Robotic Book Scanning
at the
Stanford University Libraries & Academic Information Resources:
Report on the Status of
Digitization Facilities and Services for Bound Library Materials
7 May 2003

Introduction

The Stanford University Libraries & Academic Information Resources (SUL/AIR) recently acquired a robotic page-turning and scanning device for the mass digitization of bound print materials. Called the Digitizing Line (DL), this book scanning device is the centerpiece of SUL/AIR's broad array of on-campus digitization capabilities, and an integral part of the libraries' Digital Library Program (DLP).

The implementation of a new book scanning lab that houses the Digitizing Line is part of the DLP's effort to increase online access to the vast intellectual resources of the University Libraries. The mass digitization of the libraries' bound print collections is but one component of that effort. The DLP has developed a similar lab for the mass digitization of unbound materials, which not only serves the needs of local scanning projects, but has also been used as a mobile lab for scanning valuable collections that reside in remote locations. Furthermore, the DLP is engaged in ongoing efforts to license and purchase digital content (such as e-journals and electronic texts) that support the teaching and research mission of the University.

The following describes the features of the DL, its supporting hardware and software, and the status of SUL/AIR's implementation of a new high-volume book scanning lab.

The Digitizing Line (DL)

The DL, manufactured by 4DigitalBooks in St. Aubin, Switzerland (http://www.4digitalbooks.com) is a robotic book scanner, which produces high quality digital images of bound print materials at throughput rates as high as 1160 pages per hour. The page-turning device is designed to enable automated scanning of a wide range of book sizes, from paperbacks and pamphlets as small as 5 ¾ inches long by 4 $^{3}/_{8}$ inches wide, to oversize bound volumes as large as 24 inches long by 15 $^{5}/_{8}$ inches wide. It can process a wide variety of book structures, including both paperbacks and hard covers, with adhesive, sewn or stapled text blocks. It can also process books with a variety of paper types, including thin, medium or thick paper, newsprint, uncoated paper and coated paper. Equipped with digital cameras manufactured by I2S (http://www.i2s-bookscanner.com/), the DL produces high quality black and white, grayscale and color TIFF images at resolutions of up to 600 dpi.

Supporting Hardware and Software

The DL has been installed as part of a new book scanning lab in Green Library, and is supported by a comprehensive hardware and software system. Designed in partnership with the imaging hardware and software firm Image Access of Boca Raton, FL (http://www.imageaccess.com/), this system allows for the manual creation and automated capture of descriptive, administrative

and technical metadata, and the creation of derivatives for online access in JPEG, image-only PDF, searchable PDF, and plain text formats. The workflow engine includes modules for manual entry of descriptive metadata, automatic cropping, splitting and image processing of open-book scans, manual review and treatment of images, uncorrected conversion of images to text (via optical character recognition or OCR), and creation of searchable PDFs and text derivatives. The software allows operators to associate chapters or other subsections of a book with ranges of page images for the creation of PDF bookmarks or encoded texts. Finally, the system automatically captures comprehensive technical metadata for all master TIFF images and derivative files.

Preservation and Conservation Oversight

Upon its arrival at Stanford in October 2002, the DL underwent a rigorous testing phase, involving hundreds of different books, overseen by the libraries' Media Preservation unit.[1] The Media Preservation unit was charged with defining the scope of materials that are suitable for scanning with the DL, and the degree of stress that the automated page-turning process places on bound items of different sizes, materials and structures.

In February of 2003 the DL underwent a five-day long acceptance test, during which the Media Preservation unit confirmed the functionality, durability, page-turning rates and safety of the DL for a wide range of book structures and materials. Staff observed that the stress the DL places on most bound structures is less than the stress accumulated by the one-time face down photocopying of an item. In the process of scanning with the DL, the book is held face up and pressed gently against a glass platen. A small potential exists for a page to be folded or slightly crinkled during the robotic page-turning process. If this happens, an unusual occurrence, the DL is stopped immediately and the page is flattened by hand. The fold or crinkle is typically not a hard crease. Operators have been trained to adjust settings on the DL to reduce the occurrences of folds or other damage. Presently, the DL is not used for the scanning of rare, damaged or brittle materials.

Current Production Status

Upon successful completion of the acceptance process, the DLP began a series of pilot projects to test the DL in a production environment, and help refine the software and processing workflow of the book scanning lab. The two primary pilot projects involve the scanning of titles published by Stanford's Center for the Study of Language and Information (http://www-csli.stanford.edu/) and works for the Medieval and Modern Thought Text Digitization Project (http://www-library.stanford.edu/depts/ssrg/medieval/standish.html).

These two projects involve bound print materials that fall well within the scope of acceptable materials for the DL, and represent best-case scenarios for achieving high scanning throughput and image processing rates. Books in these projects are typically hard cover with sewn-through-

[1] Media Preservation is dedicated to assuring continued access to media materials in SUL/AIR collections; among other endeavors they run an active digital imaging program to enhance access to rare materials from the libraries' holdings. Staff possess expertise in imaging technology and preservation standards for digital imaging. In addition, staff in Media Preservation have extensive experience in book conservation techniques and a detailed knowledge of early through modern book structures.

the-fold text blocks, and are between 5 by 8 inches and 9 by 12 inches in size. They are being scanned at resolutions of 300 and 400 dots per inch (DPI), and are being converted to searchable PDF Image + Text files for online access.

While the specified page-turning rates of the DL can reach up to 1160 pages per hour, actual rates are determined by a book's dimensions, the resolution at which it is scanned, the time taken for occasional operator interventions, and the loading and unloading process. Preliminary results from the pilot projects show that for average sized books scanned at resolutions of 300 DPI or 400 DPI, output rates range between 500 and 600 pages scanned per hour. Including metadata entry and manual review of image quality, the average 300 page book can be scanned and converted to a searchable PDF with approximately 40 minutes of operator attention.

The automated processes of image treatment and OCR, which require no operator intervention, can take well over an hour for a 300 page book. However, these processor-intensive and time-consuming stages occur on unmanned workstations in parallel with other stages, causing no delays in the workflow. For books with more recent imprint dates, the accuracy of the searchable text generated by uncorrected OCR has been measured at close to 99% character accuracy, which is typically sufficient for searching PDF's in which the page image is displayed.

Once the pilot projects and resultant refinements in the workflow and image processing tools are completed in early May, the first large-scale project planned for the new lab is the digitization of approximately 2,500 books published by the Stanford University Press. Several other projects are planned for the lab, including an initiative to digitize the most heavily circulating books in the libraries' collection published prior to 1923, or otherwise in the public domain. In general, responsibility for the selection of content for future digitization using this new lab will rest with the libraries' collection development staff. The libraries will make book digitization services available to faculty to support their teaching and research efforts. As the scale of book digitization projects and services grows, the DLP is considering plans to operate the lab in multiple shifts, up to 16 hours a day.

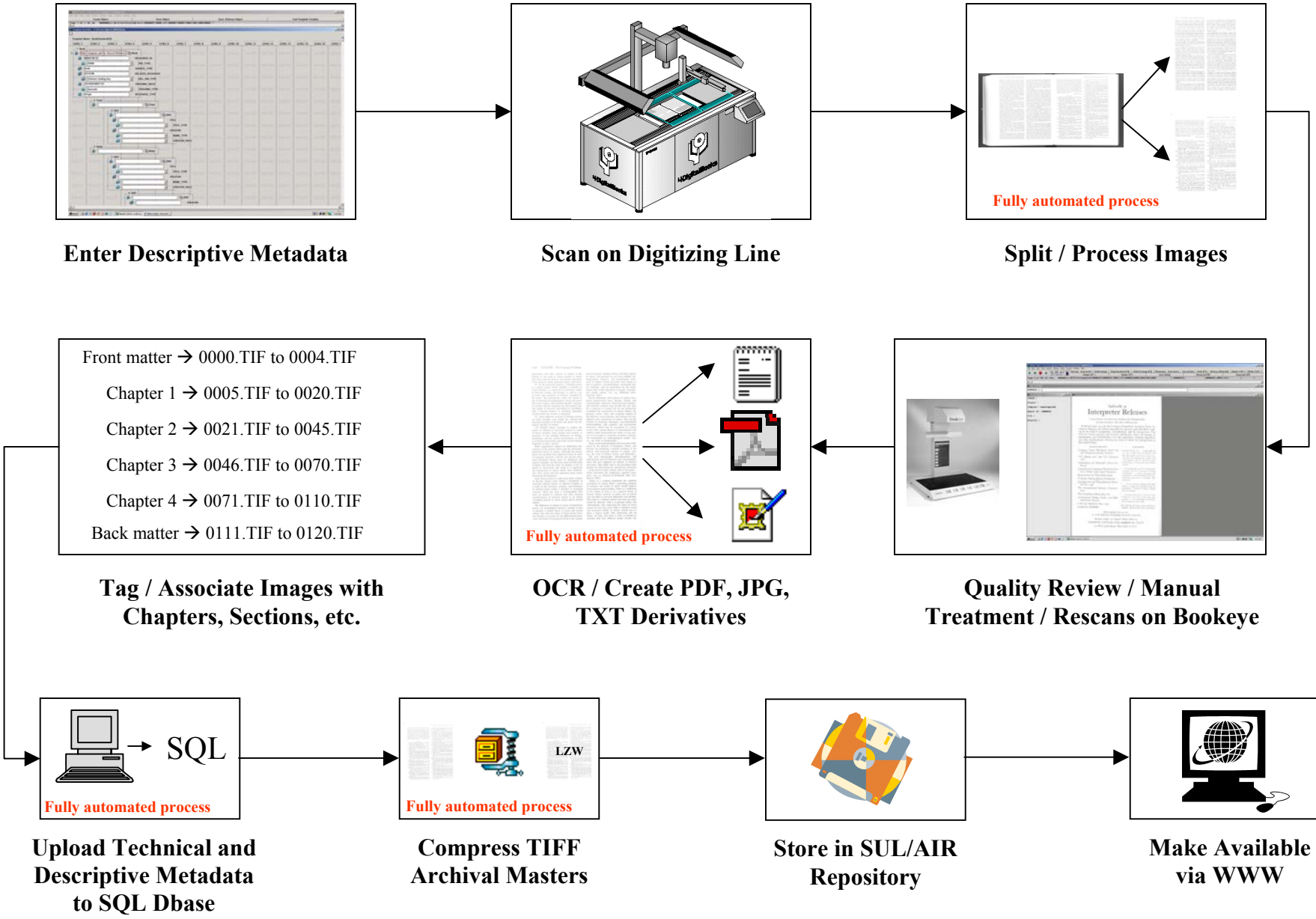
Architecture and Workflow of the Digitization Lab

Scanners have become almost as common as fax machines and photocopiers, but ordinary book scanning technology does not scale well to thousands (or millions) of pages. The DLP required an efficient and highly integrated digitization system consisting of hardware, software and workflow management processes that could fully exploit the unique capabilities of the DL. SUL/AIR's experience with large digitization projects taught us the need for a highly integrated system that manages the process of scanning, metadata creation, image clean-up, quality control, and creation of derivatives for web delivery. Such a system would need to be flexible enough to simultaneously manage multiple projects with both a diversity of materials (books, journals, newspapers, manuscripts, unbound materials) and end goals (preservation, searchable web access, print-on-demand). The system would also need to seamlessly feed content to the libraries' digital repository to ensure the preservation of the content for years to come.

With these goals in mind the DLP worked closely with Image Access and 4DigitalBooks to develop an infrastructure that would support the scanning activities of both the DL, and the libraries' other digitization facilities. The system was developed using Image Access' imaging and workflow management software, Bscan. The workflow design of the digitization process for

the DL is illustrated in **Figure 1**. The workflow utilizes both automated and manual processes, and is modular to allow for customization from project to project, and integration of new image processing technologies as they evolve. The following section describes in some detail the steps in the workflow to take a book from its selection for digitization to on-line access.

Figure 1: Scanning Workflow for SUL/AIR Digitizing Line



Digitization Workflow using the SUL/AIR Digitizing Line

Pre-indexing/Metadata Entry

Once a book (or batch of books) is selected for digitization, it is inspected for physical integrity and suitability for scanning on the DL. The first stage in processing a book is the creation of a record in the project database and the capture of descriptive metadata for the item. For items with MARC records in the SUL/AIR online catalog, this stage might consist of simply keying the title of the item and scanning the barcode on the book itself for future linkage to an existing catalog record. For items without pre-existing cataloging, the Bscan interface allows an operator to manually create a record.

The Bscan interface also allows an operator to describe the hierarchical organization of an item, similar to the process of encoding a text. Operators choose templates appropriate for a specific content type (such as book, journal, newspaper), create a structural map of the item, and type the titles and other labels for different sections. For example, for a book, this might involve typing the titles of parts, chapters, front matter and back matter. For a journal, this might involve typing information about issues in the volume, and article titles and authors. This process sets the stage for associating (or tagging) ranges of images to the different sections of a book.

Upon completion of this process, which can take an operator as little as one minute, a record is created in a database which assigns a unique object ID to the item to be scanned.

Scanning

The next stage is scanning the book. Trained operators load the book onto the DL, and use a simple touch-screen interface to adjust settings that assure the safe and effective scanning of the particular book structure. While the DL has many automatic features to measure the weight, thickness and dimensions of a book, operators can manually adjust ten different settings to accommodate many of the idiosyncrasies of bound structures.

Using the barcode tracking sheet printed during the pre-indexing stage, the operator sets the object ID as the filename prefix for all images scanned for this particular book. This links the database record to the images that are created during the scanning process.

The book is then scanned, each image capturing facing pages of the open book.

Page Splitting and Automated Image Processing

After the scanning process is complete, the single image files containing both left and right pages are split into two separate page images. This completely automated process crops any black space or other “noise” around the fore-edge of the book, and splits the image into a clean left and right page. While effective, this process is not perfect, meaning operators may need to do some additional cropping and image cleanup at a later stage. However, the automated process has proven useful in minimizing the need for manual cropping.

During this stage, additional image treatment processes, such as de-skew and "noise" removal, are possible, and are applied on a project by project basis.

Manual Review and Image Cleanup

Once the images are split and processed, they automatically appear on the screen of a quality control workstation. Quality control operators review every page in a book, checking that all images were scanned, correctly cropped and split, and meet specified quality standards. This Bscan module has manual tools for deleting, rotating, cropping and de-skewing images, and is integrated with Adobe Photoshop to allow operators to perform more sophisticated image treatment and review. The quality control workstations are also connected to a Bookeye planetary scanner (http://www.bookeyeusa.com/), which allows operators to manually rescan pages if necessary, or insert scans of fold-outs or pages the robotic scanner could not capture.

The manual review workstation also allows operators to choose the format of the derivative files that will be used for online access. For example, depending on the quality of the original images (and project specifications), the operator can choose to create black and white PDF files as opposed to grayscale.

Creation of Derivatives for Online Access

Once the quality control operators confirm that images meet specified quality standards, the batch is automatically sent to a bank of workstations dedicated to converting images to text using OCR. In our current process OCR is completely automated and does not involve the use of human operators to correct errors in text conversion. The accuracy of OCR varies greatly with the quality of the original printed page and scanned image. The extraordinarily quality of the images created by the I2S camera appear to yield high accuracy in our text conversion process. After OCR is complete, derivative files such as image-only PDF, searchable PDF, JPG and ASCII text are created.

Association of Images with Book Subdivisions

An optional stage of the process is the association of images with the different parts of a book or bound item. An operator links ranges of images to parts, chapters, or other logical subdivisions within a book. With this information, bookmarks can automatically be inserted into PDF files, separate PDFs can be created for each chapter, or the text created during OCR can be automatically encoded with TEI or some other XML encoding scheme.

Compression and Metadata Export

In a final, automated stage of the lab workflow, the master TIFF image files are compressed and the metadata is uploaded to a relational database. In addition to the descriptive metadata created in the early stages of the workflow, the software automates the creation of technical metadata for all files created, as well as the scanning process itself. Technical information is captured on the file size and dimensions of individual images, versions and specifications of all file formats,

details on all equipment and software used during the capture process, and information on times, dates and operator names for each stage in the workflow.

All of this metadata is exported to a relational database that can be viewed and edited using a simple web interface. The web interface provides a mechanism for tracking the progress of digitization efforts, and for correcting or enhancing metadata. The database was designed in such a way that all metadata can be easily transformed to XML for permanent storage in the libraries' digital repository.

Preservation and Access

Master TIFF images, derivatives created for online access, and all associated metadata are permanently stored in the Stanford Digital Repository, SUL/AIR's archive for long-term digital preservation. At this point, the content is ready to be made available to users for discovery and online access in the SUL/AIR web environment.

Contacts

Project managers at the Stanford University Libraries/Academic Information Resources:

Stuart K. Snydman
Digital Library Projects Manager
Stanford, CA 94305-6067
Tel: 650-723-4223
Email: snydman@stanford.edu
http://library.stanford.edu/depts/dlp/

Catherine A. Aster
Head, Media Preservation
Stanford, California 94305-6004
Tel: 650-725-4042
Email: caster@SUL/AIRmail.stanford.edu
http://library.stanford.edu/depts/pres/mediapres/

For more information about the 4DigitalBooks Digitizing Line contact:

Ivo Iossiger
President & CTO

4DigitalBooks™ - ASSY SA
Fin de Praz 22, CP130, 2024 St-Aubin, Switzerland
Tel: +41 / 32 835 57 75
Fax: +41 / 32 835 57 76
Email: directors@4digitalbooks.com
http://www.4digitalbooks.com/

For more information about Bscan imaging workflow software and the Bookeye planetary scanner, contact:

Kevin Deitch
Account Manager

Tim Winn
Manager of Technical Services

Image Access, Inc.
543 NW 77th Street, Boca Raton, FL 33487
Tel: 561-995-8334
Fax: 561-995-8036
Email: bscan@imageaccess.com
http://www.imageaccess.com/