Digital Curation at Work

Modeling Workflows for Digital Archival Materials

Colin Post
University of North Carolina
ccolin@live.unc.edu

Alexandra Chassanoff
Educopia Institute
alex@educopia.org

Christopher A. Lee
University of North Carolina
callee@ils.unc.edu

Andrew Rabkin
University of North Carolina
rabkin@unc.edu

Yinglong Zhang
University of North Carolina
yinglongz@unc.edu

Katherine Skinner
Educopia Institute
katherine@educopia.org

Sam Meister
Educopia Institute
sam@educopia.org

ABSTRACT

This paper describes and compares digital curation workflows from 12 cultural heritage institutions that vary in size, nature of digital collections, available resources, and level of development of digital curation activities. While the research and practice of digital curation continues to mature in the cultural heritage sector, relatively little empirical, comparative research on digital curation activities has been conducted to date. The present research aims to advance knowledge about digital curation as it is currently practiced in the field, principally by modeling digital curation workflows from different institutional contexts. This greater understanding can contribute to the advancement of digital curation software, practices, and technical skills. In particular, the project focuses on the role of open-source software systems, as these systems already have strong support in the cultural heritage sector and can readily be further developed through these existing communities. This research has surfaced similarities and differences in digital curation activities, as well as broader sociotechnical factors impacting digital curation work, including the degree of formalization of digital curation activities, the nature of collections being acquired, and the level of institutional support for various software environments.

CCS CONCEPTS

• **Applied computing → Digital libraries and archives**;

KEYWORDS

Digital curation, workflows, open-source software

1 INTRODUCTION

Libraries, archives, and other cultural heritage institutions care for collections of digitized and born-digital materials of increasing size and variety. In addition to published materials in standardized formats like electronic journals and e-books, collections include unique or special objects in a diverse array of formats: organizational records, personal or family archives, websites, audio and video recordings, and more. Digital curation, or the ongoing care and attention needed to keep objects viable for present and future use, starts from at or before the time of acquisition and continues well after the initial provision of access. For records or manuscript materials that have typically been the province of archives and special collections, a great deal of processing work needs to occur before materials are available to users. While libraries and archives have established practices and policies for managing analog materials, individual institutions and the archival field as a whole are still developing best practices, standards, and shared terminologies for processing digital materials [3, 14].

Digital curation activities are complex, involving diverse skills and techniques, software tools and systems, and a range of professional and paraprofessional staff. Due to the variety of functions involved and the ongoing nature of digital curation work, no single software environment can manage the entire scope of the stewardship of digital materials. Furthermore, digital curation responsibilities often cross departments or units, and benefit from collaboration among individuals at an institution with different areas of expertise [12, 21]. Digital curation approaches also encompass factors specific to particular institutional contexts: institutions must work within their existing processes, policies, institutional constraints, and technical platforms to marshal the necessary technologies and staff to address local needs.

This paper describes preliminary findings from the OSSArcFlow Project, a two year (2017-2019) project funded by the Institute for Museum and Library Services (IMLS) involving project members at the University of North Carolina at Chapel Hill and the Educopia Institute. The project team began working with 12 partner institutions in July 2017 to explore how three open-source software (OSS) environments (the BitCurator environment, ArchivesSpace, and Archivematica)[1] can support digital curation activities. These and other OSS tools have already been widely adopted in the cultural heritage sector. The vitality and collaborative nature of these communities of users will facilitate the further development of OSS solutions—aided by empirical research of digital curation activities and needs.

The project endeavors to foster reflection on the current state of digital curation across a variety of institutional contexts by constituting a cohort of partner institutions. The project team has worked directly with these partner institutions, and has facilitated collaboration, discussion, and sharing of resources among the partners.

[1]For more on these environments see https://bitcuratorconsortium.org/, https://archivesspace.org/, https://www.archivematica.org/.

The partner institutions represent varying sizes and types, geographic locations, nature of digital collections, available resources, and level of development of digital curation activities.[2] The project team compared similarities and differences in digital curation approaches across the various institutions. We analyzed workflow steps, the order of steps, the tools used for different steps, as well sociotechnical factors impacting the development and sustainability of digital curation work in particular institutional contexts.

The modeling and analysis is part of the broader aims of the OSSArcFlow project to advance workflows that incorporate OSS solutions to support stewardship of digital collections across a variety of cultural heritage institutional contexts. Institutions regularly report that there are both gaps and overlaps between different digital curation tools and software environments that have to be managed. Gaps between tools can make it difficult to push data through workflows. For example, the output from one tool may have to be transformed before it is compatible with the next tool. Practitioners can spend large portions of time transforming data and metadata so that it can interface with different systems. Overlaps between tools challenge curators to make decisions about when and where to complete particular functions. While the OSSArcFlow project focuses on three OSS environments, the workflow modeling and analysis has also attended to the range of other tools and systems in use, as well as factors specific to local contexts.

The OSSArcFlow project strives to advance empirical knowledge about the current state of digital curation to contribute to the development of digital curation software and practices. Although tools, techniques, and skills continue to mature, digital curation scholarship and practice is still in many ways at a nascent stage—institutions are actively developing digital curation workflows, practices, and policies, and there are notable gaps between tools. These workflows not only provide insight into how digital curation is currently being practiced and conceptualized, but they also serve as analytical tools for investigating various institutional factors.

2 RELATED WORK

As digital curation practice in cultural heritage organizations continues to mature, there is a growing need for empirical research into the myriad factors that influence this work. Tools, techniques, skills, standards and best practices for digital curation are always enacted in particular local contexts. Digital curation activities not only contend with numerous social and institutional factors like budgets and external funding sources[13], staff skills and training [16], existing hardware and software [24], and relationships between archivists and IT staff [20], but also involve the interaction between social factors and various technological systems. As Summers and Punzalan [21] describe for web archiving, archives are "sociotechnical systems in which archivists collaborate with automated agents" (823).

Scholarship has sought to identify and describe the impact of sociotechnical factors on digital curation activities in practice. Mikeal et al. [18] detail efforts to develop a statewide system for managing electronic theses and dissertations (ETDs) for the Texas Digital Library, citing many issues that arose over the course of this complex, large-scale digital library project. Principally, the team had to develop a federated system that would integrate into the existing infrastructure for several institutions with quite different student bodies and disparate processes for managing ETDs. Cocciolo [7] reflects on how differing professional identities can alter perspectives on digital curation work—and occasionally lead to clashes—for instance, between digital archivists and digital asset managers, who had conflicting notions of how the digital asset management system at their institution should be used.

Recent discussions of digital library work more generally have emphasized the need for rubrics or frameworks for systematic evaluation, both in terms of specific aspects of this work, like preserving particular media types [11], as well as holistic assessments to inform long-term planning [5, 6]. Andrews, Harker, and Krahmer [2] suggest the analytical hierarchy process for weighing many contextual factors that might impact collection development and management for an institutional repository, such as faculty attitudes toward self-archiving, institutional policies regarding open access, and technical infrastructure. For long-term digital preservation, Becker and Rauber [6] discuss an evaluative approach to multi-criteria decision-making that attempts to encompass the highly contextual factors that impact preservation planning, such as variability in quality of tools, dynamic nature of goals and constraints over time, and differing needs for various user communities.

The present research has modeled workflows as a method for both gaining a deeper empirical understanding about current digital curation work and generating documentation that communicates a high-level representation of these activities. Increasingly, modeling workflows has been recognized as a valuable approach to conceptualizing work for a variety of cultural heritage activities, from dealing with missing or damaged books [19] to managing e-resources [10]. For archival contexts specifically, Daines III [9] suggests that adopting business process management (BPM) can improve the efficiency of processing archival collections by fostering a holistic perspective that can help to identify gaps or obstacles in processes and provide insights to streamline tasks. Gustainis [15] put this approach into practice, using workflows to systematically assess the efficiency of archival processing at the Center for the History of Medicine at Harvard.

Modeling workflows can be especially useful in understanding organizational changes associated with the adoption of new technologies. As Collins [8] describes, work with digital technologies is often far less linear and visible than work with analog collections, and processes frequently cut across traditional organizational boundaries. Workflows can help to make sense of these new ways of structuring tasks and processes. Anderson [1] proposes workflows as a way to assess evolving e-resource management processes. Similarly, Dowdy and Raeford [10] use workflows to evaluate existing e-resource services in order to inform the selection and implementation of new e-resource management software.

In addition to visualizing a particular work process, modeling workflows promises other benefits for cultural heritage organizations. In case studies applying workflow modeling at Brigham Young University, Duke, and the University of Michigan, all three report that the process of actually constructing the workflow is in itself insightful [4, 9, 10], offering a space for productive reflections

[2]The visual workflow diagrams are all available at https://educopia.org/ossarcflow/.

on goals and the nature of departmental and organizational work. Workflows also serve as explicit artifacts of organizational knowledge, clarifying workflow outputs and aiding inter-departmental communication [4].

Several authors describe modeling workflows as both a research method and valuable research output for studying digital curation specifically. Gengenbach [12] represented workflows based on semi-structured interviews to illustrate how digital forensics tools and methods fit (or fail to fit) alongside existing archival practices. Whyte [25] developed workflows at the Thomas Fisher Rare Book Library to assess the feasibility of various digital curation activities. This opened up further questions as well such as how to better communicate with donors when acquiring born-digital materials or how to provide better access to born-digital materials.

The present research investigates the role of OSS in digital curation workflows. Gengenbach et al. [13] argue that OSS systems stand to play a significant part in digital stewardship, as the open-source model enables sharing resources and knowledge, facilitates collaboration, and paves the way for smaller institutions to engage in digital curation. This is illustrated by consortial arrangements around tools [22]. However, OSS projects face key challenges as well, including maintaining community support. The present research to model workflows intends to foster community building around OSS tools in two ways: by developing a greater empirical understanding of how OSS tools are currently used in digital curation activities, and by encouraging the further dissemination of workflows themselves as resources that can be shared to assist in implementing these systems at other institutions.

3 METHODS

In August 2017, the project team began working with 12 partner institutions to construct models of their workflows. The project team conducted semi-structured interviews with each of the partner institutions, in which partners discussed the steps, tools, and individuals involved in digital curation activities, as well as other factors that influence digital curation work at their institutions. Based on these interviews, the project team constructed draft workflows in three formats: a narrative description, a tabular representation, and a visual diagram made using Lucidchart, a web application especially suited for creating process-oriented diagrams. Partners were given the opportunity to review the draft workflows, and then discussed these draft workflows with the project team in follow-up interviews. The project team produced final versions of the workflows, making any changes suggested by the partners (see fig. 1).

Using these workflow outputs, the project team conducted a systematic comparison of digital curation activities across all 12 partner institutions. This comparison assessed the steps in each of the partner's workflows, the sequence of steps, how the steps were organized into broader stages, and what tools were being used for each step, particularly how the BitCurator environment, ArchivesSpace, and Archivematica fit into each of the workflows. To facilitate comparison, the project team derived standardized language to describe the workflow steps and stages.

The workflow analysis was supplemented by qualitative analysis of the semi-structured interviews and corresponding notes. The project team coded these interviews to identify the many factors impacting digital curation work and influencing decisions about particular steps and tools. As this research is a largely exploratory inquiry into a rapidly changing area of practice, qualitative codes were generated from points discussed by the interview participants, rather than trying to apply a preexisting lexicon or codebook. Using NVivo, one project member developed a codebook by making several iterative passes through interview notes. A second project member used the codebook to code a selected set of interview notes. After running an inter-coder comparison in NVivo, we identified a small number of discrepancies between the two coders. We discussed each case and refined the codebook to clarify and add several new nodes.

Digital curation scholarship and practice continues to develop, and workflows are eminently dynamic. As a result, our research into this area remains exploratory. The workflows created and analyzed for the purposes of this paper reflect the digital curation activities at the various institutions at only one particular point in time. Although the project team intentionally developed a sample of partner institutions that represent a diversity along many relevant dimensions, the sample is still relatively small. The intent of this research is not to generalize to all cultural heritage institutions, but rather to provide insights into the current nature of digital curation practice, and to lay the foundation for future empirical research.

4 PARTICIPANT INSTITUTIONS AND PARTNER SOFTWARE ENVIRONMENTS

Prior to submitting the IMLS proposal, the OSSArcFlow project team researched and recruited 12 partner institutions varying in size, geographic location, and type (see table 1).[3]

Although the partners use a variety of software tools in support of their digital curation activities, OSSArcFlow focuses specifically on the role of three OSS environments common in digital curation workflows: the BitCurator environment, ArchivesSpace, and Archivematica. To participate in the project, institutions needed to have already implemented or pledged to implement at least one of the three environments. These environments all already have strong user communities, and all hold the potential for further development, especially by exploring connections and hand-offs among these three environments. To facilitate the translation of this empirical research into ongoing development, the project team sought the participation of the communities developing and hosting the OSS environments. The project team consisted of individuals directly involved with the development and maintenance of the BitCurator environment, and the project team sought external advisory roles for Lyrasis and Artefactual.

The BitCurator environment collects a variety of open-source tools to aid in the analysis and processing of born-digital archival materials, including those acquired on removable media. Previous work has examined the role of the BitCurator environment in digital curation workflows [17]. ArchivesSpace is an environment for describing archival collections of all kinds, including analog and born-digital materials. Archivematica is a digital preservation system designed to comply with the Reference Model for an Open Archival

[3] More detailed descriptions of the state of digital curation activities at each institution can be found in the 'digital curation dossiers' created in collaboration with the project partners, which are also available at https://educopia.org/ossarcflow/.

Figure 1: Excerpt from Duke University Digital Curation Work ow

Information System (OAIS), o ering a suite of micro-services applicable to collections from ingest through to long-term storage in a repository.

5 FINDINGS

In this section, we present ndings from the systematic comparison across the 12 digital curation work ows, highlighting some of the notable similarities and di erences among the partner institutions in terms of work ow steps, step order, and tools used. Figure 2 presents a visual overview of the 12 work ows, with color codes for each of the work ow stages discussed below (see key). To provide context for the ndings and discussion that follow, the gure is intended as a reference for readers to see what digital curation activities were employed at each institution, and how these activities were ordered.

5.1 Steps Common Across Institutions

There are some broad similarities in several work ow steps across the institutions. All institutions except Odum create either an accession record or a resource record (i.e., a record for a discrete digital object like a digital photograph, as opposed to a record for an entire archival collection) for digital materials. At Odum, a similar record is created as part of a self- or guided-ingest process.

Nine institutions routinely create forensic disk images of physical media as part of their work ow. Two of the institutions that do not (AUC, Odum) primarily receive born-digital content via network transfer, rather than receiving physical storage media. While MIT does not create forensic disk images for all physical media, they do when physical media are considered to be the master copy of les, or have other high evidential value. As with AUC and Odum, though, MIT receives most of its digital material from network transfer or network drives. Eight of the nine institutions use Guymager in the BitCurator environment to create these disk images,[4] although this tool is often used alongside other disk imaging utilities like Forensic Toolkit Imager.[5] Eight institutions analyze forensic or technical information about les. Six of these institutions use tools in the BitCurator environment to accomplish these tasks. This information is gathered and analyzed for a number of di erent purposes, such as informing appraisal decisions or generating technical metadata.

All institutions create or capture descriptive metadata at some point in their work ow, and nine of the institutions use ArchivesSpace for this step. Nine of the institutions also create or capture technical metadata. Four institutions migrate metadata either across formats and/or platforms as part of this process.

All institutions maintain processed digital materials in both dedicated preservation environments and in repositories, although the capacities and capabilities of these environments vary across the partners. In some cases, the repository also serves as a preservation environment, such as the Stanford Digital Repository. In other cases, a separate environment is used speci cally for preservation, such as Preservica,[6] a commercial digital preservation platform, at KHS. In some cases, copies of materials are stored in multiple locations for preservation purposes. For example, Duke writes multiple copies to tape in addition to storing collections in the Duke Digital Repository, and are currently seeking a geographically-separate partner site to store one of these copies. Odum maintains ve copies of their data across four locations.

[4] https://guymager.sourceforge.io/
[5] https://accessdata.com/product-download/ftk-imager-version-3.4.3
[6] https://preservica.com/

Table 1: Overview of Partner Institutions

Partner	Description	OSS Environments
Atlanta University Center, Robert W. Woodru Library (AUC)	AUC is an independent academic library providing information services to a consortium of Historically Black Colleges and Universities.	BitCurator, ArchivesSpace
District of Columbia Public Library (DCPL)	DCPL is a public library system serving Washington, D.C. The Special Collections employs two digital curation librarians.	BitCurator, ArchivesSpace
Duke University Libraries	Duke is a private research university in Durham, NC. Duke employs sta responsible for digital curation across a number of library units.	BitCurator, ArchivesSpace
Emory University, Stuart A. Rose Manuscript, Archives, and Rare Book Library	The Rose Library at Emory, a private research university in Atlanta, GA, collects a range of born-digital manuscript collections.	BitCurator
Kansas Historical Society (KHS)	KHS collects materials documenting Kansas History, and serves as the o cial repository of government records.	BitCurator
Massachusetts Institute of Technology Institute Archives and Special Collections (MIT)	The Institute Archives and Special Collections serves as a repository for institutional records of MIT, a major research university in Cambridge, MA.	BitCurator, ArchivesSpace, Archivematica
Mount Holyoke College (MHC)	A member of the Seven Sisters and the Five College Consortium, MHC is a small liberal arts college in South Hadley, MA.	ArchivesSpace
New York Public Library (NYPL)	A public library system serving New York City, NYPL has three research libraries that collect archival material, as well as a department for Special Collections and Preservation Services.	BitCurator, ArchivesSpace
New York University (NYU)	A private research university in New York City, NYU acquires a variety of archival materials, including those housed in the Fales Collection.	BitCurator, ArchivesSpace, Archivematica
Odum Institute	Part of the University of North Carolina at Chapel Hill, Odum manages social science data throughout the research lifecycle.	BitCurator
Rice University, Woodson Research Center	Woodson Research Center is the Special Collections and University Archives for Rice, a private research university in Houston, TX.	BitCurator, ArchivesSpace
Stanford University	Stanford is home to 23 libraries, all 19 of those under the direction of the University Librarian collect digital resources.	BitCurator, ArchivesSpace

All institutions describe materials for some kind of discovery layer, such as a nding aid or as a resource record in a public-facing content management system. Seven of the institutions use ArchivesSpace for this function.

5.2 Steps Less Common Across Institutions

While there are broad similarities in digital curation activities across the institutions, there are also many steps that are practiced at only one or a handful of institutions. In some instances, this may re ect genuine novelty of a particular work ow; in other cases, these steps may be practiced experimentally at other institutions, but not yet formally included as part of regular digital curation work.

Only one institution (Emory) quarantines les, and only in cases when they have been acquired from a working computer, and only four (Duke, Emory, MIT, NYPL) perform virus or malware checks. All four use ClamAV in the BitCurator environment for this step. Only Duke and Emory document existing le structure of disks, and both use walk in the BitCurator environment for this.

Only Emory and NYPL explicitly conduct inventories of born-digital collections, although many other work ow steps such as documenting characteristics of physical storage media, documenting existing le structure, analyzing forensic or technical information about les could be seen as proxies for or supplements to a traditional archival inventory.

Four institutions assign unique identi ers to born-digital materials (MIT, Rice, Stanford, Odum), and two other institutions rename les using institutional naming conventions (AUC and DCPL). Only

three institutions deduplicate files as a regular part of their workflow. Emory and MIT both use FSlint in the BitCurator environment, while NYU uses Forensic Toolkit[7] to carry out this task.

5.3 Comparison of Step Order and Stages

There are similarities with how institutions generally organize their workflows, from accessioning materials to providing access, but many differences within these stages. In modeling and analyzing the visual workflow diagrams, we have organized tasks into four stages, based on how partners themselves conceptualized their workflows: pre-accessioning, accessioning, processing, and access.

5.3.1 Pre-Accessioning. There are notable differences in how institutions acquire materials, and to what extent archival units are involved in the acquisition process. All partners interact with donors in some manner before acquisition, to gain information about digital materials and to arrange the transfer of materials to the archives, but interactions range from informal conversations on an ad hoc basis (e.g. a collections curator happens to bring them in) to formalized submission systems that solicit metadata from donors. Seven partners conduct pre-acquisition assessment of materials. For instance, NYU collects forensic and technical information about files and file types, and also assists donors in reviewing materials before collections are acquired.

An important caveat to note is that the interview instrument primarily collected information on partner accessioning and processing activities. In many cases, pre-accessioning activities are handled by collections curators external to an archives unit, so pre-accessioning information is likely not exhaustive. Approaches to communicating with donors about acquiring digital materials is also an area lacking clear professional consensus; for instance, Whyte [25] points to this as a question for further research after her own institutional assessment using workflows.

5.3.2 Accessioning. The distinction between accessioning and processing can be unclear for digital collections, as digital materials often require preparatory work before traditional processing steps like arrangement and description can begin. In general, creating an accession or resource record serves as the hinge between these stages. In comparison to the other stages, there is the greatest degree of variability in what steps fall under accessioning.

At some institutions (DCPL, Duke, KHS, MIT), accessioning is relatively brief, consisting of generating an accession record and collecting metadata. Other institutions (NYPL, NYU, Stanford) carry out more extensive steps in the accessioning stage, such as inventorying the collection, documenting characteristics of physical storage media, normalizing file formats, and analyzing forensic and technical information about files. Institutions may also perform these activities, although later in the workflow. Strategies for appraising digital materials prior to acquisition and during accessioning remain an area of active discussion and development in the field.

5.3.3 Processing. Processing forms the heart of the institutions' workflows. Many of the points made above about common and less common workflow steps reflect variations across partners' approaches to processing.

Our analysis revealed some interesting observations about the differences between processing digital and analog materials. Only five of the institutions describe arrangement as part of their workflows (Duke, Emory, NYPL, NYU, Rice). The fact that relatively few of the institutions discuss arrangement as a distinct step for digital materials bears further inquiry. Potentially, this indicates an uncertainty for institutions as to how best to provide access to digital collections. If mechanisms for discovery and access are not yet formalized, there may be less understanding or clear direction to take regarding arrangement for what uses and for what users are materials being arranged?

Several institutions maintain distinct workflows for processing email (DCPL, MIT, NYPL, NYU). All four use ePADD[8] for processing email, and NYU and NYPL also use GotYourBack[9] As personal digital manuscript collections become more prominent in archives and special collections, more institutions may develop workflows specific to email, which is similar in many ways to letters, a staple of analog manuscript collections.

5.3.4 Access. In describing their workflows, all institutions stress the importance of providing access to digital collections, although there is great variability across the partners, characterized by some degree of uncertainty. Some institutions provide online access to certain collections. Others provide reading-room only access, although few have wholly formalized procedures for making digital collections locally available. As with pre-accessioning, however, the interview instrument did not solicit as much detail about providing access as about accessioning and processing stages.

5.4 Comparison of Tools Used

Institutions use a combination of OSS, commercial, and homegrown tools and systems across their workflows. This range illustrates a premise of the present research, reflecting the diversity of digital curation activities carried out and the inability of any one tool to meet all needs.

5.4.1 OSS. The use of OSS tools for digital curation among these institutions is directly influenced by the sampling, as partners participated in the project because of their interest in and commitment to using at least one of three OSS systems. However, beyond ArchivesSpace, Archivematica, and the BitCurator environment, partners use a number of other OSS or freeware tools. Many partners use Kryo flux[10] to create images of floppy disks. Bagger[11] is used by many of the partners to create bags (containers, based on conventions from the Library of Congress, that include both payload data and metadata about the payload), especially for creating OAIS-compliant information packages. Several partners use AVPreserve[12] tools for a variety of steps. As already mentioned, many partners use ePADD for working with email. Dataverse[13] is integral to much of Odum's digital curation activities. Many partners use

[7] https://accessdata.com/products/computer-forensics/ftk

[8] https://library.stanford.edu/projects/epadd
[9] https://github.com/jay0lee/got-you-back/wiki
[10] https://www.kryoflux.com/
[11] https://github.com/LibraryOfCongress/bagger
[12] https://www.weareavp.com/
[13] https://dataverse.org/

Hyrda (now Samvera)[14] or Fedora[15] as the basis for their digital repositories. These partners' use of OSS tools both within and outside the scope of OSSArcFlow demonstrates the importance of OSS for these institutions' work ows, re ecting a broader enthusiasm in the eld about the possibilities (and challenges) for using OSS tools and environments as the foundation for the stewardship of digital collections [13].

5.4.2 Commercial. Partners also use a range of commercial tools throughout their work ows, in some cases to supplement OSS tools and environments, and in other cases as the primary tools for digital curation activities.

Several partners use Forensic Toolkit Imager for disk imaging, and institutions use another Access Data product, Forensic Toolkit, for other work ow steps. For instance, NYU uses the Bookmarking feature of Forensic Toolkit for arranging les. NYPL also uses Forensic Toolkit for many other processing activities like appraisal and arrangement. Institutions often use Forensic Toolkit and Forensic Toolkit Imager alongside the BitCurator environment. Duke does use the BitCurator environment for disk imaging, but Forensic Toolkit Imager is more often used because it is easier to train students on. NYPL mostly uses Forensic Toolkit Imager for hard drives, but occasionally uses the BitCurator environment for early Mac or Linux objects. Stanford primarily uses Forensic Toolkit because the digital archivist has a personal a nity for the tool, but other sta are interested in increasing the use of the BitCurator environment because its outputs are more readily machine-actionable.

Throughout work ows, partners use Google Sheets and Google Forms to capture and track accession information and other metadata. Google Docs and Google Drive are also used by partners for managing internal documentation and communication. In addition to Google products, many partners use Amazon products; in particular, Amazon Glacier and other Amazon Web Services are frequently used to back up collections.

5.4.3 Homegrown. Complementing widely available OSS tools and commercial products, partners also use a variety of homegrown scripts and applications, either developed in-house or adapted from open-source code made available by other institutions. In some cases, these applications ful ll discrete, targeted needs. For instance, Duke uses STEADY,[16] a Ruby application, for transforming metadata, as well as an ArchivesSpace plugin developed by Harvard that allows for direct importing of spreadsheets; and MHC uses drxfer,[17] a tool for transferring digital records. In other cases, these applications are responsible for signi cant portions of digital curation activities, such as DART, a homegrown collection management system used at KHS.

6 DISCUSSION

As discussed above, there are many broad similarities as well as some signi cant di erences across the digital curation work ows at the 12 partner institutions. While the comparison of work ow steps, stages, and tools is itself instructive, the project team also endeavored to identify major sociotechnical factors impacting these digital curation activities. This section covers prominent themes emerging from our qualitative analysis of the interview data.

6.1 Work ows Under Construction

Across nearly all partner institutions, participants remark that digital curation work ows are still in active development, with many aspects incomplete or unformalized. This has impacted the ability to further develop software supporting digital curation activities. While all institutions have articulated numerous pain points involved in moving data across various systems, few are able to describe issues at the level of granularity necessary for elaborating formal software requirements.

For some, work ows remain ad hoc because they do not have a consistent volume of incoming digital materials. Without regular, sustained processing of materials, digital curation activities are only put into practice intermittently. Although work ows may remain underdeveloped because collections have been relatively small and inconsistent, both DCPL and KHS express concerns about eventually needing to scale up infrastructure to process larger collections that they anticipate in the future. This raises a `chicken and egg' conundrum: should institutions scale up work ows in anticipation of larger and more consistent acquisitions of digital materials, or should institutions wait until these collections materialize to formalize and bulk up their work ows?

Even for institutions with established digital collections, significant aspects of digital curation remain marked by uncertainty, including methods for appraising digital materials prior to acquisition or during accessioning, describing digital materials, and providing access to digital collections. While many have some work ow steps in place for these activities, partners continue to engage in productive discussions at OSSArcFlow meetings about ongoing developments in these areas. In response to this ubiquitous challenge, one of the most common strategies discussed by institutions is to formalize work ows, policies, and procedures. Several partners have cited the bene t of formalization, but actually undertaking these e orts remains aspirational. Others have codi ed certain processes in the form of manuals and guidelines. Contributing to these formalization e orts at the level of the broader profession is one of the main aims of the present research.

6.2 Work ows Tailored to Types of Materials

In many cases, the most formalized and developed work ows are those tailored to particular types of materials. While KHS describes having ad hoc work ows for born-digital manuscript materials, they maintain an established work ow for a born-digital newspaper collection that receives a consistent stream of incoming materials from a vendor. Similarly, AUC has yet to acquire signi cant born-digital manuscript collections but regularly processes ETDs. Another major driver in this regard is a split between born-digital and digitized materials. For instance, Duke maintains distinct work ows for these two kinds of materials, impacting how materials are described, who processes the collections, and how materials are made available.

Related to type of material, the source of a collection (i.e., institutional records, governmental records, or manuscript collections)

[14] https://samvera.org/
[15] https://duraspace.org/fedora/
[16] http://steady2.herokuapp.com/
[17] https://github.com/mtholyoke/drxfer

can lead to distinctly tailored work ows. Regularly ingested materials like institutional records may have established pipelines, whereas manuscript materials acquired by collections curators may not. Especially at institutions that regularly collect manuscript materials, partners described varying relationships with collections curators. As a partner at NYPL re ects, "Some curators bring me in early, some just give three record cartons full of HDDs with no records." Partners at Stanford and NYU routinely consult with donors and collections curators before materials enter the archives, while other partners mention nding digital media unexpectedly in boxes of analog materials. Many partners aspire to develop more formal mechanisms for early interaction with donors and curators.

6.3 Institutional Policies and Culture

Institutional policies and culture both within archives units and at home organizations more broadly exert in uence in many different ways. Institutional attitudes toward experimenting with technologies to address local issues is an impediment for several partners. The partner at DCPL, in particular, stresses that he has little freedom to experiment with technologies, as the organization relies heavily on vendors for IT solutions: "it is very ingrained in the culture to pay someone else to x it." As Gengenbach et al. [18] point out, administrative reservations about OSS can be a major obstacle for implementing OSS digital curation tools in library and archives contexts. Some partners describe more freedom to try out OSS tools, but note other practical limitations like minimal IT support to troubleshoot any issues. To address this situation, though, both Rice and NYPL have used the implementation of OSS tools to strategically build stronger relationships with their supportive albeit limited IT sta by directly engaging them as stakeholders throughout the process, and keeping them informed about their goals for using this software.

For Stanford and Odum, the ability to identify sensitive information is a driving concern. Stanford libraries must adhere to a university-wide data privacy policy, and Odum needs to consistently identify personally identi able information in social science research data. Both use bulk extractor in the BitCurator environment for this task, but have encountered serious di culties: Stanford has found that it requires signi cant processing power to run bulk extractor against collections hundreds of terabytes in size; Odum runs bulk extractor against large metadata les, and evaluating bulk extractor output requires them to visually scan through large spans of XML. This concern has a direct e ect on Stanford's work ow, as collections potentially containing sensitive information are stored on a dark server, storage space which lacks many of the built-in curation functionalities of the Stanford Digital Repository.

6.4 Managing Limited Resources

The imperative to carry out digital curation with limited budget, sta , time, storage space, technical support, or other resources is a pervasive factor discussed by nearly all of the partners. As digital curation necessitates ongoing investment, managing these resources plays a key role in shaping work ows. For instance, KHS can only store a set amount of material in Preservica due to limited funds for storage space. Many digital curation projects at AUC are dependent upon grant funds; similarly, the partner at Stanford mentions that many processing activities are accomplished only if a project archivist can be dedicated to the collection.

In response to this issue, many institutions pursue strategies of prioritized or tiered curation. At Emory, tiered curation is a documented approach, with detailed criteria to channel collections into one of three tiers [23]. Others employ prioritized curation in a more ad hoc fashion; for example, a partner at MHC observes that materials are processed more quickly if donated by a prominent individual. A range of factors may in uence this prioritizing, such as anticipated use or the degree of di culty involved in processing the materials. For Emory, collections with unusual le formats that will require additional work to process are placed at a higher tier. The level of priority may impact work ows in a number of ways, with some work ow steps reserved for higher tiers.

6.5 Issues with Technologies

Partners experience various technological issues that a ect digital curation activities, from particular tools not functioning as anticipated to di culties moving data between disparate systems. The latter has been a principal concern of the OSSArcFlow project, and the research has revealed many speci c points of desired interoperability that bear further investigation. Many partners have expressed desires to more easily move data generated by reports in the BitCurator environment to descriptive elds in ArchivesSpace. However, as many partners continue to grapple with approaches to representing digital materials in ArchivesSpace (and a variety of other discovery platforms), the eld may still be at too early a stage for signi cant development work in this regard.

Others describe a need for some utility to track materials as they move across work ows. A partner at MIT discusses di culties encountered by not having ready access to this information: some sta are only intermittently available, and the archivist cannot quickly direct these sta when they are able to work on digital curation. Related to this, the partner at MIT discusses challenges moving collection les in and out of Archivematica throughout the work ow, for instance if collections need to be moved into the BitCurator environment for a particular step and back into Archivematica following this step.

Several partners discuss issues associated with transitioning to new technologies, including those systems focused on by the OSSArcFlow project. Implementing new tools or systems requires periods of active development and testing, which can draw on more resources than initially anticipated, but can also open up new possibilities. A partner at MHC notes that the transition from Archivist's Toolkit to ArchivesSpace has made them aware of a need for greater exibility to deal with larger-scale work ows.

As discussed above, all partners have storage environments for processed collections, but the nature of these varies considerably. Some partners maintain dedicated repositories with built-in curation functionalities, such as running xity checks and le characterization, while other repositories lack such functionality. This directly impacts work ows, as steps not carried out by the repository may then need to be done manually or with the use of additional tools.

7 CONCLUSION

Although this research is premised upon the ability of workflow models to reflect the current state of digital curation in local contexts, perhaps the most significant finding is that there is no such thing as a single, stable, wholly descriptive 'digital curation workflow' for any of the partner institutions. However, the process of collaboratively constructing the workflow diagrams provided crucial space and time for the partners to reflect on digital curation activities at their institutions, and these workflow documents will continue to serve as resources both to the OSSArcFlow partners and to the cultural heritage field more broadly. Already, the project team has received feedback commenting on the utility of these workflow models to gain insight into how institutions are carrying out digital curation work in quite different ways.

However, the project also revealed limitations of formal representations of workflows, though these insights are instructive for further research. Especially as digital curation scholarship and practice continues to rapidly develop, these digital curation activities will likewise evolve and adapt—to new technologies, improved skills training, and different kinds of digital collections. The workflow documents may communicate coherent and fixed processes, but in reality, institutions have many workflows, all in varying states of flux. Digital curation practitioners are currently grappling with many issues, especially in regards to implementing OSS digital stewardship tools. Though these matters are far from stable, this research has demonstrated that modeling workflows can spark incisive discussions that shed new light on how practitioners are thinking and working through pressing challenges.

ACKNOWLEDGMENTS

OSSArcFlow is funded by a grant from the Institute for Museum and Library Services (IMLS grant LG-71-17-0016-17).

REFERENCES

[1] Elsa K. Anderson. 2014. Workflow Analysis. *Library Technology Reports* 50, 3 (April 2014), 23–29.

[2] Pamela Andrews, Karen Harker, and Ana Krahmer. 2018. Applying the Analytic Hierarchy Process to an Institutional Repository Collection. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL '18)*. ACM, Fort Worth, TX, 37–40. https://doi.org/10.1145/3197026.3197064

[3] Susanne Annand, Sally DeBauche, Erin Faulder, Martin Gengenbach, Karla Irwin, Julie Musson, Shira Peltzman, Kate Tasker, Laura Uglean Jackson, and Dorothy Waugh. 2018. Digital Processing Framework. https://ecommons.cornell.edu/handle/1813/57659

[4] Sarah Barbrow and Megan Hartline. 2015. Process mapping as organizational assessment in academic libraries. *Performance Measurement and Metrics* 16, 1 (March 2015), 34–47. https://doi.org/10.1108/PMM-11-2014-0040

[5] Christoph Becker, Hannes Kulovits, Mark Guttenbrunner, Stephan Strodl, Andreas Rauber, and Hans Hofman. 2009. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries* 10, 4 (Dec. 2009), 133–157.

[6] Christoph Becker and Andreas Rauber. 2011. Preservation Decisions: Terms and Conditions Apply. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL '11)*. ACM, Ottawa, Canada, 67–76. https://doi.org/10.1145/1998076.1998089

[7] Anthony Cocciolo. 2016. When Archivists and Digital Asset Managers Collide: Tensions and Ways Forward. *The American Archivist* 79, 1 (June 2016), 121–136. https://doi.org/10.17723/0360-9081.79.1.121

[8] Maria Collins. 2009. Evolving Workflows: Knowing when to Hold'em, Knowing when to Fold'em. *The Serials Librarian* 57, 3 (Sept. 2009), 261–271. https://doi.org/10.1080/03615260902877050

[9] J. Gordon Daines III. 2011. Re-engineering Archives: Business Process Management (BPM) and the Quest for Archival Efficiency. *The American Archivist* 74, 1 (April 2011), 123–157. https://doi.org/10.17723/aarc.74.1.h8159344u8331165

[10] Beverly Dowdy and Rosalyn Raeford. 2014. Electronic Resources Workflow: Design, Analysis and Technologies for an Overdue Solution. *Serials Review* 40, 3 (July 2014), 175–187. https://doi.org/10.1080/00987913.2014.950040

[11] Nicola Ferro, Gianmaria Silvello, Erik Buelinckx, Boris Doubrov, Antonella Fresa, Magnus Gaber, Klas Jadeglans, Borje Justrell, Bert Lemmens, Jerome Martinez, Victor Munoz, Sonia Oliveras, Claudio Prandoni, Dave Rice, Stefan Rohde-Enslin, Xavi TarrÃĪs, Erwin Verbruggen, Benjamin Yousefi, and Carl Wilson. 2018. Evaluation of Conformance Checkers for Long-Term Preservation of Multimedia Documents. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries (JCDL '18)*. ACM, Fort Worth, TX, 145–154. https://doi.org/10.1145/3197026.3197037

[12] Martin Gengenbach. 2012. *"The Way We Do It Here": Mapping Digital Forensics Workflows in Collecting Institutions*. Ph.D. Dissertation. University of North Carolina - Chapel Hill, Chapel Hill, NC. http://digitalcurationexchange.org/system/files/gengenbach-forensic-workflows-2012.pdf

[13] Martin Gengenbach, Shira Peltzman, Sam Meister, Blake Graham, Dorothy Waugh, Jessica Moran, Julie Seifert, Heidi Dowding, and Janet Carleton. 2016. OSS4EVA: Using Open-Source Tools to Fulfill Digital Preservation Requirements. *Code4Lib Journal* 34 (2016). https://journal.code4lib.org/articles/11940

[14] AIMS Working Group. 2012. *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*. Technical Report. https://perma.cc/JE6D-MTLT

[15] Emily R. Novak Gustainis. 2012. Processing Workflow Analysis for Special Collections: The Center for the History of Medicine, Francis A. Countway Library of Medicine as Case Study. *RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage* 13, 2 (2012), 113–128. https://rbm.acrl.org/index.php/rbm/article/view/378

[16] Erin Lawrimore. 2013. Collaboration for a 21st Century Archives: Connecting University Archives with the LibraryâĂŹs Information Technology Professionals. *Collaborative Librarianship* 5, 3 (Jan. 2013). https://digitalcommons.du.edu/collaborativelibrarianship/vol5/iss3/4

[17] Sam Meister and Alexandra Chassanoff. 2014. Integrating Digital Forensics Techniques into Curatorial Tasks: A Case Study. *International Journal of Digital Curation* 9, 2 (2014), 6–16. http://www.ijdc.net/article/view/9.2.6

[18] Adam Mikeal, James Creel, Alexey Maslov, Scott Phillips, John Leggett, and Mark McFarland. 2009. Large-scale ETD Repositories: A Case Study of a Digital Library Application. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '09)*. ACM, Austin, TX, 135–144. https://doi.org/10.1145/1555400.1555423

[19] Natalie Ornat and Renee Moorefield. 2018. Process mapping as an academic library tool: Five steps to improve your workflow. *College & Research Libraries News* 79, 6 (2018), 302–305. https://crln.acrl.org/index.php/crlnews/article/view/17004

[20] Seth Shaw, Richard Adler, and Jackie Dooley. 2017. *Demystifying IT: A Framework for Shared Understanding between Archivists and IT Professionals*. Technical Report. OCLC, Dublin, OH. https://www.oclc.org/research/publications/2017/oclcresearch-demystifying-it-shared-understanding.html

[21] Ed Summers and Ricardo Punzalan. 2017. Bots, Seeds and People: Web Archives as Infrastructure. In *Proceedings of the 2017 ACM Conference on Computer Supported Collaborative Work and Social Computing*. ACM, Portland, Oregon, 821–834.

[22] Shaun Trujillo, Meghan Bergin, Margaret Jessup, Johanna Radding, and Sarah Walden McGowan. 2017. Archivematica outside the box: Piloting a common approach to digital preservation at the Five College Libraries. *Digital Library Perspectives* 33, 2 (March 2017), 117–127. https://doi.org/10.1108/DLP-08-2016-0037

[23] Dorothy Waugh, Elizabeth Russey Roke, and Erika Farr. 2016. Flexible processing and diverse collections: a tiered approach to delivering born digital archives. *Archives and Records* 37, 1 (Jan. 2016), 3–19.

[24] Andrew Weidner, Sean Watkins, Bethany Scott, Drew Krewer, Anne Washington, and Matthew Richardson. 2017. Outside the Box: Building a Digital Asset Management Ecosystem for Preservation and Access. *Code4Lib Journal* 36 (2017).

[25] Jess Whyte. 2017. Preservation Planning and Workflows for Digital Holdings at the Thomas Fisher Rare Book Library. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries (JCDL '17)*. IEEE Press, Toronto, Canada, 323–325. http://dl.acm.org/citation.cfm?id=3200334.3200394