Automated Digital Processing at the Bentley Historical Library

Michael Shallcross Nancy Deromedi
Bentley Historical Library
1150 Beal Avenue
Ann Arbor, MI 48108-2113 U.S.A.
shallcro@umich.edu deromedi@umich.edu

Abstract

Archival processing in the digital era requires traditional steps such as appraisal, arrangement, and description in addition to procedures that ensure the authenticity, integrity, and security of content. Given the labor-intensive nature of manual procedures, the Bentley Historical Library's Digital Curation Division wrote a series of scripts that call various applications and command line utilities and thereby automate key steps in the ingest and processing of born-digital archival materials.

1. Institutional Context

Established in 1935 by the University of Michigan Regents, the Bentley Historical Library serves as the official archives of the university and documents the history of the state of Michigan and the activities of its people, organizations and voluntary associations. The library has successfully managed and preserved digital content since the 1997 accession of former University President James J. Duderstadt's digital desktop. Given the steep increase in born digital and digitized content accessioned by the library in recent years, archivists have sought more efficient and standardized processing procedures. The Andrew W. Mellon Foundation-funded MeMail Project (2010-2011) provided the library with resources to establish a workflow and corresponding policies for the ingest and processing of archival email, but a similar solution was needed for mixed digital content (i.e. Office documents, PDFs, audio and video files, images, etc.). Archivists in the library's Digital Curation Division have advanced the work of the MeMail Project in developing the AutomatedProcessor (or AutoPro), a series of inter-dependent scripts that automates key steps in preparing digital content for long-term preservation and access.

2. Digital Processing as a Concept and Approach at the Bentley Library

Archival processing in the digital era requires traditional steps such as appraisal, arrangement, and description in addition to procedures that ensure the authenticity, integrity, and security of content. "Digital processing" therefore corresponds to the "generate AIP" function of the Open Archival Information System (OAIS) Reference Model's Ingest entity. After a Submission Information Package (SIP) has been assigned an accession record, digital processing permits archivists to assume intellectual control, establish the integrity of materials, and perform preservation events (i.e. scans for viruses and personally identifiable information, conversion to preservation formats, recording of descriptive and technical metadata, etc.) that transform the SIP into an Archival Information Package (AIP). Bentley archivists initially developed a manual workflow with more than 40 discrete steps that required the operation of numerous stand-alone applications and saving tool output to various log files. In addition to being highly labor intensive and introducing numerous opportunities for operator error, this approach was daunting for staff without technical expertise. Given these challenges, the Digital Curation division developed AutoPro to fulfill two goals: (1) to make digital processing more efficient by automating key workflow steps and (2) to reduce technical barriers and thereby permit archivists to focus their energies on the traditional archival functions of appraisal, arrangement, and description.

3. Automated Processing: an Overview

AutoPro is comprised of 33 Windows CMD.EXE shell scripts that move content through an 11 step workflow and thereby simplify the operation of more than 20 applications and command line utilities. In addition to providing a framework to guide archivists through the workflow, AutoPro tracks the current processing status, generates log files for all operations, and records PREMIS preservation metadata that will be stored alongside the processed content in a preservation environment. The Windows Command Prompt and Explorer windows function as the main interfaces. Archivists must approve the successful completion of each step and may stop at any point in the workflow and resume their work at a later time.

Immediately after content is accessioned and deposited in the Bentley Library's interim repository (a secure Windows file server), AutoPro runs a virus scan (the University of Michigan employs Microsoft Forefront Endpoint Protection on all work stations) and creates a working backup so the SIP can be restored in case of an error or accidental data loss.

AutoPro then searches for archive files (.ZIP, .TAR, .RAR, etc.); if any are found, a script employs 7-Zip [1] to extract the contents to a directory named after the archive file, with the original file paths preserved. After verifying the extraction's success, AutoPro moves the archive file to a separations directory and records the operations in a log file. The newly extracted content is then searched for additional archive files, from which the contents are extracted, if necessary.

At this stage, AutoPro uses ReNamer [2] to replace spaces and non-alphanumeric characters with underscores in folder and file names and log the original and new names in a comma-separated values (.CSV) file. Next, AutoPro searches for files with missing or user-supplied extensions, identifies correct extensions with the TrID File Identifier utility [3], and verifies the results with DROID [4]. AutoPro preserves the TrID output (which includes a report on likely file types, based upon the target file's binary signature) in a log file. If an extension is successfully identified, the original and new filenames are recorded in a .CSV file.

In transforming the SIP to an AIP, the Bentley Library relies upon file format conversion as a primary preservation strategy. Based upon the Library of Congress's work on the "Sustainability of Digital Formats" [5] and documentation from the Florida Center for Library Automation and other peer institutions, the library has identified a number of at-risk (i.e. proprietary or potentially obsolete) file formats and developed conversion pathways to sustainable formats with various open source and freeware tools. AutoPro searches for these at risk formats (based upon extension) and then employs the following tools (with digital media and target format in parentheses): ImageMagick (raster images to .TIFF) [6], Ghostscript (.PS and .PDF to .PDF/A; an Adobe Acrobat Preflight droplet verifies if the original PDF meets PDF/A specifications) [7], Inkscape (vector images to .SVG) [8], ffmpeg (audio to .WAV; video to MP4 with H.264 encoding) [9], Aid4Mail (email to .MBOX) [10], and Microsoft Office File Converter (Office files to Open Office XML) [11]. These preservation versions are stored alongside the original and denoted by a suffix consisting of '_bhl-' and the CRC32 hash of the original file (i.e. oralHistoryProject_bhl-0fbc2cc7.wav). AutoPro also creates a log of all file conversions, including the original and new filenames, timestamp, and conversion software.

In order to protect the identities of record creators and limit its exposure to risk, the Bentley Historical Library has established policies in regard to personally identifiable information (PII) such as credit card numbers and U.S. Social Security numbers. AutoPro thus employs Identity Finder DLP Endpoint [12] to scan for PII. Archivists then use the Identity Finder interface to verify search results and—if true positive hits are found—redact the PII (from Open Office XML and plain text files) or assign appropriate access restrictions to the content.

Archivists then proceed to a more in-depth appraisal and arrangement of content. AutoPro loads data visualizations (such as the distribution of file extensions, date range of content, relative size of directories, etc.) produced by TreeSize Professional to better characterize and launches Quick View Plus (a file viewing program) to rapidly review a wide range of file types for description in finding aids [13]. While reviewing content with Quick View Plus or the Windows Explorer, archivists use a batch file in the right-click context menu to remove superfluous files or folders to a separations directory. Every effort is made to retain the original order of materials, but archivists may group unorganized content in directories or package content in .ZIP files to simplify the management and storage (with such actions recorded in log files). Once the arrangement is established, AutoPro calls DROID to extract technical metadata and generate an MD5 checksum for all content (including files in .ZIP archives). Archivists then use AutoPro to identify series and add descriptive and administrative metadata about the materials; the resulting XML file is used to deposit unrestricted content in Deep Blue, the University of Michigan's DSpace repository. Finally, AutoPro employs BagIt to transfer a copy of all material to a secure dark archives [14]. Once this step is accomplished, AutoPro cleans the processing directory and temporary files and archivists record the completed digital deposit in the Bentley's collections management database.

4. References

NOTE: all URLs successfully accessed 28 August 2012.

[1] 7-Zip is an open source file archiving application. For more information see http://www.7-zip.org/.

[2] ReNamer is a freely distributed file renaming tool. For more information, see http://www.den4b.com/?x=products&product=renamer.

[3] TrID is a freely distributed utility that identifies file types based upon a library of over 4,800 binary signatures. For more information, see http://mark0.net/soft-trid-e.html.

[4] DROID is a file identification tool developed by the National Archives (U.K.). For more information, see http://droid.sourceforge.net/.

[5] For more information on the Library of Congress's "Sustainability of Digital Formats," see http://www.digitalpreservation.gov/formats/index.shtml.

[6] ImageMagick is an open source raster image editor. For more information, see http://www.imagemagick.org/script/index.php.

[7] Ghostscript is an open source interpreter for the PostScript language and PDF documents that may be used to convert the latter documents to PDF/A. For more information, see http://www.ghostscript.com/.

[8] Inkscape is an open source vector graphics editor. For more information, see http://inkscape.org/.

[9] ffmpeg is freely available software used for audio and video recording and conversion. For more information, see http://ffmpeg.org/. AutoPro utilizes a Windows build available from http://ffmpeg.zeranoe.com/builds/.

[10] Aid4Mail is a proprietary email conversion program. For more information, see http://www.aid4mail.com/.

[11] Microsoft File Convertor is part of the freely available Office Migration Planning Manager. For more information, see http://www.microsoft.com/en-us/download/details.aspx?id=11454.

[12] Identity Finder Data Loss Prevention (DLP) Endpoint is proprietary software that can identify potentially sensitive information. For more information, see http://www.identityfinder.com/us/Business/IdentityFinder/EnterpriseClient.

[13] TreeSize Professional is a proprietary hard disk space and file manager and Quick View Plus is a file viewing utility. For more information, see http://www.jam-software.com/treesize/ and https://avantstar.com/, respectively.

[14] BagIt is part of an open source set of transfer tools developed by the Library of Congress. For more information, see http://sourceforge.net/projects/loc-xferutils/.