LAMP Digitization Proposal

Text below in italics is directly from the LAMP Digitization Project Principles (http://www.crl.edu/area-studies/lamp/news/proposal-guidelines).

Standard information for all proposals from the University of Florida (UF) George A. Smathers Libraries is provided below, when applicable for all projects. This information is current as of March 2014.

I. Narrative

Title and Abstract

Diario de Pernambuco Project Phase III
At the Trinidad LAMP meeting (June 16, 2012) UF proposed and received support to scan and digitize the *Diario de Pernambuco*. The University of Florida's (UF) holdings include 276 reels of microfilm dating from 1825 through 1923. These reels are the only holdings outside of Rio's Biblioteca Nacional, which was UF's original source. UF is committed to hosting the digital newspaper content. Project cost estimates to digitize UF's holdings of this important primary resource range from $52,000 to $100,000 depending on the number of pages on each reel. During the 2012 meeting in Trinidad, the membership voted to award UF $25,000 to cover the initial part of the project; another $25,000 was approved by LAMP at the 2013 Coral Gables meeting for a Phase II; this proposal respectfully requests an additional $21,360 to support Phase III completion of the remaining 81 reels and have all 276 complete with full, online open access, and digital preservation.

The following is a progress report as of March 2014: The George A. Smathers Libraries at the University of Florida has processed the remainder for Phase I-II of 195 digitized reels returned from Creekside Digital. These reels represent 20,098 issues of *Diario de Pernmabuco* newspaper dating from 1825 to 1896. Access is free and open at: http://ufdc.ufl.edu/AA00011611/

As of February 11, 2014, the 20,098 issues have been viewed 343,960 times.

UF project staff worked with the vendor to precisely expend the grant award allocation.

Content

The *Diario de Pernambuco* is acknowledged as the oldest newspaper in circulation in Latin America (see: Larousse cultural; p. 263). Digitized newspapers during the proposed timeframe will offer researchers insights into early Brazilian commerce, social affairs, politics, family life, slavery, and such; published in the port of Recife. The *Diario* contains numerous announcements of maritime movements, crop production, legal affairs, and cultural matters. The 19th century newspapers include reporting on the rise of Brazilian nationalism as the Empire gave way to the earliest expressions of the Brazilian republic. The 1910s and 1920s were years of economic and artistic change, with surging exports of sugar and coffee pushing

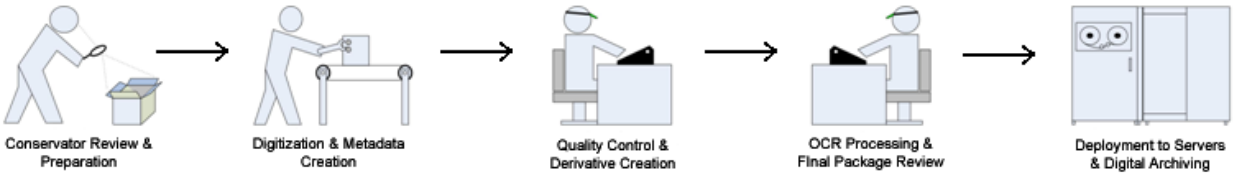
revenues which supported rapid expansions of infrastructure, popular expression, and national politics.

Copyright / Permissions

The Smathers Libraries support US Copyright Law as well as moral and cultural heritage rights, and other applicable rights. In order to support these rights for UF, partners, and constituencies, the Libraries follow a permissions-based model (http://dloc.com/AA00002865 and http://ufdc.ufl.edu/permissions). Full documentation on rights and permissions in place are maintained for all materials. If the permissions and rights in place allow the assignment of rights to LAMP, then those can be assigned. Additional information will be provided based on the specific project needs.

Conversion Procedure

The UF Digital Library Center (DLC) is a one of the largest digitization and digital curation facilities in the southeast. The DLC utilizes many types of equipment and relies on industry standards for digitization that adheres to digital preservation standards.
The common workflow is shown in the image below.



All imaging is completed in accordance with established professional standards. Imaging methods will depend on object characteristics, and follow principals and guidelines established in *Moving Theory into Practice: Digital Imaging for Libraries and Archives* by Anne R. Kenney and Oya Y. Rieger and Cornell University's *Digital Imaging Tutorial*. Imaging (i.e., scanning, text, metadata) is based on specifications previously established by UF and its partners (http://digital.uflib.ufl.edu/technologies/documentation/imaging.htm).

All objects are digitized to meet standard requirements for the item's physical format. Images are captured as uncompressed TIFF files (ITU T.6) at 100% scale. All project imaging is calibrated regularly to maintain color fidelity and optimum image results.

Equipment for digitization includes:

- Super 8K-HS digital camera (for maps, architectural drawings and other large format materials)
- CopiBook (appropriate items, up to 15 x 23 sizes)
- Flatbed scanners (Microtek 9800 XL)
- Nikon Super CoolScan 5000 ED Film Scanner and Nikon SF-210 Auto Slide Feeder (slides, scanned individually or in batches)

- Details on all available equipment are here: http://digital.uflib.ufl.edu/technologies/technologies.htm

What quality control will be used to ensure best practices are adhered to throughout the conversion process?

UF utilizes many types of equipment and relies on industry standards for digitization that adheres to digital preservation standards.

If OCR is generated, will it be edited or uncorrected?

OCR text is uncorrected.

Metadata

Metadata processing is common for all materials.

Metadata: Metadata Encoding and Transmission Standard (METS; http://www.loc.gov/standards/mets/) metadata is created using the SobekCM tools and system, which are a full suite of production, digital collection (access), and repository (preservation) tools. The production workflow is integrated with the access system for consistency. As items are processed, the metadata is enhanced automatically and manually as objects move through the imaging/curation workflows. The SobekCM system assigns a unique Bibliographic Identifier (BibID) to each object processed, and that BibID is used to track the item (see UF Metadata Information, http://ufdc.ufl.edu/sobekcm/metadata). The METS files include technical and structural data about each image, as well as descriptive and administrative information.

Any pre-existing metadata (e.g., from catalog records, finding aids, museum accession records) will be imported into the SobekCM system at the first stage, before the start of imaging. The metadata for materials is prepared by Catalogers, Archivists, Subject Matter Experts, Registrars, Curators, and others as appropriate for the project.

The SobekCM system stores all metadata in METS/MODS as well as automatically transforming and providing the metadata in MARCXML and Qualified Dublin Core, with all metadata accessible online. All materials are optimized for search engine access to ensure worldwide reach through Google and other search engines. SobekCM includes integrated support for OAI-PMH (Open Archives Initiative or OAI) to ensure all metadata is harvestable following OAI-PMH standards.

The SobekCM system specifications are optimized for data exchange for harvesting by other digital libraries such as the U.S. National Science Foundation's National Science Digital Library, the U.S. Institute for Museum and Library Services' National Leadership Grant collection, and OAIster at the University of Michigan.

Added-Value Features

Describe any proposed products beyond digital image files. For instance:

- ***Will text files be made searchable via the application of Optical Character Recognition software or double-keying?***
    - SobekCM provides full text searching within collections as well as having the collections and materials optimized for search engine access to ensure worldwide reach through Google and other search engines
- ***Will searchable text files be marked up in accordance with specific schema?***
    - TEI and other schemas are applied on a project-specific base.
- ***Will numerical files be rendered in forms suitable for statistical manipulation?***
    - SobekCM supports standardized file formats, including data sets and numerical files.
- ***Will cartographic and related materials include geospatial referencing?***
    - Yes. SobekCM supports map-based searching and browsing for all materials with geographic metadata.

Access

Describe how the users will access the data.

Delivery system:

- SobekCM, http://ufdc.ufl.edu/sobekcm/.
- ***In what format will the files be delivered?***
    - Imaged object files are delivered online in JPG, JPG2000, and JPG thumbnail images along with the OCR text files (TXT and PRO, for location of text on the image files) and with the metadata, displayed as a "citation" and also available and displayed in all metadata formats (METS/MODS, MARCXML, Qualified Dublin Core).
- ***Will the data be freely available on the internet? If not, what limitations to access will be in place for this data (and why)?***
    - All data will be freely available.
- ***What search and browse capabilities will be used to access the data?***
    - SobekCM support for all collections and items includes:
        - Full text searchable
        - Browseable - with browse views by title and thumbnail, and by new items
        - Serve text, image, multimedia, audio, video files, data sets, and more within the same collection
        - Support for multiple file types (text, image, oversized images, video, audio)
        - Powered by rich metadata support, with automatic transformations for maximum interoperability
        - Google-map based searching or map browsing
    - Custom views for specific item-types:
        - Full-screen page turner view

    - Sanborn maps
    - Image zoom and pan viewing capabilities
- ***Will the metadata allow for easy harvesting of data?***
    - Yes.
      All materials are optimized for search engine access (SEO) to ensure worldwide reach through Google and other search engines.
      SobekCM includes integrated support for OAI-PMH (Open Archives Initiative or OAI) to ensure all metadata is harvestable following OAI-PMH standards.

Archiving

Describe terms for the preservation and ongoing maintenance of content.

What is your process for sustained preservation of the files?
Will the data be archived at any location(s) other than CRL?

The University of Florida George A. Smathers Libraries are committed to long-term digital preservation of all materials in the UF Digital Collections, including the IR@UF, and in UF-supported collaborative projects as with the Digital Library of the Caribbean (dLOC). Redundant digital archives, adherence to proven standards, and rigorous quality control methods protect digital objects. The UF Digital Collections provide a comprehensive approach to digital preservation, including technical supports, reference services for both online and offline archived files, and support services by providing training and consultation for digitization standards for long-term digital preservation.

The Libraries support locally created digital resources, including the UF Digital Collections which contains over 200,000 digital objects with over 20 million files (as of September 2011). The Libraries create METS/MODS metadata for all materials. Citation information for each digital object is also automatically transformed into MARCXML and Dublin Core. These records are widely distributed through library networks and through search engine optimization to ensure broad public access to all online materials.

In practice consistent for all digital projects and materials supported by the Libraries, redundant copies are maintained for all online and offline files. The digital archive is maintained by the Florida Center for Library Automation (FCLA). Completed by the FCLA in 2005, the Florida Digital Archive (FDA) (http://fclaweb.fcla.edu/fda) is available at no cost to Florida's public university libraries. The software programmed to support the FDA is modeled on the widely accepted Open Archival Information System. It is a dark archive and no public access functions are provided. It supports the preservation functions of format normalization, mass format migration and migration on request.

As items are processed into the UF Digital Collections (UFDC) for public access, a command in the METS header directs a copy of the files to the Florida Digital Archive (FDA). The process of forwarding original files to the FDA is the key component in UF's plan to store, maintain and

protect electronic data for the long term. If items are not directed to load for public access, they do not load online and are instead loaded directly to the FDA.

How will you deliver the files to CRL?

Files to partners are regularly transferred using FTP or mailed external hard drives, with both supported and selectable by partners for best applicability for their processing.

What will you do with the original source material?

Decisions on the disposition of source material are handled by the appropriate collection manager, curator, or archivist. There are occasions when digitization for digital preservation is an absolute necessity because materials are disintegrating and cannot be preserved further in physical form. Most often, digitization for digital preservation is conducted alongside conservation of the physical materials where the materials, once conserved and if handled less frequently, will remain preserved in physical form. Because digitization for digital preservation and the ongoing work for digital curation are laborious and expensive processes, the physical objects selected for digitization are often from special and area studies collections where the physical materials are significant as artifacts and will continue to be preserved in that form.

The University of Florida has no plans at this time to de-accession the source microfilm for *Diario de Pernambuco*.

II Plan of Work

A detailed work plan should include an estimated schedule for the digitization project, broken down by the phases of the project (selection, permissions, preparation, conversion and quality control, metadata creation, delivery, preservation, etc). The work plan should also include information about the staffing needed to complete all aspects of the project.

For Phases I-II, UF worked with Creekside Digital for vendor digitization. Based on updated pricing methods and costs, UF plans to work with iArchives to complete Phase III. As with all vended digitization, UF will ensure quality control of metadata and materials, and open access and archiving of all materials from the dedicated webpage in the UFDC for this project: http://ufdc.ufl.edu/AA00011611

The stages of this project consist of:

- Microfilm preparation and shipping
- Vendor digitization
- Ingest of vended materials, which are digitized according to the National Digital Newspaper Program Specification (NDNP).

During Quarter One (Q1), the microfilm will be prepared and shipped. Depending on vendor workflows after the funds are awarded, the vendor will convert the 81 reels in several batches

scheduled across Q2-4. The vendor digitization and UF validation/ingest processes will be done for each of the batches and details on these activities are below.

In accordance with the NDNP specification, the vendor will scan from the microfilm and create derivative files according to specifications described in Appendix B (page 35) of the National Digital Newspaper Program Technical Guidelines. The film will be scanned in 8-bit grayscale with a maximum resolution between 300-400 dpi, relative to the physical dimensions of the original material. Scanning will produce an uncompressed, unprocessed TIFF 6.0 file for each newspaper page on the microfilm. In cases where newspaper titles were microfilmed with two pages per frame, the vendor will make adjustments to produce a single image file for each newspaper page. The scanned TIFF will be de-skewed and cropped to the page edge, if necessary. Prior to digitization of each reel, the vendor will scan a target. The vendor will scan a second target during the reel digitization to aid in monitoring of scan quality. Each target will be described appropriately in reel metadata.

In addition to a TIFF 6.0 file for each newspaper page on microfilm, the vendor will produce a:

- JPEG 2000 file from the TIFF 6.0 file. The JPEG2000 file will conform to the 21 specifications listed in Appendix B of the NDNP Technical Guidelines. For instance, each JPEG 2000 file will have 6 decomposition levels, 25 quality levels and a compression ratio of 8:1.
- PDF file from the TIFF 6.0 file. The PDF file will conform to the 18 specifications listed in Appendix B of the NDNP Technical Guidelines. It will have a file name corresponding to a specific page image, hidden text and metadata referring to the source publication, the date of publication, page number, the reel number and sequence order.
- One OCR text file for each newspaper page image. The text conversion process will produce files that meet the specifications listed in Appendix B of the NDNP Technical Guidelines.
  Each text file will contain:
  - Uncorrected text
  - Word-bounding boxes zoned for column recognition. Files will be free of article level segmentation.
  - Bounding box coordinate data at the word level.
  - UTF-8 characters
  - No graphic elements
- The text created through OCR will be encoded using ALTO Version 2.0. If possible, the vendor will supply confidence level data at the page, line, character, and/or word level. Additionally the vendor will seek to provide point size and font data at the character or word level.

QUALITY CONTROL, VALIDATION, LOADING & ARCHIVING

UF will remain in frequent communication with the vendor via telephone and email. After completion of microfilm scanning, file creation and metadata encoding,. Vendor staff will validate the vendor created files. Upon completion of the validation process, the vendor will send validated files on external hard drives to the Smathers Libraries where project staff will:

- Validate all deliverables (request vendor recreate deliverables that failed to validate); confirming vendor has accounted for discrepancies noted by project staff during their initial evaluation of reels
- Ensure vendor correctly used Issue Present and Page Present indicators
- Verify the four digital files associated with a newspaper page (TIFF, JP2, PDF and OCR text file) use the same file name and differ only by respective file extensions.
- Load all files online to the UF Digital Collections powered by SobekCM, with all newspaper issues together, including the issues completed in Phase I and II: http://ufdc.ufl.edu/AA00011611
- Archive all files for long-term digital preservation

III. Budget

A detailed budget should include estimated costs for the digitization project, broken down by the phases of the project. The budget should include any project support requested of LAMP, as well as expected from sources other than LAMP.

The format for this parallels the previous budgets for Phases I and II. UF will provide overall project management and production to ingest issues of *Diario de Pernambuco* to UFDC and archive to Florida Digital Archive.

Diario de Pernambuco Phase III (81 Reels)	
Expense Categories	**Expense Detail**
81 reels (1-up reels) with 650 frames each[1]	81 x 650 = **52,650 frames**
Per frame/page cost: $0.40	$0.40 x 52,650 = **$21,060**
Shipping	**$300**
Total funds requested	**Total requested: $21,360**

[1] Quote from vendor includes other costs within per page/frame (included services: segmentation at the issue level; OCR; ALTO; derivatives).