Hafsah Hujaleh and Jess Whyte

Narrowing the Lens: Preservation Assessment for Digital Manuscripts

As a response to Ben Goldman's 2011 call to action in RBM *regarding born-digital manuscripts, this article revisits the topic thirteen years later. Drawing inspiration from Goldman's work at the University of Wyoming, the authors attempt to narrow the preservation lens further by focusing on specific collections and considering factors like file types, risk, and resource availability. The authors suggest further humble, but practical, steps towards preserving born-digital materials, based on their experiences and while emphasizing the importance of contextual decision-making in the face of complex challenges.*

In 2011, Ben Goldman wrote in *RBM* on bridging the gap for born-digital manuscripts by taking immediate, practical steps to secure those materials. Goldman rallied preservationists to do something, anything, to begin work while we waited for that, "one, perfect, affordable, all-encompassing solution for electronic records."[1] For Goldman, narrowing the lens was a way to find a starting point when faced with a monumental and complex task. His approach included inventorying content, getting it out of media carriers, and finding somewhere to store it. But what are the next steps for modest progress?

Inspired by Goldman's work at the University of Wyoming's American Heritage Center, we, at the University of Toronto Libraries, narrowed the lens further by focusing on specific collections, sometimes down to the file-by-file level. We completed the inventories, retrieved the files from at-risk media, and we needed to decide how to conserve them. When looking at the whole picture—disparate file types, massive aggregate sizes, sensitive personal information, competing priorities, and limited staff time—the task appeared overwhelming. However, by tackling it one collection at a time, the work grew manageable, decisions became contextual, and care became more focused. Born-digital collections aren't monolithic and shouldn't be treated as such.

1. Ben Goldman, "Bridging the Gap: Taking Practical Steps Toward Managing Born-Digital Collections in Manuscript Repositories," *RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage* 12, no. 1 (Mar. 2011): 11, https://doi.org/10.5860/rbm.12.1.343.

Conservation strategies might vary based on factors such as age, donor, formats, risk, and resource availability, but the crucial thing is to start. This article, echoing Goldman's sentiment, aims to offer practical recommendations using the University of Toronto Libraries as a case study.

Past The Fence: Navigating Greener Terrain

Goldman commented on the state of the field in 2011, noting that, despite many institutions collecting digital manuscripts, most could not estimate their size or lacked corresponding policies for working with digital materials.[2] Probable reasons included staffing shortages, lack of expertise, funding constraints, or a reluctance within the profession to, "consider any form of digital preservation as true preservation."[3]

The subsequent decade saw a surge of collection-holders beginning this work, while also documenting their workflows and processes.[4] In 2014, version 1.0 of the BitCurator environment was released. BitCurator provided a set of open-source digital forensics and data analysis tools and aimed to support the work of transferring digital content from media and documenting that process by providing functionalities for disk imaging, metadata extraction, and analysis and focusing specifically on serving libraries, archives, and museums.[5] While not the only new tool in town, BitCurator's launch and educational efforts triggered a wave of collaboration, as institutions began inventorying their digital media backlogs, managing new acquisitions, and sharing their techniques. In addition to BitCurator, other important developments

2. Goldman, "Bridging the Gap," 12–13.
3. Goldman, "Bridging the Gap," 14.
4. John Durno, "Digital archaeology and/or forensics: Working with floppy disks from the 1980s," *Code4lib Journal*, no. 34 (2016): 1, https://journal.code4lib.org/articles/11986.; Leigh Anne Gialanella, "Disk imaging for preservation: Part 1," *Bits & Pieces*, University of Michigan Library (January, 2018), https://www.lib.umich.edu/blogs/bits-and-pieces/disk-imaging-preservation-part-1.; Caralie Heinrichs and Emilie Vandal, "Digital Preservation Guide: 3.5-Inch Floppy Disks," ISI 6354, University of Ottawa Libraries documentation (Dec 2019): 1–36, https://biblio.uottawa.ca/omeka2/linking-cultures/files/original/3bed48a813ef348a3873c0c50cb066ec.pdf.; Charles Levi, "Five hundred 5.25-inch discs and one (finicky) machine: A report on a legacy e-records pilot project at the Archives of Ontario," *Archivaria*, no. 72 (2011): 239–246, https://archivaria.ca/index.php/archivaria/article/view/13365/14674.; Matthew McKinley, "Imaging digital media for preservation with LAMMP," *The Electronic Media Review* 3 (2015): 89–96, https://resources.culturalheritage.org/emg-review/volume-three-2013-2014/mckinley/.; Sam Meister and Alexandra Chassanoff, "Integrating digital forensics techniques into curatorial tasks: A case study," *International Journal of Digital Curation* 9 (2014): 6–16, https://doi.org/10.2218/ijdc.v9i2.325.; Alice Prael, "Centralized Accessioning Support for Born Digital Archives," *Code4Lib Journal*, no. 40 (2018), https://journal.code4lib.org/articles/13494.; Alice Prael and Amy Wickner, "Getting to know FRED: Introducing workflows for born-digital content," *Practical Technology for Archives*, no. 4 (May 2015), https://hdl.handle.net/1813/76848.; Walker Sampson, "Guest post: Walker Sampson on disk imaging workflow," BitCurator Consortium (Nov 05 2015), https://bitcurator.net/category/use-cases/.; Jess Whyte, "Preservation Planning and Workflows for Digital Holdings at the Thomas Fisher Rare Book Library," in *Proceedings of the 2017 ACM/IEEE Joint Conference on Digital Libraries* (JCDL) (Toronto, ON, Canada: 2017), 1–10, https://doi.org/10.1109/JCDL.2017.7991611.
5. Christopher A. Lee, Matthew Kirschenbaum, Alexandra Chassanoff, Porter Olsen, and Kam Woods, "Bitcurator: tools and techniques for digital forensics in collecting institutions," *D-Lib Magazine* 18, no. 5/6 (2012): 14–21.

in the field include was the establishment of Library of Congress' National Digital Stewardship Alliance (NDSA) in 2010 through an initiative of the no longer active National Digital Information Infrastructure and Preservation Program (NDIIPP). Today, NDSA works with a range of organizations to make contributions to the field, including workflow development, peer support, and guidance materials.[6]

New ways of working brought new challenges. A 2020 post detailing the OCLC's efforts to gauge time requirements for processing born-digital collections revealed that, while many institutions now had dedicated staff for these collections, challenges persisted. These challenges included revising workflows, and figuring out how to navigate "the cognitive shift required to move back-and-forth between the granular file-level analysis . . . and the aggregate-level thinking required for the archival sense-making work."[7] Non-homogeneous collections, especially those from private donors, often proved more complex, requiring a deeper dive due to their diverse file formats and content types. The appraisal of born-digital records also necessitated a different mindset about privacy. While digital formats could offer broader access and analysis capabilities, they also posed challenges for maintaining the sort of "privacy through obscurity" a traditional reading room might ensure.[8]

Dealing with private and sensitive data in digital collections isn't easy. Many of these challenges come from the size and nature of digital collections themselves. In 2013, Goldman and Pyatt highlighted how archivists grappled with these issues in their article, "Security Without Obscurity: Managing Personally Identifiable Information in Born-Digital Archives."[9] In one example, staff with the STOP AIDS Project records team at Stanford University used digital forensic tools and keyword searches to pinpoint private and sensitive details, but remained uncertain about their findings, inundated with false positives, and lacking confidence in the adequacy of their efforts.[10] Digital collections are also typically larger than paper ones, making sensitive information harder to spot.[11] Sensitive information can pop up in unexpected areas of digital storage, like log files, system files, or even in supposedly deleted files still retrievable through forensic means. The people who donate or create these digital collections may not be aware of these hidden traces.

6. NDSA, "About the NDSA," accessed February 29, 2024, https://ndsa.org/about/
7. Chela Scott Weber, "Time Estimation for Processing Born-Digital Collections," *Hanging Together: The OCLC Research Blog*, 28 April 2020, https://hangingtogether.org/?p=7911.
8. Scott Weber, "Time Estimation for Processing Born-Digital Collections."
9. Ben Goldman and Timothy D. Pyatt, "Security Without Obscurity: Managing Personally Identifiable Information in Born-Digital Archives," *Library and Archival Security* 26, no. 1–2 (2013): 37–55.
10. Goldman and Pyatt, "Security Without Obscurity," 45.
11. Goldman and Pyatt, 41; Tim Hutchinson, "Protecting Privacy in the Archives: Preliminary Explorations of Topic Modeling for Born-Digital Collections," in *Proceedings of the 2017 IEEE International Conference on Big Data (Big Data)* (Boston, MA, December 2017), 2,251–2,255, p. 1.

Machine-learning tools for identifying sensitive information are gaining traction in archives. Tim Hutchinson, at the University of Saskatchewan, has written on their use since 2017.[12] This work is promising, but just as Goldman warned in 2011 in "Bridging the Gap," we should perhaps not wait for, "one perfect, affordable, all-encompassing solution,"[13] as there is "merit to the notion that technology alone cannot solve problems associated with digital preservation," since "waiting for high-end solutions has only led to neglect."[14]

The Narrow Path: Delving into the Minute

Goldman's steps to begin engaging with born-digital manuscripts were to separate out the digital materials, inventory them, find somewhere to store them, transfer digital content to storage, document your work, and then formulate policies and methods for future acquisitions.[15] Over the last eight years, the University of Toronto Libraries has been acting on Goldman's advice to get the bits off the failing media. Disks in manuscript collections were inventoried and imaged, and this work provided a baseline of bit-level preservation for content, by protecting the information encoded from the risk of single-copy failure.[16] However, more needed to be done to assess the content and its renderability. We knew what we had, but we didn't really know its condition, or how to approach assessing it.

This next way forward confirms we can read the bits in a meaningful way and identify any potential risks. The steps proposed here can be pursued to varying degrees by most organizations, including those with limited resources:

- Retrieve a working copy of the collection from preservation storage.
- Identify duplicate files.
- Identify empty files and system files.
- Identify files that require intervention to access and recommend an appropriate preservation action (e.g., creating a migrated copy in a more stable file format).
- Identify files that are not "of interest."
- Share these recommendations and discuss them with peers to determine the best approach.
- Document the decisions and work.
- Act on the chosen recommendations and rewrite the collection to preservation storage. (figure 1).

12. Hutchinson, "Protecting Privacy in the Archives."
13. Goldman, "Bridging the Gap," 11.
14. Goldman, "Bridging the Gap," 15.
15. Goldman, "Bridging the Gap," 16–20.
16. Whyte, "Preservation Planning and Workflows," 2017.

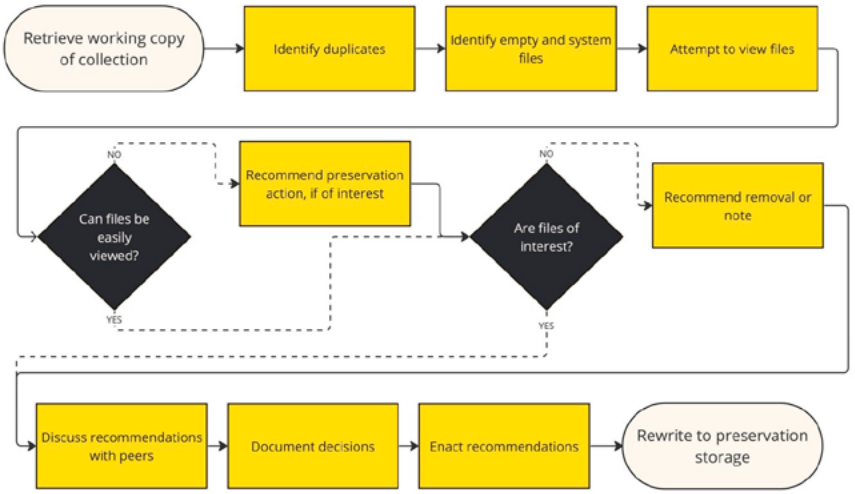

Figure 1. Example preservation assessment workflow

Retrieve a working copy of the collection from preservation storage

Assuming content has already transferred from the donor and/or off the original media to your institution's storage, pull back a working copy of those files, a "use" copy. This copy should be distinct from your primary preservation copy. You can always revert to your preservation copy in case an inadvertent alteration occurs while you are working.

Identify duplicate files

Don't do your work twice. Begin by identifying duplicate files. We use a tool called brünnhilde by Tessa Walsh[17] to compare file checksums and list duplicate files, but other tools—like Forensic Toolkit[18] or FSlint/Czkawka[19]—can also do this work. Then, we consider moving subsequent instances of duplicate files to a Duplicates directory, or deleting them, while logging that work. This reduces workload, mental clutter, and storage use. Deletion or moving may be contentious for those who wish to retain originals as-is. For those who wish to retain their duplicates, consider creating a list of duplicate files to avoid duplicating your preservation assessment work.

A challenge here are close duplicates—files that are essentially duplicates, but that would not be identified as such by a checksum comparison. Examples of this might

17. Tessa Walsh, Brunnhilde 1.9.6 (Mar 31, 2023), https://github.com/tw4l/brunnhilde.
18. Exterro, Forensic Toolkit FTK 7.6 (Exterro, 2023), https://www.exterro.com/ftk-product-downloads.
19. Czkawka is an updated and expanded upon version of FSlint. The two tools have different creators, but both make a point of referencing and acknowledging the other. Rafal Mikrut, Czkawka 7.0 (2020–), https://qarmin.github.io/czkawka; Padraig Brady, FSLint (2004–), https://github.com/pixelb/fslint.

be two copies of the same manuscript in two different file formats, or two copies, but one version includes a cover page. A similar challenge is digitization duplicates. Identifying these and deciding on an appropriate action requires familiarity with the collection, access to both the analog and digital materials, time, and ideally, internal guidelines on how to proceed.

Identify empty files and system files

Empty files are files that are zero bytes. They contain nothing. We use the command-line tool 'find' to locate empty files or directories. For example, the command 'find . -empty' will list all empty directories or files. To limit this search to just files, add '-type f.' You may have a reason for keeping empty files and directories. We typically do not, and find they create visual clutter for the preservationist. Moving them to an 'Excluded' directory or deleting them reduces that clutter.

While handling born-digital manuscripts, it is common to come across numerous system files, which often serve vital functions for software and operating systems, but may not be significant to the collection. Our typical approach relies on our intuition about what constitutes a system or software file, based on file types and locations, and we generally exclude these from our attention. This intuitive approach has served us well, but there are more structured and systematic ways to weed such files. Dalhousie University, for example, uses specialized tools that draw from the National Software Reference Library's (NSRL) Reference Data Set.[20] This dataset allows them to compare the checksums of the files in their collections against known system files. This method offers a robust and automated way of excluding recognized system files from further scrutiny or archival action. The benefit of using the NSRL's Reference Data Set is its extensive database, ensuring that a wide variety of system files, across many versions and variations of software, are effectively identified and can be set aside.[21] It's also worth noting that certain digital forensics tools come pre-integrated with such filtering features. For example, the Forensic Toolkit (FTK) incorporates "Known File Filter" capabilities.[22] These tools can automatically identify and flag system files, simplifying the process and reducing the chance of manual oversight.

20. "National Software Reference Library," National Institute of Standards and Technology (U.S.), https://www.nist.gov/itl/ssd/software-quality-group/national-software-reference-library-nsrl; Creighton Barret, "Tools for identifying duplicate files and known software files" (presentation, BitCurator Users Forum April 2017), http://hdl.handle.net/10222/72878.

21. Creighton Barret, "Digital Forensics Tools and methodologies in archival repositories" (presentation, Faculty of Computer Science, research seminar, Dalhousie University May 16 2017), http://hdl.handle.net/10222/72923.

22. Exterro, Forensic Toolkit FTK 7.6 (Exterro, 2023), https://www.exterro.com/ftk-product-downloads.

Identify files that require intervention to access and recommend an appropriate preservation action (e.g., creating a migrated copy in a more stable file format).

After identifying our primary files of interest, our next significant steps evaluate the files' accessibility.

1. Using Tools for File Identification

Our starting point is file format identification (if this hasn't already been done). For this, we use Siegfried,[23] an instrumental tool that assists us in pinpointing file formats—whether it's a WordPerfect document, a JPEG image, or a PDF. Siegfried works by using file format signatures, a unique sequence of bytes at the beginning of a file used to identify the file's format and type and comparing these signatures to existing databases such as the National Archives UK's PRONOM technical registry.[24] Siegfried fits well into our workflows and is easily automated because it is a command-line tool. We can run it on a whole collection without clicking on each file, and the output is also machine-readable. Siegfried is also built into the aforementioned brunnhilde tool. However, not every file can be precisely identified, particularly older formats that don't have a discernible signature. Yet, in these circumstances, we employ our instincts to deduce a file's type. Insights from the file's age, its contextual environment, or even its naming convention often guide our judgment.

2. Determining File Renderability

Then we ask: can the file be opened and experienced without hurdles? If the process of accessing a file is anything but straightforward, migration typically emerges as our recommended solution. Consider older documents made in software like WordPerfect or MacWrite that involve a tedious importing phase; our preference in such instances is to produce a PDF replica of the original and to identify it as a migrated version, while ensuring the original file remains untouched.

3. Tackling Complex Textual Files

Occasionally, we encounter textual files that resist access. In these scenarios, we use LibreOffice[25] and Quick View Plus[26] for viewing a diverse range of text-based file formats. If these fail, we usually turn to using hexadecimal code-to-ASCII-text converters like the "strings" tool.[27]

23. Richard Lehane, Siegfried v. 1.11.0 (Dec 2023), https://github.com/richardlehane/siegfried.
24. The National Archives UK, PRONOM Technical Registry (2002–), http://www.nationalarchives.gov.uk/pronom.
25. The Document Foundation, LibreOffice 7.0 (Berlin, Germany: The Document Foundation, 2020), https://www.libreoffice.org/download/download/.
26. Avanstar Software, Quick View Plus 2020 (2020), https://www.avantstar.com/quick-view-plus-2020.
27. The Open Group, "Strings," in *The Single Unix Specification* (2018), https://pubs.opengroup.org/onlinepubs/9699919799/utilities/strings.html.

4. Deciding on Preservation Actions

Once the files are opened, discussions can take place with the necessary parties to decide if further preservation actions should be taken to reduce overall risk. The goal is longevity and accessibility. Example preservation actions might involve creating PDF versions of an outdated WordStar document, or making a recording of a Flash animation that is difficult to play on contemporary computers.

Identify files that are not "of interest"

As we delve into individual files, our goal also includes sifting out content that may not be of particular interest or relevance to our collection's goals. Three primary criteria shape this process: relevance to the author or collection, presence of sensitive personal information, and third-party content.

Relevance largely hinges on the file's ties to the author's work or personal life. Prime examples of high-interest materials include novel drafts, personal correspondence, introspective writings, and other creative works. While the "irrelevant" category is typically system, or empty files and superfluous duplicates, there are also ambiguous cases. Take, for instance, an innocuous recipe penned by the author. Though initially brushed off as trivial, internal deliberations deemed it noteworthy, tying the recipe back to the author's lived experiences. Relevance is subjective, often shaped by human judgment and necessitating collective discussions. These discussions occur on an ad hoc basis between the preservationist and collecting archivist and throughout the assessment process.

We also use this opportunity to look for sensitive information. When a donor bequeaths their materials, the contents might be as much a mystery to them as to us. From taxes and banking details to personal contact information, the reservoir of sensitive information can be vast. Our stance: highly confidential content, such as taxes and bank statements, warrants exclusion. Meanwhile, materials like files that include an author's personal contact details are merely highlighted, serving as red flags for caution when providing external access. Finally, materials sometimes contain content unrelated to the author, or not by the author. Recognizable identifiers, like the family member's name, help us identify these and exclude them.

Share these recommendations and discuss them with peers to determine the best approach

Once the assessments are complete, we conduct a collaborative review with the preservation assessor, collecting archivist, and a preservation librarian, discussing the recommendations together. Most are blanket recommendations, applying to large swathes of file types, so the conversations can go quickly. We then enact the selected recommendations, such as migrating to other formats or deletion, before re-depositing the collection

into storage with the changes and a copy of the preservation assessment. This process aims to maintain the balance of integrity, privacy, and relevance in our collections.

Document the decisions and work

Documentation ensures that the decisions made today can be understood and followed by future conservators, scholars, and the public. By recording our choices and actions, we provide a clear trail of our thought processes, which aids in maintaining transparency, consistency, and accountability in the preservation workflow. This is especially crucial in a field where standards and best practices continue to evolve, and where the context of a collection might change over time. Proper documentation serves as a guidepost for those who will manage and use these collections in the future, ensuring that our work is comprehensible and reproducible.

Drawing from this understanding of the significance of documentation, we found inspiration in the Software-Based Art Preservation project at Tate Modern. Their thorough conservation report templates demonstrate the value of detailed record keeping. Designed to holistically document software-based artwork, these reports covered every facet from production and conservation history to intricate technical specifics, software dependencies, and more. All these were framed with a focus on the artworks' significant properties, and the intended audience.[28] While these templates were crafted for artworks, we recognized the adaptability of their approach and tailored it to suit our manuscript collections.[29] Currently, our assessments are held in a shared drive and deposited into preservation storage with the collection.

Act on the chosen recommendations and re-write the collection to preservation storage

In this last step, we enact our decisions and then rewrite a new preservation copy to our preservation storage, typically replacing the original. It's important to note here that every institution or individual may approach preservation with their own unique set of principles and comfort levels. While some might be hesitant to let go of original files or disk images in all cases, we found that maintaining multiple versions of an item or collection can introduce confusion and, in some cases, increase risk. We still maintain original manifests from acquisitions that include file modification dates and checksums, and we take care not to modify files when working with them at this stage. However, we respect the varied approaches in the community. Whatever your choice, it should reflect your best judgment and the specific needs of your collection.

28. Tom Ensom, Patricia Falcão, and Chris King, "Software-Based Art Preservation – Project | Tate," 2021, accessed March 15, 2023, https://www.tate.org.uk/about-us/projects/software-based-art-preservation.

29. Hafsah Hujaleh, "Using Innovative Methods to Rethink Preservation Assessments," (presentation, BitCurator Forum, March 27–30 2023), https://bitcuratorconsortium.org/bcf23-session-2-using-innovative-methods-to-rethink-preservation-assessments.

Limitations

There are limitations to this type of work. It is labor-intensive and requires human review, not only for file format accessibility and renderability, but also for content suitability. This means that we must prioritize collections, and the work does not scale well. Additionally, it is labor-intensive for large-scale, data-dump-style donations, where an entire hard drive is donated and accessioned.

Much of the assessment relies on judgement. Therefore, errors may occur. This could be reduced by consistently seeking clarification from colleagues, but that, in turn, may increase the time it takes to complete the assessment. Another limitation related to human error is the fear of deletion. The subjective nature of the work will at times give rise to moments of hesitation and/or uncertainty when deciding whether a file should be deleted or excluded. This most commonly occurs in collections of authors that are more well known, where there is an assumption that everything they provide could be of interest. In these cases, there is a fear that holding institutions cannot delete or discard any files, solely because of who they are.

If We Do Nothing, Most of it Will Probably Be Fine

In 2011, Goldman wrote, "leaving unattended the born-digital materials already entrusted to us—leaving disks in boxes—is guaranteed failure. If we were going to ignore the material, we might as well not have acquired it,"[17] while warning that the approach he proposed would not be sufficient in the long run. We have tried here to meet Goldman's challenge, inspire another step on the road, and to achieve progress even when it is modest. Goldman got us to a point where if we do nothing, we'd probably be fine, but it's still nice to go a little further.

The journey of preservation is not a solitary one. By sharing our methodologies and insights, we hope to foster a spirit of collaboration and collective growth within the community.[30] As we emphasized, each organization and individual will have its unique approach, reflecting their context and constraints, and those approaches may change with each unique collection. What remains consistent is the passion for safeguarding our digital heritage and ensuring its accessibility for future generations. For those looking for inspiration and peers, we also recommend the Digital Preservation Coalition, whose *Handbook* includes an interactive decision tree for the selection of digital materials,[31] and the BitCurator Consortium, whose annual and accessible Forum brings together colleagues from all levels of experience. Moving forward, we're exploring the potential of e-discovery tools and AI/ML tools to bolster our work.

30. Hujaleh, "Using Innovative Methods."
31. Digital Preservation Coalition, *Digital Preservation Handbook, 2nd Edition*—Interactive Assessment: Selection of Digital Materials for Long-term Retention (2015), https://www.dpconline.org/handbook/organisational-activities/decision-tree/interactive-assessment.

We are also developing scripts to test the opening and conversion of common file formats, like detecting and automating PDF conversions with a headless LibreOffice converter. Another key recommendation we see progress on is to encourage donors to be more selective in their donations. Instead of entire hard drives, we prefer donors to curate and organize specific files, making our tasks as preservationists more manageable and focused.

The essence of our message in this article is about action and progress. By sharing our approach, we also invite others to share their techniques. Preservation is a collective effort, and by pooling our knowledge, we can navigate its challenges more effectively.