, others with two. We estimate about one million pages of text which equates to roughly 100-125 gigabytes of data.

SCANNING

Pilot projects with digital imaging suggest that it can be combined with preservation microfilm to ensure both longevity and fluid access. The preservationist community is in general agreement that, at the present time, we should not rely on digital files for long-term preservation. Since we have already produced preservation microfilm for almost all the materials under consideration, our project can particularly focus on issues of electronic access without compromising LAMP's mandate for long-term preservation. The documents we propose to digitize are overwhelmingly textual, with occasional charts and tables and relatively rare illustrations. The proposed scanning procedure should capture their content.

INDEXING

We also intend to explore the feasibility of more specialized access to parts of the state presidential messages. Ann Hartness's Subject Guide to Statistics in the Presidential Reports of the Brazilian Provinces, 1830-1889, published in 1977, provides detailed indexing to this subset of our holdings. (See Appendix A for samples of this work.) We have obtained copyright clearance for using this work. We will digitize the Guide and create electronic links between the index and the original texts.

Internal indexing will be performed by the agency preparing the electronic files, with advice from the Project Manager and LAMP's Project Committee. In the case of serial documents, index each separate report as well as the chapter within each report. The chapter level may have between eight to twelve subdivisions. These sections would as a general rule correspond to administrative subdivisions of the parent ministry.

STORAGE AND INTERNET ACCESS

The digital files will be mounted at the Center for Research Libraries. These files will be accessed through their client/ server (UNIX running on a Sun SPARCSTATION email/gopher/ WWW server through a Compaq 486 Novell file server) and also made available on CD-ROMS. The most likely possibility is for the selected agency performing the scanning to prepare the electronic products. Because CD-ROM mastering costs are high, we will prepare CDs for only part of the digitized materials. Any savings realized in other aspects of the project will be used to prepare more CD-ROMS.

SERVICES TO BE CONTRACTED

The following tasks, generally described above, will be contracted to companies or service bureaus. The grant runs through August 1, 1995. Please provide a cost and a schedule for the completion for each task below. Also, provide references for previous digitizing projects. Microfilm samples will be provided upon request.

1. Scan Brazilian documents from preservation microfilm.
   Data format: Bit-map, Raster data
   File format: TIFF-B
   Digital resolution: 300. We will initially experiment with a resolution of 300 dpi. Higher resolutions will be employed if and as the materials so require- -for instance if very small typefaces must be captured.
   The quality of the scanned image will be established by the project committee with recommendations from the contracting company.
   The scanning should be completed for all materials that are now available on microfilm within six months of the signing of a contract.

2. The information will be transferred to CRL on a 4 millimeter DAT tapes that each stores 8 gigabytes of information.
   A. Compression: CCITT, Group 3 or 4

3. Index the above image files.

An index will be created and linked to the raster data images and completed concurrently with the scanning.
Access points (at least to the volume level for the presidential and ministerial reports) will be established by the project committee. C. The indexing of the provincial papers will be from Ann Hartness' Subject guide to statistics in the presidential reports of the Brazilian provinces, 1830-1889. This guide indexes titles of the reports by state, and indexes the subject content of each report.

4. Provide or recommend index/ retrieval software that will provide searching capabilities of the documents.
Compatible with CRL's Sun SPARCSTATION email/gopher/WWW server or its Compaq 486 Novell platform.
Internet compatible.
Software must be able to search the document structure established during the scanning and indexing.
Software must allow for the selection and printing of pieces of the materials.
Software must be accessible through the Internet from MAC and DOS based PCs.

5. Produce CD-ROMS, bundled with search software, of selected sections of the above data chosen by the project committee.
4 3/4 in.; single density.
Produced according to current library standards.

Please return three copies of your proposals by Oct 26th to the following address:

Scott Van Jacob
Spahr Library
Dickinson College
Carlisle, PA 17013-2896
(717) 245-1866
FAX (717) 243

Addendum:

Please note the following changes to the above document:

1.The files will be stored on an optical storage device (opto-magnetic technology) at CRL. (CRL has looked at a demonstration of this technology.)
2.The vendor will deliver the electronic files to CRL on 4 mm Dat tapes.

3.Update to 2.A.: The preferred Compression is CCITT, Group 4.

4.Update to 4.A.: A., Compatible with CRL's Sun SPARCSTATION email/gopher/VAM server is connected to a Compaq 486 Novell platform. The software should run on UNIX and be compatible from MAC, DOS, and Windows 3.1 based PCs.

11. BIBLIOGRAPHY

Arms, Caroline R. *Access Aids and Interoperability*. National Digital Library Program, Library of Congress. August 18, 1997. [http://memory.loc.gov/ammem/award/docs/interop.html#ead] (July 2, 1999)

Conway, Paul. *Conversion of Microfilm to Digital Imagery: A Demonstration Project: Performance Report on the Production Conversion Phase of Project Open Book*. New Haven, CT: Yale University, 1996.

Bates, Marcia J. "Indexing and Access for Digital Libraries and the Internet: Human, Database, and Domain Factors." *Journal of the American Society for Information Science*. Vol. 49, no. 13 (1998) pp. 1185-1205.

Croft, W. Bruce. "What Do People Want from Information Retrieval? (The Top 10 Research Issues for Companies that Use and Sell IR Systems)." *D-Lib Magazine* (November 1995) [http://www.dlib.org/dlib/november95/11croft.html]

Deal, Carl W. "The Latin American Microform Project: A Model for Cooperation," *SALALM Papers*. Vol. 31 (1986) pp. 277-282.

The Digital Library Toolkit. 2nd Edition. Palo Alto, CA: Sun Microsystems, 2000.

Feather, John and Paul Sturges. Eds. *International Encyclopedia of Information and Library Science*. London, England: Routledge, 1997.

Development of the Encoded Archival Description Document Type Definition. [http://www.loc.gov/ead/eadback.html] (July 1, 1999)

Hartness, Ann. *Subject Guide to Statistics in the Presidential Reports of the Brazilian Provinces, 1830-1889*. Austin, Texas: Institute of Latin American Studies. The University of Texas at Austin, 1977.

Kenney, Anne. "Digital-to-microfilm conversion: an interim preservation solution." *Library Resources & Technical Services* (October 1993): 380-402 and (January 1994): 87-95;

-----. *Moving theory into practice: digital imaging for libraries and archives*. Mountain View, CA : Research Libraries Group, 2000.

----- and Stephen Chapman. *Digital Imaging for Libraries and Archives*. Ithaca, NY: Department of Preservation and Conservation. Cornell University Library, 1996.

----- and Lynne K. Personius. *The Cornell/Xerox/Commission on Preservation and Access Joint study in digital preservation: report, phase 1, January 1990-December 1991: digital capture, paper facsimiles and network access*. Ithaca, NY : Cornell University Library, 1992.

Lessons Learned from the LoC/Ameritech Digital Library Competition. November 1998. [http://memory.loc.gov/ammem/award/lessons.html] (May 2, 2001)

MacDougall, Susan. “Rethinking Indexing: the Impact of the Internet.” *Australian Library Journal*. (November 1996). pp. 281-85.

McClung, Patricia A. *Digital Collections Inventory Report*. Washington, D.C.: Council on Library Resources and the Commission on Preservation and Access, 1996.

Missingham, Roxanne. “Indexing the Internet: Pinning Jelly to the Wall?” *LASIE: Information bulletin of the Library Automated Systems Information Exchange*. Vol. 3, no. 3 (Sept. 1996) pp. 32-43.

Pitti, Daniel V. “Encoded Archival Description: the Development of an Encoding Standard for Archival Finding Aids.” *American Archivist*. Vol. 60, no.3 (Summer 1997) pp. 268-283.

Sundt, Christine L. “The Quest for Access to Images: History and Development.” *Advances in Librarianship*. Vol. 22 (1998) pp. 87-104.

Van Jacob, Scott. “Six Ways from Sunday: Approaches to Indexing Digital Text Images.” *Computers and Humanities*. Vol. 33 (1999) pp. 383-407.

Waters, Donald J. *From Microfilm to Digital Imagery: On the Feasibility of a Project to Study the Means, Costs and Benefits of Converting Large Quantities of Preserved Library Materials from Microfilm to Digital Images: A Report of the Yale University Library to the Commission on Preservation and Access*. Washington, DC: The Commission, 1991.