PROJECT BACKGROUND:

In 2010, the UNT Libraries Digital Projects Unit (DPU) partnered with the Greater Western Library Alliance (GWLA) to manage files and host technical reports that could not be managed by Google and Hathi Trust as part of their Technical Reports Archive and Image Library (TRAIL) initiative. The DPU has continued to work with the University of Arizona (UA) to digitize reports that have foldouts and non-consistent page sizes.

Staff members at UA gather together copies of the reports that will be digitized as part of the TRAIL project and disbind most of the reports before shipping them to the DPU. Each batch has roughly 4-6 boxes containing 200-300 reports, although the number varies depending on the size of the publications. The shipments also include MARC record files for the reports that are converted to the xml format used for metadata in the DPU. Within the DPU, any final processing/disbinding of the physical copies is completed and then the reports are digitized and uploaded according to internal standards and workflows.

HARDWARE/SOFTWARE:

Equipment	Fujitsu fi-4340C Image Scanner (Duplex scanner) Plustek OpticBook 3600 Scanner (Flatbed scanner) Zeutschel Omniscan 10000TT (Planetary scanner)
Processing Software	Scan Tailor (open source) Photoshop CS5 PrimeOCR
Metadata Software	jEdit (open source)

PROCESS OVERVIEW:

Stage One: Inventory

1. Report identifiers are noted in a text file, sorted by box numbers.
2. Items are sorted based on which ones still need to be disbound.
3. Students do pre-disbinding work on items that need to be cut.
4. Reports are cut and then routed to the scanning queue.

Stage Two: Scanning

1. Disbound pages are scanned on a duplex scanner.
2. Fold-out pages are flagged; filler pages are scanned in their places.
3. Files are renamed using a magick-numbering scheme and the report is checked for missing pages.
4. Oversized pages and pages to rescan (color covers, grayscale illustrations, cut-off pages, etc.) are marked with sticky notes.
5. Marked pages are scanned on a large flatbed scanner and the new files replace the unusable pages; fold-out pages are marked for further processing.
6. If fold-out pages are too large for the flatbed, they are scanned on the Zeutschel planetary scanner and the files are marked for further processing.
7. Excessively large foldouts are scanned in pieces and then digitally "stitched" together in Photoshop.

Stage Three: Processing

1. An initial quality-control check is done to ensure that the scanning workflow was followed correctly.
2. Oversized pages are processed manually in Photoshop: they are deskewed and resized to reflect the size of the physical page.
3. The first page of every report is processed manually in Photoshop.
4. The regularly-sized image files for each report are processed using Scan Tailor; the program deskews the text, cleans the pages, resizes each image to the same physical size, and compresses the files.
5. Pages are rotated as needed to ensure proper text orientation.
6. Oversized foldouts are re-integrated with the other digital files.
7. A final quality-control check is completed.
8. PDF files are created for each report and file types are sorted into an established folder structure to handle multiple formats.
9. The reports are flagged for OCR in the workflow; we use PrimeOCR for the technical reports.

Stage Four: Metadata

1. Staff members at UA package MARC records that match the reports and send them with each batch.
2. The MARC records are converted into UNTL xml files.
3. Staff members in the DPU do initial editing of values to improve consistency and conformity to metadata standards.
4. A student copies xml text into a web-based metadata form and edits the record for completeness and compatibility with UNT standards.
5. Metadata goes through a quality-control check and any necessary revisions are made to the records.

Stage Five: Upload

1. Reports are flagged for processing and bundling for upload.
2. Digital copies are available at http://digital.library.unt.edu/explore/collections/TRAIL/browse/