Scanning to PDFA: Building a digital collection for access AND preservation

Gail Clement[1], Derek Halling, Nancy Burford, Esther Carrigan[2], and Heather K. Moberly[3]

[1] Digital Services & Scholarly Communication, Evans Library, Texas A&M University; [2] Medical Sciences Library, Texas A&M University; [3] Oklahoma State University Libraries

PROJECT OBJECTIVES

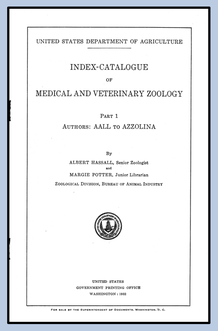
✓Improve digital open access to a significant resource for research and education in the veterinary sciences

✓Preserve valuable and rare library materials for the long term

✓Increase staff knowledge and experience in developing and carrying out digital reformatting projects

✓Work collaboratively with partners within and outside the Institution

✓Develop and document effective procedures and workflows for use in other digitization projects


UNITED STATES DEPARTMENT OF AGRICULTURE

INDEX-CATALOGUE OF MEDICAL AND VETERINARY ZOOLOGY

PART 1
AUTHORS: AALL to AZZOLINA

By
ALBERT HASSALL, Senior Zoologist
and
MARGIE POTTER, Junior Librarian
ZOOLOGICAL DIVISION, BUREAU OF ANIMAL INDUSTRY

UNITED STATES GOVERNMENT PRINTING OFFICE
WASHINGTON : 1910

Title page from one volume in the historic serial to be digitized


Cartload of volumes ready to be scanned

BACKGROUND

As part of a demonstration project to encourage the digitization and preservation of veterinary grey literature, the Texas A&M University Medical Sciences Library has partnered with Oklahoma State University Libraries to digitize the *Index-Catalogue of Medical and Veterinary Zoology*, a multilingual periodical published by the US Government Printing Office. This historical compendium of the parasitological literature is a key resource of importance to researchers in re-emerging diseases and global animal health. The compilation of content began in 1892, and resulted in over 100 separate publications comprising over 20,000 pages.

With generous grant support from the National Library of Medicine, the Library has digitized 67 publications as of March 10, 2010. These are being made openly accessible via the Texas A&M Digital Repository and will be preserved in the TDL PresNet system. Digital reformatting as PDF/A files provides a single file for each volume.

PRODUCTION REQUIREMENTS & SOLUTIONS

Characteristics of the source materials led to a specific combination of approaches for capture and conversion, as follows:

CHARACTERISTIC	APPROACH	DECISION
Paper soft-bound materials eligible for flattening	Scan volumes as page spreads on high quality flatbed scanner with oversized dimensions	EPSON Expression 10000XL flatbed scanner
All content is textual (no images) with average font size equal to 1 mm	Capture smallest significant detail as represented by finest strokes sampled in characters from various alphabets	Benchmarking based on several strokes: •cross-hash in 'e'; •spacing in dots of umlaut; •diagonal stroke in Cyrillic 'shch' Optimal resolution = 400 dpi
Discolored,, low contrast paper with faded ink	Scan in grayscale or color mode to capture tonal variations	8 bit grayscale sufficient without overloading OCR
Textual content represented in multiple languages and alphabets	Optical Character Recognition (OCR) software must recognize and encode other languages	ABBY FineReader 10 selected for its support of mutlilingual texts
Variability in page layouts, with single and double columns	Optical Character Recognition (OCR) software must handle zoning effectively	ABBY FineReader 10 selected for its auto-zoning feature
Complexities of textual content	Use uncorrected OCR	Deliver to users as PDF with uncorrected text behind page images
Format suitable for both Web delivery and digital archiving	Save and deliver as PDF/A-1b (ISO 19005-1:) ….	Adobe Acrobat 9.0 selected for its support for authoring, converting, and validating PDF/A files
Easy, free access to files online for users worldwide	Make available through Institutional Open Access Repository with harvest via OAI to Internet Search Engines	Upload to TAMU Digital Repository (D-Space vers.1.5.x)
Long-term preservation of digital files	Files in TAMU Repository are backed up and stored on redundant servers offsite	Archived in Texas Digital Library PresNet (currently under development)

QUALITY ASSURANCE MEASURES

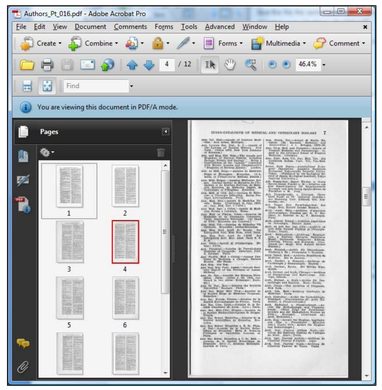
✓Standard page dimensions and aspect ratios preserved at scan time

✓Page images visually inspected for completeness and proper ordering in Windows Explorer/ABBYY

✓Page spreads split and de-skewed as a batch process in ABBYY

✓Recognized (uncorrected) text saved as PDF with text behind page images in ABBYY.

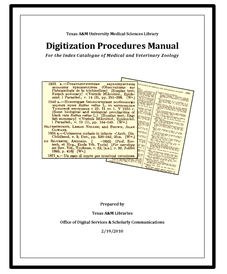
✓PDF converted to PDF/A-1b format and validated in Acrobat PreFlight.



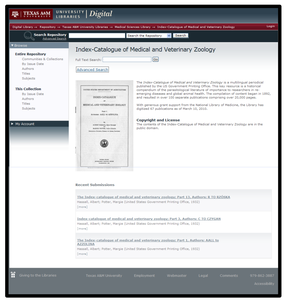
Volume scanned, OCR'd, and converted to PDF/A-1b format.

PROJECT OUTPUTS & OUTCOMES


NEW CAPACITY in the form of equipment and training for digital reformatting of special library materials



ENHANCED CAPACITY in the form of practices and procedures for digital capture and conversion



IMPROVED ACCESS for users worldwide through the digital republishing of rare and fugitive literature.

SIGNIFICANT RETURN ON INVESTMENT achieved through COLLABORATION. Texas A&M Medical Libraries partnered with Oklahoma State University Libraries to identify and consolidate source materials and secure grant funds for scanning equipment and project technicians. In-house expertise in digital collections development and existing digital repository infrastructure were leveraged to produce and sustain the digital collection over time.

Image created by Wordle (http://www.wordle.net/).

RELEVANT RESOURCES

ABBYY FineReader 10.0, http://finereader.abbyy.com/

Adobe Acrobat Pro 9.0, http://www.adobe.com/products/acrobatpro/

International Organization for Standardization, ISO 19005-1: *Document Management - Electronic document file format for long term preservation - Part 1: Use of PDF 1.4* (PDF/A-1)

Kenney, Anne R. and Stephen Chapman. *Digital Imaging for Libraries and Archives*. Ithaca, NY: Cornell University Library, 1996.

National Information Standards Organization, NISO. *A Framework of Guidance for Building Good Digital Collections,* 2007. http://framework.niso.org/node/5

Oklahoma State University Libraries, http://www.library.okstate.edu/

Texas A&M Libraries, http://library.tamu.edu/

Texas A&M Digital Repository, http://repository.tamu.edu/

Texas Digital Library PresNet, http://www.tdl.org/2010/04/presnet-development-sprint-underway/

FOR FURTHER INFORMATION

Gail Clement
Associate Professor & Outreach Librarian
Digital Services and Scholarly Communication
Texas A&M Libraries, College Station, TX 77843
Email: gclement@tamu.edu Phone: 979.862.1635

Nancy Burford
Associate Professor and Resources Management Librarian
Medical Sciences Library
Texas A&M Libraries, College Station, TX 77843
E-mail: nburford@tamu.edu Phone: (979) 845-1820