A Novel Workflow for Large Scale Thesis Digitization

TCDL 2016

A Novel Workflow for Large Scale Thesis Digitization

Texas State University
Albert B. Alkek Library

Jeremy Moore - Digital Media Specialist

Jason Long - Programmer Analyst

Todd Peters - Head, Digital & Web Services

TEXAS STATE UNIVERSITY
The rising STAR of Texas

Why Digitize?

- Faculty interest
- Space issues

Library infrastructure renovation- Archives losing ~500 linear feet

ETDs required 201 linear feet of shelf space

3 copies -preservation microfilm copy, Archives hard copy, and a hard copy in the stacks.

TEXAS STATE UNIVERSITY
The rising STAR of Texas

How Many?

- Starting with theses that have a silver halide microfilm preservation copy and second hard copy to debind
- Approximately 5,000 theses between 1930 - 2010

Since 2005, some theses have been digitized and uploaded into our local Institutional Repository.

More recently, the Graduate College is requiring that all thesis and dissertation be submitted electronically.

TEXAS STATE UNIVERSITY
The rising STAR of Texas

How to make available?

- The copyright and permissions issue is unresolved.
- No public access at this time.
- Plan for the future and create digitization workflows that will facilitate batch uploading into the Institutional Repository later.

TEXAS STATE UNIVERSITY
The rising STAR of Texas

Novel Process?

1. Prepare physical and digital space
2. Pre-Process physical items
3. Digital Capture
4. Post-Process physical and digital items

TEXAS STATE UNIVERSITY
The rising STAR of Texas

Prepare physical and digital space

TEXAS STATE UNIVERSITY
The rising STAR of Texas

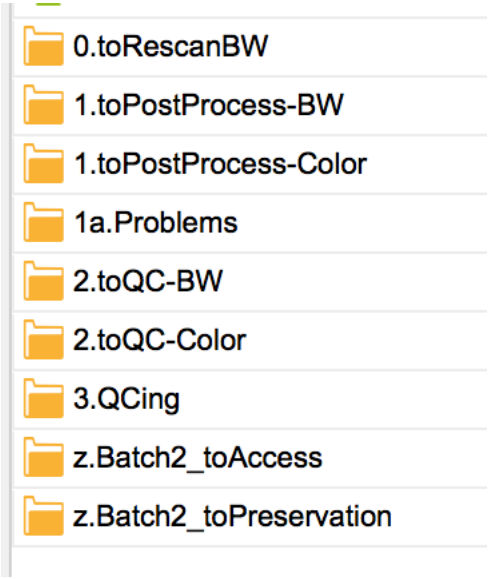
Algorithmic Project Design

- Discrete micro-processes apply to physical as well as digital
  - Check
  - Move
  - Check
  - Store
  - Check

TEXAS STATE UNIVERSITY
The rising STAR of Texas

Organizing Space

- Physical and digital space should aid, not hinder the process



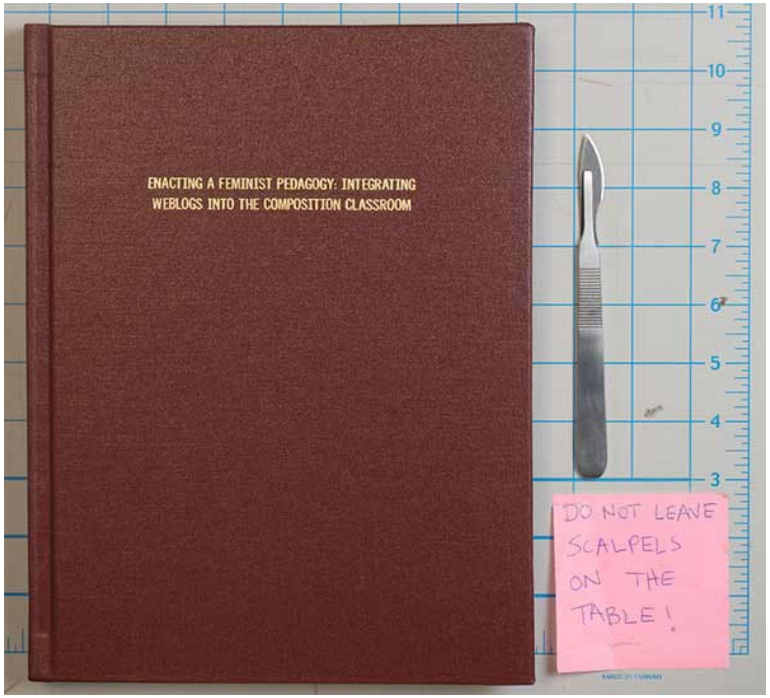
TEXAS STATE UNIVERSITY
The rising STAR of Texas

Pre-Process physical items

TEXAS STATE UNIVERSITY
The rising STAR of Texas

Sacrificing Theses

- **Binding, printing, paper quality changes with each thesis**
  - 8.5 x 11 in. in library buckram
  - Large margins
  - Printed on one side; simplex
- Scalpels > Exacto knives
  - Thank you to Jessica Phillips, Head of Preservation at UNT Libraries
- **Binding, printing, paper quality changes with each thesis**



TEXAS STATE UNIVERSITY
The rising STAR of Texas

Chop! Triumph 4315 Stack Cutter

- Inspect pages before and after chopping
  - **Binding, printing, paper quality changes with each thesis**

- Cutting theses to 8.11 in. wide because we had to pick something
  - **Binding, printing, paper quality changes with each thesis**

TEXAS STATE UNIVERSITY
The rising STAR of Texas

Digital Capture

TEXAS STATE UNIVERSITY
The rising STAR of Texas

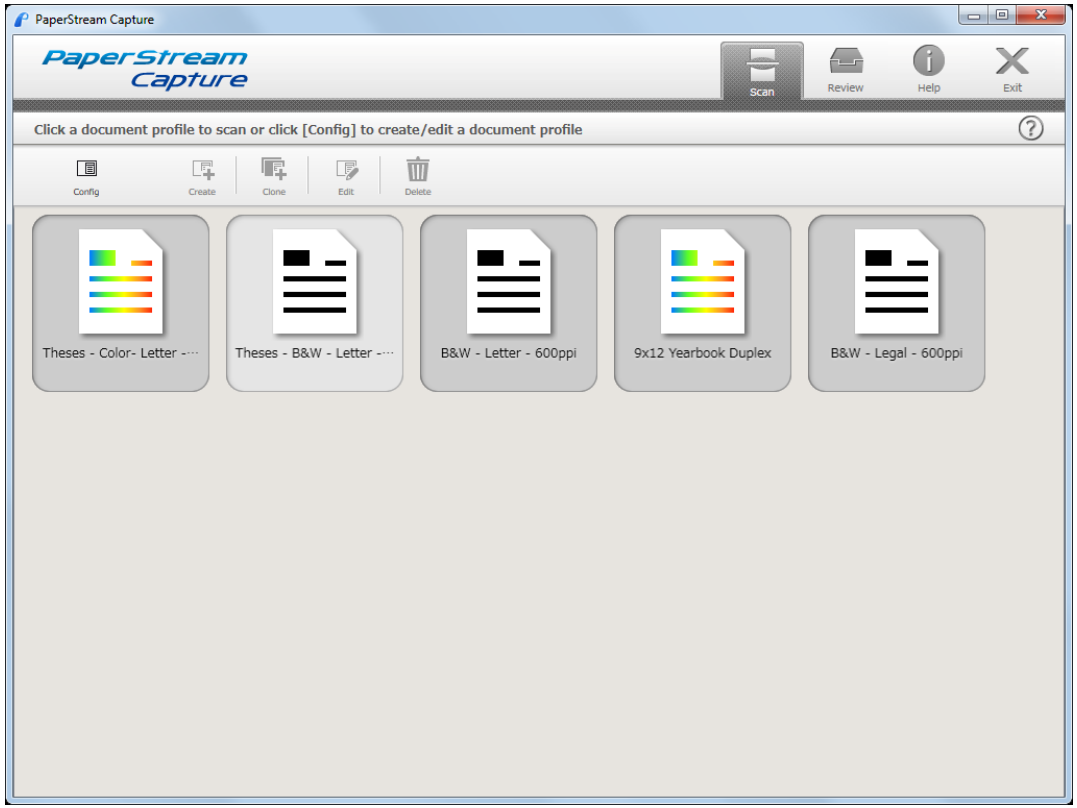
Scanning

- Fujitsu fi-6670
  - Straight-through feed path
  - Software allows for scanning once and outputting multiple files
  - **Binding, printing, paper quality changes with each thesis**
- MarcEdit Z39.50 client to download Catalog record
- Name scan folder and MARC record the same:
  - lastName_firstName_year.file Format

TEXAS STATE UNIVERSITY
The rising STAR of Texas

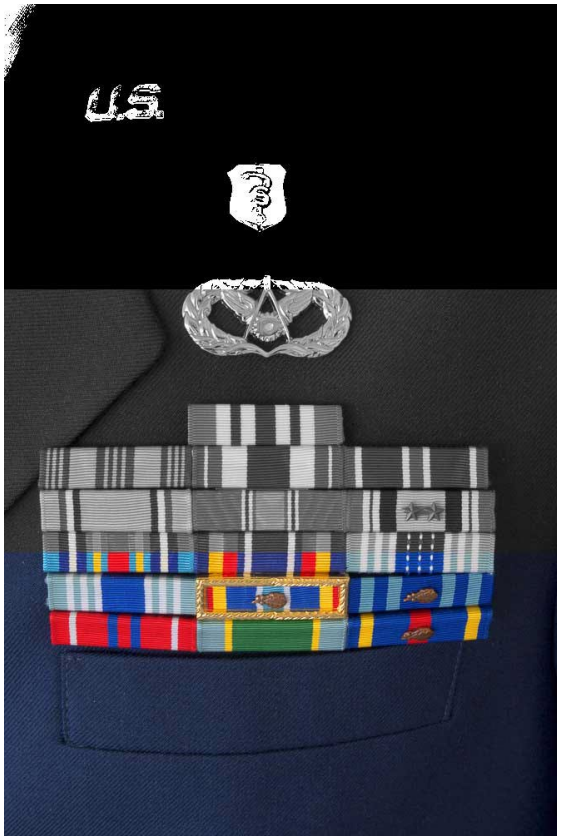
B&W vs Color

- B&W, or bitonal, is *Only Black and White*
- *Color* is a catch-all term for *Not Bitonal*



1 bit vs 8 bit vs 24 bit
or
B&W vs Gray vs Color

- Black and White (Bitonal) – 1 bit
  - 600 ppi
  - ~4 MB uncompressed TIFF
  - <1 MB group4 TIFF ***SPACE SAVINGS***
- Grayscale – 8 bit
  - 400 ppi
  - ~15mb uncompressed TIFF
- Color – 24 bit (3 x 8 bit)
  - 400 ppi
  - ~45 MB TIFF uncompressed



TEXAS STATE UNIVERSITY
The rising STAR of Texas

Post-Process physical and digital items

TEXAS STATE UNIVERSITY
The rising STAR of Texas

It’s about the end product

- It’s about the end product
  - It’s about the end product
    - It’s about the end product
      - It’s about the end product
        - It’s about the end product
          - It’s about the end product
            - It’s about the end product
              - It’s about the end product
                - Details matter

TEXAS STATE UNIVERSITY
The rising STAR of Texas

Whatever Works (within reason)

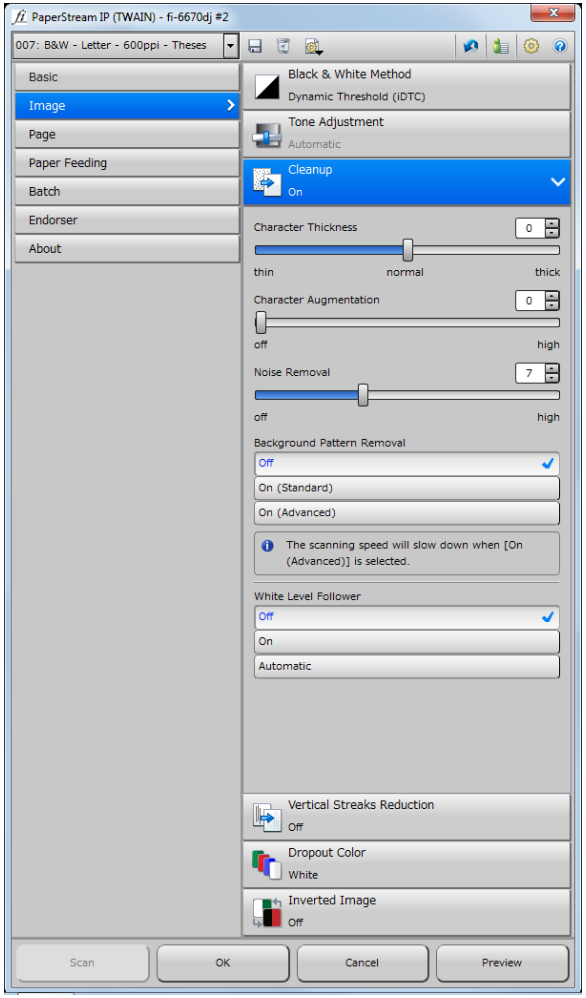
- BASH Shell scripts that will automatically copy, convert, move, name, rename, deskew, resize folders and files
- Different scripts to process the B&W and Color theses
- Color pages have a different process than Bitonal
- Adobe Photoshop and/or ImageMagick to process pages as necessary
- ABBYY Finereader Hot Folders to batch create and OCR PDFs from Preservation TIFFs

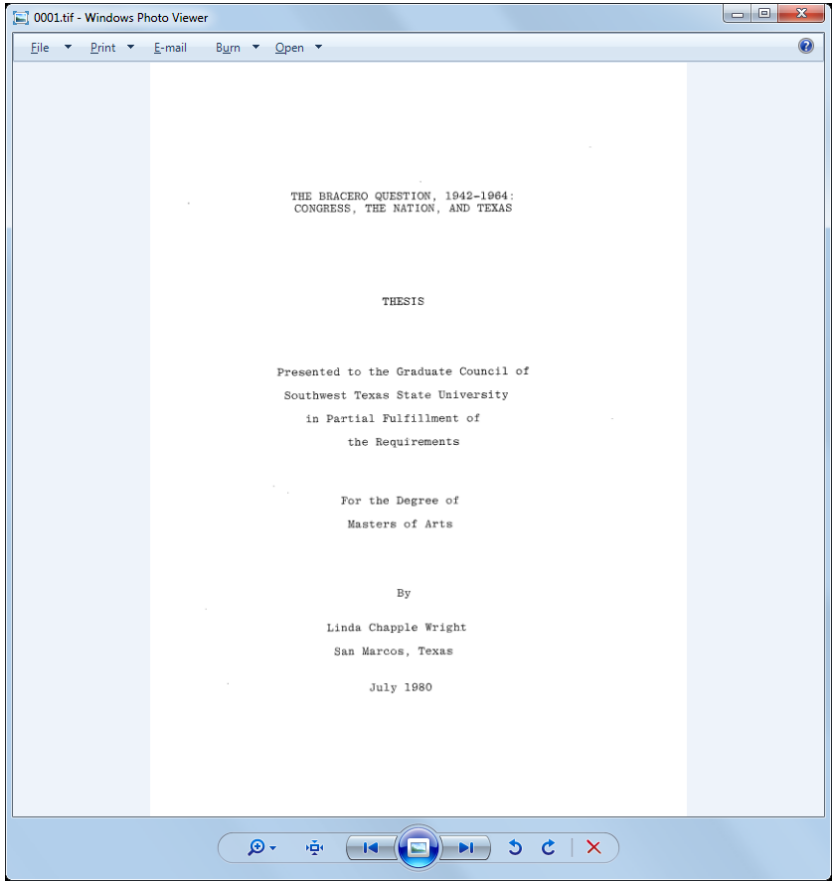
TEXAS STATE UNIVERSITY
The rising STAR of Texas

```
#!/bin/bash

# ProcessBWTheses.sh
# jeremy.moore@txstate.edu
# 01-27-2016

for dir in *; do
        cd "$dir"
        echo "$dir"
        mkdir -pv "$dir"
        x=1
        for img in *.tif; do
                mv "$img" `printf "%s/%s_%04d.tif" \
                        "$dir" "$dir" $x`
                x=$(($x+1))
        done
        cp `printf "%s.mrc" "$dir"` "$dir"
        cd "$dir"
        echo "Processing images"
        find . -name "*.tif" -print0 | xargs -0 -I {} -n 1 -P 4 \
                mogrify -deskew 40 -gravity east \
                -extent 5100x6600 -compress group4 "{}" 2>/dev/null
        echo "Images processed"
        cd ../..
done
```

TEXAS STATE UNIVERSITY
The rising STAR of Texas

Software Agnosticism





TEXAS STATE UNIVERSITY
The rising STAR of Texas

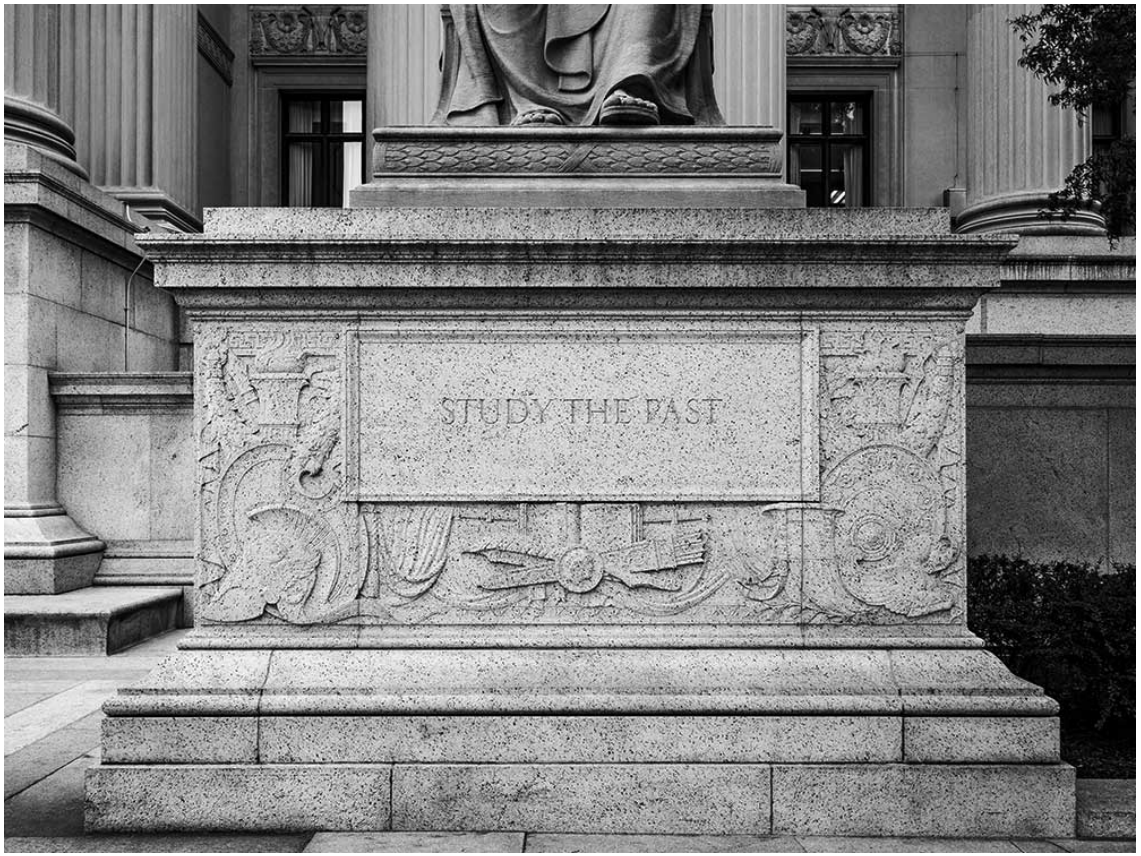
Cast of Characters

- Joan Heath, Assistant Vice President & University Librarian
- Kristine Toma, University Archivist
- Megan North, Assistant Archivist, University Archives
- Ray Uzwyshyn, Director, Collections and Digital Services
- Jeanne Hazzard, Digital Collections Librarian
- Stephanie Towery, Copyright Officer
- Aaron Sinkar, Assistant Director, Administrative Services
- Misty Hopper, Head, Cataloging and Metadata Services
- Selene Hinojosa, Collection Development Librarian
- Paivi Rentz, Library System Coordinator
- Amy Eoff, Database Management Services Assistant
- Cliff Wood, Library Facilities Assistant
- Joshua Van Meter, Library Facilities Assistant
- Digital Imaging Technicians
  - Oscar Martinez, Project Lead
  - Luke Sebree
  - Ashton Woodward
  - Vincente Rangel
  - Wenlan Cai (now graduate research assistant with University Archives)

TEXAS STATE UNIVERSITY
The rising STAR of Texas

Project Management

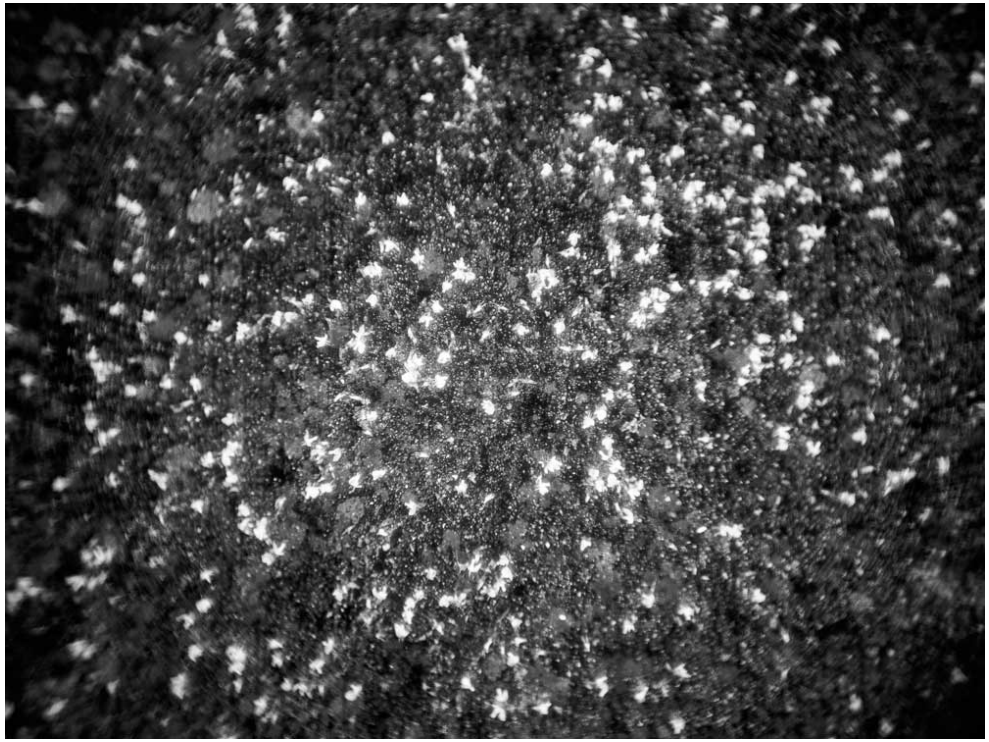
- Fail quickly and often to find strong/weak links in the production chain



TEXAS STATE UNIVERSITY
The rising STAR of Texas

Training vs. Teaching

- Everything is the same until it isn't and everything is unique, except it's not.
  - **Binding, printing, paper quality changes with each thesis**



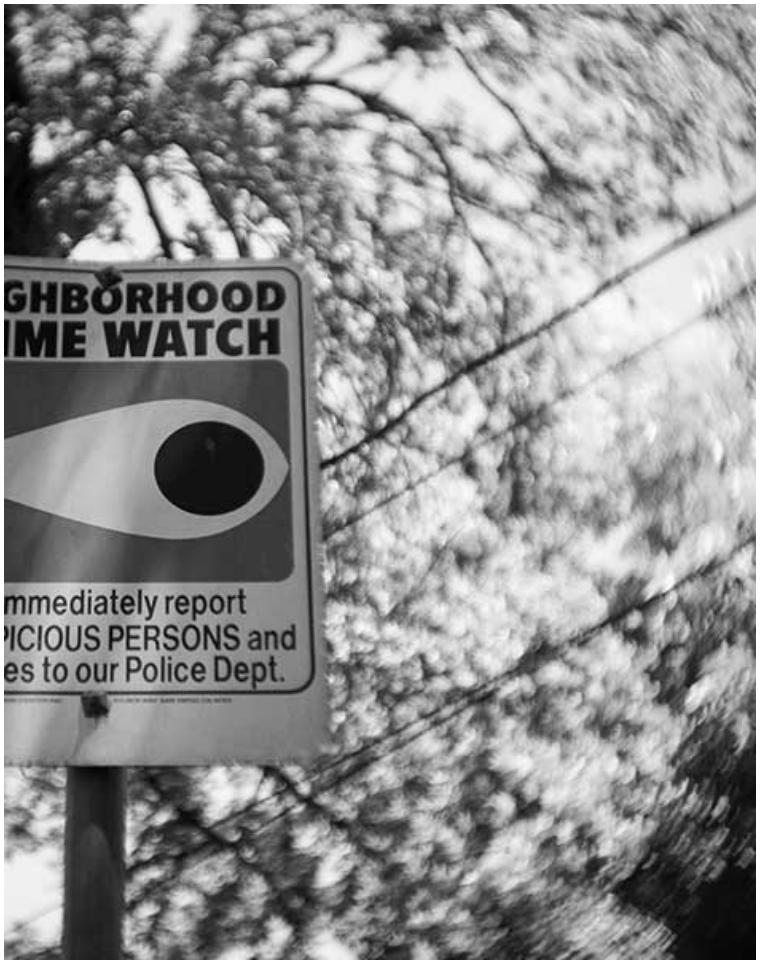
TEXAS STATE UNIVERSITY
The rising STAR of Texas

Fail fast = Skip problems

- No Ctrl+Z for most of the process

Leverage Strengths

- Human
  - Kinesthetic
  - Aesthetic
  - Intuitive
  - Creative
  - Pattern recognition
  - Reliable (with oversight)
- Computer
  - Works weekends/holidays
  - Repetitive
  - Pattern recognition
  - Reliable (with oversight)



TEXAS STATE UNIVERSITY
The rising STAR of Texas

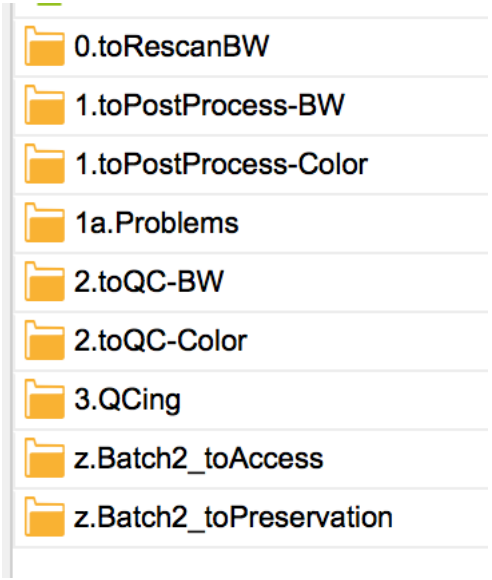
Provide Training Wheels

- Start with looking and seeing: Quality Control
- PHP Web Application for sorting Color pages from B&W

TEXAS STATE UNIVERSITY
The rising STAR of Texas

Take them away!

- **Binding, printing, paper quality changes with each thesis**
- Personal paths to the end product, not a checklist of steps
- 'Problem' volumes are now opportunities for creativity and growth



TEXAS STATE UNIVERSITY
The rising STAR of Texas

Quality Control for Documents

TEXAS STATE UNIVERSITY
The rising STAR of Texas

Questions?

TEXAS STATE UNIVERSITY
The rising STAR of Texas