ARTICLE

Navigating Uncharted Waters

Utilizing Innovative Approaches in Legacy Theses and Dissertations Digitization at the University of Houston Libraries

Annie Wu, Taylor Davis-Van Atta, Bethany Scott, Santi Thompson, Anne Washington, Jerrell Jones, Andrew Weidner, A. Laura Ramirez, and Marian Smith

ABSTRACT

In 2019, the University of Houston Libraries formed a Theses and Dissertations Digitization Task Force charged with digitizing and making more widely accessible the University's collection of over 19,800 legacy theses and dissertations. Supported by funding from the John P. McGovern Foundation, this initiative has proven complex and multifaceted, and one that has engaged the task force in a broad range of activities, from purchasing digitization equipment and software to designing a phased, multiyear plan to execute its charge. This plan is structured around digitization preparation (phase one), development of procedures and workflows (phase two), and promotion and communication to the project's targeted audiences (phase three). The plan contains step-by-step actions to conduct an environmental scan, inventory the theses and dissertations collections, purchase equipment, craft policies, establish procedures and workflows, and develop digital preservation and communication strategies, allowing the task force to achieve effective planning, workflow automation, progress tracking, and procedures documentation. The innovative and creative approaches undertaken by the Theses and Dissertations Digitization Task Force demonstrated collective intelligence resulting in scaled access and dissemination of the University's research and scholarship that helps to enhance the University's impact and reputation.

INTRODUCTION

To answer the call of implementing University of Houston (UH) Libraries strategic plan to position the Libraries as a campus leader in research and transform library space to reflect evolving modes of learning and scholarship, the UH Libraries launched a cross-departmental task force in 2019 charged with digitizing the University's extensive print theses and dissertations collection. Providing online access to newly digitized theses and dissertations boosts the reach and impact of our institution's research and scholarship while expanding available space for computing,

Annie Wu (awu@uh.edu) is Head of Metadata and Digitization Services and the Ambassador Kenneth Franzheim II and Mrs. Jorgina Franzheim Endowed Professor, University of Houston Libraries. **Taylor Davis-Van Atta** (tgdavis-vanatta@uh.edu) is Director of the Digital Research Commons, University of Houston Libraries**. Bethany Scott** (bscott3@uh.edu) is Head of Preservation and Reformatting, University of Houston Libraries. **Santi Thompson** (sathompson3@uh.edu) is Associate Dean for Research and Student Engagement and the Eva Digital Research Endowed Library Professor, University of Houston Libraries. **Anne Washington** (washinga@oclc.org) is Semantic Applications Product Analyst, OCLC. **Jerrell Jones** (jjones46@uh.edu) is Digitization Lab Manager, University of Houston Libraries. **Andrew Weidner** (andrew.weidner@bc.edu) is Head of Digital Production Services, Boston College Libraries. **A. Laura Ramirez** (alramirez@uh.edu) is Senior Library Specialist, University of Houston Libraries. **Marian Smith** (mrsmith8@uh.edu) is Digital Photo Tech, University of Houston Libraries.

https://doi.org/10.6017/ital.v41i3.14719

technology, and faculty and student learning and research activities. A study by Bennett and Flanagan revealed the positive impact and benefits of online dissemination of theses and dissertations, including enhanced discoverability by Google's strong indexing capabilities, significant increase in the usage of the works, and an overall enhancement of the reputation of an institution.[1] Encouraged by the positive outcomes and supported by funding from the John P. McGovern Foundation to initiate this project, the Theses and Dissertations Digitization (TDD) Task Force developed a phased project plan and utilized creative, automated processes and methods to execute it. This article articulates the TDD project planning and the innovative work undertaken by the task force to achieve efficiency in making our print theses and dissertations readily available to new readerships around the world.

LITERATURE REVIEW

Over the past several decades, research libraries have been building programs around digitization and open access repository infrastructures, largely aimed at expanding their digital collections and engaging communities with newly available research materials. For some, part of their programming has included projects that digitize their institution's legacy print collections of theses and dissertations. The review below explores literature on the mass digitization process, including institutional case studies, guidance documents, legal and policy papers, and local documentation developed as libraries have planned and implemented these projects.

Any library tackling a retrospective thesis and dissertation project needs a framework for determining the copyright status of these works *en masse.* Perhaps it is no coincidence, then, that copyright concerns are the most heavily documented aspect of the process. Clement and Levine provide the definitive work to date on copyright and the publication status of theses and dissertations written in the United States before 1978. Their study asserts that "pre-1978 American dissertations were considered published for copyright purposes by virtue of their deposit in a university library or their dissemination by a microfilm distributor."[2] They go on to write that, "for copyright purposes, these were acts of publication with the same legal effect as dissemination through presses, publishers, and societies."[3] They suggest that libraries should investigate the copyright status for theses and dissertations authored between 1909 and 1978 (typically found on the title page and verso); if there is no copyright notice, then the thesis or dissertation is likely in the public domain and eligible for digitization and public release without permission. Moreover, even those works that have a printed copyright notice might have fallen out of copyright if they were not renewed after 28 years for the same length of time.[4]

Broad guidance and best practice for copyright status and other matters of process around theses and dissertations is provided in *Guidance Documents for Lifecycle Management of ETDs*, which acknowledges that legal services may be required for some retrospective thesis and dissertation digitization projects, especially "before scanning without the permission of former students."[5] The authors assert that information professionals should investigate any "appropriate access options" with institutional legal expertise before engaging in a retrospective digitization project and articulate the two most commonly encountered copyright scenarios: "[either] former student authors may not allow the reproduction and open dissemination of their work, or unauthorized copyrighted material was used in the original theses and dissertations."[6] Strategies that might be employed to determine copyright status include "consulting with legal counsel at one's institution to see where it stands on this issue; negotiating with commercial entities that make such content

available at a price so that institutions can have some control over it for the purpose of broader access; and working with groups such as alumni associations, colleges, departments, and graduate schools to establish contact with thesis and dissertation authors for securing their permission to digitize, and render available online, their past scholarship."[7]

On the question of public access to newly digitized works, the *Guidance Documents* detail the implications of the "transition from print to electronic," which "has led to increased scrutiny over who will be allowed to access the electronic versions and how widely they will be disseminated."[8] When there is any legal doubt, there are many reasons for libraries to exercise caution and restrict access to electronic theses and dissertations; that said, "research available on the Web immediately upon submission of the final, approved thesis can prove advantageous to the newly-degreed student, the institution, and other researchers."[9] Again, consulting legal officers and the original authors, if possible, remains the consensus approach to establishing a strategy for access to digitized theses and dissertations.

The *Guidance Documents* also touch on the thorny issue of digitizing theses and dissertations that contain third-party content. They summarize the history and routine application of the fair use doctrine in both the creation and dissemination of scholarly works but provide little firm guidance on the matter.[10] Indeed, after reviewing the entire body of literature on retrospective thesis and dissertation projects, this remains a practical challenge that any library undertaking a mass digitization project must consider and the associated risks must be accounted for.

In recent years, several case studies have documented institutions' efforts to digitize and make more widely available legacy theses and dissertations. Of the institutions that the TDD Task Force reviewed for the environmental scan, none of their case studies attempts an exhaustive documentation of end-to-end workflows and processes developed to execute the task; most focus on particularly difficult questions inherent to the process.

Martyniak provides a rationale for the University of Florida's (UF) retrospective scanning project and details their process for contacting authors before works were scanned.[11] The workflow outlines several points of contact with authors to obtain signed distribution agreements, as well as UF's approach to automate this process as much as possible. Notably, the distribution agreement form and correspondence templates are provided as appendices to the article.[12]

As part of this retrospective digitization project, UF also released a scanning policy that articulates their approach to determining the copyright status of works and their resultant practice.[13] This policy document is an excellent example of an institution's implementation of Clement and Levine's research described above.

Likewise, Mundle describes the methods used by Simon Fraser University (SFU) to establish its approach to the issues of copyright status and access, ultimately resulting in a public thesis access policy and procedures for contacting authors whenever possible to offer them the ability to opt their work out of the project.[14] Unlike the UF, SFU began scanning before any explicit permission had been obtained from authors. SFU also shares their use of scripts to automate the ingest of metadata from original MARC records into their DSpace repository.[15]

Piorun and Palmer, meanwhile, focus on an analysis of the time and cost associated with digitizing 300 doctoral dissertations for a newly implemented institutional repository at the University of

Massachusetts Chan Medical School.[16] Piorun and Palmer detail the library's process for obtaining cost comparisons from external vendors as well as estimated costs, including labor, associated with undertaking the task in house.[17] Issues of workflow, policy development, and permissions are also addressed with an emphasis on developing accurate and streamlined methods of processing works; however, Piorun and Palmer conclude, "regardless of the amount of planning and thought that goes into a project, there is always the possibility that each record or file will need to be reworked."[18]

Shreeves and Teper discuss theses and dissertations' complicated status as grey literature and the University of Illinois Urbana-Champaign (UIUC) Library's digitization project, which they describe as "less of a collection management or preservation issue and more as an effort to tackle broader scholarly communication and outreach issues."[19] After consulting with university legal counsel, digitized works were ingested to the UIUC institutional repository as a restricted (campus-only access) collection. As authors provide consent, access to their work is broadened to the public.

Worley demonstrates that, according to an analysis of circulation numbers, works that are accessible electronically are used dramatically more than print copies, serving as rationale for undertaking digitization of student works.[20] They provide significant detail around Virginia Polytechnic Institute and State University's process to establish file specifications for its digitization process, and image quality/resolution and file format selection are discussed in some detail, with helpful visual examples.[21]

These case studies are particularly valuable in that they provide evidence and cautionary tales around how local contexts have made a difference in copyright and workflow issues. This case study contributes to the existing body of literature by attempting to provide an exhaustive, end-to-end description of the retrospective digitization process—from copyright evaluation, to physical handling, to digitizing with an eye to access controls and digital preservation concerns. Furthermore, our approach to digitization at scale incorporates automation at several points throughout the workflow, representing a production improvement to the decade-old case studies we reviewed.

PROJECT PLANNING AND EXECUTION

Digitizing a large corpus of print theses and dissertations is a complex process touching areas of equipment, copyright policy, workflows for different sections of the process, progress tracking, preservation, and communication. To handle such a multifaceted project, the TDD Task Force designed a plan that divided the project activities into three phases (see table 1). Phase one is dedicated to tasks of preparation such as the environmental scan, copyright permission investigation, digitization equipment purchasing, and print theses and dissertation inventory. Phase two includes activities such as digitization and metadata workflow development, documentation, project tracking, ingestion, and preservation of digitized files. Phase three is mainly for promotion and communication to our researchers on the availability of our digitized theses and dissertations collection. Task force members volunteered to serve in subteams for identified specific tasks in each project phase.

Table 1. Phased planning for the TDD Project

Theses and dissertation digitization task force planning		
Phases	**Task force activities**	**Subteams (*subteam lead)**
Phase one: preparation	Environmental scan	Jerrell, Anne, Crystal, Santi*, Bethany
	Physical theses/dissertations inventory/retention	Bethany
	Copyright permissions and policies	Bethany, Annie, Taylor*
	Purchase equipment	Jerrell*, Crystal
Phase two: workflow development	TD digitization workflow	Jerrell, Crystal*
	TD metadata workflow	Anne*, Taylor, Annie
	TD ingest and publishing workflow	Andrew, Taylor
	TD progress tracking	Annie, Andrew
	Preservation/storage strategy	Bethany*, Santi
Phase three: communication/promotion	Promote DTD to colleagues and researchers	Taylor*, Santi
	Communicate progress to staff and users	Annie*, Santi*
	Develop training materials for stakeholders	Anne*, Crystal*

Phase One: Preparation

A subgroup of the TDD Task Force conducted an environmental scan of similar theses and dissertations digitization approaches previously used by other institutions. The lead for the subgroup created a Google sheet that all group members used to document information found in published literature, public documents, and institutional websites. The lead assigned group members to review information from institutions with publicly available data, including: the University of Florida, the University of North Texas, the University of Illinois Urbana Champaign, Brigham Young University, William and Mary University, Texas A&M University, the University of Arizona, the California Institute of Technology, the Massachusetts Institute of Technology, Iowa State University, Xavier University, Texas Tech University, and the University of British Columbia.

Group members noted relevant information pertaining to a variety of topics focused on theses and dissertations digitization. One of the most prominent was the institution's response to copyright permissions. The group tried to determine if the institution required author permission before releasing a digitized thesis or dissertation (the "opt in" option), or incorporated policies and procedures that prioritized taking down digitized theses and dissertations once requested by the author (the "opt out" option). They observed software and hardware specifications used by other institutions—critical data that would inform the technology needed to complete a project of this scale. The group documented the key components of the digitization and metadata workflows, including roles and responsibilities, sequencing of actions, and the implications that policies and procedures had on the process. This data helped the group understand what gaps, common problems, and emerging best practices existed. Finally, the group reviewed physical retention and preservation strategies articulated by institutions to ensure it understood the long-term stewardship hurdles and requirements for analog and digital material.

Based on the assessment of the 19,800 UH theses and dissertations identified for inclusion in the project, the digitization subgroup members determined that several scanners would be required for agility in digitization production. The TDD digitization workflow was designed so that this project could run effectively, in parallel, with existing digitization projects regardless of the need for some theses to be scanned on existing equipment. Automatic document feed (ADF) scanners were a strategic choice for the rapid scanning of disbound items. Two Canon DR-G2110 ADF scanners were purchased for the project. These scanners were chosen for their scanning speed, scanning quality, ease of use, onboard image preprocessing, and reasonable price point. The Canon DR-G2110 can handle a large page stack, approximately 500 pages. Theses and dissertations can be scanned on the longer or shorter dimension, which allows faster scanning times. Among many innovative features, this duplex scanner simultaneously digitizes both sides of a document, rotates the pages based on text orientation, and auto-crops through preprocessing during the scanning process. This Canon ADF solution makes more image postprocessing automation possible since the resulting scans match our output expectations with minimal user input.

Other scanning options were needed for a smaller subset of theses and dissertations that could not be disbound. The digitization team leveraged an existing Zeutschel OS 12002 planetary scanner for items that could not be disbound. An existing Plustek Opticbook (PO) A300 Plus was used for items with foldouts containing graphs, maps, and illustrations that measure beyond 11 inches on the longest dimension. Additionally, a Plustek Opticbook 3800L was purchased to accommodate fragile US letter–sized pages that are not suitable for ADF scanning. Thin or heavily waxed papers typically do not stand up well to the fast-moving rubber rollers and other internal scanning mechanisms. While the PO 3800L provides a much longer scanning time than the PO A300, both scanners can scan into the page gutter of bound materials, a useful feature for items with insufficient margins.

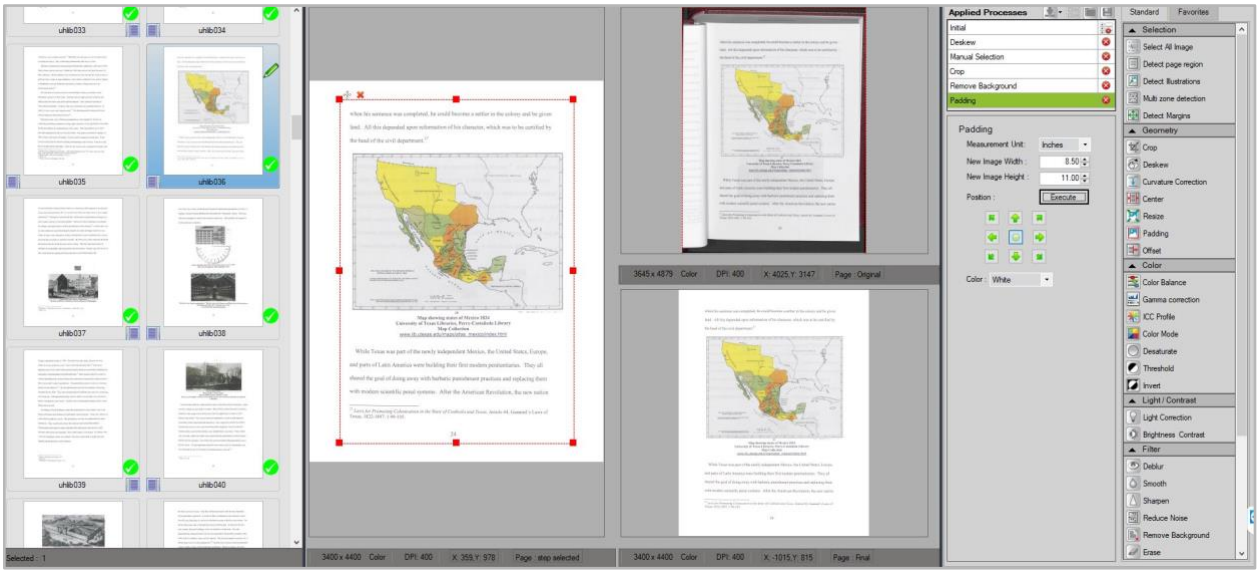

Figure 1. Image processing workflow testing on a thesis in LIMB Processing. The green check marks on the left indicate that a page has been processed correctly.

The Canon ADF scanner operates through two pieces of software working concurrently, Canon CaptureOnTouch V4 Pro and Kofax VRS (an additional product supplied by the task force's scanner vendor). Some image processing settings are applied in Kofax VRS, which communicates with CaptureOnTouch V4 Pro. Both pieces of software were bundled with purchases of the two Canon ADF scanners. LIMB Processing by i2S was also purchased for the project. LIMB Processing is a powerful mass image processing product that operates through user-built processing workflows that can be applied to multiple folders, creating standardized output suitable for automation. The LIMB software can transform an imperfect scan into a fully processed, clean derivative with minimal user input, which is especially useful for transforming legacy image data. ABBYY FineReader Server 14 is used to provide quality optical character recognition (OCR) data and features efficient tools for automation, allowing for large OCR processing jobs to be queued and run recursively with minimal user intervention (see fig. 1). With these powerful tools, UH Libraries has been able to leverage our existing scanners, new scanners, and advanced software to plan for the timely capture of nearly three million pages of content.

The number of theses and dissertations required the implementation of a semiautomatic disbinding system. The Spartan 185A Paper Cutter from the Challenge Machinery Company was purchased to ensure the replication of many clean binding removals. Options from several manufacturers were considered for these needs but the Spartan 185A offered the cutting power needed to cut millions of pages over the life of the project (see fig. 2). The cutter features several safeguards that protect the operator, such as the lowering of a protective acrylic guard and the requirement of two hands, away from the blade, to lower the blade automatically. UH Libraries chose a local cutter blade replacement company that services the equipment quarterly. In addition to the cutter, supplies for binding removal and physical volume management were needed, such as:

- X-Acto knife and/or utility blades
- Recycling bins
- Table brooms and dustpans
- Disbinding tables
- Cutting mats
- Standing mats
- Letter and legal-size folders
- Folder holders
- Surface cleaning materials
- Carts/book trucks

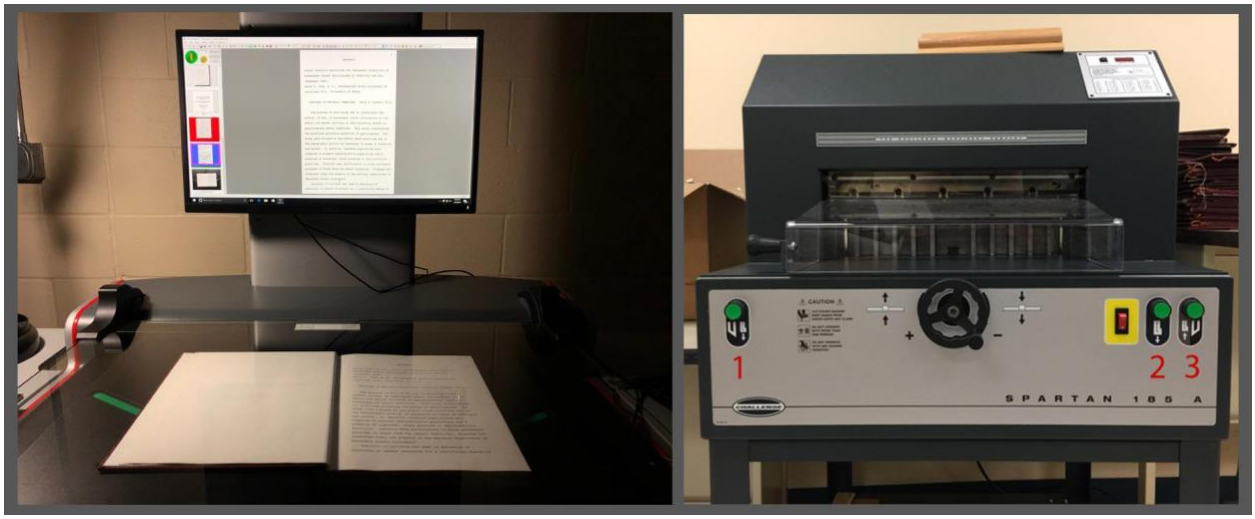

Figure 2. (L) Thesis scanning test on the Zeutschel OS 12002; (R) Challenge Spartan 185A Cutter with red numbers indicating blade lowering and cutting safety button order.

The physical retention of volumes was considered in the context of the overall preservation of the theses and dissertations collection, including the digital preservation approach to the TDD project. The UH Libraries holds two copies of a student's thesis and dissertation. After consulting with stakeholders throughout the library—such as the university archivist, the dean of libraries and associate deans, and Access Services' shelving team for shelf space/storage in different areas of the library building—the task force decided to retain one bound copy of each thesis or dissertation. Additional copies will be weeded from the general collection, and the best copy for digitization will be disbound for feeder scanning using the equipment described above. When only one copy of a thesis or dissertation exists in the collection, it will be scanned using a scanner that will not destroy or damage the binding. The retained theses and dissertations collection will be housed in UH Libraries Special Collections in the secure and climate controlled closed stacks.

Once the TDD Task Force settled on this retention strategy, the digital projects coordinator, a member of the task force who represents Special Collections, conducted a full shelf-read of the theses and dissertations already housed in Special Collections. Using a master tracking spreadsheet that was generated from catalog reports for project tracking and pulling, a small team of student workers reviewed over 20,000 volumes to identify missing titles, titles with multiple copies that can be weeded from Special Collections, and copies with label and/or cataloging errors. Missing titles were transferred from the general stacks to Special Collections, and the items were reshelved in chronological order.

A more extensive shelving shift still needs to be completed to move volumes to accommodate additions and finalize the shelf location for all items in this collection, which will no longer be growing or because all theses and dissertations at UH are submitted electronically as of 2014. As part of the shifting project, the items also need to be checked in and/or have their location codes changed in the catalog to reflect their new permanent home in Special Collections.

Phase Two: Workflow Development

The theses and dissertations digitization workflow starts with pulling physical volumes from shelves. The Task Force generated a report of all UH theses and dissertations and sorted them by call number so that student workers can pull these volumes from the General Stacks in call number order. After the pulled volumes' records are withdrawn from the catalog system, they are shelved by call number order in the "Ready for Digitization" section of the TDD shelf in the library basement, close to the Digitization Lab.

Volumes are pulled from a section of the library stacks dedicated to the TDD project and loaded onto book carts for transfer to the physical volume processing room. Using a custom-built processing table, covers are removed with utility knives and discarded. The text is placed in a folder with a pre-printed label indicating the OCLC number and call number of the volume. The spine of each volume is removed with a Spartan guillotine. The completely disbound volumes, housed in labeled folders, are then moved on book carts to the scanning room.

Prior to scanning, physical volumes are grouped in batches of approximately 50 and a text file is created that lists each OCLC number in a batch, one per line. A simple executable file reads the text file and creates a batch directory. The batch directory is labeled with the current date in YYYYMMDD format and contains a folder for each scanned volume. The scanned volume folders are labeled by OCLC number and contain a *metadata.txt* file that records the volume's descriptive metadata from the UH catalog system in YAML format: a data carrier that is easily readable for both humans and machines.

Scanning is performed with one of the two Canon DR-G2110 high-speed feed scanners controlled by Kofax VRS and CaptureOnTouch V4 Pro software. Before a volume is placed in the scanner, it is checked to ensure that the binding has been completely removed, that there are no pages that have been glued in after binding, and that there are no onion skin pages, irregular page sizes, inserts, or foldouts. If necessary, additional scans for delicate onion skin pages, inserts, or foldouts are performed on a flatbed scanner. Page images are output as 300 dpi grayscale or color TIFFs, and first pass quality control for completeness, page legibility, and rotation, and cropping is performed in CaptureOnTouch.

After page images have been captured, a batch is loaded into LIMB for final processing. Scanned volumes are again checked for completeness, legibility, and orientation. Text pages are processed as 300 dpi bitonal TIFFs. Pages with grayscale or color images are processed as such. When batch processing is complete, the documents' processed signature pages, which include names and signatures of the author, advisor, and committee members, are separated out so they are not included in the final version published online. This step protects the privacy of individuals by not sharing their signature openly over the internet.

The TDD project uses ABBYY FineReader Server 14 to generate full text PDFs and a plain text file for each scanned volume. The data in each scanned volume directory undergoes transformations both before and after the OCR processing. The transformations are accomplished with the TDD Workflow Utility, a Ruby command line application. Before running a batch through OCR, the Archive Digitized Batch function moves the high-resolution master TIFFs to an archive directory and formats the batch directory for the input that ABBYY expects. After OCR processing, the Archive OCR Batch function moves the derivative TIFFs used as OCR input to the archive directory

as well. In a final process before sending the batch to the Metadata Unit, the Process OCR Batch function adds descriptive metadata to the embedded EXIF metadata of each PDF document with ExifTool for improved accessibility.

The TDD Task Force sought to align print materials' metadata standards with the existing metadata standards applied to electronic theses and dissertations in the University's institutional repository, largely based on the *Dictionary of Texas Digital Library Descriptive Metadata Guidelines for Electronic Theses and Dissertations, v.2*.[22] Early on in the project, the metadata subteam reviewed thesis and dissertation records in the institutional repository (IR) as well as MARC catalog records in UH Libraries' library services platform, with special emphasis on the metadata elements used, to identify alignments and gaps. After analysis, the team established the crosswalk from MARC to the qualified Dublin Core profile in the IR (see table 2).

In July 2019, UH Libraries migrated to the Alma library services platform. Prior to this migration, the task force exported TDD MARC records from UH Libraries' former library services platform, Sierra, and crosswalked into Dublin Core metadata fields using the freely available software MARCEdit. Data was further normalized using OpenRefine. At this early stage, OpenRefine proved to be a valuable tool for batch editing and formatting metadata and identifying legacy terms or missing data. Once the crosswalked data was cleaned up and put into place, standard values for all records were added (see table 3).

Table 2. Metadata crosswalk from MARC to Qualified Dublin Core

Metadata field	MARC field	Qualified Dublin Core element
OCLC number	001, 035 $a	dc.identifier.other
Call number	099	[N/A, Admin use only]
Author name	100 $a	dc.creator
Title	245 $a $b	dc.title
Thesis year	264 $c	dc.date.issued
Degree information	500, 502 $a	thesis.degree.name
Subject	6xx fields	dc.subject
Department	710 $b	thesis.degree.department

During the ongoing processing of digitized materials and as part of the quality control, each volume's metadata is evaluated against its corresponding metadata record and edited when necessary. In an effort to enrich the metadata available to users and increase visibility of the volumes, information not typically provided in the MARC records, such as thesis committee chairs, other committee members, and abstracts, are added to the records using Dublin Core contributor (dc.contributor.committeeMember) and abstract (dc.description.abstract).

Table 3. Standard values added to all records

Qualified Dublin Core element	Value
dc.format.mimetype	application/pdf
dc.type.genre	"Thesis" or "Dissertation," as applicable
thesis.degree.grantor	University of Houston
dc.type.dcmi	Text
dc.format.digitalOrigin	reformatted digital

In the interest of closely observing copyright best practices, members of the TDD Task Force, including the digital projects coordinator and the director of the Digital Research Commons, created copyright review guides and applicable rights statements.

Under these guidelines, theses and dissertations are considered under copyright if a copyright notice appears on volumes created in 1977 and earlier, if the item was created between 1978 and February 28, 1989, and if it was registered with the US Copyright Office within five years of its creation, or if it was created on March 1, 1989 or later. Inserts and other research material provided in the volumes are similarly considered for copyright evaluation during the copyright review process.

Once a volume has been evaluated for copyright status, an out-of-copyright or in-copyright statement is assigned. In alignment with the UH Libraries' mission to provide valuable research and educational materials, digitized volumes and metadata records are then ingested into the institutional repository.[23] In this stage of the process, out-of-copyright volumes are made available as open access materials. Due to inherent limitations, in-copyright volumes are access restricted and available solely to the University community.

When content is ready for ingest, volumes are moved to the ingest folder and placed in staging directories based on rights status: open access or in copyright. The ingest process is the same for both types of content, but in-copyright content requires additional post-ingest processing, so ingest batch folders are labeled according to rights status for clarity. The TDD Workflow Utility's Prepare Ingest Package function is used to create ingest packages in an input format expected by the SAF Creator, a utility for preparing DSpace batch imports in the Simple Archive Format .[24] PDF files are copied and renamed in the format LastName_Year_OCLCNumber.pdf, a CSV file is created with descriptive metadata for the batch, and the original files and metadata are moved to an archive directory. The SAF Creator is then used to create an SAF ingest package that is imported into DSpace.

Limiting access to copyrighted content was a necessary component of the project that took some time to solve. The team investigated creating a separate collection for the in-copyright content with access limited to users logged in with UH credentials. The downside to this approach was that the content within the restricted collection was not discoverable to users who were not

logged into the IR. In the end, the TDD Task Force worked with the Texas Digital Library, a consortial nonprofit organization that hosts UH Libraries' DSpace repository, to enable restricted access using bitstream authentication with Shibboleth. This allows the task force to ingest all TDD project content into a single collection and apply UH authentication to copyrighted PDF documents. In this manner, descriptive metadata for all documents is discoverable, but access to the document content is only available to members of the UH community.

Applying authentication to bitstreams in the DSpace administrative interface is a tedious process involving numerous clicks and dropdown menu selections. Selenium IDE, a browser plug-in designed for automated web development testing, is used to automate that process in the Firefox web browser. After an in-copyright batch has been ingested, the TDD Workflow Utility's Prepare Selenium Script function is used to create an automation script for Selenium. When loaded in the Firefox Selenium add-on, the script automatically applies the bitstream authentication steps in the browser for each volume in the batch.

The TDD workflow comprises detailed tasks carried out at different units in the library in a sequential routine as an assembly line. TDD activities flow from pulling volumes from shelves to disbinding, scanning, image quality control and OCR, metadata creation and copyright evaluation, and digitized files ingestion into the DSpace system. As the TDD Task Force worked collaboratively to develop and confirm workflows for this complicated process, they documented each section of the workflow in the one-stop TDD workflow Google document for easy access and transparency of the overall process.[25] The TDD working group members notify each other at completion of tasks at each section.

To better track each thesis and dissertation as it moved through the digitization, metadata, and copyright verification workflows, the task force developed an Excel spreadsheet tracking system.[26] This tracking system lists UH Libraries' theses and dissertation titles, their OCLC numbers, dates, and call numbers. It records the TDD volumes pulled from shelves, digitization completed, digitization batch, borrower notes, metadata completed, and other notes. This tracking system provides a channel for the team members to inform each other of completed tasks at each unit and to communicate issues in the working process (see fig. 3).

					1600		265	791
OCLC #	Date	CALL #	TITLE	Location 2	Pulled Complete	Digi Batch	Borrower Notice	Digi Complete
11827785	1949	Thesis 150 1949.P37	An experimental study of the effect of hy	anths	x	20191009	x	x
13800917	1949	Thesis 320 1949.A53	An inquiry into low-cost housing for low	anths	x	20191009	x	x
13748263	1942	Thesis 370 1942.G37	A literature study of St. Matthew's gosp	anths	x	20191009		x
13925309	1942	Thesis 370 1942.P53	Pirates of Hawaii.	anths	x	20191009		x
13645811	1942	Thesis 370 1942.T56	Some concepts found in Putzke and Wy	anths	x	20191009		x
13645737	1945	Thesis 370 1945.I53	A study of the club program in Houston	anths	x	20191009		x
13645730	1945	Thesis 370 1945.I534	A review of literature in the field of health	anths	x	20191009		x
13748509	1947	Thesis 370 1947.G67	A study of the home environment, backg	anths	x	20191009	x	x
13645642	1947	Thesis 370 1947.M35	An investigation of engineering school a	anths	x	20191009	x	x
13651543	1947	Thesis 370 1947.S63	The trend toward the development of sta	anths	x	20191009		x
11855306	1947	Thesis 370 1947.W47	A survey of prose literature for the first g	anths	x	20191009	x	x
13678632	1948	Thesis 370 1948.B87	An analysis of the equity of the census	anths	x	20191009	x	x
13645741	1948	Thesis 370 1948.J63	A comparison of the physical education	anths	x	20191009		x
11845032	1948	Thesis 370 1948.P47	A study of leisure time and recreational	anths	x	20191009		x
13645898	1948	Thesis 370 1948.S32	The light of liberty,: a guide to better hun	anths	x	20191009	x	x
13645626	1949	Thesis 370 1949.B72	A study of current public school building	anths	x	20191009	x	x
13645658	1949	Thesis 370 1949.B87	A quantitative study between the compo	anths	x	20191009	x	x
17370612	1949	Thesis 370 1949.E44	A sociometric study of group relations in	anths	x	20191009		x
13645831	1949	Thesis 370 1949.G47	A listing of science experiences for fourt	anths	x	20191009		x
13645703	1949	Thesis 370 1949.L36	A superintendent's handbook for school	anths	x	20191009		x
11845053	1949	Thesis 370 1949.V47	Psychometric patterns as a diagnostic a	anths	x	20191009		x
13645504	1949	Thesis 370 1949.W37	A commercial occupational survey of on	anths	x	20191009		x
13650119	1947	Thesis 380 1947.M57	A proposed plan of instruction in busine	anths	x	20191009	x	x
13650129	1948	Thesis 380 1948.O93	A proposed course of study for legal ste	anths	x	20191009		x

Figure 3. A screenshot of a portion of the TDD tracking system.

Phase Three: Promotion, Communication, and Next Steps

It is important to have strategies for TDD promotion and communication to raise awareness of the online availability of the University's legacy theses and dissertations. The TDD Task Force brainstormed elements such as audience, channels, and timeline for TDD communication. Theses and dissertation authors and campus users are the two main groups the task force plans to target in its promotion and communication plan. To attract audience attention, the TDD Task Force will design an online flyer/postcard for dissemination. They are currently collaborating with the UH Libraries director of communication, the UH Alumni Office, the UH Graduate Office, and the UH Division of Research to distribute messages to targeted audiences. The task force will communicate TDD digitization progress as they reach important milestones, including the completion of pre-1978 volumes, then at the increments of 10,000 and 15,000 volumes, and once all volumes have been digitized and deposited to the repository.

With the disbanding, digitization, and metadata workflows firmly in place, the TDD Task Force commenced the process of generating digitized versions of UH's theses and dissertations in 2020. While this process will continue over the next several years, the Task Force will also focus on refining policies and workflows around its copyright and digital preservation activities.

The TDD Task Force has developed a draft copyright policy development document, which outlines copyright determination decisions and access controls for content deemed in copyright. The task force is currently consulting with UH general counsel to ensure its recommended copyright approaches are in concert with University best practices.

At the same time, the task force is developing digital preservation procedures to ensure the long-term access, storage, and preservation to digitized theses and dissertations. The group has made some foundational decisions to date. Since one physical copy of each title will be retained, allowing for future higher-resolution rescanning if needed, the task force determined that the preservation master file for each digitized thesis or dissertation will be one PDF. This will allow the UH Libraries to greatly reduce the ongoing storage costs associated with digitally preserving the TDD collection. Throughout 2023, the task force will be exploring ways to sync TDD content to its current digital preservation workflow process, including submitting content to UH Libraries' Archivematica instance for preservation curation services such as file fixity checks and normalization, and transferring preserved TDD content to cloud storage for distributed digital preservation.

Prior to ingesting any content into the institutional repository, the team reached out to the UH's electronic and information resources accessibility (EIRA) coordinator for feedback on the accessibility of the PDF documents produced by ABBYY. The EIRA coordinator recommended encoding our PDFs as PDF/A-1a, a standard designed for preservation and accessibility, and introduced the team to the accessibility tools available in Adobe Acrobat. The Adobe Acrobat Accessibility Checker has been useful for identifying and addressing accessibility issues with the PDFs that we are producing.

UH Libraries web accessibility standards strive to comply with the World Wide Web Consortium's (W3) Web Content Accessibility Guidelines (WCAG). Combined with the feedback from the UH's EIRA coordinator, the current output was reviewed against these accessibility checklists, and areas needing improvement were identified. After several adjustments, the newest output for the project passes a majority of Adobe Acrobat's Accessibility Checker accessibility parameters, with further investigation planned to address weak points moving forward.

CONCLUSION

The TDD project at UH Libraries provides an in-depth view of the planning and workflow processes needed to launch a retrospective thesis and dissertations digitization effort in an academic library setting. Collaborating across UH Libraries departments, the TDD Task Force designed a phased approach to identify technology and resources needed to undertake the project, to develop policies, procedures, and workflows to guide the work to its completion, and to communicate about the scope, purpose, and progress of the project to internal and external stakeholders. Throughout the planning and development phases, the task force leveraged automation, bibliographic data reuse, and project management tracking to achieve workflow objectives efficiently and responsibly. With the project well underway, the task force will continue refining its processes and working across UH Libraries and campus units to ensure it complies fully with copyright and digital preservation best practices. Through these ongoing efforts, the TDD Task Force is ensuring that the original research and scholarship contained in thousands of theses and dissertations are more accessible than ever before—broadening the reach and impact of UH graduates well into the future.

FUNDING

This project was funded by the John P. McGovern Foundation.

ACKNOWLEDGMENTS

The authors dedicate this work to the memory of their colleague and TDD Task Force member Crystal Cooper.

ENDNOTES

[1] Linda Bennett and Dimity Flanagan, "Measuring the Impact of Digitized Theses: A Case Study from the London School of Economics," *Insights: The UKSG Journal* 29, no. 2 (2016): 111–19, https://doi.org/10.1629/uksg.300.

[2] Gail Clement and Melissa Levine, "Copyright and Publication Status of Pre-1978 Dissertations: A Content Analysis Approach," *portal: Libraries and the Academy* 11, no. 3 (2011): 825, https://doi.org/10.1353/pla.2011.0031.

[3] Clement and Levine, "Copyright and Publication Status," 825.

[4] Clement and Levine, "Copyright and Publication Status," 826.

[5] Xiaocan (Lucy) Wang, "Guidelines for Implementing ETD Programs—Roles and Responsibilities," in *Guidance Documents for Lifecycle Management of ETDs*, eds. Matt Schultz, Nick Krabbenhoeft, and Katherine Skinner (2014): sect.1, p. 14, https://educopia.org/guidance-documents-for-lifecycle-management-of-etds.

[6] Wang, "Guidelines," 1-17.

[7] Patricia Hswe, "Briefing on Copyright and Fair Use Issues in ETDs," in *Guidance Documents for Lifecycle Management of ETDs*, eds. Matt Schultz, Nick Krabbenhoeft, and Katherine Skinner, (2014): sect. 3, p. 12, https://educopia.org/guidance-documents-for-lifecycle-management-of-etds.

[8] Geneva Henry, "Guide to Access Levels and Embargoes of ETDs," in *Guidance Documents for Lifecycle Management of ETDs*, eds. Matt Schultz, Nick Krabbenhoeft, and Katherine Skinner, (2014): sect. 2, p. 1, https://educopia.org/guidance-documents-for-lifecycle-management-of-etds.

[9] Henry, "Guide to Access Levels," 2-1.

[10] Hswe, "Briefing on Copyright," 3-9–3-13.

[11] Cathleen L. Martyniak, "Scanning Our Scholarship: The University of Florida Retrospective Dissertation Scanning Project," *Microform and Imaging Review* 37, no. 3 (2008): 122–24, https://doi.org/10.1515/mfir.2008.013.

[12] Martyniak, "Scanning Our Scholarship," 127–29.

[13] "Retrospective Dissertation Scanning Policy," (2011), University of Florida, accessed January 1, 2022, https://ufdc.ufl.edu/AA00007596/00001.

[14] Todd Mundle, "Digital Retrospective Conversion of Theses and Dissertations: An In House Project," in *Proceedings of Eighth Symposium on Electronic Theses and Dissertation* (Sydney, Australia, 2005): 3–4.

[15] Mundle, "Digital Retrospective Conversion," 3.

[16] Mary Piorun and Lisa A. Palmer, "Digitizing Dissertations for an Institutional Repository: A Process and Cost Analysis," *Journal of the Medical Library Association: JMLA* 96, no. 3 (2008): 223–29, https://doi.org/10.3163/1536-5050.96.3.008.

[17] Piorun and Palmer, "Digitizing Dissertations," 224.

[18] Piorun and Palmer, "Digitizing Dissertations," 227.

[19] Sarah L. Shreeves and Thomas H. Teper, "Looking Backwards: Asserting Control over Historic Dissertations," *College and Research Library News* 73, no. 9 (2012): 532–33, https://doi.org/10.5860/crln.73.9.8830.

[20] Gary M. Worley, "Dissertations Unbound: A Case Study for Revitalizing Access," in *Proceedings of the 10th International Symposium on Electronic Theses and Dissertations* (Uppsala, Sweden, 2007).

[21] Worley, "Dissertations Unbound," 3–6.

[22] *Dictionary of Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations, Version 2.0*, (2015), http://hdl.handle.net/2249.1/68437.

[23] To access Cougar ROAR, see https://guides.lib.uh.edu/roar.

[24] SAF Creator is a tool developed by James Creel at Texas A&M University. For more, see https://github.com/jcreel/SAFCreator.

[25] See the TDD Google document: https://docs.google.com/document/d/18gyJQ6isn7qsuelo1Z3b7BTMXlxnchmVqp8rHqUZY8g/edit?usp=sharing.

[26] See the complete TDD tracking system: https://docs.google.com/spreadsheets/d/1TeHAgVcqw6WB3N5cDAUlbTLwZqzsTwDbLtIAPD1oAN0/edit?usp=sharing.