"SIUE Digitization Lab" by Marcella Lees used under CC0

BUILDING DIGITAL ARCHIVES WITH MINIMAL STAFFING:
EQUIPMENT RECOMMENDATIONS, PRICING, SCANNING, AND POST PROCESSING

PRESENTERS: KARMINE BEECROFT, ILLINOIS STATE UNIVERSITY AND ELLIE HARMAN, ILLINOIS STATE UNIVERSITY

Zoom's live transcript is available on the Show Captions menu.

BUILDING DIGITAL ARCHIVES WITH MINIMAL STAFFING

Equipment Recommendations,
Pricing, Scanning,
and Post-processing

PRESENTED BY

Karmine Beecroft
Ellie Harman

OUR AGENDA TODAY

Getting started

Equipment

Our lab

Imaging standards

Image capture

Post processing

COMMON MISCONCEPTIONS

Digitization = preservation

Digital files require a much higher level of intervention than most analog items. Despite our best efforts, not all digitized materials will survive longer than their physical counterparts due to bit rot/corruption, file type/rendering system obsolescence, disasters, and human error.

If it's important, it's already been digitized

The piecemeal digitization of GLAMR collections has made off-line cultural heritage resources even more invisible. Users have become increasingly archives-illiterate and less willing to travel to conduct research. There is an assumption that information is accessible upon demand by default and if it isn't online it either doesn't exist or is of lower research value.

Digitizing (to archival standards) is cheap and easy

Mass digitization of cultural heritage materials requires specialized and expensive equipment, file storage solutions, and delivery platforms, as well as painstaking and time-intensive labor by trained professionals.

Eventually, everything will be digitized

In addition to the labor and costs associated with digitization, we also have to take into consideration legal restrictions, such as copyright laws, and ethical concerns, like privacy, provenance, and the purpose of digitization.

"WHY CAN'T YOU JUST SCAN IT ALL AND PUT IT ONLINE?"

1. Costs

a. Equipment & software
b. Labor
   i. Image capture
   ii. Processing & file management
   iii. Metadata
   iv. Maintenance & migration
c. File storage & backup
d. Delivery platform
e. Energy usage

2. The Law

a. Copyright
b. Contracts
   i. Vendor partnerships
   ii. Donor agreements
c. Privacy
   i. HIPAA
   ii. FERPA
   iii. etc.

3. Ethics

a. Potential for harm
b. Exploitation of marginalized communities
   i. Digital colonialism
   ii. Paywalls
c. Moral rights
   i. of creators
   ii. of subjects
   iii. the right to be forgotten
d. Context of creation
   i. Format integrity
   ii. Expectation of practical obscurity

THE DIGITIZATION PROCESS

Set-up
Conversations with curators; collection assessment; schema creation

Capture/ conversion
Scanning and/or photography

Post- processing
File management; quality control; creating access derivatives

Description
Applying structured metadata to create findable, useful digital objects

Ingest
Uploading digital objects to public online database

Digital Preservation
Back-up, migration, and monitoring/fixity checks

GETTING STARTED

PLANNING QUESTIONS

- What institutional backing is available to help us create a sustainable digitization program?
    - How will it be staffed? How will this be reflected in people's job descriptions?
    - How will it be continuously funded?
- What limits are we willing/empowered to place on the program?
    - Will we accept digitization-on-demand requests?
    - What turn-around time will we communicate to stakeholders?
    - What level of pre-processing and metadata must be completed before we accept materials for digitization?
- How will we prioritize materials for digitization?
- What formats are we most interested in digitizing?
- What is more cost-effective to outsource versus digitize in-house?
- What workflows will we employ to communicate and transfer materials between stakeholders?
- How will we track our progress on projects?
- How will we evaluate the success of our digitization projects?

YOU WILL NEED

- A dedicated space
    - Able to be secured
    - No food, sealed drinks at your discretion
    - Temperatures between 64°F and 72°F (colder is better than hotter) and 30%-60% relative humidity
    - Low dust environment (no carpet, dust frequently with microfiber cloths)
- Capture equipment (scanner/camera)
- Workstation (sturdy desk, adequate surface area, computer, monitor(s), keyboard, mouse)
- Capture, processing, and file management software
- File storage and backup solutions
- Hosting platform
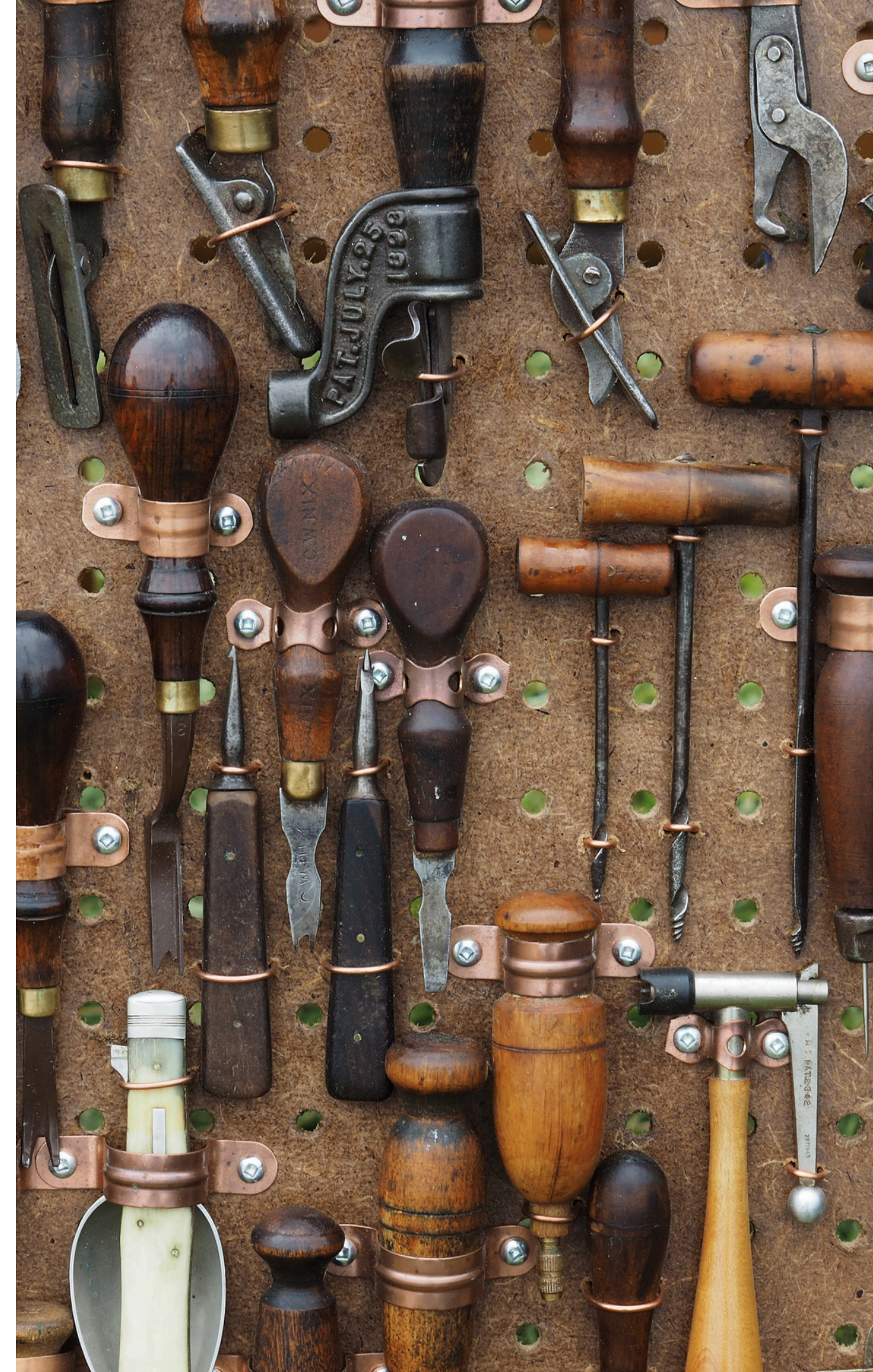
WHAT TO LOOK FOR IN A SCANNER

- Optical resolution (not interpolation) of at least 600 ppi (2400 ppi for negatives and slides)

- Ability to specify color space and file format

- Ability to turn off point-of-capture "enhancements" (smoothing, brightening, straightening, background removal, moiré correction, color "restoration," etc.)

- Can be calibrated by user

- Included or add-on transparency adaptor

- Avoid sheet-feed scanners and all-in-one devices!
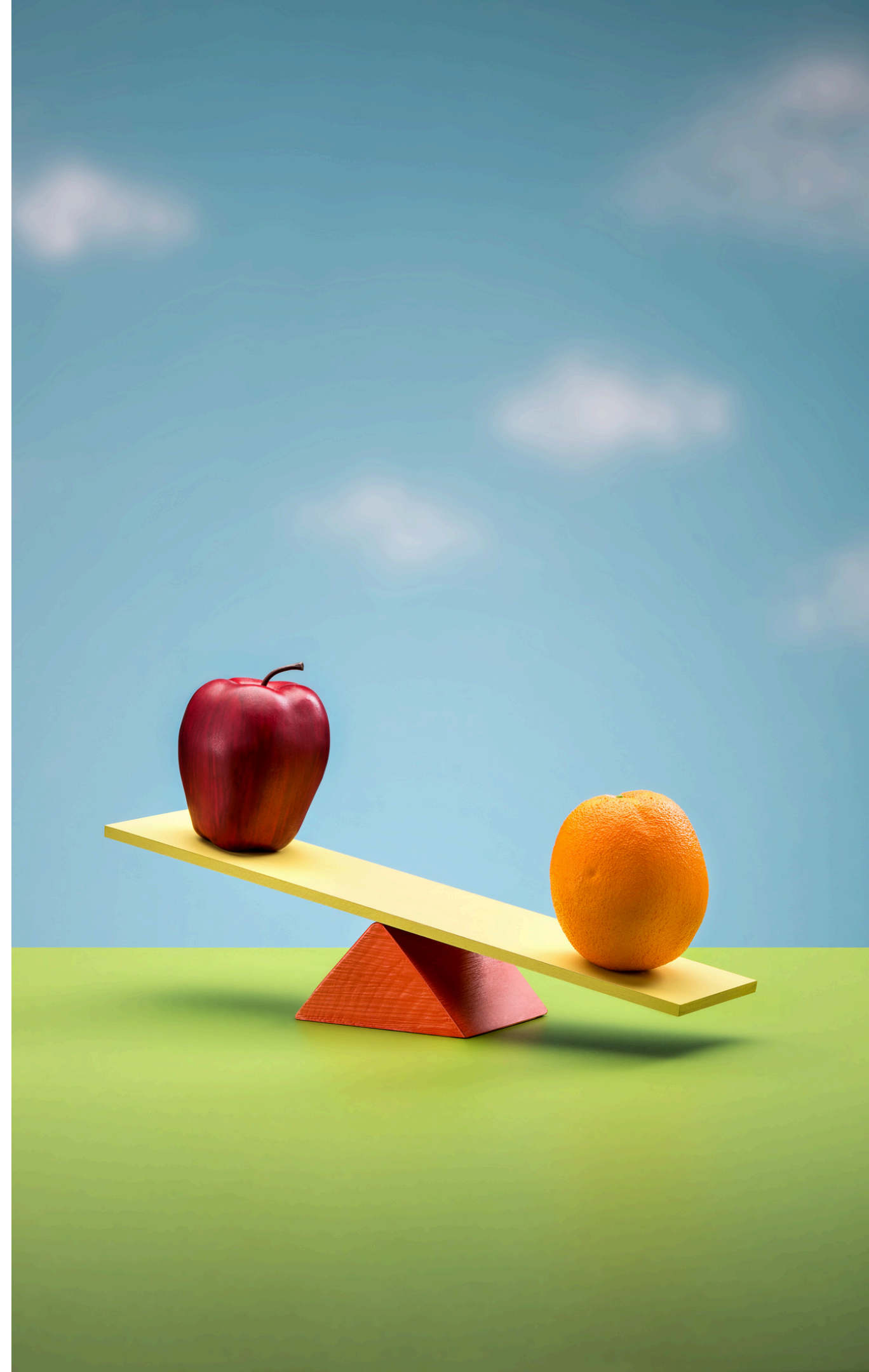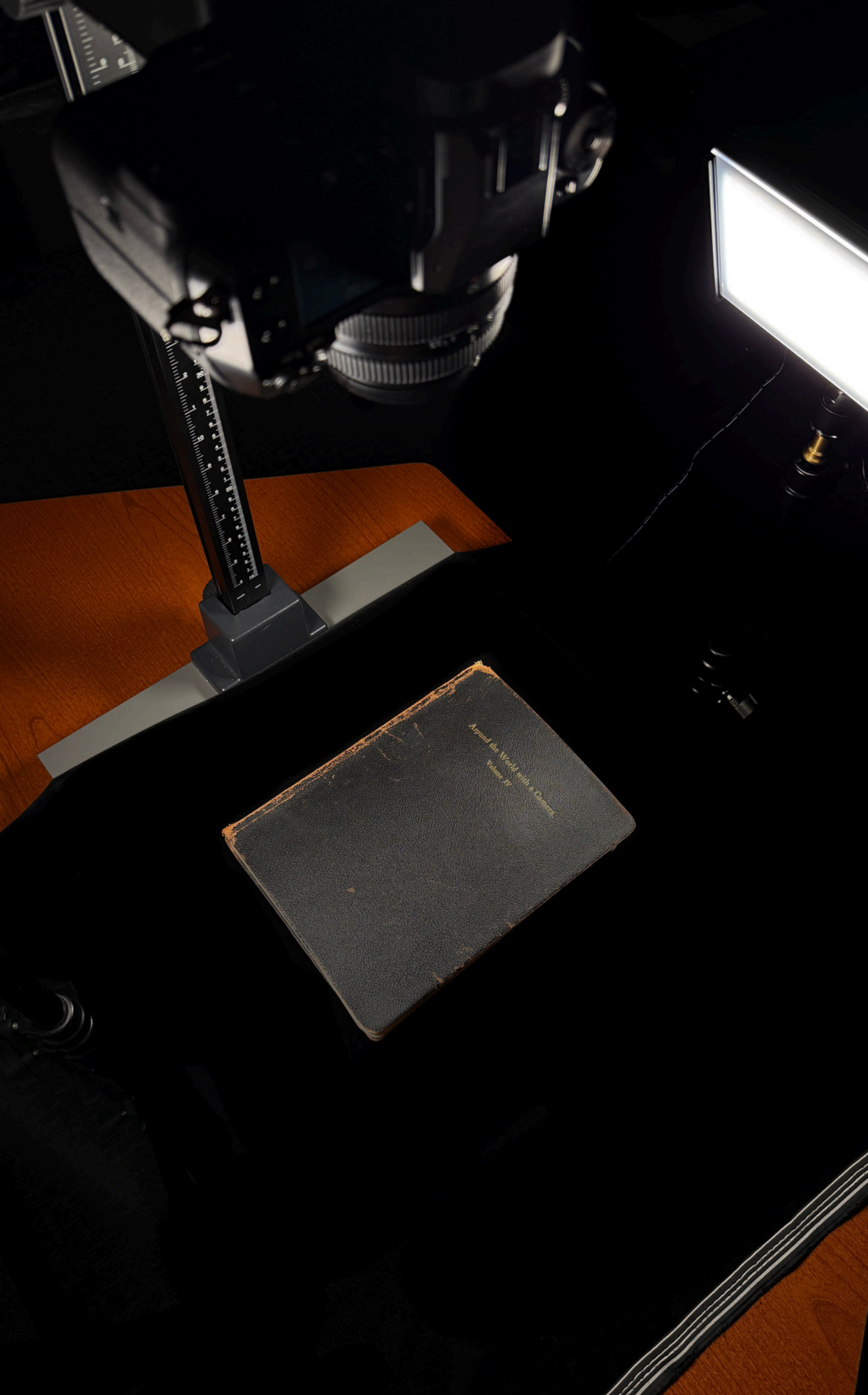
WHAT TO LOOK FOR IN A CAMERA

- Higher megapixels (MP): for greater resolution
  - Medium format is ideal, but comes at a higher price point

- Durability: built to handle a large number of exposures

- Focus and color accuracy

- Full manual control

- Variety of quality lenses

*The camera world is moving toward mirrorless... though not necessarily better or worse for digitization purposes, it is important to consider that many DSLR options are being discontinued

ECONOMIZING

- Not every project needs the highest possible imaging specifications (saves time and server space)

- Explore consortial and institutional resource-sharing

- Student employees, interns, and volunteers can provide cheap labor... in a pinch

- Prices on big ticket items/subscriptions can often be negotiated (look at cost-savings for multi-year agreements)

- Explore planning for digitization costs in donor agreements, fund drives, and endowments

- Dedup and practice responsible file management

- Buy earlier models or certified refurbished equipment (keh.com)

OUR SHOP
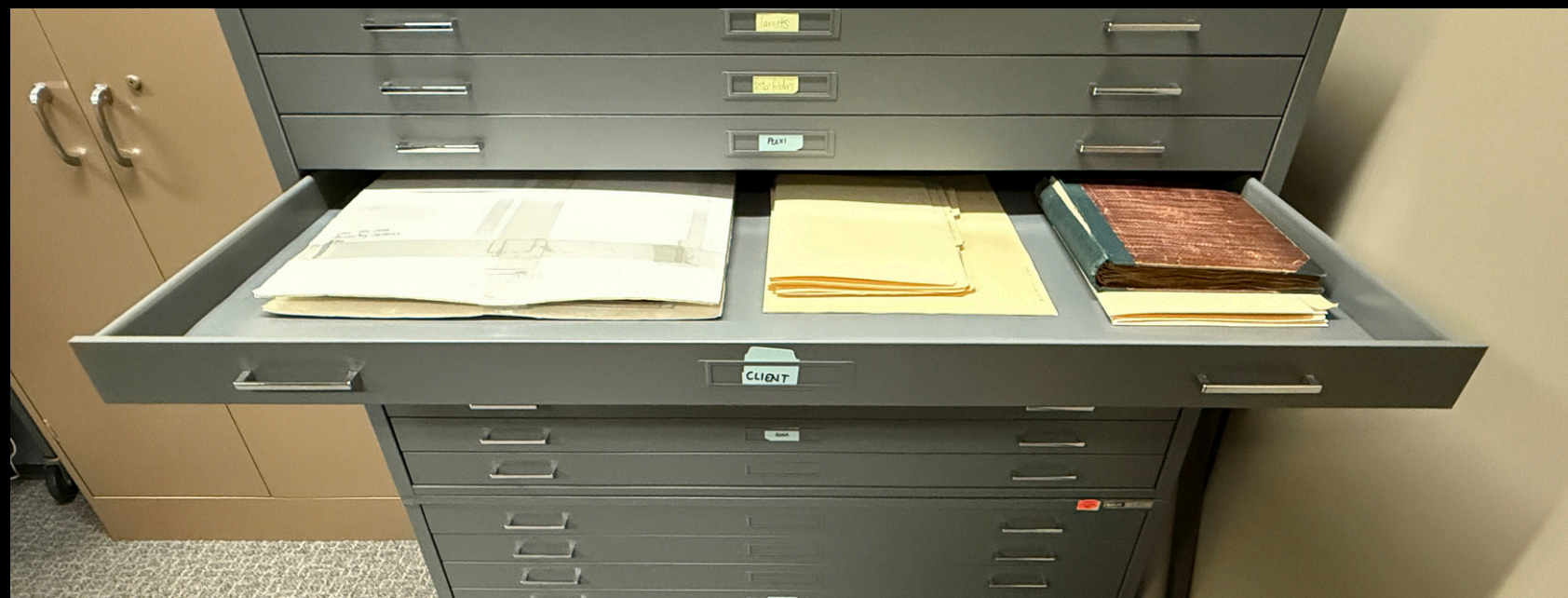
FLATBED SCANNERS

We have three Epson Expression series XL graphic arts/archival flatbed scanners - these form the backbone of our operation and are almost always in use during business hours.

Price point: $1200-$3700

PLANETARY SCANNER

Our i2S Quartz A0 SupraScan is an overhead scanner with a 51" x 36" platen. It is useful for scanning posters, artwork/prints, maps, blueprints, newspapers, and oversize books (legers, plats, etc.) that can be opened 180 degrees to lie flat.

Price point: $100K+, as well as $12K/year for service & warranty

EIZO COLOREDGE MONITORS

Professional quality monitors for image editing.
Supports 48-bit color display (1.07 billion colors) and can be exactly calibrated.
Price point: $600-$6000

DELL PRECISION SERIES TOWER WORKSTATIONS

Professional quality computers with dedicated graphics cards, large local storage capacity, and powerful multitasking abilities - essential for bulk processing large files.
Price point: $500-$5000

*If you choose to use Apple products, make sure all capture and editing software is compatible

REPROGRAPHIC PHOTOGRAPHY

The Fujifilm GFX100S medium-format camera is used for both reprographic and studio photography in our digitization center. With a copy stand and LED lights, we are able to capture a wide range of standard-sized flat papers and bound volumes.

STUDIO PHOTOGRAPHY

Our studio set-up consists of a standing backdrop, multiple flashes on tripod stands, various flash modifiers (umbrellas, diffusers, grids, etc.), and the Fujifilm GFX100S on a tripod. This studio is used for large and 3-dimensional items.

PHOTOGRAPHY PRICE POINTS

- **Fujifilm GFX100S Camera:** $3500-5000*
  - +3 lenses (24mm, 50mm,85mm) $4600+
- **Kaiser Copy Stand:** $550
- **Ikan LED Lights:** $200 each
- **Ikan LED Light Stands:** $55 each
- **Velvet (backdrop):** $60

- **Bowens Flash Heads (Gemini series):** est. $130 each
- **Impact Light Stands:** $80 each
- **Vello Flash Trigger/Receiver:** $60
  - +$40 each for additional receivers
- **Impact Umbrellas:** $15-20 each
- **Camera Tripod:** $50+
- **Impact Backdrop Support:** $200
- **Seamless Backdrop Paper:** $80 each

*Version I is discontinued but can be bought used, Version II is the higher price point

SOFTWARE & PLATFORMS

Paid products

- CONTENTdm: ~$16K/year for platform + 200GB storage

- bepress: $45.5K/year

- FromThePage: $3600/year

- Amazon S3 Glacier Deep Archive: $30/month ($0.00099 per GB)

- SilverFast
  - Software: $350
  - Calibration targets: $70-$120

- Adobe creative suite license (Photoshop, Bridge, Lightroom, Premiere Pro): $720/year

- Adobe Acrobat license: $240/year

- ABBYY FineReader PDF license: $70-165/year

Free/low cost

- Native scanner interface

- GIMP, Photopea*, Photoscape X* (Photoshop alternatives)

- Darktable, RAWTherapee (Lightroom alternatives)

- TreeSize, Karen's Directory Printer

- Teracopy

- Bulk Rename Utility

- Internet Archive
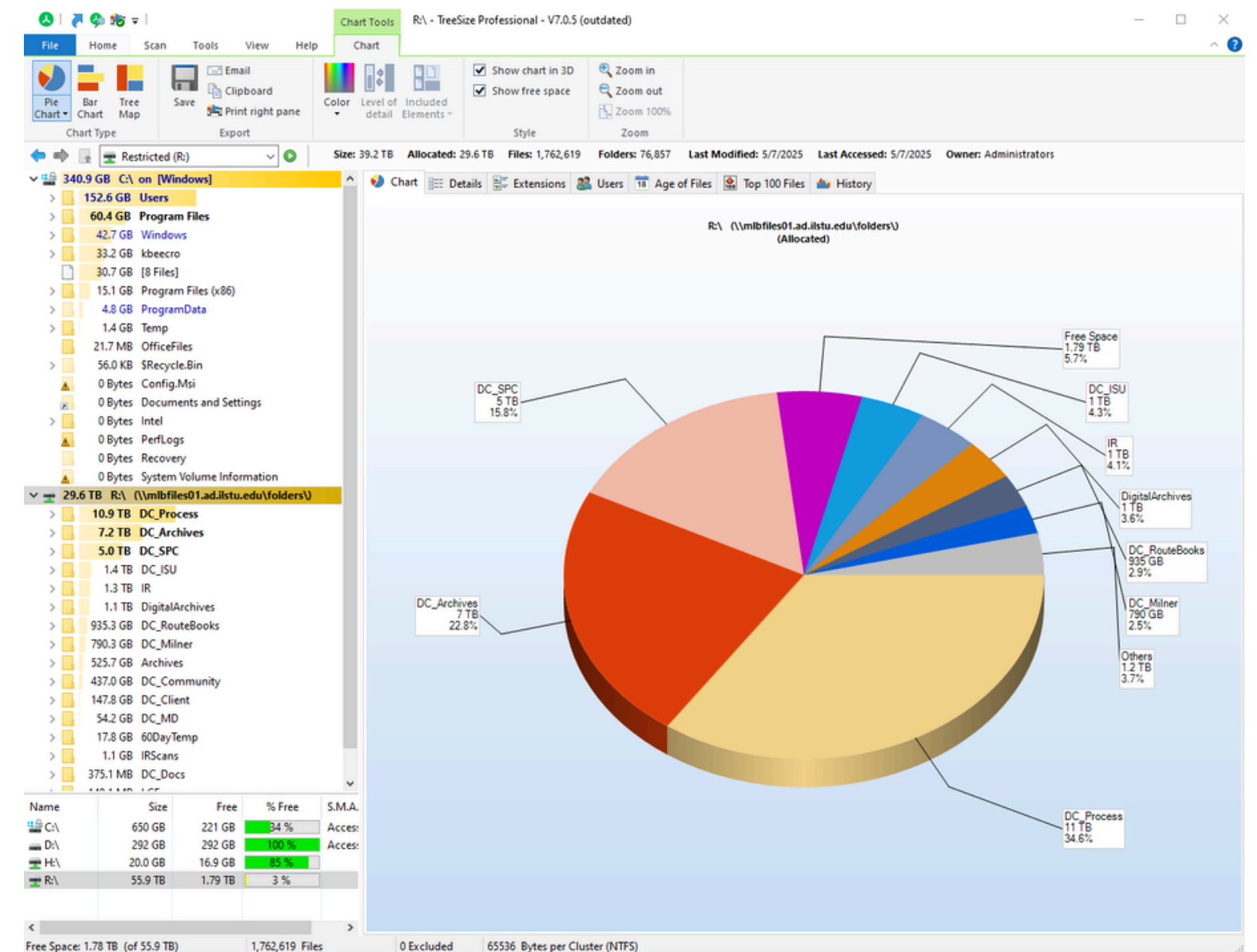
IMAGE CAPTURE

FILE STRUCTURE

share (top-level folder)

    collection (subfolder level 1)

    readme.txt (image processing specs, project info)

        deriv (subfolder level 2)

            jpg (subfolder level 3)

                box# (subfolder level 4)

                    folder# (subfolder level 5)

                        document# (subfolder level 6)

                          (files)

            tif_ed (subfolder level 3) <-- sometimes we only have jpeg derivatives

                box# (subfolder level 4)

                    folder# (subfolder level 5)

                        document# (subfolder level 6)

                        (files)

        source (subfolder level 2)

            box# (subfolder level 3)

                folder# (subfolder level 4)

                    item# (subfolder level 5)

                    (files)

FILE NAMING

- Pick whatever convention you want - just be descriptive and consistent
- Ensure that filenames are unique
- Best practice is to mirror the arrangement of the physical collection
- Include leading zeros to ensure all number sections have the same number of digits for correct sorting
- Keep filenames relatively short
- Acronyms and abbreviations are fine - as long as you can remember what they stand for
- USE ISO 8601 convention for dates (yyyy-mm-dd)
- Hint: use a/b or r/v (recto/verso) instead of f/b (front/back)

Examples:

wgss_b01_f430_d02_p24_v
smith-hutchins_hist_of_bloomington_221
colby_letter_1901-12-02_envelope
rbbb_program_1892c_06

OUR IMAGING SPECIFICATIONS

- File format: TIFF master files, JPEG derivatives

- 16-bit grayscale or 24-bit color

- 300 or 600 ppi

- Color space: Adobe RGB (1998)

FADGI Technical Guidelines

IMAGING PREPARATION

- Review physical materials
  - Identify condition issues, route for conservation treatment if needed
  - Remove fasteners (replace with Plastiklips when finished)
  - Divide materials based on necessary capture equipment

- Consult curator(s) and other stakeholders re: desires for finished product

- Create tracking spreadsheet and file directory

- Record specifications/instructions for those who will be effecting capture and save them in a shared location

HANDLING

- Avoid:
  - Cotton gloves (wear nitrile gloves with photo prints and negatives)
  - Lotions
  - Long acrylic nails
  - Nail polish
  - Jewelry
  - Trailing sleeves
  - Fabrics that shed (and lint roll that pet hair!)
- Do not open bound volumes wider than they "want" to go - flatbed scanners are often not appropriate for books
- Take care to ensure that original arrangement is maintained
- Avoid putting pressure on folds or existing tears when handling thin or brittle paper

SCANNING PROCESS

WORKFLOW

CLEAN THE PLATEN

Rub the scanner glass with a clean microfiber cloth to remove dust and smears. Do not use any cleaning products on the platen.

PLACE DOCUMENT FACE DOWN

Do not snug document into corner of platen - position it slightly away from the edges

INCLUDE TARGET(S), RULER, &/OR BACKGROUND PAPER

CLOSE LID OR PLACE WEIGHTS

Smooth out creases or folds, hold the document tight against the platen surface

PRESCAN & DRAW FRAME

No need to capture the whole platen each time

AUTOFOCUS & SCAN

A 600 ppi full-color image takes over a minute to scan - budget labor and communicate timelines accordingly

Color target
(face down)

Ruler

weights
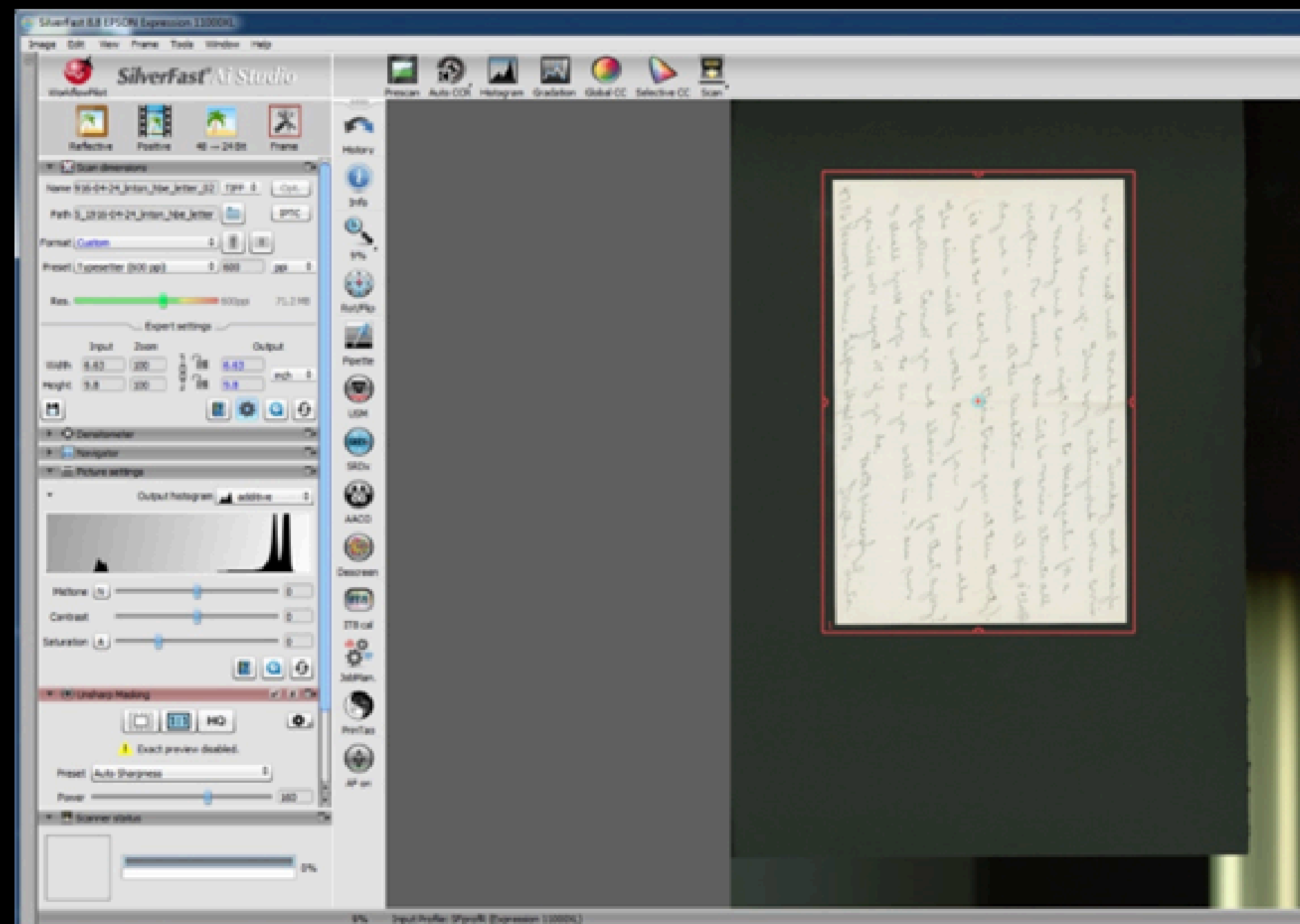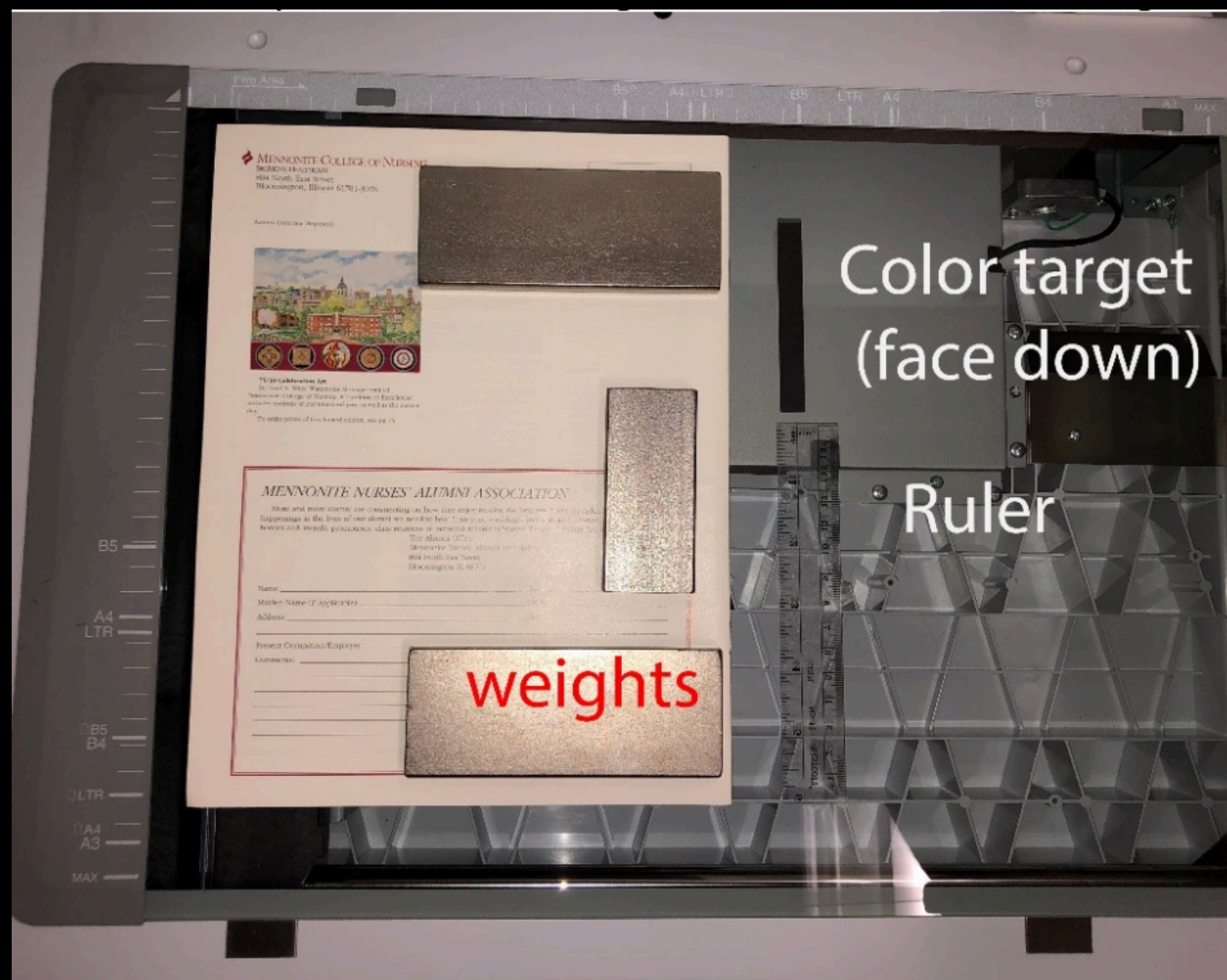
PHOTOGRAPHY PROCESS

WORKFLOW

CONSULTATION

Consult our conservator about condition, handling, and preparation of items before digitization

SET-UP

Prepare lights, camera, and any supporting materials for the object (backdrop, foam book supports, plexi, weights, mannequins, etc.).

IMAGE CAPTURE

This often involves taking multiple images of each page or angle (3D) to ensure proper focus and detail for post-processing. Lighting may also need adjusted regularly during this time.

POST-PROCESSING

Archival master files are stored and derivatives are created and edited using Adobe Photoshop.

FILE MANAGEMENT

Final derivative jpeg files are created, file-naming is completed, and the metadata team is notified of an object or collection's completion.

POST-PROCESSING

QUALITY CONTROL

- Correct specifications - PPI, color/grayscale, bit depth, file type, border/no border, background color

- Orientation

- Conforms to file naming convention

- Includes target(s) and/or ruler if specified

- Not skewed

- No folds or wrinkles in paper, page is as flat as possible

- Gutter text is visible (bound volumes)

- In focus

- No glitches, dead sensors, or distortions

Photo by Arno Senoner on Unsplash

DERIVATIVE CREATION

Basic (preserves image integrity):
- Cropping
- Straightening
- Downsampling/WxH resolution
- Changing file format

Alterations:
- Adjusting levels (black/white/midpoint)
- Adjusting curves (contrast)
- Silvering correction
- Color restoration
- Sharpening
- Shadows/highlights
- Moiré correction
- Background removal
- Photomerging

AUTOMATION

Photoshop
- Actions/batches
- Image processor module

Bridge
- Bulk IPTC metadata application

Bulk Rename Utility
- Renumber files
- Correct errors or omissions

Acrobat/ABBYY
- Batch create PDFs
- OCR

TRANSCRIPTION

Necessary for accessibility, searchability, and text analysis

OCR (Optical Character Recognition)
- Can be manually adjusted in both Acrobat and ABBYY, but ABBYY's editor module is much more powerful
- Not suitable for many historical typefaces, hit-or-miss with typewritten documents (bleed), often has trouble with columns and rows

Crowdsourced manual transcription
- Can be remarkably quick - when you capture the volunteers' attention
- Requires time and effort to maintain volunteer base

Emerging AI solutions for handwritten materials
- Often not trained/tested on historical documents
- Does best when recognizing one person's handwriting
- Getting better all the time

Fullscreen Layout ▾

Mark as blank ☐ Needs Review ☐ Preview Save
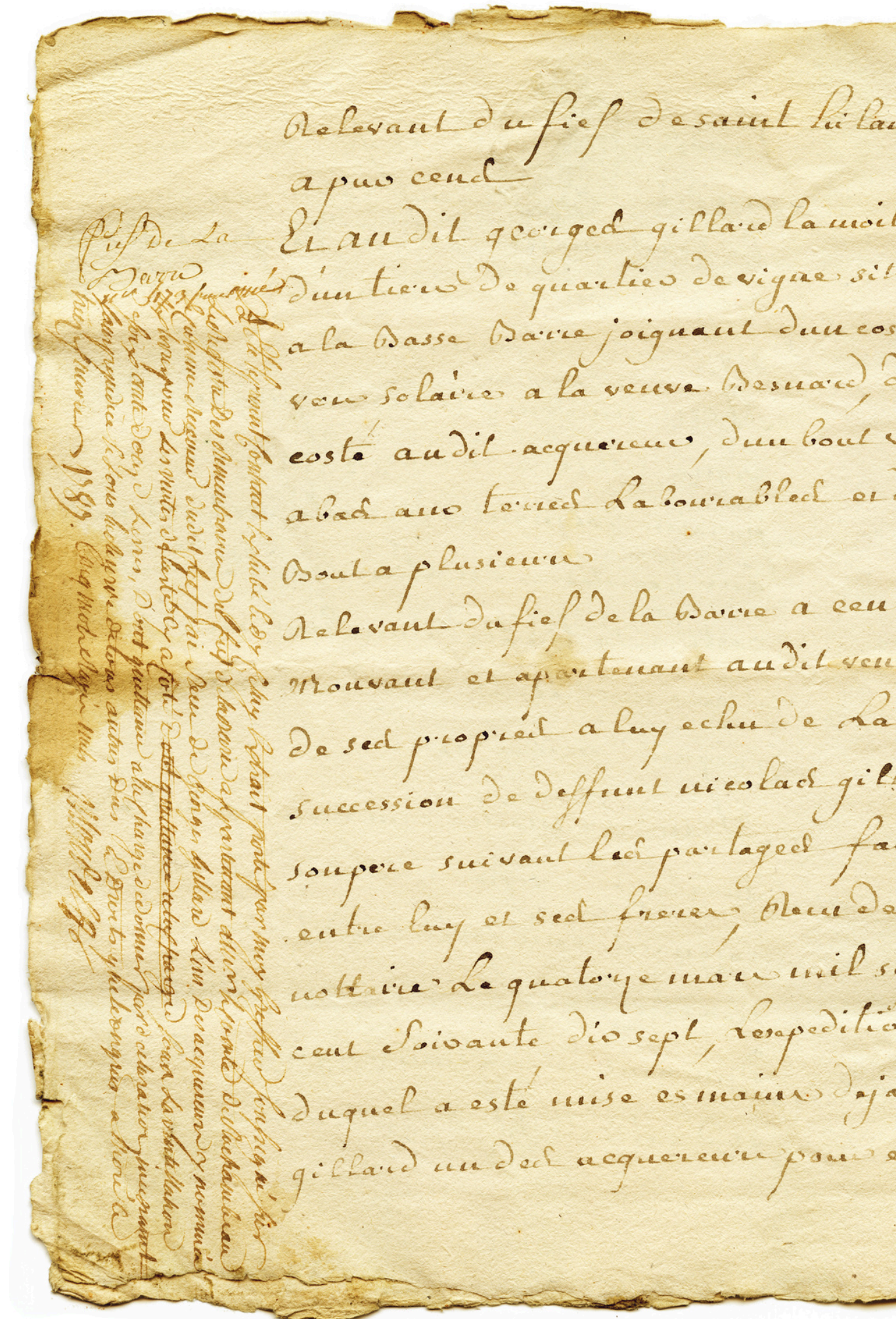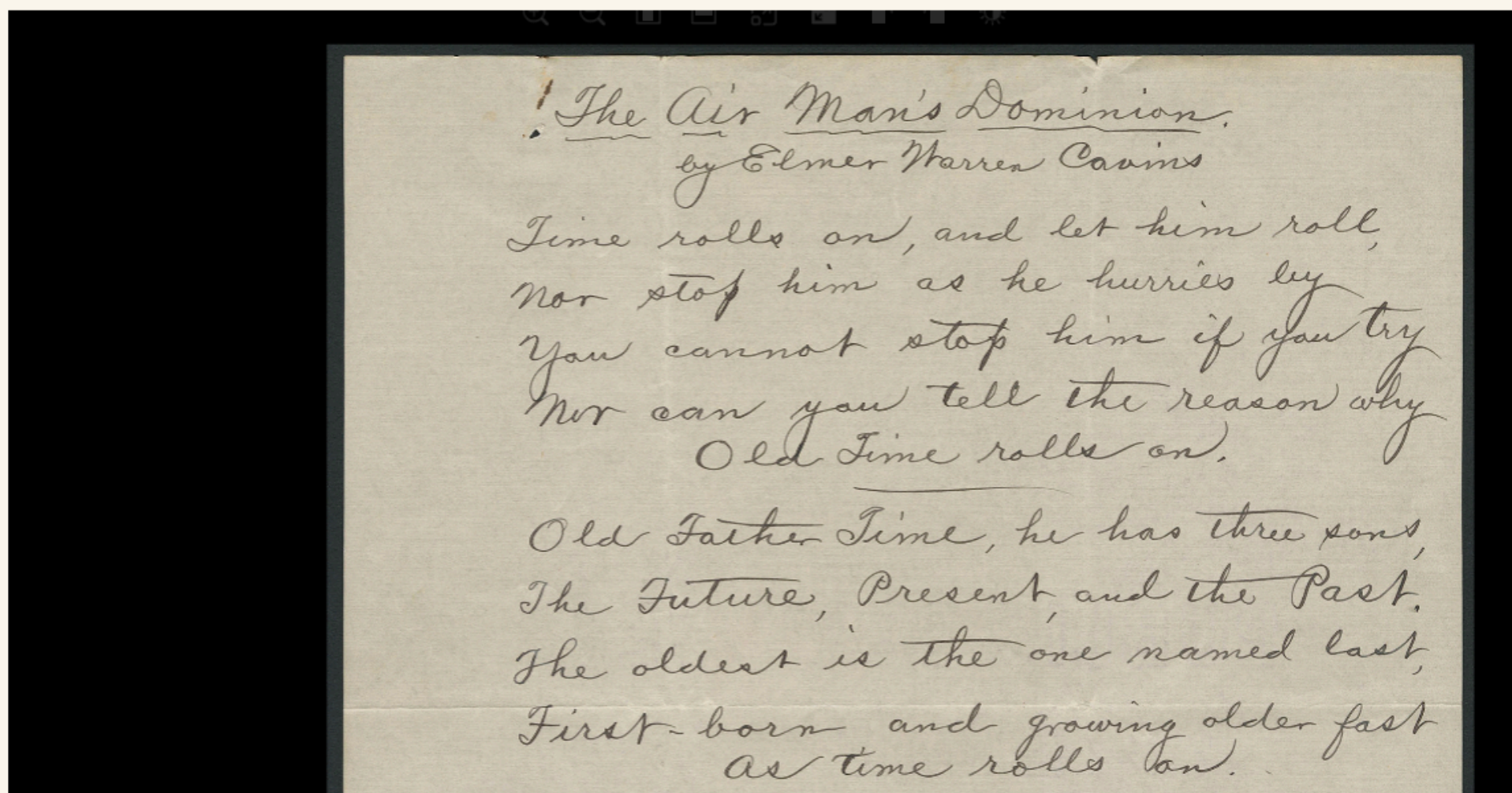
1 [Underlined: The Air Man's Dominion]
2 by Elmer Warren Cavins
3 Time rolls on, and let him roll,
4 Nor stop him as he hurries by
5 You cannot stop him if you try
6 Nor can you tell the reason why
7   Old [underlined: Time rolls] on.
8 Old Father Time, he has three sons,
9 The Future, Oresent, and the Past.
10 The oldest is the one named last,
11 First-born and growing older fast
12   As [underlined: time rolls] on.
13 Time is the father of Events,
14 Events occur in time, you see;
15 No other possibility
16 This side of vast eternity
17 [Underlined: When time shall] cease.

Transcription Conventions

- *Line Breaks:* Hit `Enter` once after each line ends. Two returns indicate a new paragraph, whether indicated by a blank line or by indentation in the original.

- *End-of-line hypenation:* Reunite both parts of the word on the top line without the hypen.

- *Spelling:* Spell words as they appear in the page image. If a word is misspelled or abbreviated you may add the correct spelling after it in [square brackets] if you wish.

- *Justification:* All text is left-justified by default. It is not possible to center or right-justify text; please do not attempt to reproduce those effects with spaces or tabs.

- *Columns:* If a page has multiple columns, transcribe the columns beneath each other moving from left to right. For example, in a two-column layout the text in the column on the right would appear below the text in the column on the left.

- *Capitalization and punctuation:* Use capitalization and punctuation as it appears in the page image.

- *Other formatting:* If words or phrases appear with formatting such as italics, underlining, or cross outs, you can indicate that inside square brackets. Ex.: [underlined: our own]. Please don't add html code to indicate formatting in the text. The html tags are not compatible with our digital collections platform and interfere with text analysis tools.

- *Possibly correct transcriptions:* If you think that you have identified the correct word but are uncertain, write it in square brackets with a question mark. Ex.: [Christianity?].

- *Illegible text:* If you cannot read a word or phrase, write [illegible] where the text appears in the document.

THANK YOU!

KARMINE BEECROFT (THEY/THEM)
DIGITIZATION CENTER COORDINATOR

kbeecro@ilstu.edu

ELLIE HARMAN (SHE/HER)
DIGITIZATION TECHNICIAN

eaharma@ilstu.edu

File naming/structure guidelines
bit.ly/file_mgmt

Safe handling guidelines

bit.ly/safe_handling

Transcription guide
bit.ly/transcription_guide