Building a digital archive for decaying paper documents

6 January 2020, by Daniel Genkins

Converting aging paper documents to digital archives can be a painstaking effort. Credit: Slave Societies Digital Archive, CC BY-ND

Paper documents are still priceless records of the past, even in a digital world. Primary sources stored in local archives throughout Latin America, for example, describe a centuries-old multiethnic society grappling with questions of race, class and religion.

However, paper archives are vulnerable to flooding, humidity, insects, and rodents, among other threats. Political instability can cut off money used to maintain archives and institutional neglect can transform precious records into moldy rubbish.

Working closely with colleagues from around the world, I build digital archives and specialized tools that help us learn from those records, which trace the lives of free and enslaved people of African descent in the Americas from the 1500s to the 1800s. Our effort, the Slave Societies Digital Archive, is one of many humanities projects that have accumulated substantial collections of digital images of paper documents.

The goal is to ensure this information—including some from documents that no longer exist physically—is accessible to future generations.

But preserving history by taking high-resolution photographs of centuries-old documents is only the beginning. Technological advances help scholars and archivists like me do a better job of preserving these records and learning from them, but don't always make it easy.

An archive in Cuba contains paper treasures that are hard to use and study – even in person. Credit: Slave Societies Digital Archive, CC BY-ND

Collecting documents

Since 2003, the Slave Societies Digital Archive has collected more than 700,000 digitized images of historical records documenting the lives of millions of Africans and people of African descent in North and South America.

Members of the core team, from universities in the U.S., Canada, and Brazil, travel to project sites throughout Latin America, where they train local students and archivists to digitize ecclesiastical and government records from their communities. We give these communities the cameras, computers

and other hardware they need to digitally preserve documents piled in the corners of 18th-century church basements, or about to be discarded by space-crunched municipal archives.

We also teach them a crucial skill for archiving and retrieval: how to create metadata, the descriptive information to help people find what interests them—like whether a document is a marriage certificate or a baptism record, and what year and town it's from. Good metadata allows visitors to the project website to, for example, search for all baptism records from 17th-century Colombia.

From digitization to preservation

Over time, we've gotten much better at digitizing documents. In older images, it's not uncommon to see the photographer's finger straying in from the side of the frame. Some of those older images are stored as relatively low-resolution JPEG files, a format that compresses the image file size by deleting some data when it's saved. Most of those files are still completely legible even when a viewer zooms in, but some are not and will need to be digitized again in the future.

A lot of people get involved, both teaching and learning how to properly photograph documents. Credit: Slave Societies Digital Archive, CC BY-ND

Our more recent preservation adheres to the rigorous standards of the British Library, which funds much of our work. Those images are taken in very high resolutions and stored in multiple file formats including TIFF, which remains the archival standard.

Transforming a collection of digitized images into a true digital archive is a time-consuming and detail-oriented effort. Early in this process, we ran into a curious problem involving photographs taken during our first few digitization efforts. Modern software frequently misinterpreted the orientation of these images, giving us pages rotated 90 degrees to the right or left or even completely upside down. In cases where an entire volume was rotated in the same incorrect way, it could be fixed automatically, but others with a range of errors had to be corrected by hand to let researchers work more easily with the material.

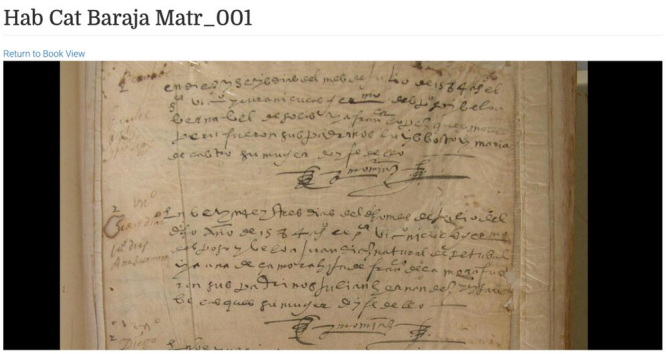
We've also found that data file names can cause problems. Many cameras assign images default names—like DSCN9126.jpg—that aren't useful for figuring out what the pictures are. We have to rename each image in a standard way that indicates how it fits into our collection.

For the time being we've chosen simply to number images sequentially within each volume; another reasonable option would be to prefix each of these numbers with an ID referring to the volume the image comes from.

These aren't major hurdles, but they and others along similar lines take some time to figure out and address properly. But this effort pays off when people hoping to explore the collection have an easier time finding and using our images.

With care, digital preservation can bring new life to crumbling documents. Credit: Slave Societies Digital Archive, CC BY-ND

Our path has led to the cloud, computers in technology companies' massive server-warehouse buildings that we access remotely to store and retrieve information. At the moment, multiple copies of our entire dataset are stored on servers on opposite sides of North America. As a result, we're far less likely to lose our data than at any previous point in the project's history.



If you can read this, you're very highly trained. Credit: The Conversation screenshot of Slave Societies Digital Archive file, CC BY-ND

Where to store them?

Once we've captured the images, we need to save them somewhere.

At present, the Slave Societies Digital Archive collection is close to 20 terabytes—roughly the space needed to store all the text in the Library of Congress.

Few institutions have the resources, personnel or expertise needed to store humanities data at such large scales. Data storage isn't exorbitantly expensive, but it's also not cheap—especially when the data needs to be accessed regularly, as opposed to being stored in a static backup or archival copy.

For many years, the Vanderbilt University Library hosted the data, but we outgrew what that organization could afford. We had been backing up many of our most important records on the Digital Preservation Network, a consortium of universities that pooled resources to fund a reliable digital storage system for scholarly production. But that organization shut down in late 2018 after consulting with each member organization to ensure that no data would be lost.

Opening access

Storing these records in secure systems is another part of the equation, but we also need to make sure that they're accessible to the people who want to see them.

Our documents, typically written in archaic Spanish or Portuguese, are very hard to read. Even native speakers need special training to decipher what they say.

For several years, we've been producing manual transcriptions of some of our most noteworthy records, such as a volume of baptisms from late 16th-century Havana. But that takes 10 to 15 minutes per page—meaning that transcribing our entire collection would take more than 100,000 hours.

Other projects have used volunteers to do similar work, but that approach is less likely to be the

solution for our archive because of the linguistic skills required to read our documents.

We are exploring automating the transcription process using handwriting recognition technology. Those systems need more work, particularly when dealing with centuries-old handwriting styles, but some researchers are already making progress.

We are also looking at ways to identify the people and places mentioned in our records, making them searchable and connecting them to other similar datasets.

As we and other researchers connect our work, the stories contained in these old documents will come to life and bring new insight to modern scholars.

Provided by The Conversation