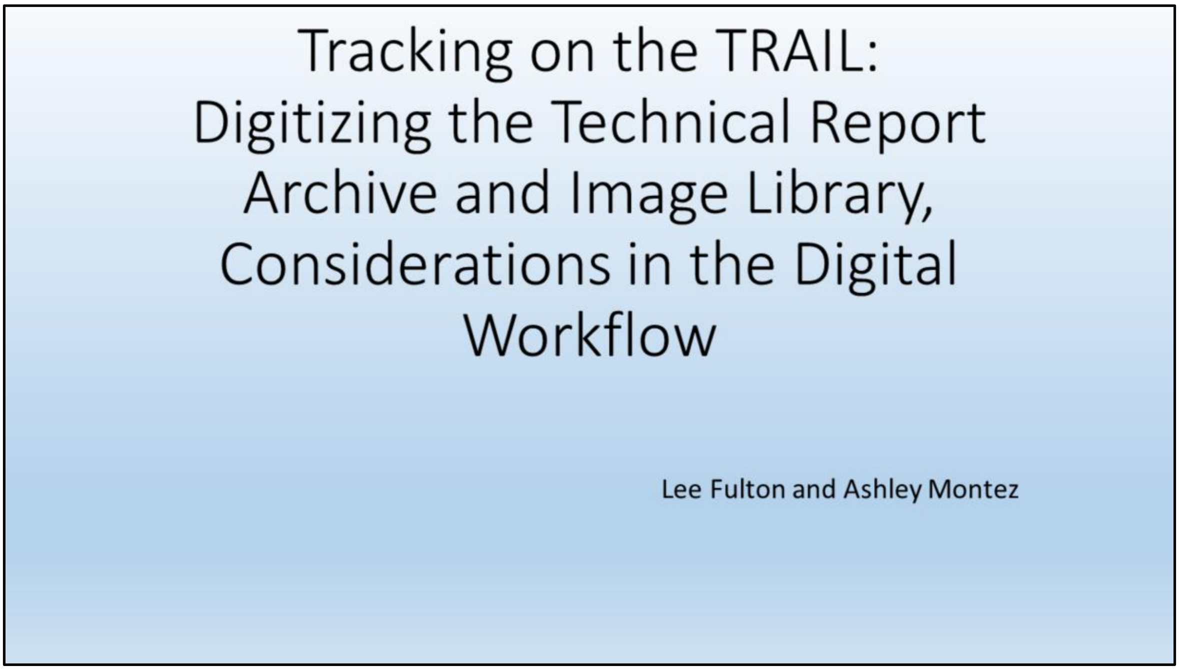
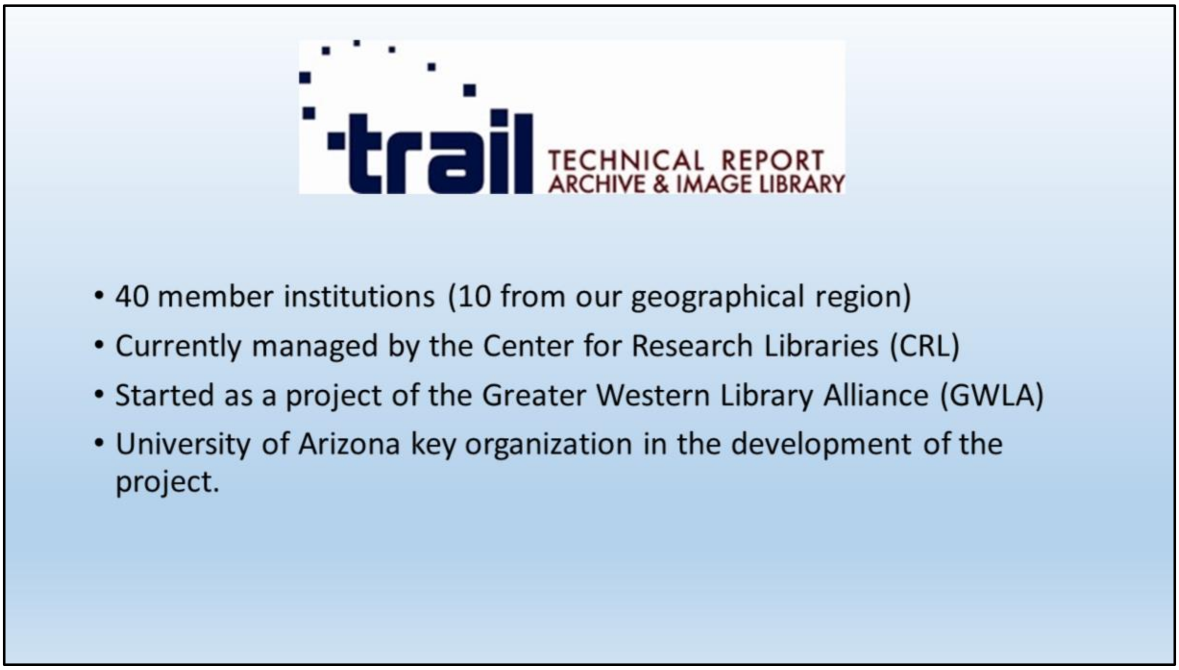


Tracking on the TRAIL: Digitizing the Technical Report Archive and Image Library, Considerations in the Digital Workflow

Lee Fulton and Ashley Montez

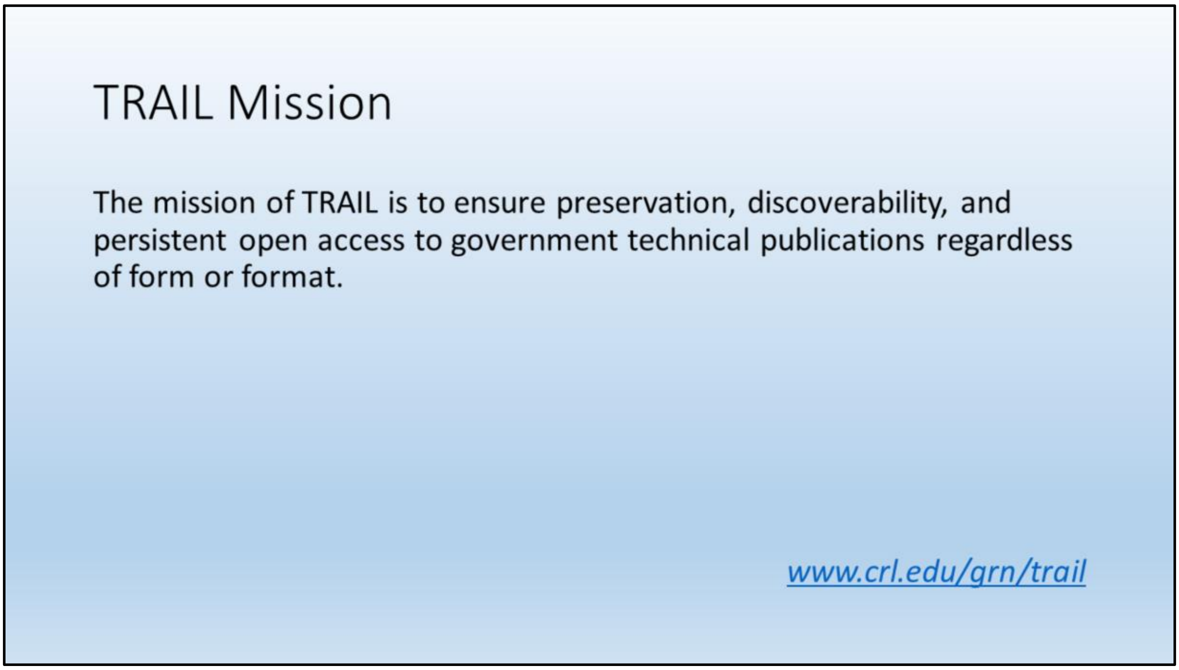
Ashley Montez is a Student Assistant in the Digital Projects Lab at the University of North Texas and is pursuing her undergraduate degree in business administration. I am Lee Fulton, a Library Specialist in the same organization. I graduated with a Masters of Information Science with an emphasis on information organization from the former School of Library and Information Sciences, now the College of Information at UNT. We are here today to talk about the work we do on our grant project.



trail TECHNICAL REPORT ARCHIVE & IMAGE LIBRARY

- 40 member institutions (10 from our geographical region)
- Currently managed by the Center for Research Libraries (CRL)
- Started as a project of the Greater Western Library Alliance (GWLA)
- University of Arizona key organization in the development of the project.

A little background first. TRAIL, or the Technical Report Archive & Image Library, originated in 2006 with a call for proposals by the Greater Western Library Alliance. The University of Arizona proposed to develop a collaborative project with the Center for Research Libraries to provide open access to federal technical reports, primarily publications prior to 1976. Thus TRAIL was born. As of April of this year, Indiana University became the 40th institution to join TRAIL. Interestingly, our geographical region represents 25% of membership with the University of Arkansas, Baylor, University of Houston, Rice, Oklahoma State, UT San Antonio, Texas A&M, Texas Tech, and the University of New Mexico joining the University of North Texas.



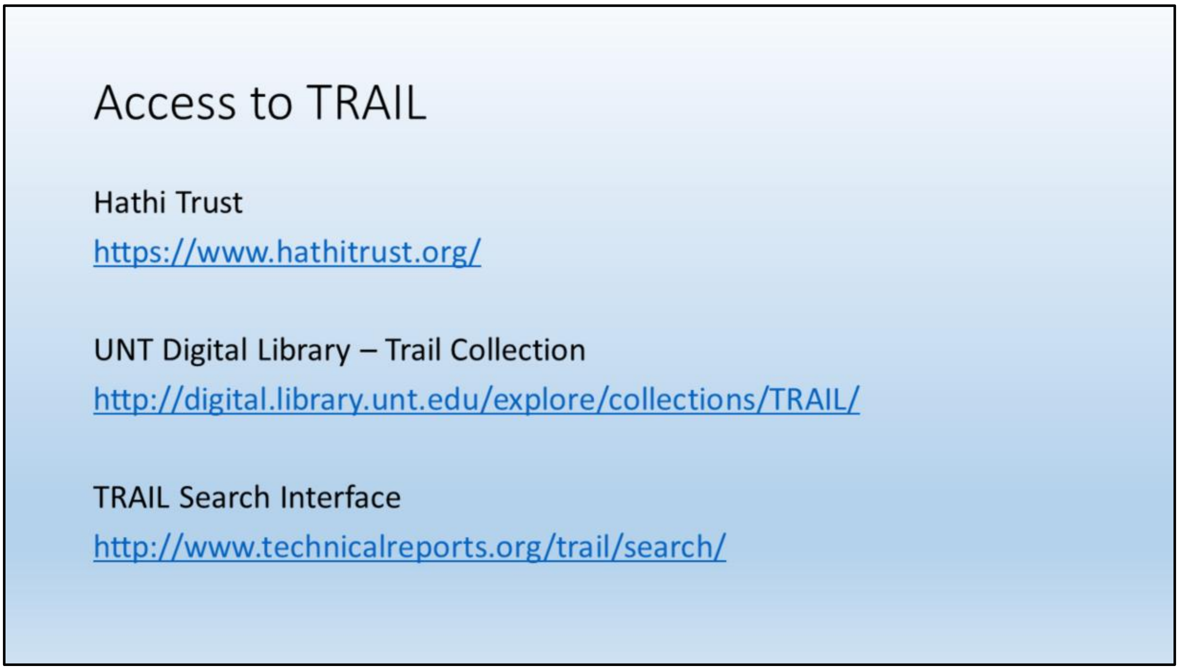
TRAIL Mission

The mission of TRAIL is to ensure preservation, discoverability, and persistent open access to government technical publications regardless of form or format.

www.crl.edu/grn/trail

A vast amount of technical report literature is available at your local major research library, but those issued prior to the advent of the Internet, say 1993 or so, may not be retrievable due to a lack of, or minimum level of cataloging and indexing. So the collaborative effort is to identify content that merits the effort necessary to move it onto the Internet.

Where can you find TRAIL documents?



Access to TRAIL

Hathi Trust

https://www.hathitrust.org/

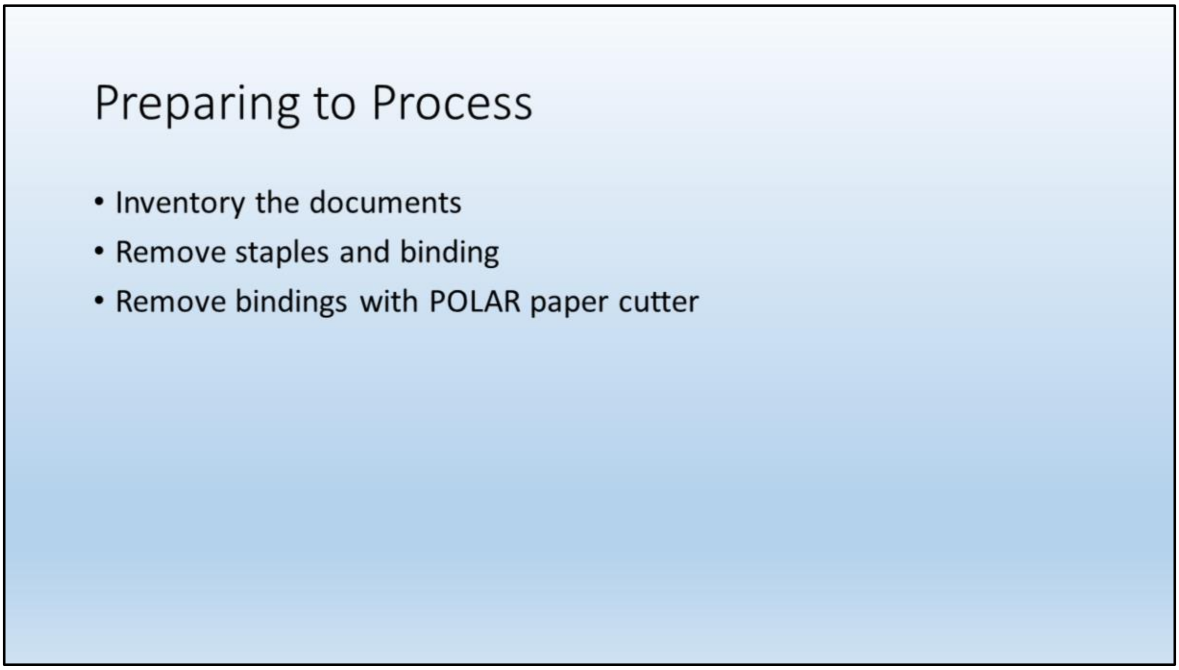
UNT Digital Library – Trail Collection

http://digital.library.unt.edu/explore/collections/TRAIL/

TRAIL Search Interface

http://www.technicalreports.org/trail/search/

The first link is to our collection of TRAIL which, the last time I browsed, contains around 17,000 separate resources. The second link is to the TRAIL Search Interface maintained by the University of Washington Libraries and is provided as a link in most university catalogs for users trying to access TRAIL content. According to their website, they have more than 40,000 reports available. The third link is to the Test (or the Pilot) Site for TRAIL access at the University of Hawaii. I don't recommend trying to retrieve TRAIL documents from the Hawaii website, because they provide access to exactly 909 reports, so it was not maintained for long, although it does provide the link to the University of Washington website. I only mention this one because when I was doing research on TRAIL, several universities, that will remain nameless, provide this link in their catalogs.

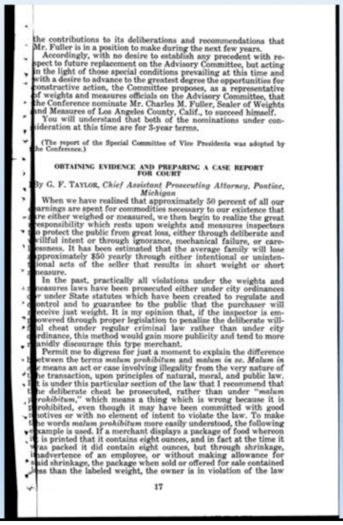


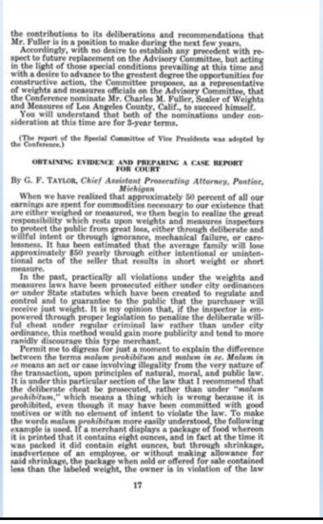
Preparing to Process

- Inventory the documents
- Remove staples and binding
- Remove bindings with POLAR paper cutter

We receive shipments in boxes and inventory each document in notepad file according to a unique identifier which has been assigned by the University of Arizona.
Then we extract all of the staples and rip sections of the document from the binding in preparation for cutting on the POLAR, a high-speed cutter located in our Preservation department. The POLAR is an invaluable time-saving piece of equipment. The day before yesterday, I spent 2 hours pulling bindings apart and extracting staples from 30 documents of varying sizes. The POLAR cut off the remaining bindings in about 10 minutes. And that is typically a ¼ inch, but sometimes less depending upon the document.

Overcut





the contributions to its deliberations and recommendations that Mr. Fuller is in a position to make during the next few years.

Accordingly, with no desire to establish any precedent with respect to future replacement on the Advisory Committee, but acting in the light of those special conditions prevailing at this time and with a desire to advance to the greatest degree the opportunities for constructive action, the Committee proposes, as a representative of weights and measures officials on the Advisory Committee, that the Conference nominate Mr. Charles M. Fuller, Sealer of Weights and Measures of Los Angeles County, Calif., to succeed himself.

You will understand that both of the nominations under consideration at this time are for 3-year terms.

(The report of the Special Committee of Vice Presidents was adopted by the Conference.)

OBTAINING EVIDENCE AND PREPARING A CASE REPORT FOR COURT

By G. F. TAYLOR, *Chief Assistant Prosecuting Attorney, Pontiac, Michigan*

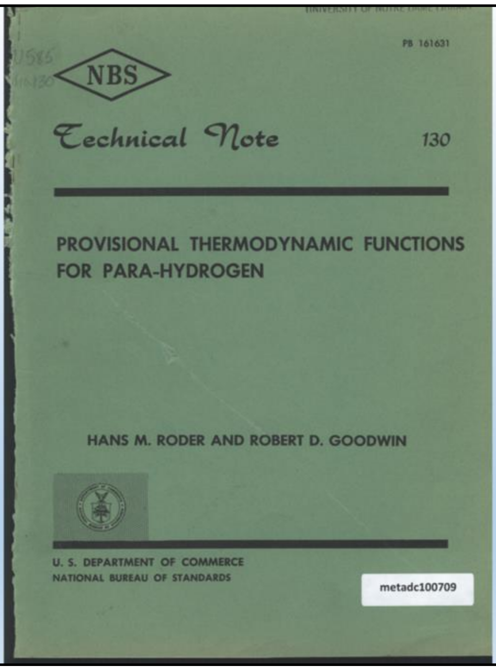
When we have realized that approximately 50 percent of all our earnings are spent for commodities necessary to our existence that are either weighed or measured, we then begin to realize the great responsibility which rests upon weights and measures inspectors to protect the public from great loss, either through deliberate and willful intent or through ignorance, mechanical failure, or carelessness. It has been estimated that the average family will lose approximately $50 yearly through either intentional or unintentional acts of the seller that results in short weight or short measure.

In the past, practically all violations under the weights and measures laws have been prosecuted either under city ordinances or under State statutes which have been created to regulate and control and to guarantee to the public that the purchaser will receive just weight. It is my opinion that, if the inspector is empowered through proper legislation to penalize the deliberate willful cheat under regular criminal law rather than under city ordinance, this method would gain more publicity and tend to more ranidly discourage this type merchant.

Permit me to digress for just a moment to explain the difference between the terms *malum prohibitum* and *malum in se*. *Malum in se* means an act or case involving illegality from the very nature of the transaction, upon principles of natural, moral, and public law. It is under this particular section of the law that I recommend that the deliberate cheat be prosecuted, rather than under "*malum prohibitum*," which means a thing which is wrong because it is prohibited, even though it may have been committed with good motives or with no element of intent to violate the law. To make the words *malum prohibitum* more easily understood, the following example is used. If a merchant displays a package of food whereon it is printed that it contains eight ounces, and in fact at the time it was packed it did contain eight ounces, but through shrinkage, inadvertence of an employee, or without making allowance for said shrinkage, the package when sold or offered for sale contained less than the labeled weight, the owner is in violation of the law

17

Here is the result of an overcut on the POLAR due to the margins in this document not being uniform. I had 3 other casualties like this on the following pages where the margins just started sliding to the left, so being double-sided, I had 8 bookedge scans to do, because you don't want to run a damaged page through a duplex scanner. And this is why we always check each document after cutting so that we can perform surgery if necessary, and this entails scotch tape and Photoshop work. I was disappointed in my performance, but 4 pages out of several thousand is not too bad. This by the way was a National Bureau of Standards document, and I can tell you that the margins were not standard.



PB 161631

NBS

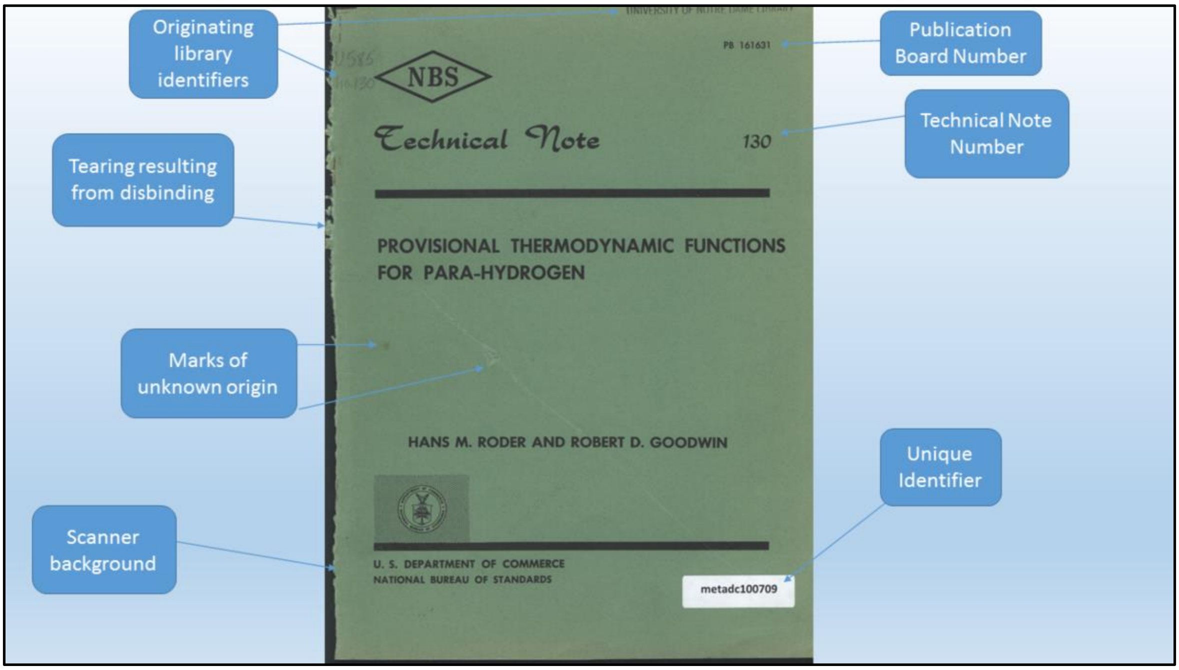
Technical Note 130

PROVISIONAL THERMODYNAMIC FUNCTIONS FOR PARA-HYDROGEN

HANS M. RODER AND ROBERT D. GOODWIN

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

Here is a typical document cover after being disbound.



UNIVERSITY OF NOTRE DAME LIBRARY

U586
no.130

PB 161631

NBS

Technical Note 130

PROVISIONAL THERMODYNAMIC FUNCTIONS
FOR PARA-HYDROGEN

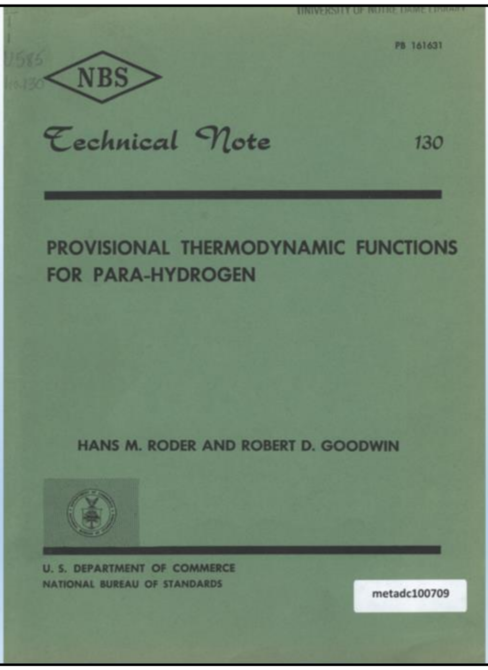
HANS M. RODER AND ROBERT D. GOODWIN

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

metadc100709

Most of these government documents follow this scheme, with the title, author, originating agency, and series (in this case National Bureau of Standards Technical Note #130). PB 161631 identifies this document as coming under the purview of The Publication Board, which evolved into the National Technical Information Service (NTIS) in 1964 and which is still with us today as an agency of the US Dept of Commerce. These attributes of the document are of interest to the metadata portion of the workflow which is beyond the scope of this presentation.

What is of interest to us is the unique identifier at the bottom right is actually a sticker that has been affixed at the University of Arizona upon receipt of the document from the originating library, in this case the University of Notre Dame. The originating library identifiers were placed there upon receipt of the document from the government. The tearing on the binding on the left side is indicative of some strong glue, alternatively, many documents can be disbound with no tearing. The stray marks of unknown origin are typical for an old document, this being from 1961.



PB 161631

U585
no.130

NBS

Technical Note 130

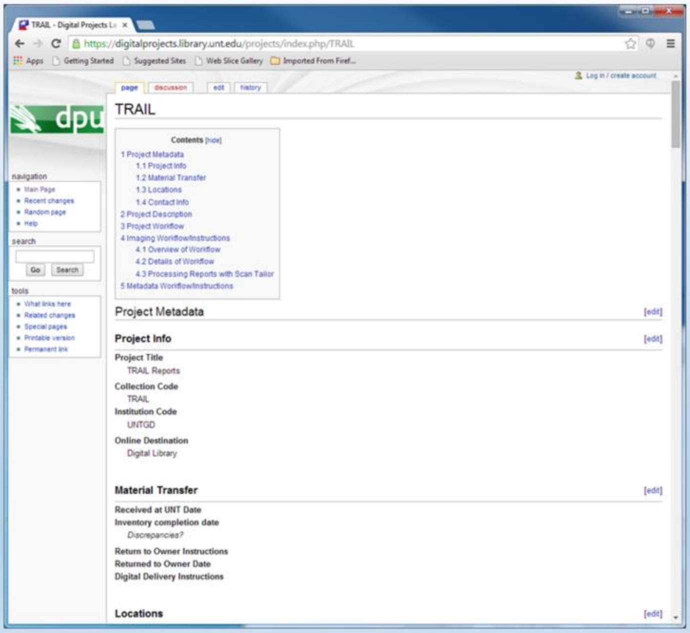
PROVISIONAL THERMODYNAMIC FUNCTIONS FOR PARA-HYDROGEN

HANS M. RODER AND ROBERT D. GOODWIN

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS

metadc100709

Here is the cover after Photoshop editing. Ashley is now going to discuss our QC process.



TRAIL

Contents [hide]

1 Project Metadata
- 1.1 Project Info
- 1.2 Material Transfer
- 1.3 Locations
- 1.4 Contact Info

2 Project Description

3 Project Workflow

4 Imaging Workflow/Instructions
- 4.1 Overview of Workflow
- 4.2 Details of Workflow
- 4.3 Processing Reports with Scan Tailor

5 Metadata Workflow/Instructions

Project Metadata [edit]

Project Info [edit]

Project Title
TRAIL Reports

Collection Code
TRAIL

Institution Code
UNTGD

Online Destination
Digital Library

Material Transfer [edit]

Received at UNT Date

Inventory completion date
Discrepancies?

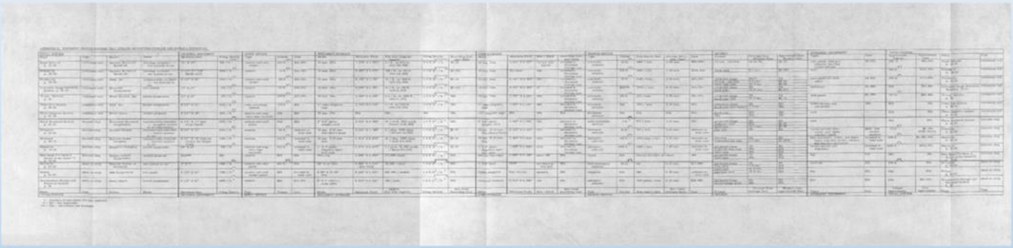
Return to Owner Instructions

Returned to Owner Date

Digital Delivery Instructions

Locations [edit]

We utilize an intranet wiki for students’ reference in step-by-step instructions.



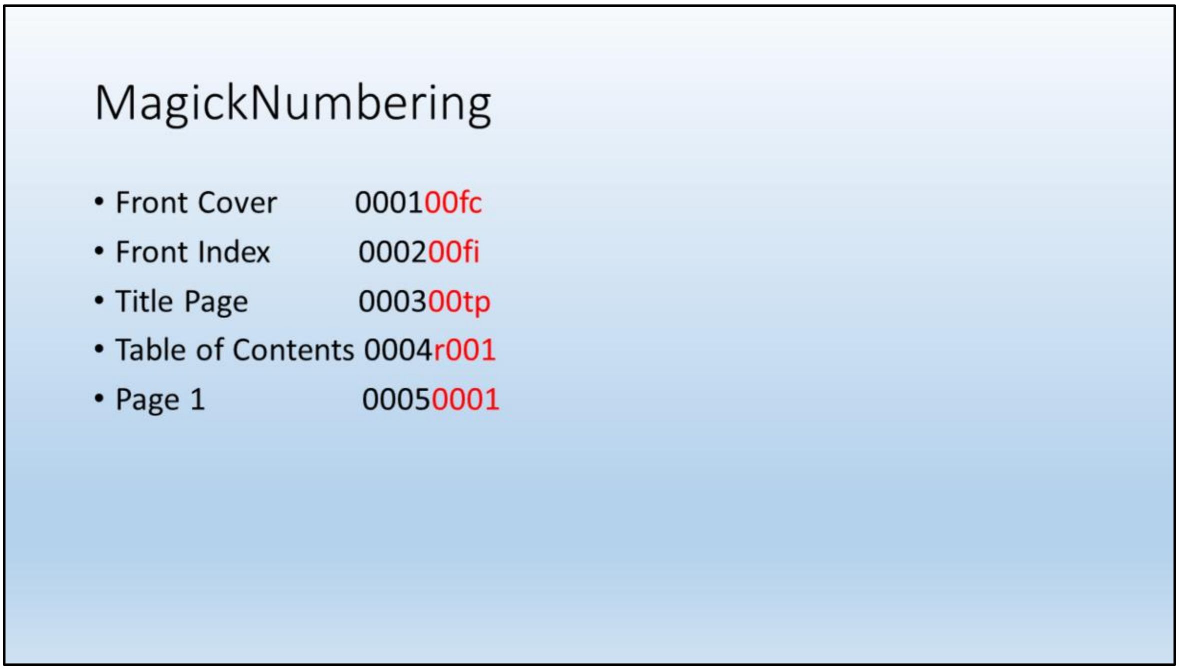
The important thing to remember is that most every TRAIL document contains at least one oversized page which folds out. These can range from anywhere from slightly larger than your standard 8.5 X 11.0 page to several feet long with multiple folds. This example is a standard 11 inches or so in height, but is 3 ½ feet in width.

We tag each oversized page with a post-it-note for easy accessibility when scanning these later. These may be scanned on either a large bookedge or on the Zeutschel. Extremely large pages require scanning on the SupraScan, our newspaper scanner. Despite the size of these scanners, many pages require stitching later on. Stitching is the combining of 2 or more pages in Photoshop so as to recreate the actual oversized page.

Duplex Scanning

- Fujitsu-6670
- Fujitsu-6140Z

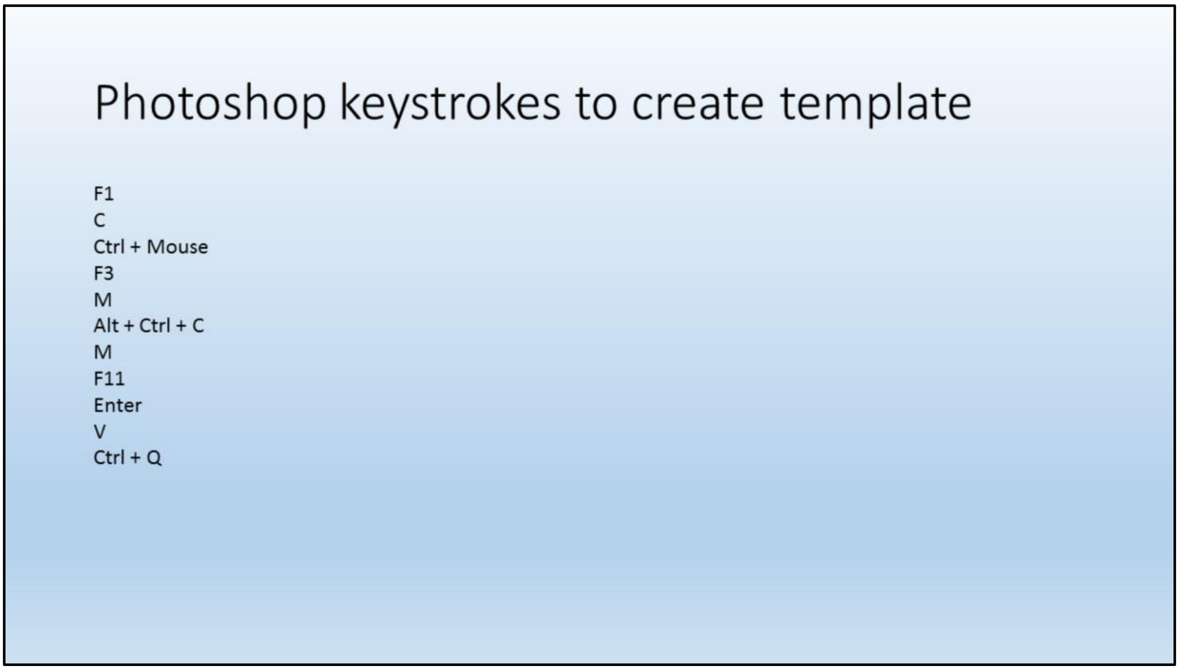
Preference is to scan on the F-6670 because of its speed and its autocropping. It is also capable of scanning oversized pages of around 30 inches in width.
The 6140Z requires page sizing prior to every document scan, or you can forget and create more work for yourself later on. Grayscale images need to be marked for rescanning at a different scan setting and will be discussed later in this presentation. Covers and any photos in color need to be scanned on one a flatbed scanner. Because of the age of these documents, we have run across very few color photos. The majority we found were in one U.S Fish and Wildlife Service document from 1981, a more recent document than we are accustomed to.



MagickNumbering

• Front Cover	000100fc
• Front Index	000200fi
• Title Page	000300tp
• Table of Contents	0004r001
• Page 1	00050001

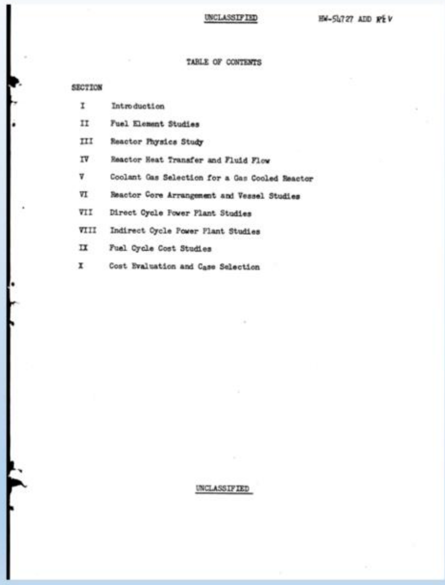
After scanning, we are ready to number the pages. We use an 8-digit numbering system for each image/page. The first 4 numbers are the numerical order in which the page falls. The second set of 4 numbers indicate the actual number of the page itself. The covers, front and back, and the title page use acronyms to designate them. Roman numeral pages are designated with an r.
It is easiest to number the pages in a document in Adobe Bridge.



Photoshop keystrokes to create template

F1
C
Ctrl + Mouse
F3
M
Alt + Ctrl + C
M
F11
Enter
V
Ctrl + Q

Photoshop is the first step in the QC process. This is the sequential list of Photoshop keystrokes necessary to render the template ready for further processing. The keystrokes are beyond the scope of this presentation, however it is important to note that creating custom function keys like F1, F3, and F11 is an essential step in speeding up the QC process. For our uses, we designate F1 as grayscale and F3 as bitmap so as to easily switch between the two formats as necessary. Also important is becoming comfortable with using short-cut keys in general with the ultimate goal of only using the mouse for selecting content for editing. With enough practice, a user can create a template without reference to this checklist.

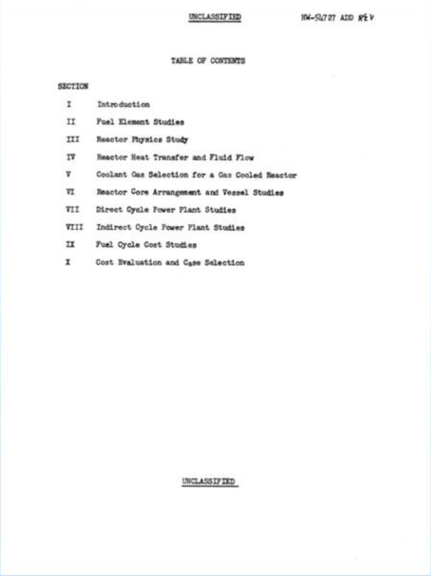


UNCLASSIFIED HW-54727 ADD REV

TABLE OF CONTENTS

SECTION

UNCLASSIFIED



UNCLASSIFIED HW-54727 ADD REV

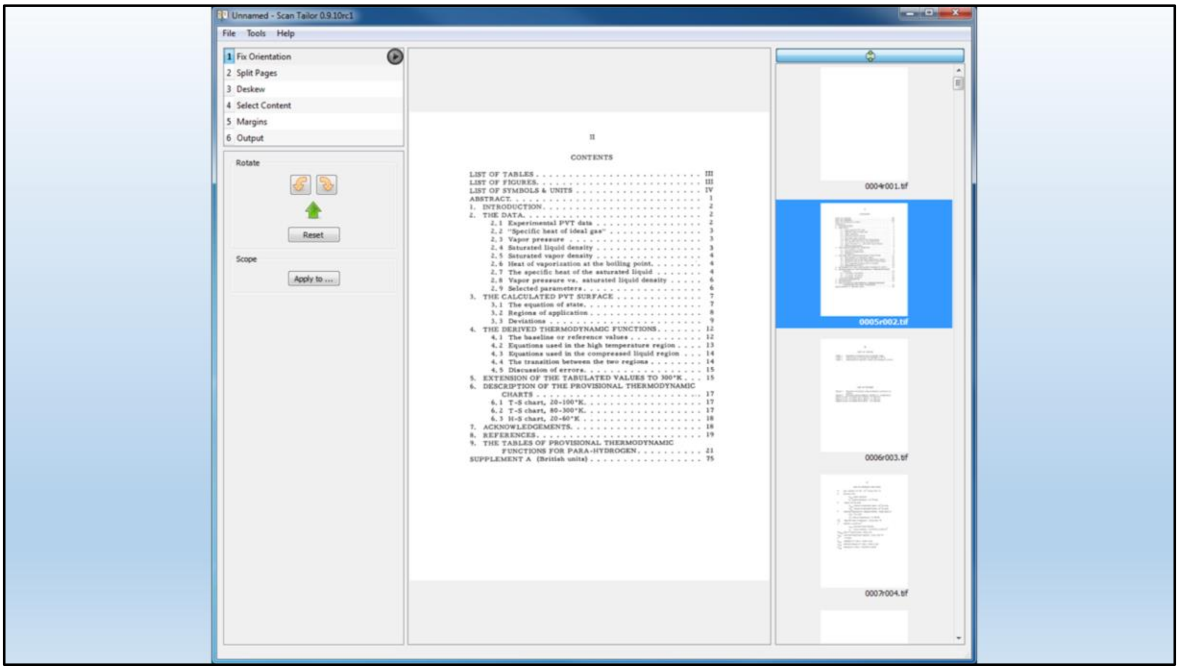
TABLE OF CONTENTS

SECTION

I Introduction
II Fuel Element Studies
III Reactor Physics Study
IV Reactor Heat Transfer and Fluid Flow
V Coolant Gas Selection for a Gas Cooled Reactor
VI Reactor Core Arrangement and Vessel Studies
VII Direct Cycle Power Plant Studies
VIII Indirect Cycle Power Plant Studies
IX Fuel Cycle Cost Studies
X Cost Evaluation and Case Selection

UNCLASSIFIED

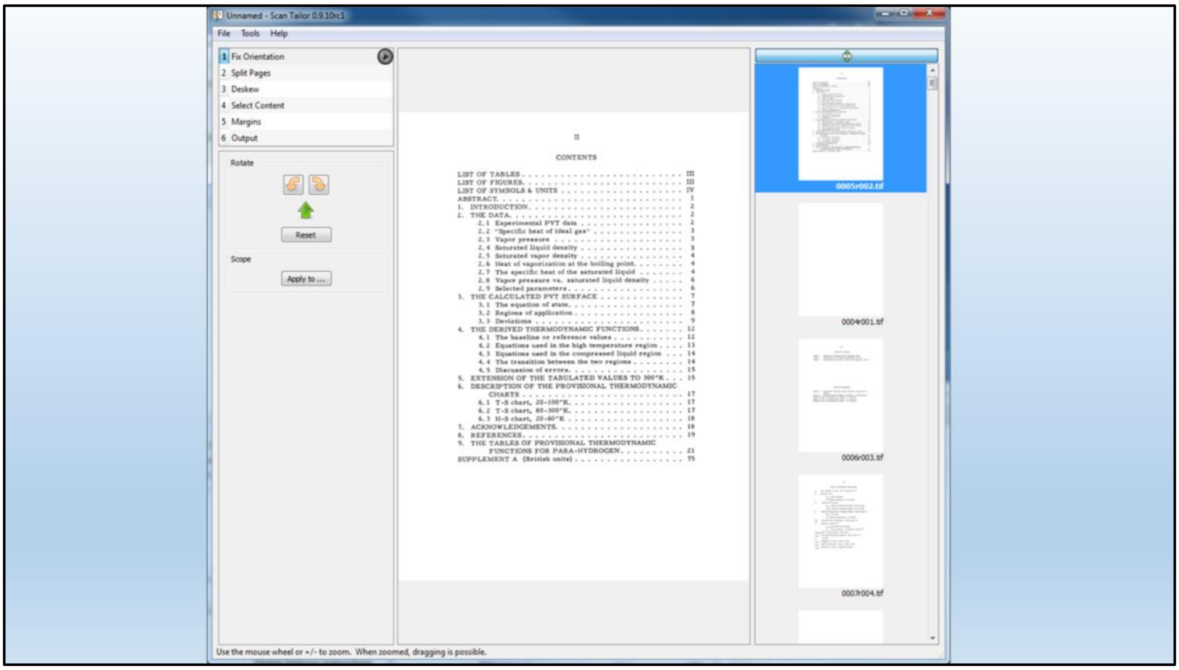
Here is the before and after for our template.



II

CONTENTS

An open source program, ScanTailor is utilized for text documents to create uniform outputs and capture only the relevant content. The reason we create a template is because ScanTailor allows the user to base the size of all the pages in a document on the first page in order. We usually choose the table of contents page and move it to the beginning. Now all pages following will be the same size after ScanTailor is finished processing.

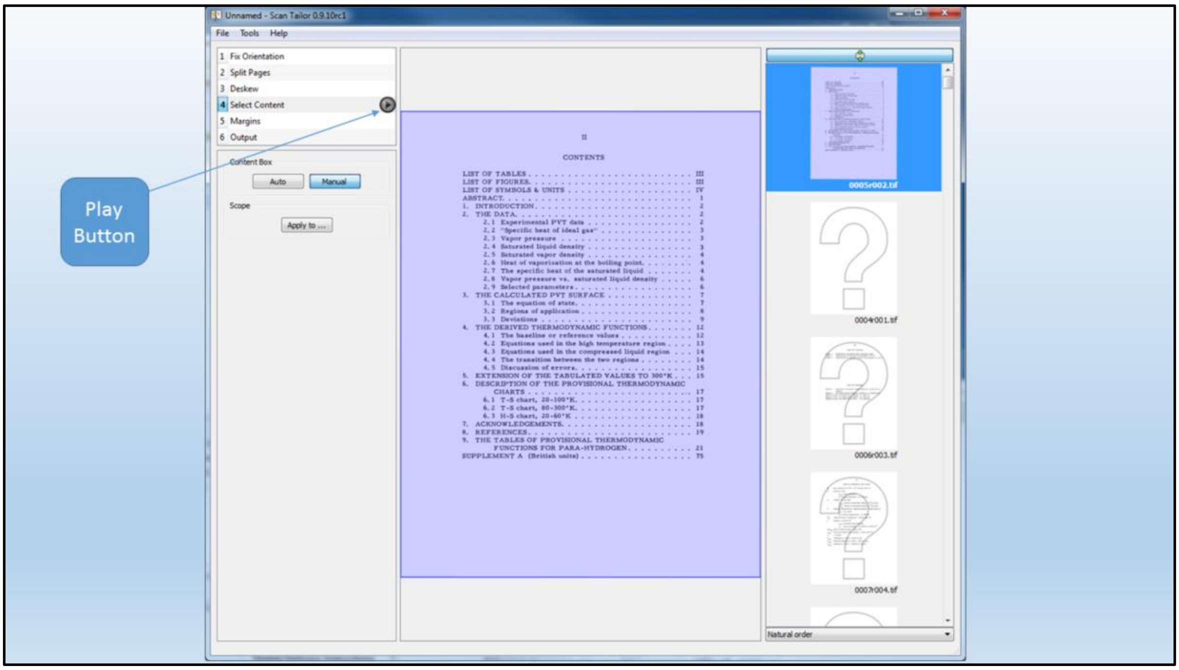


II

CONTENTS

Use the mouse wheel or +/- to zoom. When zoomed, dragging is possible.

Here we have placed our template first in order and are now ready to begin processing.

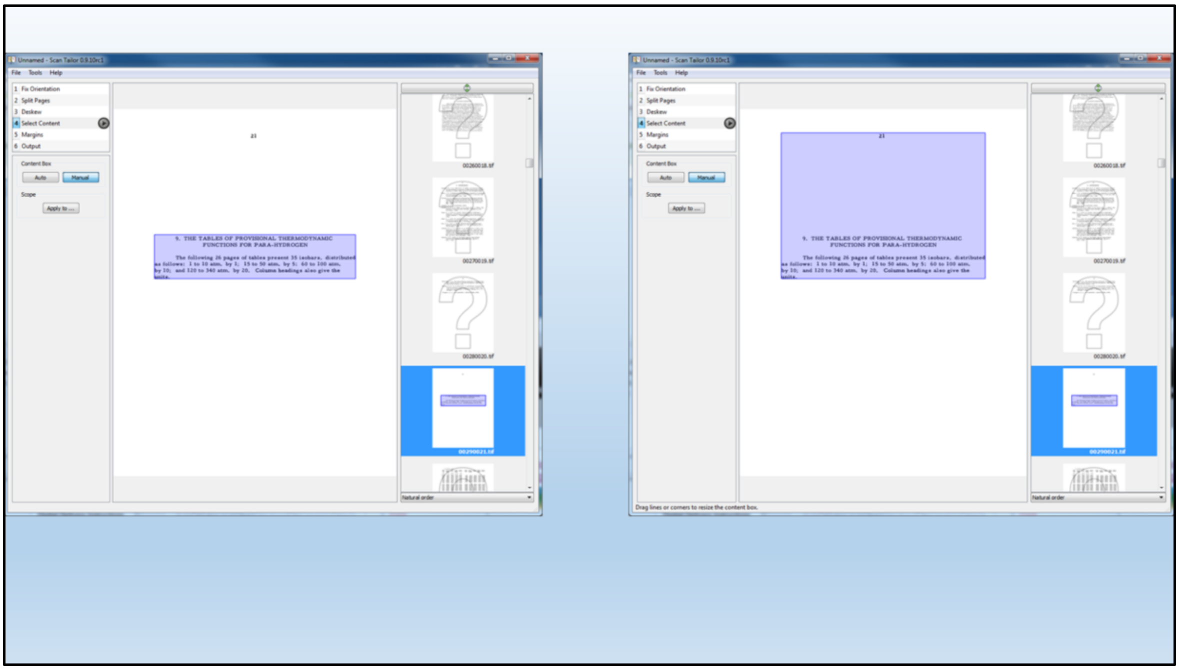


Play Button

II

CONTENTS

Step 2 is the same for all documents and we just ensure that we have a manual setting. Step 3 deskews the document so that each page is corrected for any crookedness. Step 4 begins with selecting the whole template page as shown, then we hit the play button.



21

9. THE TABLES OF PROVISIONAL THERMODYNAMIC FUNCTIONS FOR PARA-HYDROGEN

The following 26 pages of tables present 35 isobars, distributed as follows: 1 to 10 atm. by 1; 15 to 50 atm. by 5; 60 to 100 atm. by 10; and 120 to 340 atm. by 20. Column headings also give the units.

21

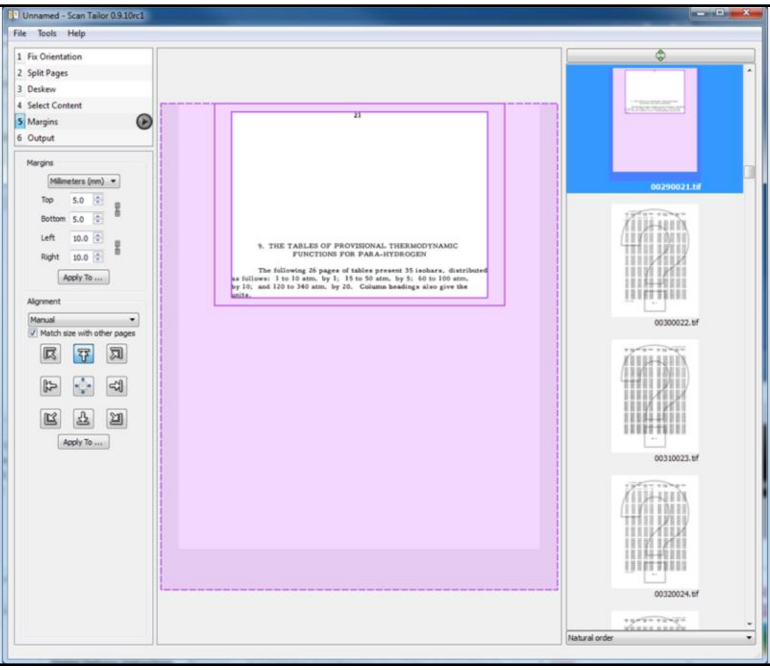
9. THE TABLES OF PROVISIONAL THERMODYNAMIC FUNCTIONS FOR PARA-HYDROGEN

The following 26 pages of tables present 35 isobars, distributed as follows: 1 to 10 atm. by 1; 15 to 50 atm. by 5; 60 to 100 atm. by 10; and 120 to 340 atm. by 20. Column headings also give the units.

Drag lines or corners to resize the content box.

The most time-consuming step in ScanTailor depends upon the document itself. ScanTailor recognizes content on a page, but sometimes this includes stray marks and staple holes. Conversely, if the page number is enough distance from the body of the words on the page then it may not be picked up by ScanTailor and we have to physically drag the content box to include it. This can be one of the most tedious steps in our QC process.

Margins



Margins

Millimeters (mm)

Top 5.0

Bottom 5.0

Left 10.0

Right 10.0

Alignment

Manual

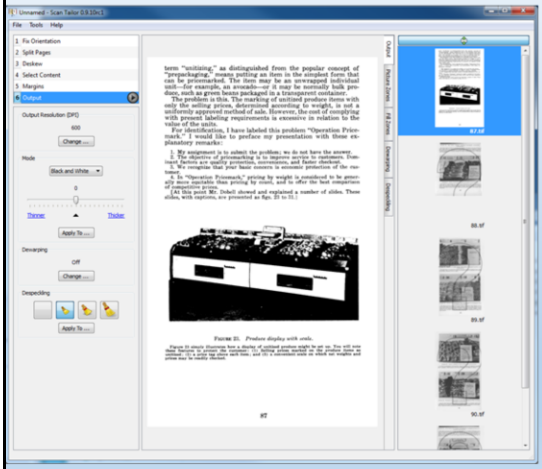
☑ Match size with other pages

21

9. THE TABLES OF PROVISIONAL THERMODYNAMIC FUNCTIONS FOR PARA-HYDROGEN

The following 26 pages of tables present 35 isobars, distributed as follows: 1 to 10 atm. by 1; 15 to 50 atm. by 5; 60 to 100 atm. by 10; and 120 to 340 atm. by 20. Column headings also give the units.

Margins are usually straightforward if you set ScanTailor to identify the original margins. However, a document with numerous grayscale images will require physical measurement with a ruler as ScanTailor does not recognize the original margins on a grayscale image.



Unnamed - Scan Tailor 0.9.10rc1

Output Resolution (DPI): 600

Mode: Black and White

0

Dewarping: Off

Despeckling

term "unitizing," as distinguished from the popular concept of "prepackaging," means putting an item in the simplest form that can be pricemarked. The item may be an unwrapped individual unit—for example, an avocado—or it may be normally bulk produce, such as green beans packaged in a transparent container.

The problem is this. The marking of unitized produce items with only the selling prices, determined according to weight, is not a uniformly approved method of sale. However, the cost of complying with present labeling requirements is excessive in relation to the value of the units.

For identification, I have labeled this problem "Operation Pricemark." I would like to preface my presentation with these explanatory remarks:

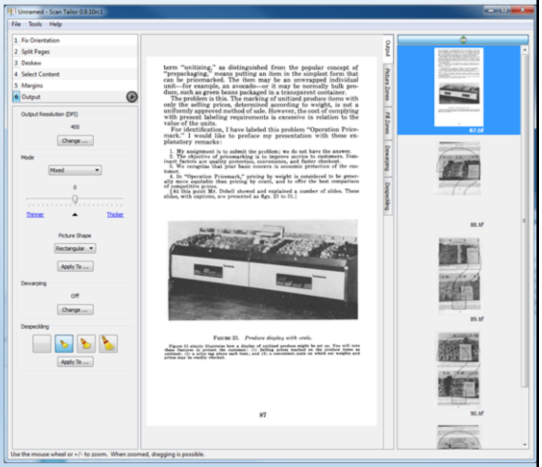
1. My assignment is to submit the problem; we do not have the answer.
2. The objective of pricemarking is to improve service to customers. Dominant factors are quality protection, convenience, and faster checkout.
3. We recognize that your basic concern is economic protection of the customer.
4. In "Operation Pricemark," pricing by weight is considered to be generally more equitable than pricing by count, and to offer the best comparison of competitive prices.

[At this point Mr. Dobell showed and explained a number of slides. These slides, with captions, are presented as figs. 21 to 31.]

FIGURE 21. *Produce display with scale.*

Figure 21 simply illustrates how a display of unitized produce might be set up. You will note these features to protect the customer: (1) Selling prices marked on the produce items as unitized; (2) a price tag above each item; and (3) a convenient scale on which net weights and prices may be readily checked.

87



Unnamed - Scan Tailor 0.9.10rc1

Output Resolution (DPI): 400

Mode: Mixed

0

Picture Shape: Rectangular

Dewarping: Off

Despeckling

term "unitizing," as distinguished from the popular concept of "prepackaging," means putting an item in the simplest form that can be pricemarked. The item may be an unwrapped individual unit—for example, an avocado—or it may be normally bulk produce, such as green beans packaged in a transparent container.

The problem is this. The marking of unitized produce items with only the selling prices, determined according to weight, is not a uniformly approved method of sale. However, the cost of complying with present labeling requirements is excessive in relation to the value of the units.

For identification, I have labeled this problem "Operation Pricemark." I would like to preface my presentation with these explanatory remarks:

1. My assignment is to submit the problem; we do not have the answer.
2. The objective of pricemarking is to improve service to customers. Dominant factors are quality protection, convenience, and faster checkout.
3. We recognize that your basic concern is economic protection of the customer.
4. In "Operation Pricemark," pricing by weight is considered to be generally more equitable than pricing by count, and to offer the best comparison of competitive prices.

[At this point Mr. Dobell showed and explained a number of slides. These slides, with captions, are presented as figs. 21 to 31.]

FIGURE 21. *Produce display with scale.*

Figure 21 simply illustrates how a display of unitized produce might be set up. You will note these features to protect the customer: (1) Selling prices marked on the produce items as unitized; (2) a price tag above each item; and (3) a convenient scale on which net weights and prices may be readily checked.

87

Use the mouse wheel or +/- to zoom. When zoomed, dragging is possible.

In step 6, these same grayscale images must be rendered into 400 DPI and a mixed color mode. Now we are ready to hit the play button and send the output back into our directory for final QC in Adobe Bridge.

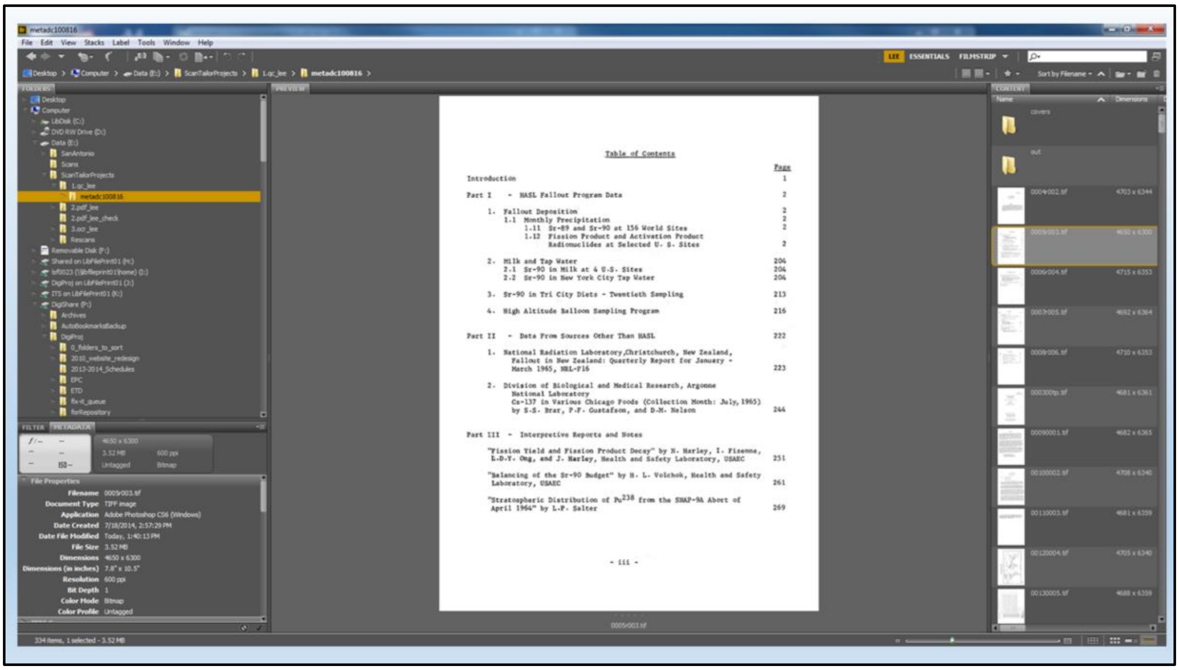


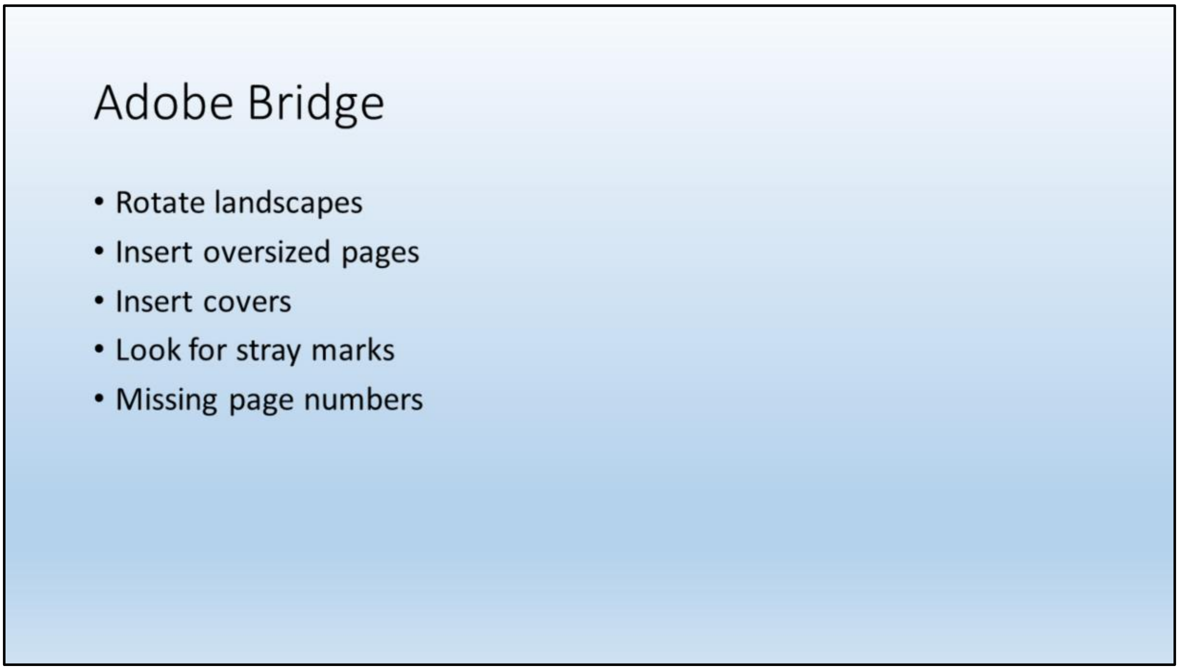
Table of Contents

- iii -

File Properties

Filename	0005r003.tif
Document Type	TIFF image
Application	Adobe Photoshop CS6 (Windows)
Date Created	7/18/2014, 2:57:29 PM
Date File Modified	Today, 1:40:13 PM
File Size	3.52 MB
Dimensions	4650 x 6300
Dimensions (in inches)	7.8" x 10.5"
Resolution	600 ppi
Bit Depth	1
Color Mode	Bitmap
Color Profile	Untagged

We use Adobe Bridge for document management after scanning and processing is complete.

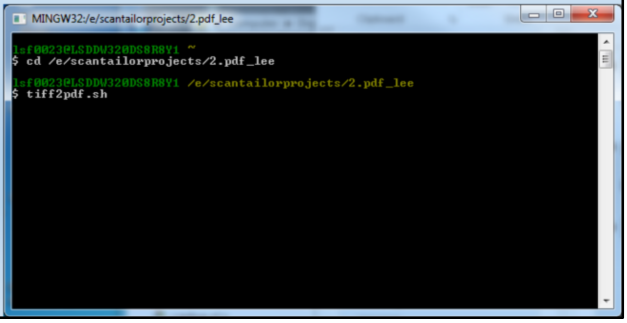


Adobe Bridge

- Rotate landscapes
- Insert oversized pages
- Insert covers
- Look for stray marks
- Missing page numbers

This is our final step before pdf creation. ScanTailor does not handle landscape orientation so at this point we rotate those pages that are supposed to be landscape. Since we want covers and oversized pages in their original size, we insert them after ScanTailor, as you will remember the template will render all other pages the same size after processing. Finally, we are looking for any oversights from the ScanTailor phase. Lee will now talk about the final QC step.

- Last step in this phase of TRAIL workflow
- Check that all pages are there, in order, and readable
- Must use win32 application tiff2pdf

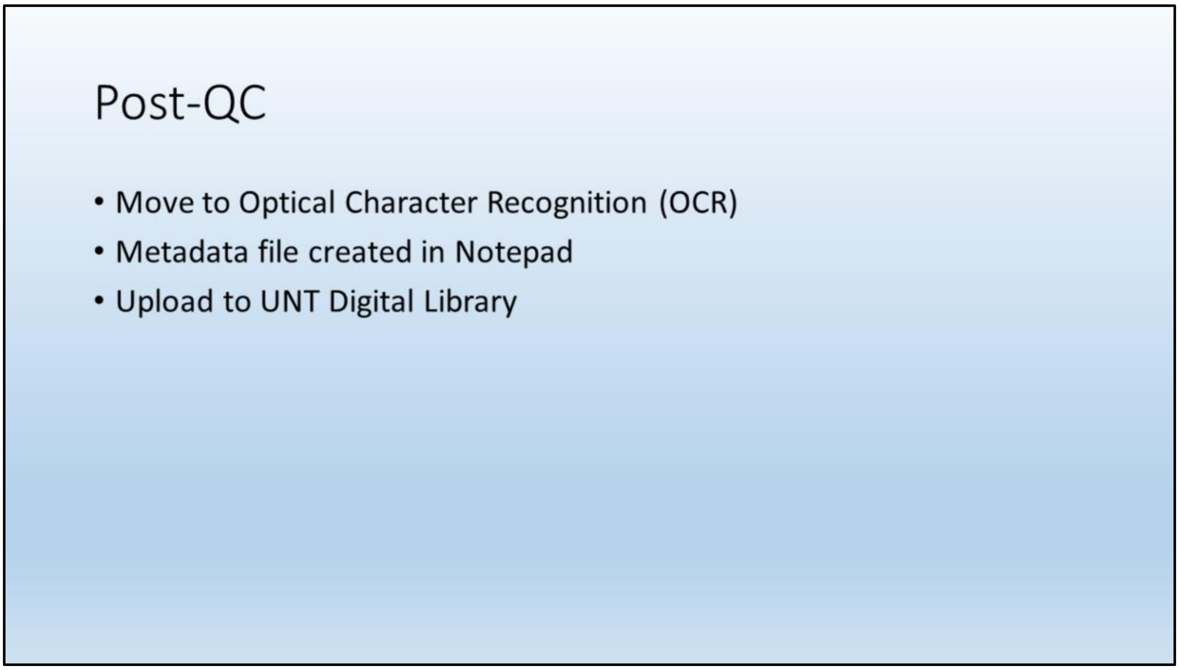


```
lsf0023@LSDDW320DS8R8Y1 ~
$ cd /e/scantailorprojects/2.pdf_lee

lsf0023@LSDDW320DS8R8Y1 /e/scantailorprojects/2.pdf_lee
$ tiff2pdf.sh
```

I look at the final pdf as a new user who has retrieved this document from an database search, or I try to forget that I have been looking at for way too long already.

A note on pdf creation. We do not create pdfs the way most people do. Since upgrading to Adobe XI, we discovered that it does not put pages in order according to our MagickNumbering system that Ashley told you about earlier. A former co-worker found an open source application online that utilizes the win32 application and can process many files sequentially, and is much faster than creating a pdf within the Acrobat program. There are still a few computers in the lab that have Acrobat X and will correctly order our documents, but you if you have multiple documents ready to be turned into pdf, you must run from computer to computer to run multiple Adobe applications so that you do not crash one computer trying to do them all. And it takes a lot of time. So what happens after we create the pdf?

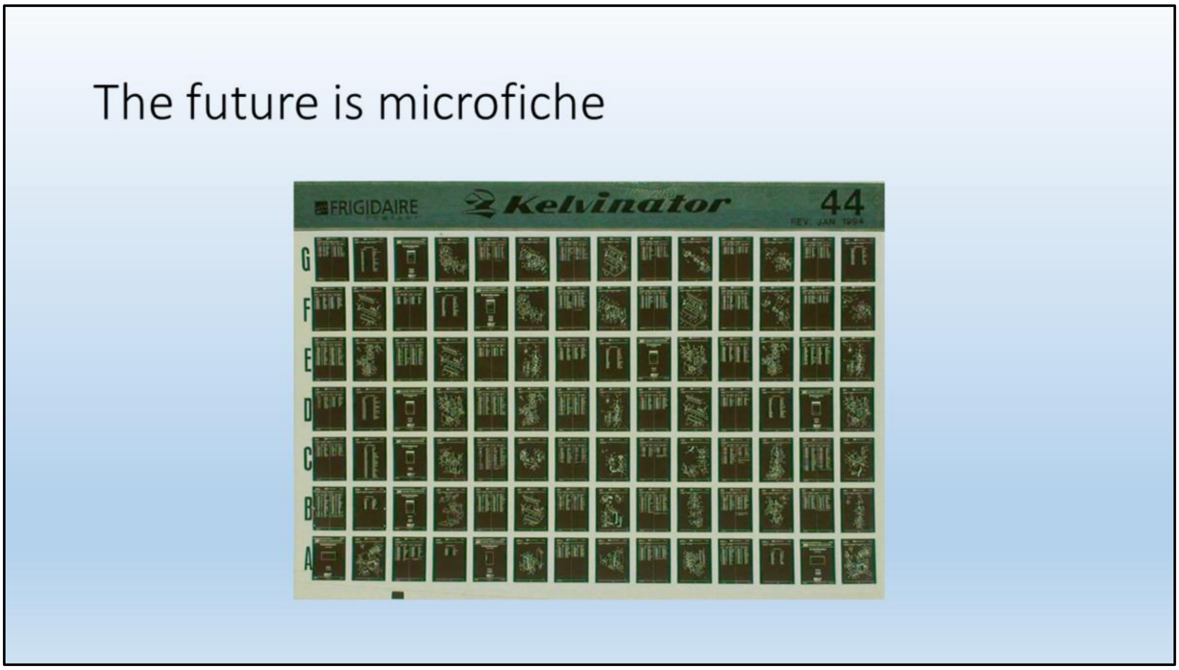


Post-QC

- Move to Optical Character Recognition (OCR)
- Metadata file created in Notepad
- Upload to UNT Digital Library

Someone recently asked me what is my favorite step in the workflow and I said I enjoyed creating the template in Photoshop, essentially the first step in the QC process. I stick with that answer, but I really like completing a document and moving it into the OCR folder on our directory. After the characters in the document have been recognized, the document is moved to the metadata phase in preparation for upload to the UNT Digital Library. We will be working on the metadata phase next month as our supply of documents to scan and process is down to about one box of 30 or so documents. Which brings us to the immediate future for TRAIL.

What does the future hold for TRAIL? I think I am probably the first person in a long time to say this, but…the future is…



The future is microfiche

…microfiche, at least for UNT’s portion of TRAIL. We received a box from Arizona on Monday containing 6 smaller boxes containing over 1000 separate resources on microfiche. This was not a surprise, I was ready for this, having already been informed this is the direction we are taking right now. Some consist of one card, some are several cards, dating from 1975 to 1996, which will be by far the most recent documents that we have worked with. These shall be digitized by an outside contractor and returned to us for Quality Control. I am interested to see what resulting adjustments will occur to our workflow. Of course the quality of the original image will determine if this results in a faster workflow due to the absence of scanning. It will be a major adjustment not being able to immediately rescan an image.

Overall, I can say with a fair amount of confidence, that we have barely delved into the supply of old government documents that need wider dissemination.

Thank you for your time and we hope you enjoyed our presentation.