Winter 2015

Time is on my side : harnessing the power of automation for efficient archival workflows.

Heather Fox
University of Louisville, heather.fox@louisville.edu

Sarah Dorpinghaus
University of Kentucky, sarah.dorpinghaus@uky.edu

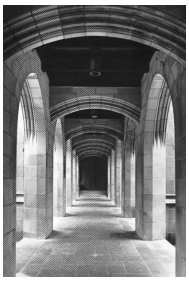
Original Publication Information

This article was originally published in *Kentucky Libraries*, volume 79, number 1, Winter 2015.

TIME IS ON MY SIDE: HARNESSING THE POWER OF AUTOMATION FOR EFFICIENT ARCHIVAL WORKFLOWS

BY SARAH DORPINGHAUS
DIGITAL ARCHIVIST, SPECIAL COLLECTIONS RESEARCH CENTER
UNIVERSITY OF KENTUCKY LIBRARIES

HEATHER FOX
ARCHIVIST FOR METADATA & SCHOLARLY COMMUNICATIONS, ARCHIVES & SPECIAL COLLECTIONS
UNIVERSITY OF LOUISVILLE



In the 21st century, the idea of utilizing technology to increase efficiency is nothing new. However, many librarians and archivists lack specialized training in computer programming, file management, and other advanced technological skills. Instead, our professional training has focused on more traditional aspects of librarianship like the reference interview, collection development, and subject analysis. Although these remain important components of serving our users, learning new ways to streamline workflows can free up time we spend on tasks that can be automated, and allows us more time to concentrate on specialized aspects of our work that only we, as trained archivists and librarians, can accomplish. Regardless of institutional resources, simple techniques can make a significant impact on processes and productivity. This article provides examples of how archivists at the University of Louisville and the University of Kentucky have successfully integrated automation techniques into their collection management workflows, and illustrates the impact these techniques can have, regardless of institutional resources. The first will zoom in on how one archivist without a dedicated programming staff saved hours of time using the command line interface (CLI), while the second describes a collaborative effort between archivists and programmers to address a digitization workflow inefficiency.

ONLY THE LONELY

Digital Initiatives is a two-person unit within the University of Louisville Archives and Special Collections library. Although we do not have a dedicated programming staff, we often collaborate with our Office of Libraries Technology on large projects requiring installation of specialized software and data migration. When necessary, we employ technology skills in our own personal arsenals to manage smaller tasks. Faced with wrangling data about an important photograph collection documenting the history of Louisville that existed in multiple locations and applications, and without full-time, dedicated technology support, I used a couple of simple tricks to make a daunting task manageable.

In early 2014, I streamlined the workflow for metadata creation for an ongoing project to digitize and upload images from the Caufield & Shook collection, comprised of roughly 500,000 images from one of Louisville's premier photography studios active in the late nineteenth century through the 1960s. By implementing some little tricks learned on the job, I was able to automate the creation of a master database of item-level records in Microsoft Access, allowing us to track our progress and scan, describe and upload images in this huge collection. Furthermore, tab-delimited text files of the metadata created from this database can be exported into CONTENTdm, our digital content management system, for online display. Understanding how to export and import text files allows the transfer of data between different applications. The delimiter retains the data structure (columns with field names, one row per record) without the encoding particular to the application originally populated by the data.

Scanning of the collection had begun in 2000, but without a systematic workflow. Legacy Word documents, Excel spreadsheets and databases tracking the first images were dispersed in multiple iterations in directories (or file folders) spread across the networked servers, with no centralized means of tracking which images had been scanned and uploaded. The goal was to centralize existing metadata, then find places within the work-

flow where I could take advantage of the student workforce, so I could focus on the parts of the process that required my professional skills.

By taking advantage of CONTENTdm's export function, I exported a tab-delimited text file of existing metadata records from the online collection that was then imported into the blank master database. Records for the over 6000 undescribed scans in the master file directory then needed to be created. Luckily, I had recently learned how to export existing file name and other metadata from a directory into an Excel spreadsheet using the command line interface (CLI). This simple and powerful method of communicating with the computer's operating system using text-based commands obviated the need to cut and paste 6000 filenames into the master database one-by-one.

Most of us regularly use a graphical user interface (GUI) and keyboard shortcuts to complete simple tasks, like cutting and pasting. GUIs rely on images to communicate between the computer and the user. For example, imagine the ribbon of functions across the top of Microsoft Word: the copy function is represented by two sheets of paper side by side, and the past function is represented by a clipboard. Keyboard shortcuts achieve the same ends with combinations of keystrokes: pressing the "control" key followed by the "x" key on a PC cuts the selected text; "control" followed by "v" pastes the text where the user points the cursor. But, how does one complete this task 6000 times in an efficient way? The answer is to learn simple CLI commands.

The command line interface facilitated data extraction from the master directory in a matter of seconds, a task that would have required hours of time, had I been forced to cut and paste filename by filename (see fig. A). Once in a structured format, the data could be manipulated within Excel using the "split cell" and "insert" functions, isolating the information that needed to be retained, and exporting the information into a tab-delimited file. The records could then be ingested into the new master database, removing duplicate records by running an access query, and the master database was complete.

I now regularly use my CLI skills to extract filenames from batches of newly scanned images added to the collection. With roughly

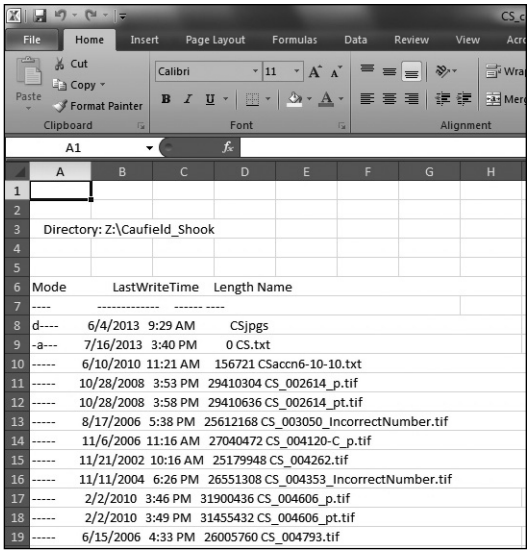


Figure A

498,000 images left to upload, any amount of streamlining helps. Assuming responsibility for my workflow and applying a new technology skill allowed me to "spin straw into gold." Imagine what possibilities exist when teamwork is added to the mix.

WITH A LITTLE HELP FROM MY FRIENDS

The University of Kentucky Special Collections Research Center (SCRC) is fortunate to have several programmers on staff who, to put it briefly, work with code to build tools that help others perform their work more efficiently and effectively. After three years of collaborating with SCRC programmers, two basic tenets have been ingrained into my approach to workflows. First, every task that can be reasonably automated should be automated. A tool that takes three days of development time is worth creating if it saves six days of labor, whether it is staff or student time that is saved. Second, workflows should be continually assessed and revised to increase efficiency. This evaluation can easily be done during recurring departmental meetings. For example, the agenda for SCRC bi-monthly imaging meetings always includes a brief review of weaknesses in our workflows and discussion of potential tools that could reduce the arduousness of a task.

A better understanding of the role of programmers and recognizing the need for automation opened the door to conversations that led to the eventual creation of several important tools utilized in SCRC imaging and digital preservation workflows. To better grasp the impact of these tools, one must understand the context of our imaging work. For

over five years, SCRC has dedicated its imaging efforts to large-scale digitization of manuscript collections. Rather than select and digitize specific pieces, we target collections of high research value and scan every item. All descriptive metadata are pulled automatically from the corresponding Encoded Archival Description (EAD) finding aid, and no further description is created. The goal is to recreate an online version of the brick-and-mortar reading room environment in which a researcher would find him/herself, if working with a collection in person. To achieve this, the digital surrogates are organized in a directory structure that mirrors the physical containers of boxes and folders.[1] Therefore, SCRC imaging specialists are more concerned with the physical arrangement of a collection than the intellectual arrangement, because they start digitization with box 1, folder 1 and work their way through to the final box.

One of the first workflow inefficiencies we addressed through automation is the building of directories. Every box and folder (and sometimes every item) requires the creation and naming of a directory to match the physical container nesting. For example, box 8, folder 2, item 1 of collection 1998ms004 has a corresponding directory named "1998ms004_8_2_1". This requires a lot of clicking and typing for our imaging specialists, and leaves the door wide open for human error. We approached our programmers with this issue and they responded by creating a tool named "Gossamer"[2] that analyzes the .xml finding aid, and uses the EAD-coded containers to build directories in a nested structure matching the physical arrangement of the collection (see fig. B). Finding aids are run through Gossamer before a collection is digitized so that imaging specialists can simply populate the empty directories as they digitize the collection. In addition to saving staff time, Gossamer has eliminated human error while building directories. And, as an unexpected plus, a quick scan of the directories helps detect container numbering errors in the finding aid.

Figure B

Unfortunately, most libraries and cultural heritage institutions do not have programmers on staff to build time-saving tools, but there are other avenues for implementing automation in workflows, even without a programmer. After identifying your needs, start by searching online for existing tools and software, many of which are available at little or no cost. SCRC uses several programs of this ilk to perform simple tasks like batch renaming or converting files. As mentioned earlier in this article, basic command line prompts and regular expressions can also be used to make batch changes to metadata, directory structures, and more. Bertram Lyons, a senior consultant at AVPreserve, recently released "An Introduction to Using the Command Line Interface to Work with Files and Directories" for both Mac and Windows OS that is freely accessible via the AVPreserve website.[3] Seeking assistance from IT staff or other colleagues is another option, especially for those unsure of their technological skills. Professional organizations such as *Code4Lib*, *Digital Library Federation*, and *Society of American Archivists* encourage members to consult with the collective knowledge of the profession via listservs. All of these options can significantly improve existing workflows by automating laborious and repetitive tasks, with just a little time and research to find the right tool for your needs.

WANNA BE STARTIN' SOMETHIN'

Ready to launch your command line interface? Already searched online for "automation tools"? You are well on your way! However, consider the following before searching for and implementing automation tools. First, be sure the correct workflow impediment is identified, so that the desired outcome can be articulated clearly. That is, do not be overly concerned with the means to the end; focus on *what* needs to be accomplished rather than *how* it will be accomplished. Also, this is not the time to let your perfectionist side take over. Utilizing automation tools requires flexibility and the acceptance that there may be some minor errors, in trade for saving significant time. Be prepared to make mistakes… because you will. That being said, it is best to test out a new automation tool on a *copy* of the selected files or dataset. That way you can laugh at your mistakes, rather than gasp in horror or

curse in confusion. And lastly, although many helpful tools exist, some tasks are truly easier and more effective to complete by hand.

These recommendations are not meant to discourage, but rather to pave a smooth(er) road for successful inclusion of automation in your existing workflows. This is an ever-growing community and the tools are continually improving, so jump in and enjoy the ride! Or perhaps we should say "C:\Users\you> chdir hooray_for_automation"!

Sarah Dorpinghaus
sarah.dorpinghaus@uky.edu

Heather Fox
heather.fox@louisville.edu

FOOTNOTES

1 This directory structure and corresponding ingest and digital preservation process was developed by Dr. Michael Slone and Eric Weig, 2011

2 The Gossamer tool was developed by Dr. Michael Slone, 2012.

3 Lyons, Bertram. "An Introduction to Using the Command Line Interface to Work with Files and Directories" *AVPreserve*. Web. Oct. 2014