FEATURE

Born-Digital Preservation: The Art of Archiving Photos With Script and Batch Processing

by Sharon Bradley, Rachel Evans, and Leslie Grove

Vol. 40 No. 5 — July/August 2020

With our IT department preparing to upgrade the University of Georgia's Alexander Campbell King Law Library (UGA Law Library) website from Drupal 7 to 8 this fall, a web developer, an archivist, and a librarian teamed up a year ago to make plans for preserving thousands of born-digital images. We wanted to harvest photographs housed only in web-based photo galleries on the law school website and import them into our repository's collection. The problem? There were five types of online photo galleries, and our current repository did not include appropriate categories for all of the photographs. The solution? Expand our archives photo series structure in Digital Commons, write and run scripts to automate the gathering of image file URLs and basic metadata, and then clean up the spreadsheets to batch load it all into the collection.

> We are very excited about the variety of ways this content can be used by site visitors to paint a more detailed picture of the history of our institution and the school's intellectual life.

IMPETUS FOR THE PROJECT

For background, the UGA Law Library was one of the early adopters of Digital Commons software. Digital repositories were initially promoted as a mechanism to capture the scholarly output of an institution, and our library's initial approach to capturing scholarly output was focused heavily on faculty publications.[1] Relatively early on, we decided to capture all intellectual activity at our institution. This would include any items created by law school departments or about the law school, expanding authorship beyond faculty members. The law library has spent 15 years expanding the repository collection's scope (**digitalcommons.law.uga.edu**). It currently includes a robust set of Georgia historical materials, strategic planning documents, and even archives and special collections items. Next came the photographs.

Initially, the photograph collection in our repository contained only historical images. Prints of speakers and events from the physical archives were scanned and uploaded individually. Photographs of our vast portrait collection and other items from our art collection were added. These digitized images became part of the school's online archives, expanding their accessibility via the repository. The majority of the born-digital photographs in our repository were of recent speakers, used to highlight and support the related entries in our named lecture series. In general, this online photograph collection remained small. The photographs added were all tied to our special collections and limited to what the special collections librarian was able to scan, load, and catalog individually.

Peripherally, the school's office of communications and public relations was taking photographs at many of the events the library was archiving in the repository. These born-digital photographs (from 2001 to the present) were uploaded to the law school's public web server and stored in a photo gallery (**law.uga.edu/photo-gallery-archive**). Over the years, the website's framework evolved, and various iterations of photo galleries were used by that office to present small collections of images to the public following law school events. During this time, the

scope of the events the photographs captured expanded to include student-related tournaments (such as mock trial and moot court competitions), important visits from state judges and Supreme Court justices, and even annual homecoming celebrations attended by prominent graduates of the institution.

Inspired by a session ("Automating Processing and Intake in the Institutional Repository with Python"[2]) at the 2019 CALIcon (Computer Assisted Legal Instruction Conference), the three of us seized an opportunity to strengthen our relationship with both the office of communications and public relations and the IT department by facilitating the archiving of the photo gallery's digital media. The project would preserve the media files and metadata and alleviate the IT team's concerns surrounding migrating that content as it upgraded the CMS. In addition, the files and data would enhance many of the existing collections in our institutional repository.

BORN-DIGITAL PROBLEMS AND SOLUTIONS

The 2019 CALIcon session that inspired us focused on automating repository intake of PDFs with Python. The idea behind that workflow informed our own. The three of us cooked up a similar design for batch-loading data, adapting it to digital media files instead of PDFs.

Since all of the photo galleries were hosted on our own server, it meant we could look directly into the file structure on the server to locate the photos, along with parsing the HTML. We were fortunate that each photo already had its own URL. The HTML (whether "scraped" or read directly from files) would be processed to extract the metadata needed for our record fields. The result was a spreadsheet with four columns, corresponding to the minimal fields for these item records in our repository:

- Title [title]
- Date [publication_date]
- Abstract [abstract]
- URL [fulltext_url]

We were lucky all photo gallery iterations included the information needed for these four required fields. One of the lessons learned early on was that we needed to create a framework in which to load items that was neither too detailed nor too granular. The librarian and archivist in our team of three reviewed the existing galleries and developed an initial revised organization scheme. The programmer would create scripts to provide the data in a tab-delimited file. The librarian would be the human component for spot-checking in Excel and sorting the results into various sheets organized by the existing and new gallery series in the repository.

Our programmer decided to use PHP instead of Python (as advocated at CALIcon), due to her additional experience with that language. She relied on **php.net** as a refresher for syntax rules and **stackoverflow.com** as a troubleshooting guide. It is her opinion that even a novice programmer could approach the task using PHP or any modern scripting language. Because of the variety of tools used when the photographs were originally posted on our website—and the evolution of the website between 2001 and the present—one encompassing solution would not be practical. Our programmer wrote three scripts to deal with the five different types of photo gallery pages that we wanted to harvest.

The digital archives includes both scanned and born-digital photos.

Harvesting so many different gallery formats required several scripts. The color stripe represents the number of each type. Yellow, green, and orange indicate three kinds of HTML formats.

A snippet of harvesting code our programmer used to automatically pull the data and image URLs

We test-drove Google's AI tool for analyzing and indexing photos.

Statistics on the three scripts used to harvest the five gallery types

AI CONSIDERATIONS

Another presentation ("From Concept to Concrete: Teaching Law Students about AI[3]) at the 2019 CALIcon informed us further. The session highlighted the image-processing capability of Google's free AI tools, such as the Vision API (**cloud.google.com/vision**). It is capable of detecting faces and objects, reading text that is found in an image, and making some good guesses about keywords that might be used to describe the image. After testing several images, we determined that, although the results were impressive, they were not as useful as the titles and text that could be fetched from our own HTML.

We did, however, find another use. Google Vision's text-recognition capabilities allowed us to quickly rename the photos taken of our incoming 1L (first-year law) class. A photo of each student holding a sign with his or her name and ID number was followed by the official directory photo. The second photo was named according to the text found in the first photo. Aside from a few amusing anomalies, in which the AI thought the print on a student's shirt was a series of repeating letters, the strategy was a success. It saved us hours of manual work.

THE HARVESTING RESULTS

The organizing and planning communication and workflows between departments, including testing small batches of photos from the scripts in our sandbox Digital Commons site, took a couple of months in fall 2019 and early spring 2020. However, the actual harvesting of the photographs took our programmer only a few days of work. For the numerous photos in gallery sets, a number was added to the end of titles. In all, she retrieved the following:

- 793 photographs with script one (from 2015, 50 blue Drupal-based galleries)
- 1,273 photographs with script two (from 2013 and 2014, 88 red Flash-based galleries)
- 10,014 photographs with script three (from 2001 to early 2013, 465 yellow, green, and orange, HTML-based galleries)

This brought the grand total of individual images for the librarian to 12,080 born-digital media files, from a total of 603 galleries.

COLLABORATING TO DEFINE THE COLLECTION

Of course, not all 12,000-plus images will be added to the institutional repository. Part of our early leg work in coordinating with the office of communications and public relations included developing guidelines that the library could follow for identifying what images we would archive and what images we wouldn't. Particularly in the earliest days of digital photo galleries, there were photosets of more intimate gatherings (for example, a library holiday party from 2002 and an employee's baby shower). For many photosets, it was obvious that the content would not fit the scope of the repository and would not enhance any other existing series. However, there were other subjects that were murkier and that we needed more direction on how to handle (for example, retirement party photographs: Do we keep all employee retirement photos or only prestigious faculty retirements?). We devised a set of categories and subcategories in consultation with the director of the office of communications and public relations (see page 27).

To get to that point, we created a shared Google Sheet for tracking progress of the project overall and reviewing the various galleries to make category assignments. This step was tedious, but it proved especially helpful for the librarian who ultimately did the cleanup, sorting, and batch-uploading of the photographs and metadata. To help manage our expectations and stay ahead of our timeline for August 2020, we set milestones approximately once a month in fall 2019 and early spring 2020 to evaluate our progress. We realized early on that the decisions about collection names would likely change as the project came closer to completion, as would the number of galleries in the repository (either by consolidating similar series or creating new ones).

Luckily, through our work with other areas of our repository, we know that Digital Commons makes this manageable, allowing us to expand or contract the organizational scheme without too much extra effort. The

collect function can help with this work. We also plan to cross-list some of the image content in multiple galleries using the group function. We will certainly be employing the batch-revise function in the coming months to update and reload sets of content to clean up or include new data in fields for multiple items. Additionally, as the project comes to a close, working with so much content over a short period of time has allowed our librarian to become more familiar with the items, our school's typical events, and the related terminology. Moving forward, we will be using this valuable information to construct metadata filters. That way, as new images are loaded individually or in smaller batches, much of the collecting, grouping, and cross-listing work will be automatic based on keywords in titles, abstracts, and tagging fields.

PAINTING THE BIGGER PICTURE

At the time of this writing, our project is nearly complete. All photograph image links and metadata have been gathered into three master spreadsheets, and the librarian is about halfway through the process of removing the rows of images and data we do not want to preserve and batch-loading the ones we are archiving. In the end, we expect that roughly half of the images (about 6,000) will be kept. We will retain a backup for a period of time before the image galleries are gone (just in case we realize we actually do want to keep something our original guidelines did not permit). We also rely heavily on Archive-It (**archive-it.org**), a monthly crawl service that preserves our institution's web content (**archive-it.org/home/ugalaw**) via the Internet Archive (**archive.org**) as a second-tier fail-safe.

We are very excited about the variety of ways this content can be used by site visitors to paint a more detailed picture of the history of our institution and the school's intellectual life. We hope to continue expanding the avenues by which users discover content through linking related records together. We envision a vast web of content across our collections, including books or articles written by faculty members, conference programs where they spoke, and photographs of them interacting with students at a homecoming celebration.

Acknowledgments

The authors would like to thank John Beatty, Jesse Bowman, and Stephan Martone for their inspiring presentations in June 2019. Without their groundwork, this project would not have been possible. We would also like to thank Heidi Murphy (director of communications and public relations at the University of Georgia School of Law) for collaborating with us and the rest of the School of Law web team and Digital Commons team for their continued feedback and motivation throughout this yearlong effort.

RESOURCES

1. Donovan, James M. and Watson, Carol A., "White Paper: Behind a Law School's Decision to Implement an Institutional Repository" (2008). Available online:**digitalcommons.law.uga.edu/law_lib_artchop/15**.

2. Beatty, John, "Automating Processing and Intake in the Institutional Repository With Python" (2019). Available online: **2019.calicon.org/node/1/sessions/automating-processing-and-intake-institutional-repository-python**.

3. Bowman, Jesse and Martone, Stephan, "From Concept to Concrete: Teaching Law Students about AI" (2019). Available online: **http://2019.calicon.org/node/1/sessions/concept-concrete-teaching-law-students-about-ai**.

Sharon Bradley *(**bradley_s@law.mercer.edu**) is the digital and scholarly resources librarian at the Mercer University School of Law Library and former special collections librarian at the University of Georgia's Alexander Campbell King Law Library (UGA Law Library).*

Rachel Evans *(**rsevans@uga.edu**) is the metadata services and special collections librarian at the UGA Law Library.*

Leslie Grove *(**lgrove@uga.edu**) is the senior web developer at the University of Georgia School of Law. The three began this project in July 2019 and aim to complete it in July 2020.*
