

iPRES 2024 • iPRES 2024 Proceedings

Scaling Up Digital Preservation Workflows With Homegrown Tools and Automation

Shelly Black¹ Brian Dietz¹ Trevor Thornton¹

¹NC State University Libraries

Published on: Aug 30, 2024

URL: <https://ipres2024.pubpub.org/pub/icwasqi0>

License: [Creative Commons Attribution-ShareAlike 4.0 International License \(CC-BY-SA 4.0\)](https://creativecommons.org/licenses/by-sa/4.0/)

Abstract – At NC State University Libraries, the Special Collections Research Center leverages an integrated system of locally developed applications and open-source technologies to facilitate the long-term preservation of digitized and born-digital archival assets. These applications automate many previously manual tasks, such as creating access derivatives from preservation scans and ingest into preservation storage. They have allowed us to scale up the number of digitized assets we create and publish online; born-digital assets we acquire from storage media, appraise, and package; and total assets in local and distributed preservation storage. The origin of these applications lies in scripted workflows put into use more than a decade ago, and the applications were built in close collaboration with developers in the Digital Library Initiatives department between 2011 and 2023. This paper presents a strategy for managing digital curation and preservation workflows that does not solely depend on standalone and third-party applications. It describes our iterative approach to deploying these tools, the functionalities of each application, and sustainability considerations of managing in-house applications and using Academic Preservation Trust for offsite preservation.

Keywords – [digital archiving](#), [digital preservation](#), [automation](#), [open-source](#), [application development](#).

This paper was submitted for the iPRES2024 conference on March 17, 2024 and reviewed by Jan Hutar, Remco van Veenendaal, Karin Bredenberg and 1 anonymous reviewer. The paper was accepted with reviewer suggestions on May 6, 2024 by co-chairs Heather Moulaison-Sandy (University of Missouri), Jean-Yves Le Meur (CERN) and Julie M. Birkholz (Ghent University & KBR) on behalf of the iPRES2024 Program Committee.

Introduction

Between March 2018 and March 2024, the NC State University Libraries (Libraries) has ingested over 87 terabytes of digitized and born-digital special collections assets into its self-hosted digital preservation system. Since August 2020, those assets have been replicated in entirety in Amazon storage through the Academic Preservation Trust (APTrust) digital preservation service. The processes of doing so have relied largely on applications developed in-house using open-source technologies. These applications, used to manage our digitization and born-digital archival processes, incorporate several automated steps and have been integrated into preservation workflows. This has resulted in staff being able to prepare assets for preservation more efficiently, as the ingest process no longer depends on manual intervention.

BACKGROUND

The Libraries¹ has a long standing commitment to technical development to support resource access and discoverability, as well as working with open-source technologies. Among others, applications developed by the Libraries that have been implemented at other organizations include Suma, “a tablet-based toolkit for collecting, managing, and analyzing data about the usage of physical spaces” [1]; QuickSearch, “a customized federated search tool with a bento-box results interface” [2]; and, to a lesser extent, Lentil, “a Ruby on Rails

Engine that supports the harvesting of images from Instagram,” [3] which served as the underlying architecture for the “My #HuntLibrary” project [4] and other social media harvesting projects, and has been sunset due to changes in the Instagram API. While the Libraries has supported open-source projects, development is always opinionated, meaning it must first and foremost fit the needs of in-house users, and not all applications developed in-house have been deemed appropriate for open sourcing.

The work discussed in this paper includes contributors from several departments, but the main contributors and maintainers of applications and workflows are the Libraries’ Special Collections Research Center (SCRC) and Digital Library Initiatives (DLI) departments. The materials referred to are exclusively those managed by the SCRC.² The Libraries established the SCRC [5] in the mid-1980s. The department manages the Libraries’ rare and unique collections, including archival records, manuscript materials, and rare books. The physical collection is approximately 30,000 linear feet of manuscript and archival materials, and over 16,000 rare books; the digital collections include over 1.5 million digitized assets and over 1 million born-digital assets. The department also manages a public services point and conducts outreach and instruction services. The SCRC currently has 12 full-time permanent staff, one full-time term-limited staff, and two dozen undergraduate and graduate student staff. Within the department, the Collections Stewardship and Discovery unit has staff whose primary responsibilities include digitization, born-digital processing, and digital preservation services, as well as physical collections management. The staff responsible for digital archival materials currently consists of one manager and two full-time staff, one of whom manages digitization and one of whom manages born-digital archival processes. This unit has included term-limited project librarians and Libraries Fellows [6] (term-limited half-time appointment in the department), all of whom have contributed significantly to the development of the processes that will be discussed in this paper. In summer 2022, a Libraries Fellow went from half-time to full-time appointment, and in fall 2022, that appointment became permanent, filling a new digital archivist role. The program also consistently employs many graduate students assigned to digitization and born-digital processing projects.

DLI is one of two technically-focused departments in the Libraries, the other being Information Technology (IT). While IT manages major business processes and hardware, including server infrastructure and storage, DLI is focused largely on web application development, including the Libraries search and discovery environment [7]. DLI staff consists of 11 full-time employees and one half-time Libraries Fellow. DLI and the SCRC have a history of collaboration on application development to support SCRC workflow needs, dating back to at least 2008. DLI staff working on SCRC applications have expanded and contracted over the years. While there have been many current and previous developers in DLI who have contributed to the applications and processes discussed in this paper, currently the digital discovery and preservation portfolio is managed largely by less than two full-time equivalent staff.

As staff in DLI and the SCRC successfully built a digital program and a corresponding accumulation of assets, administrative support for the preservation of digital special collections materials grew [8]. The Libraries had

already committed to the preservation of e-journals through membership in LOCKKS, CLOCKKS, and Portico, and in 2011, became a charter member of APTrust [9].³ Libraries staff have contributed to APTrust's development through consortial meetings regarding priorities, features, and functionality. Between 2010 and 2018, much of the work that led to the development of a digital preservation system was performed collaboratively under the auspices of the Libraries' Digital Collections Technical Oversight Committee and Digitization and Digital Curation Working Group. This included the creation of a "value matrix," through which curatorial staff could assign scores to materials to determine what level of preservation services would be required (as well as a revised version of the document),⁴ and a self-assessment using the National Digital Stewardship Alliance's Levels of Digital Preservation 1.0 matrix [10]. Staff from DLI and the SCRC began developing specs for a system in 2017, and development was completed in 2018. The Libraries began ingesting digital assets into its homegrown digital preservation system in March 2018, and its first ingest into APTrust occurred in August 2020.

IMPLEMENTATION

The in-house applications that have a role in our preservation services include a preservation application, SCPS (Special Collections Preservation System, pronounced "scoops"); Wonda for orchestrating digitization workflows, including preservation ingest workflows; and DAEV (Digital Assets of Enduring Value, pronounced "Dave") for orchestrating born-digital workflows, which also includes ingest workflows.⁵ Each application is integrated with ArchivesSpace, an open-source application for managing and describing collections, which we began using in 2014 [11].

For technical development, DLI values iteration, automation, and microservice architectures. In the case of SCRC applications, staff have continued to refine and extend existing software to include additional automated processes. The applications used in our current digitization workflows incorporate previous iterations of those that have been in use for over a decade; the application used in our born-digital workflow is on its third iteration.

The applications are hosted internally on Red Hat Enterprise Linux 9 running an Apache web server and Passenger application server. They were developed using the Ruby on Rails framework. Databases are managed through a MariaDB cluster, Solr is used for indexing and search, and Redis is used to manage background job queues, enabling automation. API integration facilitates communication between the applications. Shared storage infrastructure consists of virtual file servers for working storage and preservation staging. DLI and SCRC staff mount Network File System (NFS) storage across several servers and Mac, Windows, and Linux workstations. In the first half of 2019, DLI and SCRC staff worked closely with IT staff to configure, test, and deploy a storage environment that would support an ecosystem with multiple mount points.

Wonda

Wonda is an “orchestration” system that brings together multiple tools used for digitization, automating derivative creation, and publication. It launched in late 2019 [12]. Before Wonda was introduced, workflows for supporting access to digitized materials were semi-automated [13]. Digitization project personnel could batch create metadata records through CSV uploads to our digital collections metadata and asset management system (Special Collections Asset Management System, or SCAMS); project managers could create access derivatives by running a script on a server; and the publication component was automated, based on the existence of a metadata record and access derivative with a matching filename stem.

Today Wonda is most commonly used to facilitate digitization for online access, where a staff person scans items from an archival or manuscript collection that are described in ArchivesSpace.⁶ Using a Rails Engine gem developed in-house, Wonda provides an interface for selecting an archival object record from ArchivesSpace via search or direct selection by URI [14].⁷ The archival object URI and basic metadata are stored in Wonda's database, establishing a link between the digitization project and the ArchivesSpace record. Upon starting a project and creating a resource,⁸ Wonda will generate a resource ID, to be used as the base filename for digitized objects; create a directory in working storage where assets will be saved; and, assuming the resource is intended to be made available online, will create a stub record in SCAMS using metadata imported from ArchivesSpace.

Different user roles are built into Wonda, including “admin,” “project manager,” and “technician,” and users can assign projects to each other. After a “technician” completes scanning and saves the preservation scans to working storage, a “project manager” is notified that the metadata and scanned assets are ready for quality control. The completion of quality control triggers a series of automated tasks managed by Wonda. This includes making access derivatives and generating video captions using AVPD, a locally developed tool. After derivatives have been created, Wonda sends a request to another homegrown tool, Ocracoke, to perform OCR on the JPG2000s for full-text search. Each tool is hosted on its own server, which accesses the preservation assets via a mounted file server. Once these steps are complete, Wonda notifies SCAMS that derivative processing is complete and to make the resource publicly available in the digital collections platform.

At this point, the “project manager” is notified that the project is available on the public digital collections platform and is ready for further quality control. Upon successful review, the “project manager” indicates to Wonda to initiate the process of ingesting the assets into SCPS. Lastly, Wonda automatically generates two digital object records in ArchivesSpace, one for the public access version and one for the preservation package. The former digital object record is represented in the collection guide as a link to the associated resource in the digital collections platform. Wonda will also write to the associated archival object record “Conditions Governing Access” and “Existence and Location of Copies” notes, the contents of which are determined by whether the resource was made available online. After a two week period and there is certainty that assets have been ingested into SCPS and APTrust, Wonda moves working storage files into a directory for deletion.

By orchestrating what were formerly manual tasks, Wonda has reduced potential errors and the time it takes to manage derivatives and metadata, efficiently publishing derivatives and transferring the original scans to preservation storage. It has also made archival description more consistent. The modular design of Wonda enables it to be implementation-agnostic with regard to its component applications. In other words, components can be added or replaced as needed, provided that they conform to the requirements for communicating with Wonda. For example, AVPD, which is locally developed and is used to process and manage access derivatives for images and AV, recently replaced two applications that did this task for images and AV separately. Wonda also provides SCRC staff with useful metrics by reporting on the total numbers of completed projects, published resources, and files created based on a date range.

DAEV

DAEV is a web application that supports the assessment, packaging, and description (shorthand to “processing”), as well as preservation ingest, of born-digital materials. It guides “technicians” in using open-source command line tools, records actions taken during processing, generates preservation metadata, and provides integration with ArchivesSpace. Iterations of DAEV have been in production since 2015. Processing workflows are defined in YAML files, and instructional text for workflows, including commands to be run in a terminal, are contained in Markdown files and are created and maintained by SCRC staff.

When creating a new project in DAEV, a “technician” user searches for archival object records from ArchivesSpace and retrieves descriptive metadata—functionality that uses the same gem as Wonda. Next, the “technician” selects a storage media type, such as an optical disc. DAEV prompts them to select a workflow, such as disk imaging or packaging the files into a tarball. Different tools are used depending on the workflow. For example, Brunnhilde is run on directories or disk images, and calls: Siegfried, a signature-based file format identification tool; fiwalk, a SleuthKit tool that generates Digital Forensics XML; and tree, a recursive directory listing tool. Other tools common to multiple workflows include ClamAV for virus scanning and bulk_extractor for identifying personally identifiable information. GNU utilities, such as tar and md5sum, are also used. The initial version of DAEV included instructions for using GUI applications in the BitCurator virtual environment [15]. In 2018, staff began using command line utilities instead [16], and DAEV was updated to reflect this change. The use of variables in the Markdown files allows staff to write a single set of commands that DAEV customizes for each project; for example, the string “{{ working_storage_path }}” in a Markdown file would be converted to “\$HOME/born_digital/daev_123_456” in the command that displays on the screen to the user. “Technicians” can then copy and paste that command into the terminal. Thus, DAEV has enabled full-time and student employees who are less familiar with the command line to learn and use these tools.

DAEV was designed to satisfy the needs of SCRC’s burgeoning born-digital program. As practices have evolved, version 2 was written to include new functionality and increased automation, and entered production in late 2023. While “technicians” had to create the directories containing assets before, the new DAEV

automatically creates these in working storage. The application also automates the transfer of virus scan, file format, and other reports from a “preliminary” to “final” directory. Whereas version 1 required manual ingest into SCPS, version 2 introduces the ability to assign projects to users and automates ingest just like Wonda. Once a “project manager” indicates that a project has passed quality control, DAEV initiates ingest, then creates a digital object record in ArchivesSpace and associates it with the archival object record. Version 2 also writes a “Conditions Governing Access” note to the archival object record, formerly a manual step. DAEV assigns a UUID (Universally Unique Identifier) to the resource, which is used in the digital object record and in SCPS. Like Wonda, DAEV also moves working directories into a directory for deletion two weeks following ingest.

The latest version of DAEV also includes expanded workflows, and commands have been updated so that processors run most tools using a Docker container [17]. Because a container cannot access a Mac or Windows host device, tools run on the host include ddrescue for disk imaging and cdparanoia for ripping audio. Because DAEV includes predefined workflows, staff anticipate adding new ones as more kinds of storage media, such as internal hard drives, are acquired.

SCPS

SCPS is a web application that manages the long-term preservation of digital assets in the Libraries’ local primary storage and APTrust. SCPS has been in production since 2018.

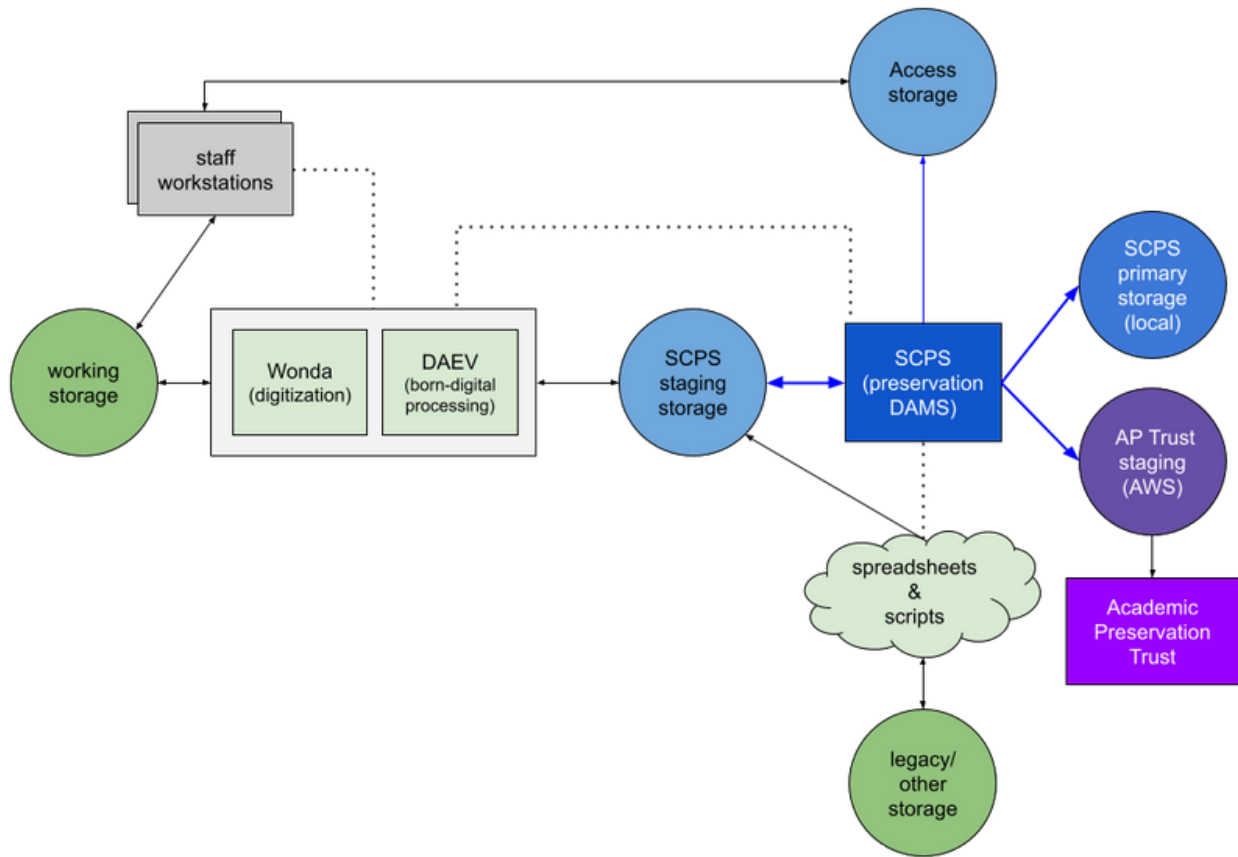


Figure 1: Diagram showing integration of SCPS with other applications and file servers

As depicted in Fig. 1, newly produced packages are ingested into SCPS from Wonda and DAEV automatically, and those from legacy storage have been ingested through a scripted process. “Project managers” are expected to ensure, during quality review, that packages conform to the established guidelines for creating submission information packages (SIP) for SCPS. The Libraries has established a relatively flexible definition of what a SIP can contain. For assets created through digitization, we expect archival-quality assets conforming to established standards. For those created through the born-digital archival process, we expect a set of files (i.e., some form of the content that was donated or transferred to the Libraries); a set of reports created by staff (e.g., virus and PII scan results); a processing report, serialized as JSON, that documents the nature of the content and the actions taken by a processor; and an assets report, which documents all of the files in the SIP.

In addition to the assets in the SIP, which will be represented as is and without transformation (i.e., migrated) in the archival information package (AIP), the archival package record includes a UUID, number of assets, size, creation timestamp, source (e.g., collection name), as well as data related to APTrust ingest, including storage tier, ID, and checksum verification timestamp.⁹ Once in locally-maintained preservation storage, SCPS performs regular, scheduled checksum validations on assets to ensure file fixity. An asset record includes a

hash value, checksum verification timestamp, as well as file name, UUID, ingest timestamp, size, and MIME type.

Following ingest into SCPS, the files are prepared for ingest into APTrust using the BagIt for Ruby library (BagItSpec v0.97). SCPS then creates a record of upload and confirms the ingest via subsequent automated requests to the APTrust API.

SCPS allows for the retrieval of dissemination information packages (DIP), as well. Like SIPs, the definition for DIPs is flexible, and SCPS allows staff to retrieve an entire DIP or component files. A DIP for born-digital packages can include the “content” as well as the reports; for digitized materials, likewise, it can be a single file or as many as have been requested (e.g., as part of a researcher duplication request). If requested, staff may transcode the archival copy to a lower quality (e.g., full-resolution TIFF to a lower resolution JPEG). In some instances, transcoded files may be added to the related AIP. SCPS supports staff discovery through collection name, package identifier, or filename for when staff need to retrieve copies for use on the physical reading room laptop or in the virtual reading room [17]. When requested files are ready, SCPS notifies the staff user by email.

SCPS also supports reporting by providing the total number of ingests by date range and the number of ingests by asset types, such as born-digital, still image, video, and audio. Additionally, it reports on the number of ingests in APT’s Core-Service and Glacier Deep.

APTrust

NC State is among APTrust’s largest depositors. For preservation storage, APTrust uses Amazon Web Services (AWS).¹⁰ SCPS uploads the prepared bags into an APTrust “receiving bucket” in AWS storage using AWS SDK (software development kit) for Ruby. Upon ingest, born-digital assets are transferred to APTrust’s Core-Service storage, which maintains three copies in Amazon S3 Standard located in Virginia and three more in S3 Glacier Deep Archive located in Oregon. Digitized assets are uploaded to the lower-cost S3 Glacier Deep Archive, which maintains three copies in a single region.¹¹ Currently, APTrust performs fixity checks every 90 days [18].

NC State was the first member to put APTrust’s data restoration plan [20] to the test after a significant data loss. In June 2021, Libraries IT staff inadvertently deleted 35 terabytes of data on a locally hosted preservation storage volume, as well as its backup [8]. Because SCPS stores its inventory and checksums separately from preservation packages, the SCPS developer was able to identify which files were missing. IT was able to recover some assets from local datastores, except for 16 terabytes which needed to be retrieved from AWS. APTrust had not performed a large data restore before, but assisted the Libraries in fully recovering data in six weeks and at a rate that did not incur AWS egress charges. Following a security audit and post-mortem with APTrust, NC State made improvements in documentation and procedures.

DISCUSSION

For more than a decade, the Libraries has been developing its integrated and open-source technical infrastructure as part of its digital curation and preservation strategy. With the exception of ArchivesSpace and APTTrust, this paper documents and offers a model that does not rely on using third-party or commercial solutions, some of which are not turnkey and still require in-house expertise [20]. The authors acknowledge that the Libraries' level of access to technical support is exceptional. Across the United States, universities have outsourced or centralized IT services and staff to improve operational efficiencies, standardize workflows, and improve cybersecurity. This has stifled service maintenance and innovation at academic libraries, which are organizations with specific technology needs [21], [22]. While the authors do not foresee these changes happening at NC State in the immediate future, the Libraries is part of a large public university where extrinsic factors have an impact on institutional priorities and budget. As part of the Libraries' iterative approach to managing in-house systems and tools, assessment is warranted and reveals potential challenges—albeit familiar to many systems and solutions—associated with the customized ecosystem.

Sustainability of managing in-house applications

While some Libraries applications have been designed with wider distribution in mind, DLI has taken an opinionated approach to building Wonda, DAEV, and SCPS. These applications were designed for the Libraries' specific context and needs, as opposed to being adaptable by other institutions. For this reason, the source code for the applications is located in NC State's private GitHub repository. There is considerable labor required of free and open-source software (FOSS) maintainers in responding to issues or pull requests [23]. Making Wonda, DAEV, and SCPS publicly available would not be practicable at current staffing levels. Thus, while the applications, as well as their infrastructure, use open-source tools, they themselves are not free and open-source. A consequence of this development strategy is that the applications are not part of a community and lack external stakeholders. This can lead to duplicated efforts and ultimately a lack of project memory [24].

The Libraries has been able to commit human and technical resources to SCRC applications despite changes in top-level leadership and department heads, as well as updated strategic plans. The applications are a result of a collaboration with responsive in-house developers and systems administrators who have made improvements to the ecosystem over time. However, DLI collectively manages many custom-built applications, vended systems, and their dependencies. This requires balancing maintenance, updates, and sunseting of existing tools and applications while also building new ones. Wonda, DAEV, and SCPS are currently in maintenance mode and are supported by 2 full-time equivalent developers, but at times have been supported by only one. Should SCPS become unsupported, the Libraries would need to consider alternatives, such as licensing a vended service or relying solely on APTTrust. In the case of the latter we could adopt DART for ingest, as APTTrust is repository agnostic and advocates for open-source tools [25]. However, there are no third-party alternatives for Wonda and DAEV.

Sustainability of APTrust

APTrust has experienced steady growth in its number of members and ingests [26]. Stewardship continuity is one of APTrust's core values [27]. Consequently, APTrust established a robust operational reserve [28] and created a succession plan for sunsetting or moving host institutions [29]. The sustainability and financial health of APTrust largely depend on two variables: 1) membership dues, which support AWS storage fees, and 2) the University of Virginia, which provides an organizational home and subsidizes staff salaries. Should APTrust cease to exist, require a new host, switch from AWS to a different storage provider, or lose the Libraries as a member, SCPS could be configured to distribute assets elsewhere.

Things take time

The Libraries has been actively preserving digital special collections assets since 2018. However, the organization has been creating and acquiring such assets for over 20 years.¹² Many things account for this timeline. It takes time to develop the level of trust required to obtain support, in staff time and money, to establish a digital preservation program. At the Libraries, there is not a single staff person dedicated entirely to digital preservation; there is no "Digital Preservation Librarian" or team. Digital preservation is a collaborative effort between several staff persons, departments, and committees. Additionally, designing, developing, testing, and deploying applications is also time intensive. DLI's portfolio includes applications for other Libraries departments. Furthermore, the opening of a new technology-rich library building requiring intensive IT support and the COVID-19 pandemic (2020-2021) shifted our application development timelines.

Conclusion

The Libraries' integrated ecosystem of technical services applications were more than a decade in the making, and are built on top of previous versions of applications still in use. Wonda, DAEV, and SCPS have enabled SCRC to more effectively steward digital archival assets for long-term preservation, ingesting over 87 terabytes of data into locally managed preservation storage, as well as into APTrust. Of equal importance, these applications have made it easier for SCRC to meet its mission in serving current researchers who access our digital assets online or in the reading room. Workflows, including ingest into local and distributed preservation storage, have been expedited thanks to automation and integration between applications.

While Wonda, DAEV, and SCPS have made our work more efficient, allowing us to commit time to new projects, there is still legacy data to manage. The Libraries ingested approximately 30 terabytes of this data into SCPS and APTrust during the first two years of the pandemic. The large majority of this data was "low hanging fruit," material that was easily identified and described, though this still required many hours to complete. What remains are approximately 10 terabytes of data that will require a significant amount of analysis, possibly new archival description and file arrangement, and ingest that combines both manual and scripted tasks.

As we continue to mature in this area of work, we will likely revise our decisions and workflows, which will result in requests for new and expanded features in SCPS, as well as Wonda and DAEV. Given the Libraries' iterative approach to development, we are hopeful that these applications will continue to adapt to sustain our growing collections, as well as effectively support staff needs and researcher use.

Footnotes

1. NC State University, as part of the University of North Carolina (UNC) System, is a public, land-grant R1 research university founded in 1887 and located in Raleigh, North Carolina. Nearly 38,000 students attend NC State, with about 26,000 undergraduate students. NC State is the largest campus in the UNC System. The Libraries, with a staff of just under 200 full-time staff members, excluding vacant, affiliate, and student positions, includes two main libraries, three branch libraries, and has a collection that includes 3.2+ million titles, 122,000+ serial subscriptions, 14.5+ million uses of the collection in 2022-23, and 3.6+ million full-text article downloads in 2022-23 (<https://www.lib.ncsu.edu/do/collections>). ↵
2. The Libraries also preserves electronic theses and dissertations (16,801 packages comprising over 41 gigabytes), but those are not managed by the SCRC and will not be considered in this paper). ↵
3. The Libraries had also subscribed to DuraCloud services, though under utilized them, and, by 2013, was also a member of the now defunct Digital Preservation Network (DPN). ↵
4. This was written before a technical solution was in place in the hopes that it would inform that solution. The matrix was ultimately deemed more complicated than necessary and was not implemented as is. The Libraries currently distinguishes levels of preservation services based on the difficulty of replicating lost assets. ↵
5. Our technical services ecosystem also includes a digital collections metadata and asset management system for managing metadata of digitized resources, a public facing digital collections platform (<https://d.lib.ncsu.edu/collections>), and an archival collection guides platform (<https://www.lib.ncsu.edu/findingaids>). Although Wonda and DAEV are integrated with those applications, they are not integrated with SCPS, and thus, a detailed treatment of them is beyond the scope of this paper. ↵
6. Wonda may also be used for items with catalog records, as well as for items without an access component and that go directly to preservation storage. ↵
7. The feature also allows one to select a record from the Libraries catalog and also create records using an arbitrary identifier, used for resources not described in a Libraries datastore. ↵
8. In Wonda, a resource is an item that is digitized and described as a single thing, such as a photograph or report. In DAEV, a resource includes reports and the archival assets, such as a disk image or tarball, comprises a single intellectual unit, and is described as an archival object record at the archival file level. ↵

9. The data necessary to create METS and PREMIS documents exist in local data stores, and expressed in administrative web forms. In 2019 work began to map this data to METS and PREMIS documents, but that work has not been prioritized or resumed, because our assessment was that no immediate problem would be solved. Data stores do contain necessary identifiers (e.g., resource IDs, URIs) to associate related records with one another. [↵](#)
10. In December 2023, at the APTrust 2023 Member Meeting (<https://aptrust.org/resources/member-meetings/2023-member-meeting>), it was shared that the Libraries had contributed more data by storage capacity than any other member organization (minutes of the meeting are restricted to APTrust members). In an email from Nathan Tallman, Executive Director of APTrust, to the APTrust Advisory Council, sent March 4, 2024, he shared that, while the Libraries no longer held this position, it was in the top three contributors counting by storage (second), objects (second), and files (third). [↵](#)
11. Electronic theses and dissertations are also ingested into Amazon S3 Standard. [↵](#)
12. Before 2018, the Libraries performed back-up of these resources, and institutional memory was relied on for tracking them. [↵](#)

References

1. NC State University Libraries. “Suma.” Accessed: Mar. 5, 2024. [Online]. Available: <https://www.lib.ncsu.edu/projects/suma> [↵](#)
2. NC State University Libraries. “QuickSearch.” Accessed: Mar. 5, 2024. [Online]. Available: <https://www.lib.ncsu.edu/projects/quicksearch> [↵](#)
3. GitHub. “NCSU-Libraries/lentil.” Accessed: Mar. 5, 2024. [Online]. Available: <https://github.com/NCSU-Libraries/lentil> [↵](#)
4. NC State University Libraries. “My #HuntLibrary.” Accessed: Mar. 5, 2024. [Online]. Available: <https://www.lib.ncsu.edu/projects/my-huntlibrary> [↵](#)
5. NC State University Libraries. “Special Collections Research Center.” Accessed: Mar. 5, 2024. [Online]. Available: <https://www.lib.ncsu.edu/scrc> [↵](#)
6. NC State University Libraries. “Fellows Program.” Accessed: Mar. 5, 2024. [Online]. Available: <https://www.lib.ncsu.edu/fellows> [↵](#)
7. NC State University Libraries. “DLI Projects and Initiatives.” Accessed: Mar. 5, 2024. [Online]. Available: <https://www.lib.ncsu.edu/dli/projects> [↵](#)

8. J. Sexton and B. Daigle. (Dec. 2022). Data Loss and Recovery: Strategies for Organizational Change. Presented at CNI Member Meeting. [Online]. Available: <https://www.cni.org/wp-content/uploads/2022/12/Data-Loss-and-Recovery-Bradley-Daigle.pdf> ↵
9. APTrust. “Background.” Accessed: Mar. 5, 2024. [Online]. Available: <https://aptrust.org/about/background> ↵
10. M. Phillips, J. Bailey, A. Goethals, and T. Owens. “2013 NDSA Levels of Digital Preservation.” OSF. Accessed: Mar. 5, 2024. [Online]. Available: <https://osf.io/9ya8c/> ↵
11. ArchivesSpace. Accessed: Mar. 5, 2024. [Online]. Available: <https://archivesspace.org> ↵
12. T. Thornton. (Feb. 2020). Developing an integrated technical infrastructure for archives at NC State. Presented at Lighting the Way: A National Forum on Archival Discovery and Delivery. [Online]. Available: <https://purl.stanford.edu/qb052kj0451> ↵
13. B. Dietz and J. Ronallo. (2011). Automating a Digital Special Collections Workflow Through Iterative Development. Presented at ACRL 15th National Conference. [Online]. Available: <http://hdl.handle.net/11213/18156>. One author represents the SCRC, the other DLI. ↵
14. GitHub. “archivesspace_browser_rails.” Accessed: Mar. 5, 2024. [Online]. Available: https://github.com/NCSSU-Libraries/archivesspace_browser_rails ↵
15. GitHub. “BitCurator / bitcurator-distro.” Accessed: Mar. 5, 2024. [Online]. Available: <https://github.com/BitCurator/bitcurator-distro> ↵
16. B. Dietz. “BitCurator: Beyond Environment.” Accessed: Mar. 5, 2024. [Online]. Available: <https://bitcuratorconsortium.org/bitcurator-beyond-environment-2> ↵
17. S. Black, B. Dietz, and Farrell. Virtualization for Processing and Accessing Digital Archives. Presented at the 19th Int. Conf. on Digital Preservation, Champaign-Urbana, 2023. [Online]. Available: <https://hdl.handle.net/2142/121116> ↵
18. APTrust. “Running Fixity Checks in the Cloud.” Accessed: March. 5, 2024. [Online]. Available: <https://aptrust.org/2024/01/24/running-fixity-checks-in-the-cloud> ↵
19. B. J. Daigle, “Mutually Assured Preservation: Fostering Active Preservation Practice Through 'Fire Drills,’” Int. J. of Digit. Curation, vol. 15, no. 1, 2020, doi: 10.2218/ijdc.v15i1.724 ↵

20. O. Y. Rieger, R. C. Schonfeld, and L. Sweeney. “The Effectiveness and Durability of Digital Preservation and Curation Systems.” Ithaka S+R. Accessed: March. 5, 2024. [Online]. Available: <https://doi.org/10.18665/sr.316990> ↵
21. J. Crum. “IT Centralization: Impact on Academic Libraries.” LITA Blog. Accessed: March. 5, 2024. [Online]. Available: <https://litablog.org/2018/12/it-centralization-impact-on-academic-libraries> ↵
22. University of Central Florida Libraries. “Annual Report 2018-2019.” Accessed: March. 5, 2024. [Online]. Available: <https://stars.library.ucf.edu/lib-docs/163> ↵
23. N. Eghbal, Working in Public: The Making and Maintenance of Open Source Software. South San Francisco, CA, USA: Stripe Press, 2020. ↵
24. G. Hanganu and R. Gardler. “Planning for sustainability.” OSS Watch. Accessed: Mar. 5, 2024. [Online]. Available: <http://oss-watch.ac.uk/resources/planningsustainability> ↵
25. DART Documentation. “Getting Started.” GitHub. Accessed: Mar. 5, 2024. [Online]. Available: https://aptrust.github.io/dart-docs/users/getting_started ↵
26. APTrust. “2023 Member Meeting.” Accessed: Mar. 5, 2024. [Online]. Available: <https://aptrust.org/resources/member-meetings/2023-member-meeting> ↵
27. B. Daigle. “APTrust Strategic Plan and Core Principles Version 2.0.” Libra Open. Accessed: Mar. 5, 2024. [Online]. Available: <https://doi.org/10.18130/0n5a-4b87> ↵
28. APTrust. “Budget Reports.” Accessed: Mar. 5, 2024. [Online]. Available: <https://aptrust.org/about/budget-reports> ↵
29. B. Daigle. “APTrust Succession Policy (Version 3.0).” Accessed: Mar. 5, 2024. [Online]. Available: <https://doi.org/10.18130/fmbe-a977> ↵