

## 6.3 Appendix C

### *Hardware & Software for Quality Control Inspection*

#### **Computer workstations**

- Each with the capacity to perform quality control of images, text and to archive files. Each is relatively current, operating at speeds at or greater than 2.2 GHz Pentium 4 processors or better – many with dual processors; a minimum of 1GB RAM, and 120 GB hard-drives, and Plextor CD/DVD writers. Most workstations run Windows XP (sp2), though some still run Windows NT 4.0. Their monitors are calibrated weekly if not more frequently.
- The (Image) Quality Control Unit currently operates 5 dedicated computer workstations.
- The Text Conversion and Mark-up Unit currently operates 5 dedicated computer workstations, with dual monitor configurations optimized for side-by-side review of digital page image and text.
- And, both Units routinely access to the workstations of the Digital Imaging Unit when they are not scheduled for production. The Digital Imaging Unit operates 15 computer workstations, many with dual monitor configurations.
- The Digital Imaging Unit also maintains a dedicated workstation for microfilm scanning, cf, Mekel 525 GS, below.
- Additionally, the Copy Control/Tracking Unit operates 4 dedicated computer workstations and has access to the workstations of the Analog Imaging Unit when they are not otherwise scheduled for use.

#### **Mekel 525GS – Gray-scale Microfilm Scanner**

- <http://www.mekel.com/prod03.htm>
- The Mekel 525GS is a robust, potentially high-speed gray-scale microfilm scanner, capable of 300 dpi uncompressed-TIF out-put. N.B. Gray scale out-put is native JPEG, which staff converts immediately to uncompressed TIF while the image remains on screen. (Rather than reopening the image after saved, this routine retains optimal as-scanned image qualities.)
- We do not propose to use this unit for project production. In addition to the gray-scale issue, this unit runs in production mode only with considerable set up and difficulty.
- We propose to use the Mekel 525GS to image selected frames of microfilm reels in order to create control sets for vendor information and quality control. Control sets will be used for a variety of purposes:
  - To visually inspect second-generation negative microfilms produced for digitization from stored camera-master first-generation negative microfilms. The configuration of the Mekel 525GS is optimal for inspection of preservation microfilm as outlined by RLG *Preservation Microfilming Handbook* (Mountain View, CA : Research Libraries Group, 1992).

N.B. Erich Kesse, Director, Digital Library Center and Principal Investigator for this project, was one of the editors of the *Handbook's* technical specifications.

## **Appendix C**

### ***Hardware & Software for Quality Control Inspection***

(continued)

#### **Mekel 525GS – Gray-scale Microfilm Scanner** (continued)

- To determine image imperfections not readily identified by standard microfilm inspection procedures. Digitization, in our experience, is an excellent means of identifying illumination imbalance and related issues that impact digital image quality. Imperfections will be noted on the report sent to the microfilm digitization vendor. Our vendors, ByteManagers and iArchives, operate microfilm digitization hardware with software controls that mitigate know illumination issues.
- To process page images, using our installation of the PrimeRecognition (see, PrimeRecognition, below) and UF DLC Zoning application to pretest text conversion and mark-up. Pretests will set benchmarks for vendor text product quality.

#### **UF DLC Tracking Database**

- The Tracking Database is the DLCs work queue and product record database. It manages all aspects of each of our digital projects and will be used in Florida's National Digital Newspaper Project.
- The application stores data in Microsoft SQL tables and its interface and behaviors are programmed in C# for the .NET framework, v1.1.
- In addition to collection data and queuing work, it also generates packing lists, statistical reports, etc.

#### **UF DLC QC Application**

- The Quality Control (QC) application is a locally programmed GUI that generated JPEG thumbnails and JPEG2000 derivatives and presents them in sequential order for inspection. (N.B. *Actual JPEG2000 compression/quality, optimization, tile size, etc. can be set as instructed.*)
- In addition to visual inspection, the application allows its user to attach or confirm structural metadata (both physical – e.g., page and section numbering – and intellectual – e.g., chapter headings, article titles, etc.), to accept or decline images, and to perform basic image manipulation or correction (e.g., rotation, etc.)
- The application report rates of acceptance with and without correction and rates of decline together with detailed error findings.
- The DLC QC application will be used to review, accept or decline vendor image product.

#### **Adobe PhotoShop CS**

- Adobe PhotoShop CS is used in conjunction with the DLC QC Application to perform basic manipulation and correction, as well as to generate JPEG2000 derivative versions.
- Programming effort is currently underway to move away from Adobe PhotoShop toward open source software (likely either IrfanView or GIMP together with their JPEG2000 plug-ins).

## **Appendix C**

### ***Hardware & Software for Quality Control Inspection***

(continued)

#### **Adobe PhotoShop CS (continued)**

- We hope to compile the DLC applications suite for distribution to other digital library programs and hope that participants in the National Digital Newspaper Program might agree to alpha-testing, as a means of asserting a modicum of vendor independence that might, in turn, reduce costs. [N.B. *This open source/alpha testing plan is not budgeted as part of this proposal.*]

#### **UF DLC Zoning Application**

- A locally programmed application, similar to PrimeZone, a plug-in application for PrimeRecognition optical character recognition (OCR) software.
- DLC Zoning improves accuracy by identifying columns and complex layout structures. N.B. PrimeRecognition has automatic zoning capability, but it is our general experience that the addition of Zoning control improves OCR out-put.
  - One of our objectives for this project is to determine the extent to which such a tool is necessary and to review cost over increased accuracy.
- The UF DLC Zone application is being developed as an open source application that can be given to other institutions. It will be alpha-tested by the University of Central Florida and the University of the Virgin Islands in FY2004-2005.
- The application is different from PrimeZone in that, like the UF DLC QC application, it allows the attachment of structural metadata to zones and can be used in concert with PrimeRecognition to produce intelligent mark-up.

#### **PrimeRecognition**

- PrimeRecognition (<http://www.primerecognition.com>) optical character recognition (OCR) software is run by the Digital Library Center on a dedicated server. The application is configured with PrimeOCR, PrimeView, and PrimeVerify, using six (6) voting OCR engines.
- We propose to use PrimeRecognition, together with the UF DLC Zoning application, to establish a control set of files for text conversion against which to bench-mark vendor text product.
- N.B. Our vendor, iArchives uses similar software with multiple OCR engines. Rather than voting and selecting the best fit, its OCR application provides alternate selections parenthetically following the best fit. In order to be compliant with Library of Congress specification for this project, we have instructed the vendor to set this feature off.

#### **RecordNow MAX**

- RecordNow MAX is the CD/DVD burning application of choice by the Digital Library Center.
- It creates MD5 checksums prior to burn and verifies the burn against stored checksums to ensure accurate burn.
- All data (image, text, etc.) is burnt by the DLC to gold based media, whether CD or DVD, using Plextor writers.

## **Appendix C**

### ***Hardware & Software for Quality Control Inspection***

(continued)

#### **RecordNow MAX** (continued)

- To ensure that disk spin cycles are not detrimental to burn quality, CDs do not burn above 8X and DVDs do not burn above 4X.

#### **UF DLC FileSort Application**

- FileSort is the DLC application that again calculates the MD5 checksum of files archived to CD/DVD and stores it together with other file information (name, size, format, version, creation date, write method, media etc.) either extracted from the file header or supplied by the application's user.
- FileSort saves this information in an independent Microsoft SQL database, backed-up nightly.
- FileSort acts on stored information to queue and trigger archive maintenance: inspection and migration. And, it can be used to assist in the inspection process generating new MD5 checksums for long-stored CD/DVDs and comparing them to stored MD5s generated when the file was originally archived.

#### **FCLA MXF Client** or UF DLC interface based on the client's DTD and controls

- The Florida Center for Library Automation (FCLA) MXF Client is a METS compatible file exchange format used to ship file/metadata packages between the Digital Library Center, where they are created, and FCLA where they are both deployed and, again, archived.
- In this project, we propose to use the Client or a new interface now being programmed by the DLC programmers for FCLA as a replacement to the MXF Client as a method of shipping packages between us.
- Packages shipped to FCLA will be reviewed for quality.  
(Copy will be archived there and another copy will be mounted in the planned PALMM Florida Newspapers Collection, before being sent on to the Library of Congress for deployment in the National Digital Newspaper Project.)

#### **Microsoft Office Suite**

- The Microsoft Office 2003 suite will be used to generate Word reports, Access database records/tables, Excel spreadsheets, and PowerPoint presentations as necessary.

#### **Adobe Acrobat Exchange Professional**

- Adobe Acrobat will be used to open and review the quality of vendor PDFs and their hidden text.
- Adobe Acrobat also will be used to generate PDF files.

## **Appendix C**

### ***Hardware & Software for Quality Control Inspection***

(continued)

#### **Microsoft SQL** and other data control and programming software

- Microsoft SQL will be used to generate SQL records/tables as necessary.
- Microsoft SQL underpins the majority of our internal data stores, including those used by the Tracking Database and the FileSort application
- Systems programmers assigned to the Digital Library Center support use a variety of other Microsoft programming tools, including: the .NET framework v1.1 and Microsoft Studio .NET Professional. The majority of programming is done in #C for the .NET framework, that a number of applications are programmed in PERL.