THE PORTAL TO TEXAS HISTORY
texashistory.unt.edu

Expanding the Scale of PDF Preservation to Accommodate a State Press Association through the Texas Digital Newspaper Program

TEXAS DIGITAL NEWSPAPER PROGRAM

UNIVERSITY LIBRARIES UNT

Ana Krahmer, Digital Newspaper Program Coordinator

ana.krahmer@unt.edu

Overview

- What is TDNP?
- Initial PDF Newspaper Projects
- The Texas Press Association Archive
- Technology and Standards

What is TDNP?

About us

What is TDNP?

- Dedicated to preserving Texas newspapers, from any time or place, for any title.
- Thus far, we host nearly 3 million pages of newspapers, dating from 1829 to present.

TJP

TEXAS JEWISH POST ✡ SINCE 1947

INSIDE

Subscribe today!
subscriptions@texasjewishpost.com

Israel at 62, P.2-3
Yom HaZikaron, Yom HaAtzmaut in the Metroplex

Quick Look

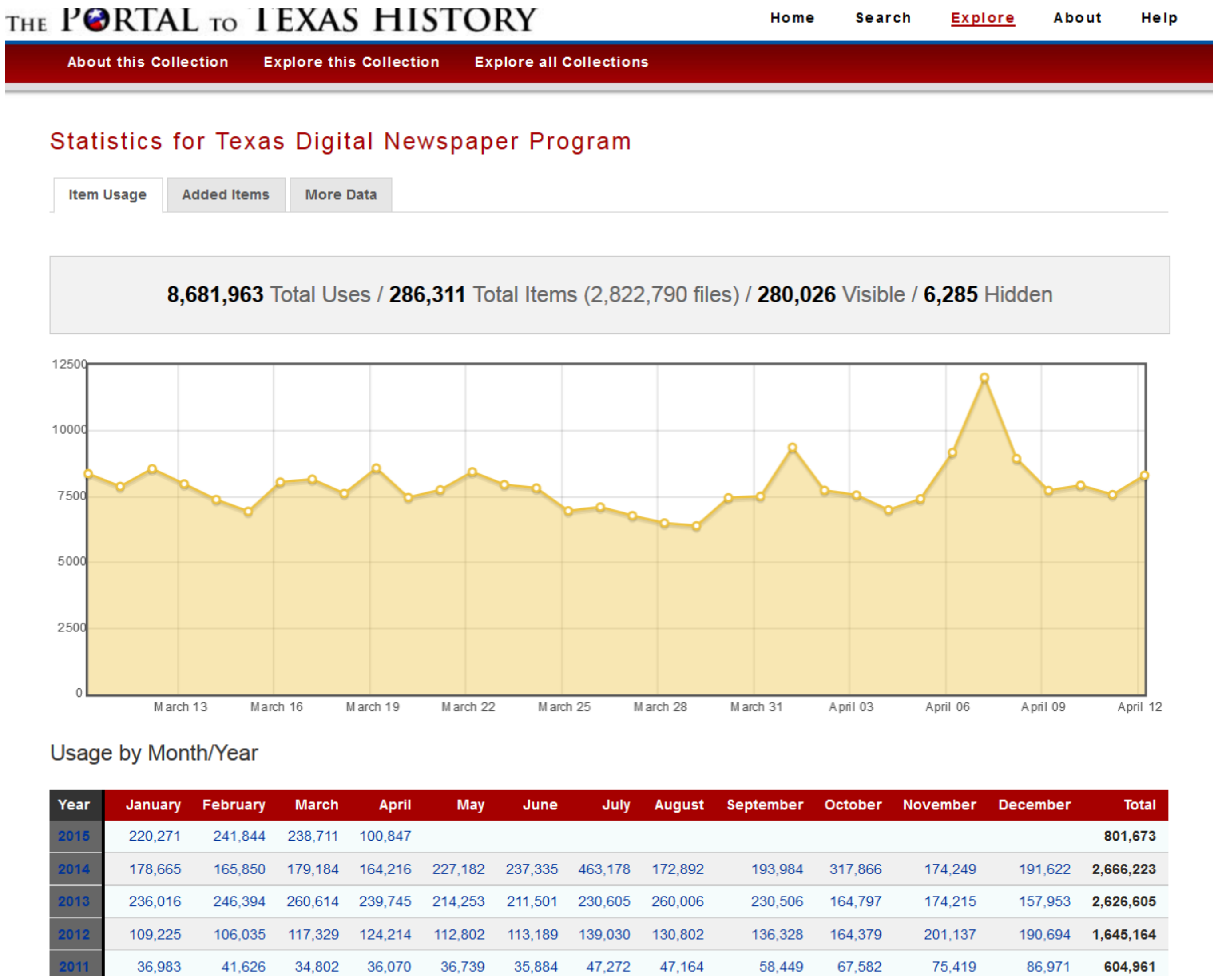
4	News Analysis
5	Opinion
6	Dallas Doings
10	Seniors
19	Simchas
20	D'var Torah
21	Services
22	Calendar
23	Lifelong Learning
24	Around The Town With Rene
26	Columnists
27	Jewish Foodie

Thursday, April 15, 2010 • 1 Iyar 5770 • Vol. 64 No. 15 • tjpnews.com $1.00

One Night,
One Community,
One Cause event:
'Death of a Salesman,'
May 13

Performance will benefit Akiba, Levine and Yavneh academies as well as the Youtheatre program of the JCC P.14-15

JEWISH FOODIE: CHICKEN SALAD, P.27
Sweet or savory, it's perfect for warm-weather entertaining

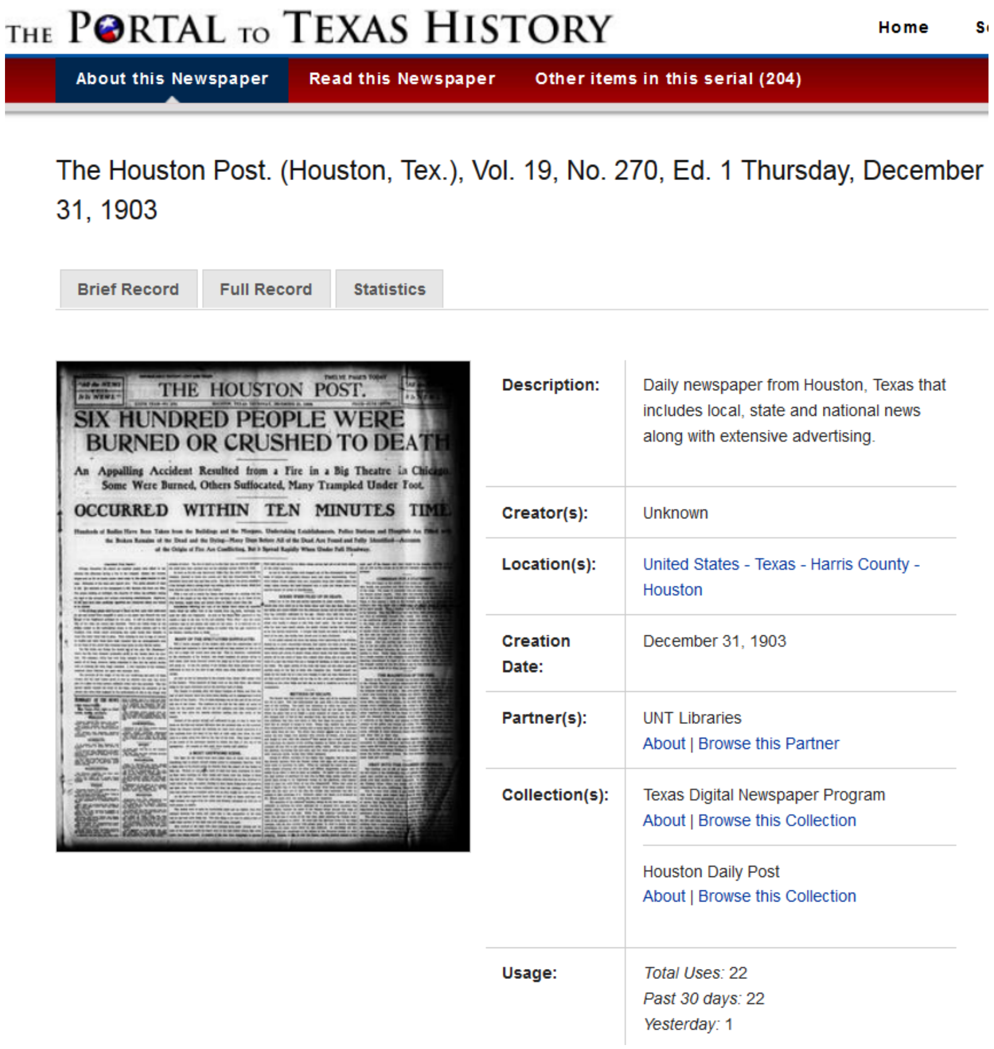


Statistics for Texas Digital Newspaper Program

8,681,963 Total Uses / **286,311** Total Items (2,822,790 files) / **280,026** Visible / **6,285** Hidden

Usage by Month/Year

Year	January	February	March	April	May	June	July	August	September	October	November	December	Total
2015	220,271	241,844	238,711	100,847									**801,673**
2014	178,665	165,850	179,184	164,216	227,182	237,335	463,178	172,892	193,984	317,866	174,249	191,622	**2,666,223**
2013	236,016	246,394	260,614	239,745	214,253	211,501	230,605	260,006	230,506	164,797	174,215	157,953	**2,626,605**
2012	109,225	106,035	117,329	124,214	112,802	113,189	139,030	130,802	136,328	164,379	201,137	190,694	**1,645,164**
2011	36,983	41,626	34,802	36,070	36,739	35,884	47,272	47,164	58,449	67,582	75,419	86,971	**604,961**


THE PORTAL TO TEXAS HISTORY
Home
About this Newspaper
Read this Newspaper
Other items in this serial (204)
Brief Record
Full Record
Statistics
THE HOUSTON POST.
SIX HUNDRED PEOPLE WERE BURNED OR CRUSHED TO DEATH
OCCURRED WITHIN TEN MINUTES TIME

The Houston Post. (Houston, Tex.), Vol. 19, No. 270, Ed. 1 Thursday, December 31, 1903

Description:	Daily newspaper from Houston, Texas that includes local, state and national news along with extensive advertising.
Creator(s):	Unknown
Location(s):	United States - Texas - Harris County - Houston
Creation Date:	December 31, 1903
Partner(s):	UNT Libraries About | Browse this Partner
Collection(s):	Texas Digital Newspaper Program About | Browse this Collection Houston Daily Post About | Browse this Collection
Usage:	*Total Uses:* 22 *Past 30 days:* 22 *Yesterday:* 1

ana.krahmer@unt.edu

What is TDNP?

- Dedicated to preserving Texas newspapers, from any time or place, for any title.
- Thus far, we host nearly 3 million pages of newspapers, dating from 1829 to present.

TJP

TEXAS JEWISH POST ✡ SINCE 1947

INSIDE

Israel at 62, P.2-3
Yom HaZikaron, Yom HaAtzmaut in the Metroplex

Subscribe today!
subscriptions@texasjewishpost.com

Quick Look

4 News Analysis
5 Opinion
6 Dallas Doings
10 Seniors
19 Simchas
20 D'var Torah
21 Services
22 Calendar
23 Lifelong Learning
24 Around The Town With Rene
26 Columnists
27 Jewish Foodie

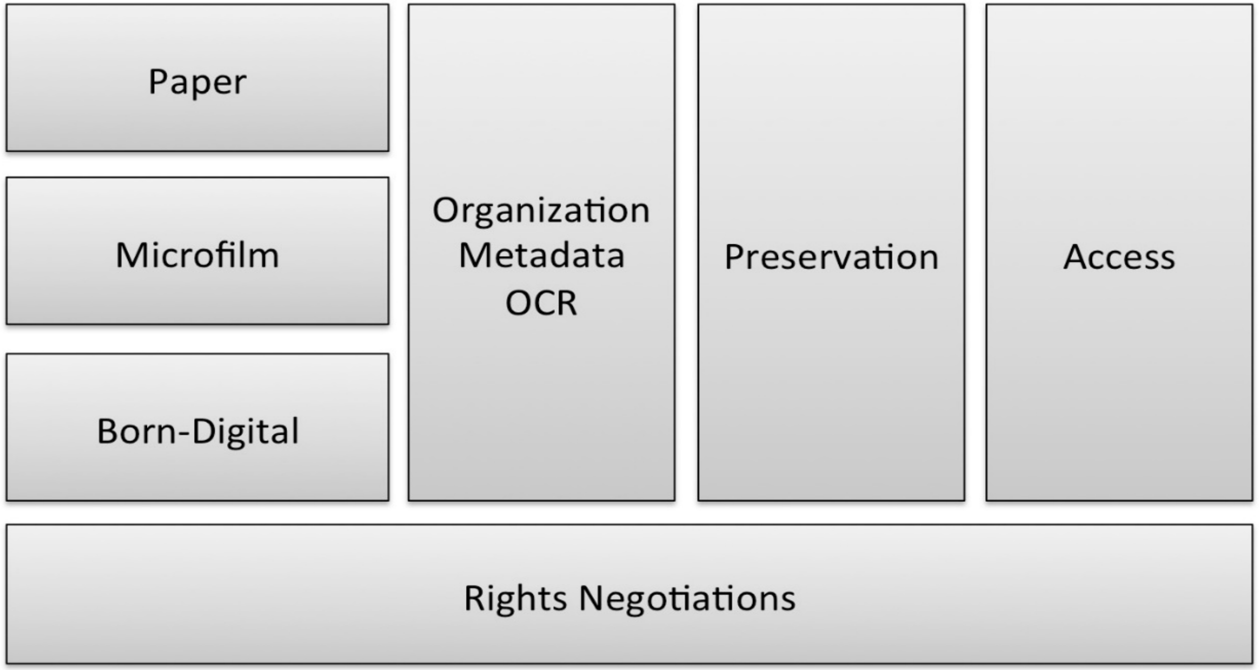
Thursday, April 15, 2010 • 1 Iyar 5770 • Vol. 64 No. 15 • tjpnews.com $1.00

One Night,
One Community,
One Cause event:
'Death of a Salesman,'
May 13

Performance will benefit Akiba, Levine and Yavneh academies as well as the Youtheatre program of the JCC P.14-15

JEWISH FOODIE: CHICKEN SALAD, P.27
Sweet or savory, it's perfect for warm-weather entertaining

Workflow



Paper

Microfilm

Born-Digital

Organization
Metadata
OCR

Preservation

Access

Rights Negotiations

Initial PDF Newspaper Projects

Preservation on The Portal to Texas History

Initial PDF Newspaper Projects

- Began working with PDFs in 2010.
- Preserved PDF issues after receiving grants to digitize earlier, analog issues.
- Earliest PDF issue is from 18 March 1998 (University of Dallas).
- PDFs were acquired from publishers.
- Permissions granted by publishers.

Initial PDF Newspaper Projects

JOSIAH MONROE, BEARS HEADED FOR PRIME TIME • SPORTS, B1

The Bastrop Advertiser

www.bastropadvertiser.com

Texas' Oldest Weekly Newspaper Since March 1, 1853

Semi-Weekly Since Sept. 5, 1977

Volume 157, Number 21

THURSDAY, MAY 13, 2010

50¢

INSIDE

BRANCHING OUT

MURKY WATERS

WATCH BOTH WAYS

POLICE BLOTTER

WEATHER

INDEX

We Recycle

ELECTION RESULTS

Shakeup in water district

Tahitian Village residents speak loudly with votes for change

Auction helpers

COMMUNITY

Relay For Life event celebrates survivors

CRIME

Suspect in child porn case returned to Texas

COURTS

Capital murder trial moved to Bastrop County

INSIDE THIS EDITION

Attorney General issues opinion in water district dispute

STORY ON PAGE 9

Canadian Wildcats rout River Road in season opener

STORY ON PAGES 12-13

IN DISTRICT COURT NEWS:

Jury delivers guilty charge in child sexual assault trial

STORY ON PAGE 7

$1

The Canadian RECORD

VOLUME 120 NUMBER 35

THURSDAY 02 SEPTEMBER 2010

COMING HOME: A Vietnam Story

Sweetwater Reporter

DEDICATED TO PROUDLY DELIVERING LOCAL NEWS SINCE 1881

Sunday, January 3, 2010

www.sweetwaterreporter.com

Vol. 112, Issue 041 • $1

What's Inside

Donnie Rannefeld announces intent to file for Nolan County Judge

Students participate in Angel Tree project

Donnie Rannefeld

Nominations sought for Big Country Athletic Hall of Fame

Life

FFB wishes public a Happy New Year

Viewpoints

The Grinch Who Stole New Year's

Inside

Masons donate to WTCAC

NOTICE

Youth to perform theatrical production at MLK Banquet

Many attend RPMH annual Christmas party

Initial PDF Newspaper Projects

- Flexibility of Portal to embargo gives publishers confidence.
- Example: *Cherokeean Herald*
- When publishers see one successfully-preserved title, they show interest.

thecherokeean.com

Cherokeean HERALD

CELEBRATING 160 YEARS

Texas' Oldest Weekly Newspaper

Vol. 161 ■ No. 18 ■ 16 pgs.

WEDNESDAY
June 23, 2010
Rusk, Texas ■ 50 cents

DAVE CAMPBELL'S TEXAS FOOTBALL

Making the cut in 2010?

Trio of Cherokee County teams look to pile up wins this football season

Photo composite:
Alto: LeAnn Jones
Rusk: Lynn Littlejohn
Jacksonville: Mike Cummins

From left, Alto quarterback Javante Jones, Rusk quarterback Trey Thrash and Jacksonville quarterback Ryan Black look to lead their teams to playoff berths in their respective districts. Alto is ranked seventh in the preseason state 1A Division I poll and is touted as a possible state semi-finalist. Jacksonville is picked to finished fourth in 14-4A and reach the playoffs for the third consecutive year. Rusk is picked to finish fourth in 17-3A.

Alto selected #7 in 1A Division I; J'ville picked to finish 4th in 14-4A, Rusk picked 4th in 17-3A

Rusk Eagles

See TEXAS FOOTBALL, pg. 6A

Warden accepts promotion

Todd Foxworth named warden at Michael Unit near Palestine

Warden Todd Foxworth

See WARDEN FOXWORTH, pg. 6A

County chooses Tyler architects

County commissioners accept Sinclair & Wright Architects

See ARCHITECTS, pg. 6A

CITY OF ALTO

Council, DPS compromise on 18-wheeler weigh-ins

See ALTO COUNCIL, pg. 6A

AISD board discusses preliminary TAKS scores

ALTO GRADUATE GETS 'STUCK AT PROM' WITH SCHOLARSHIP CONTEST pg. 8A

INSIDE ■ LETTERS TO THE EDITOR 2A ■ VIEWPOINTS 3A ■ DEATHS 4A ■ CLASSIFIEDS 5-8B

SCHOOLS ■ WORDS OF WISDOM FROM RUSK'S VALEDICTORIAN AND SALUTATORIAN pg. 4B

WEATHER

30% chance of rain
THURSDAY
87 high – 74 low

Texas Press Association Archive

Preserving Recent Texas History

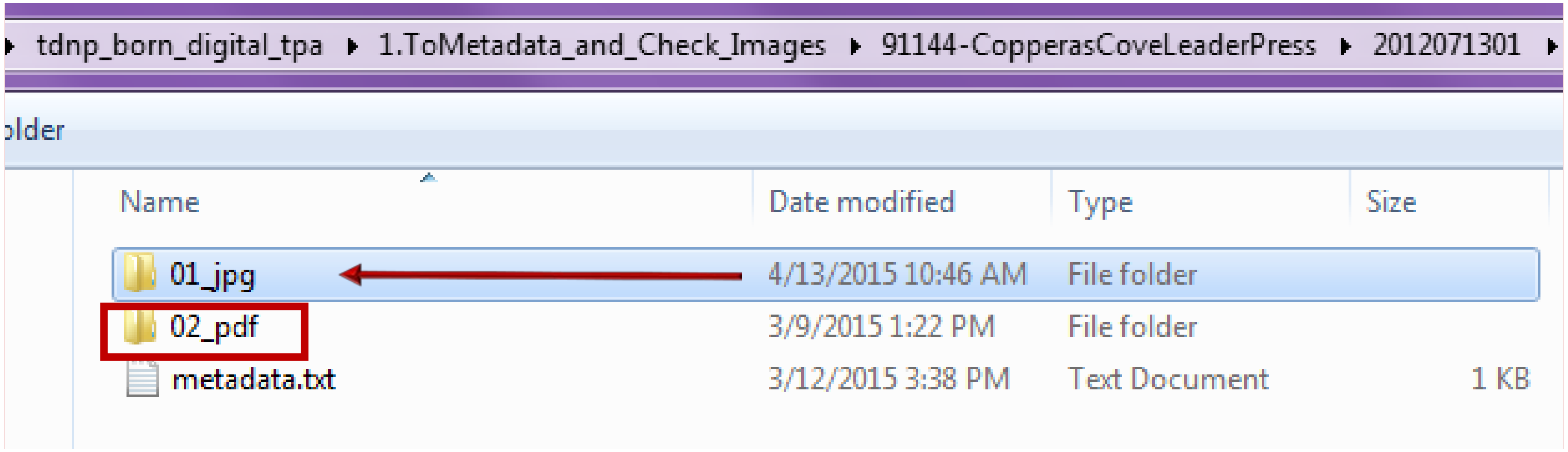
ana.krahmer@unt.edu

TPA Archive Partnership

- Collaboration with the Texas Press Association and NewzGroup out of Missouri.
- 12TB of PDF newspapers, prepared for preservation through batch processing of PDF content.
- Range from 2010-August 2014.
- Scheduled transfer of 2014-2015 newspapers from NewzGroup.

TPA Archive Partnership: File Workflow

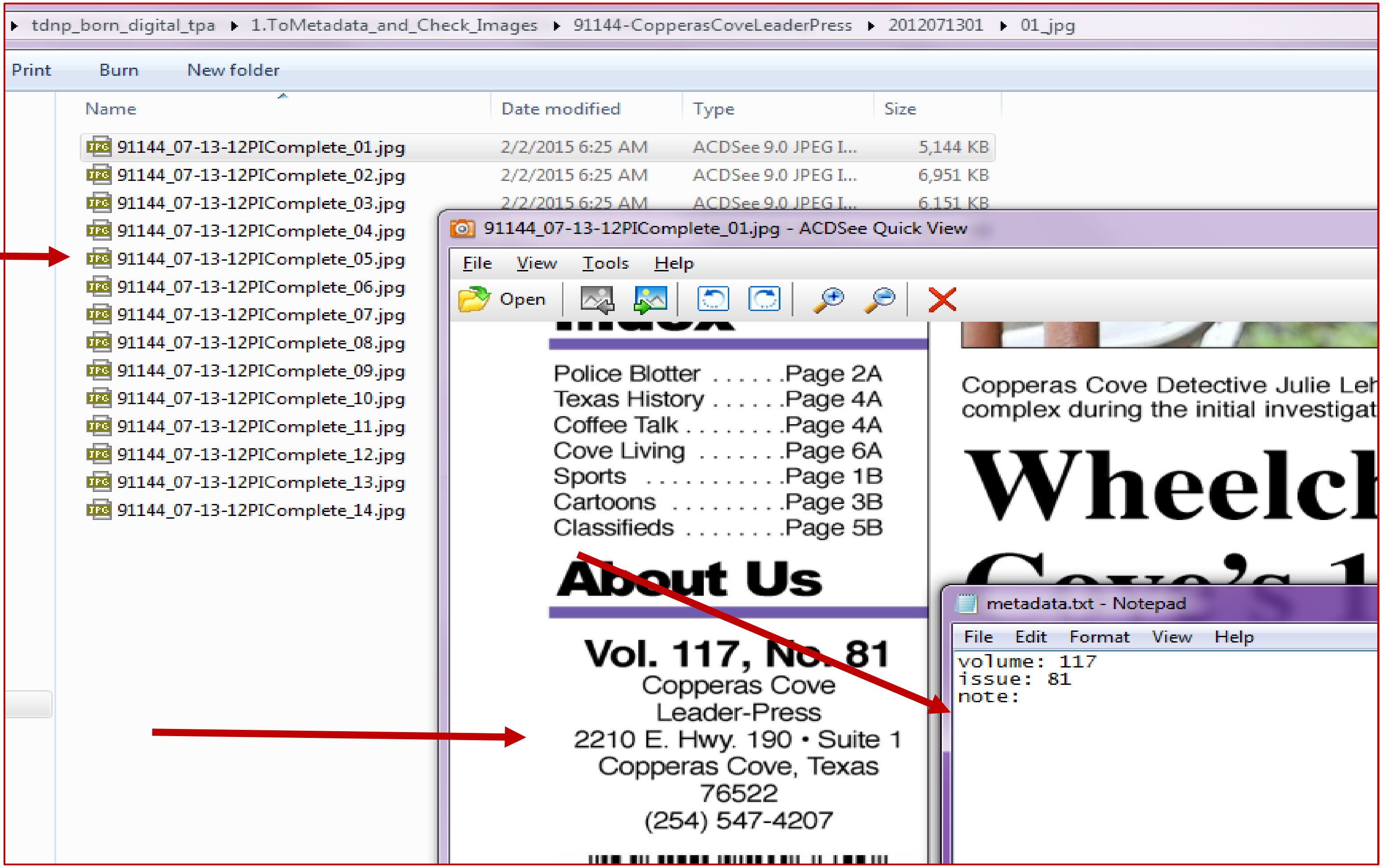
- QC work performed after batch processing.
- Add initial layer of metadata, pre-OCR.



tdnp_born_digital_tpa ▸ 1.ToMetadata_and_Check_Images ▸ 91144-CopperasCoveLeaderPress ▸ 2012071301 ▸

Name	Date modified	Type	Size
01_jpg	4/13/2015 10:46 AM	File folder	
02_pdf	3/9/2015 1:22 PM	File folder	
metadata.txt	3/12/2015 3:38 PM	Text Document	1 KB

TPA Archive Partnership: File Workflow


tdnp_born_digital_tpa ▸ 1.ToMetadata_and_Check_Images ▸ 91144-CopperasCoveLeaderPress ▸ 2012071301 ▸ 01_jpg
Print
Burn
New folder
91144_07-13-12PIComplete_01.jpg - ACDSee Quick View
File
View
Tools
Help
Open
metadata.txt - Notepad
File
Edit
Format
View
Help

Name	Date modified	Type	Size
91144_07-13-12PIComplete_01.jpg	2/2/2015 6:25 AM	ACDSee 9.0 JPEG I...	5,144 KB
91144_07-13-12PIComplete_02.jpg	2/2/2015 6:25 AM	ACDSee 9.0 JPEG I...	6,951 KB
91144_07-13-12PIComplete_03.jpg	2/2/2015 6:25 AM	ACDSee 9.0 JPEG I...	6,151 KB
91144_07-13-12PIComplete_04.jpg			
91144_07-13-12PIComplete_05.jpg			
91144_07-13-12PIComplete_06.jpg			
91144_07-13-12PIComplete_07.jpg			
91144_07-13-12PIComplete_08.jpg			
91144_07-13-12PIComplete_09.jpg			
91144_07-13-12PIComplete_10.jpg			
91144_07-13-12PIComplete_11.jpg			
91144_07-13-12PIComplete_12.jpg			
91144_07-13-12PIComplete_13.jpg			
91144_07-13-12PIComplete_14.jpg			

Police Blotter . . . . . .Page 2A
Texas History . . . . . .Page 4A
Coffee Talk . . . . .Page 4A
Cove Living . . . . . . .Page 6A
Sports . . . . .Page 1B
Cartoons . . . . .Page 3B
Classifieds . . . . .Page 5B

About Us

Vol. 117, No. 81
Copperas Cove
Leader-Press
2210 E. Hwy. 190 • Suite 1
Copperas Cove, Texas
76522
(254) 547-4207

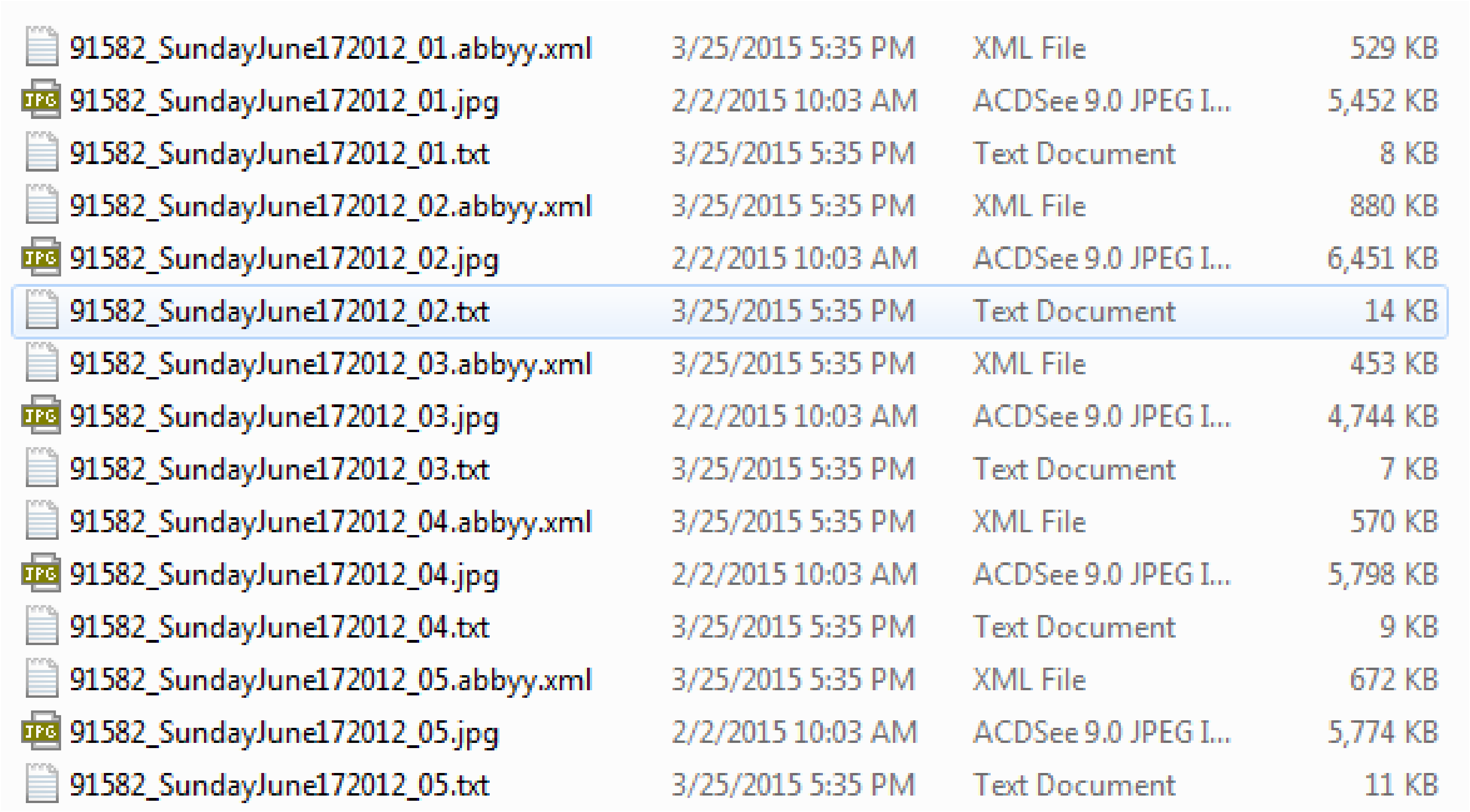
Copperas Cove Detective Julie Leh
complex during the initial investigat

Wheelch

Cove's 1

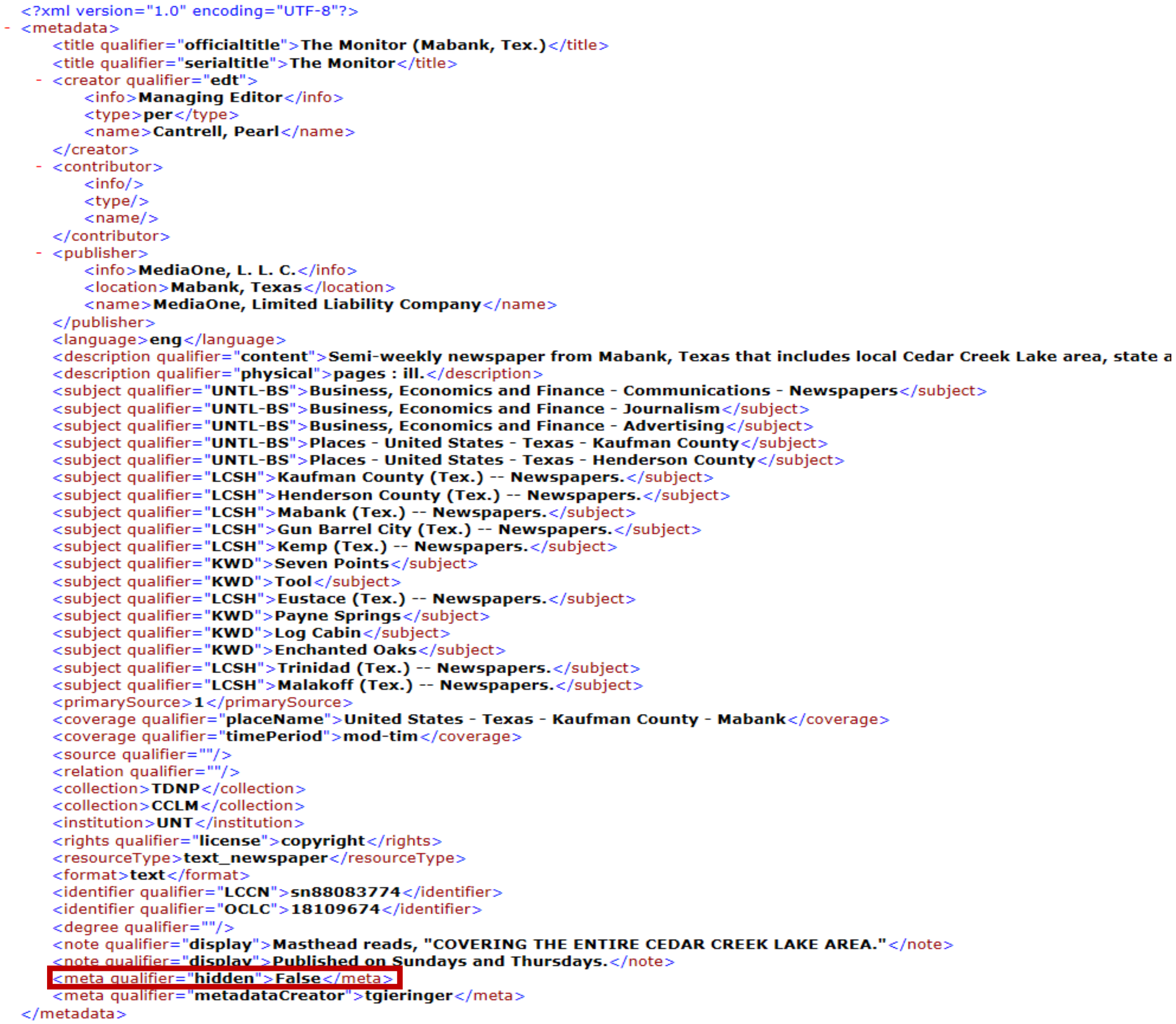
```
volume: 117
issue: 81
note:
```

TPA Archive Partnership: File Workflow



91582_SundayJune172012_01.abbyy.xml	3/25/2015 5:35 PM	XML File	529 KB
91582_SundayJune172012_01.jpg	2/2/2015 10:03 AM	ACDSee 9.0 JPEG I...	5,452 KB
91582_SundayJune172012_01.txt	3/25/2015 5:35 PM	Text Document	8 KB
91582_SundayJune172012_02.abbyy.xml	3/25/2015 5:35 PM	XML File	880 KB
91582_SundayJune172012_02.jpg	2/2/2015 10:03 AM	ACDSee 9.0 JPEG I...	6,451 KB
91582_SundayJune172012_02.txt	3/25/2015 5:35 PM	Text Document	14 KB
91582_SundayJune172012_03.abbyy.xml	3/25/2015 5:35 PM	XML File	453 KB
91582_SundayJune172012_03.jpg	2/2/2015 10:03 AM	ACDSee 9.0 JPEG I...	4,744 KB
91582_SundayJune172012_03.txt	3/25/2015 5:35 PM	Text Document	7 KB
91582_SundayJune172012_04.abbyy.xml	3/25/2015 5:35 PM	XML File	570 KB
91582_SundayJune172012_04.jpg	2/2/2015 10:03 AM	ACDSee 9.0 JPEG I...	5,798 KB
91582_SundayJune172012_04.txt	3/25/2015 5:35 PM	Text Document	9 KB
91582_SundayJune172012_05.abbyy.xml	3/25/2015 5:35 PM	XML File	672 KB
91582_SundayJune172012_05.jpg	2/2/2015 10:03 AM	ACDSee 9.0 JPEG I...	5,774 KB
91582_SundayJune172012_05.txt	3/25/2015 5:35 PM	Text Document	11 KB

TPA Archive Partnership: File Workflow



```
  <?xml version="1.0" encoding="UTF-8"?>
- <metadata>
    <title qualifier="officialtitle">The Monitor (Mabank, Tex.)</title>
    <title qualifier="serialtitle">The Monitor</title>
  - <creator qualifier="edt">
        <info>Managing Editor</info>
        <type>per</type>
        <name>Cantrell, Pearl</name>
    </creator>
  - <contributor>
        <info/>
        <type/>
        <name/>
    </contributor>
  - <publisher>
        <info>MediaOne, L. L. C.</info>
        <location>Mabank, Texas</location>
        <name>MediaOne, Limited Liability Company</name>
    </publisher>
    <language>eng</language>
    <description qualifier="content">Semi-weekly newspaper from Mabank, Texas that includes local Cedar Creek Lake area, state a
    <description qualifier="physical">pages : ill.</description>
    <subject qualifier="UNTL-BS">Business, Economics and Finance - Communications - Newspapers</subject>
    <subject qualifier="UNTL-BS">Business, Economics and Finance - Journalism</subject>
    <subject qualifier="UNTL-BS">Business, Economics and Finance - Advertising</subject>
    <subject qualifier="UNTL-BS">Places - United States - Texas - Kaufman County</subject>
    <subject qualifier="UNTL-BS">Places - United States - Texas - Henderson County</subject>
    <subject qualifier="LCSH">Kaufman County (Tex.) -- Newspapers.</subject>
    <subject qualifier="LCSH">Henderson County (Tex.) -- Newspapers.</subject>
    <subject qualifier="LCSH">Mabank (Tex.) -- Newspapers.</subject>
    <subject qualifier="LCSH">Gun Barrel City (Tex.) -- Newspapers.</subject>
    <subject qualifier="LCSH">Kemp (Tex.) -- Newspapers.</subject>
    <subject qualifier="KWD">Seven Points</subject>
    <subject qualifier="KWD">Tool</subject>
    <subject qualifier="LCSH">Eustace (Tex.) -- Newspapers.</subject>
    <subject qualifier="KWD">Payne Springs</subject>
    <subject qualifier="KWD">Log Cabin</subject>
    <subject qualifier="KWD">Enchanted Oaks</subject>
    <subject qualifier="LCSH">Trinidad (Tex.) -- Newspapers.</subject>
    <subject qualifier="LCSH">Malakoff (Tex.) -- Newspapers.</subject>
    <primarySource>1</primarySource>
    <coverage qualifier="placeName">United States - Texas - Kaufman County - Mabank</coverage>
    <coverage qualifier="timePeriod">mod-tim</coverage>
    <source qualifier=""/>
    <relation qualifier=""/>
    <collection>TDNP</collection>
    <collection>CCLM</collection>
    <institution>UNT</institution>
    <rights qualifier="license">copyright</rights>
    <resourceType>text_newspaper</resourceType>
    <format>text</format>
    <identifier qualifier="LCCN">sn88083774</identifier>
    <identifier qualifier="OCLC">18109674</identifier>
    <degree qualifier=""/>
    <note qualifier="display">Masthead reads, "COVERING THE ENTIRE CEDAR CREEK LAKE AREA."</note>
    <note qualifier="display">Published on Sundays and Thursdays.</note>
    <meta qualifier="hidden">False</meta>
    <meta qualifier="metadataCreator">tgieringer</meta>
  </metadata>
```

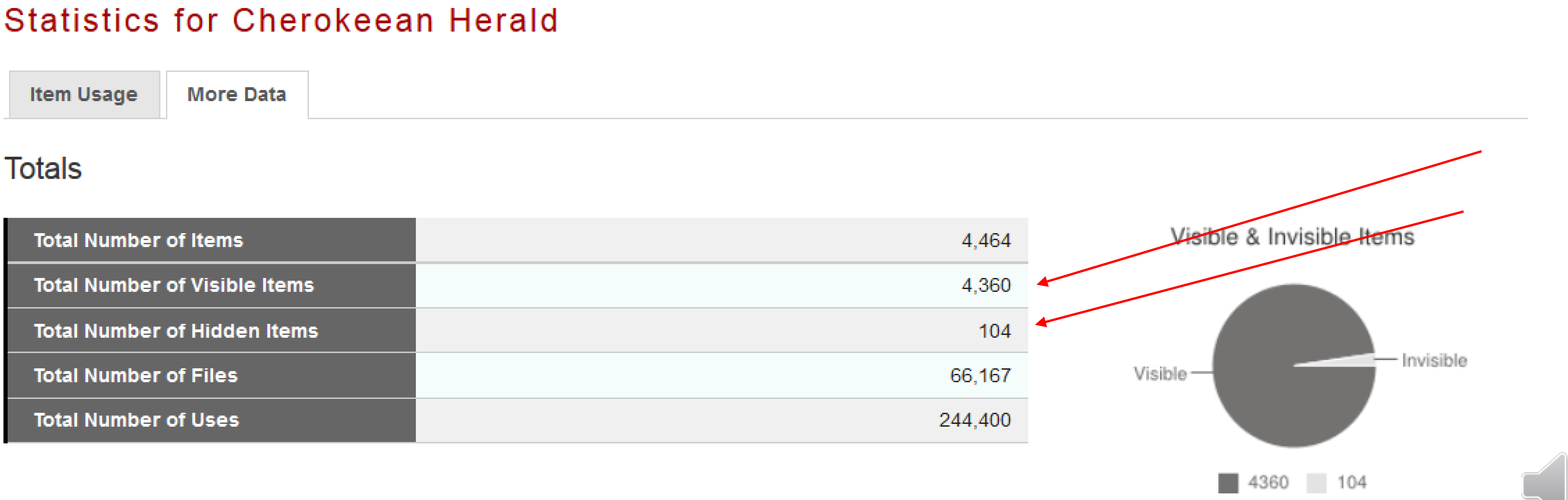
- Example of batch metadata for PDF issues
- Applied in XML file to sets divided by year and/or by content changes (managing editor, publisher, masthead, etc.)
- Newspapers uploaded with embargo are marked as “hidden”=TRUE.

TPA Archive Partnership: Permissions

- Publishers are busy, hard people to catch.
- When they respond, they respond with interest, with the exception of large-city dailies.
- Embargos have ranged from the most recent 6 months to 3 years.
- The Texas Digital Newspaper Program holds a membership in the Texas Press Association.
- Krahmer attends all TPA annual conventions and summer leadership meetings, along with as many regional Press Association meetings as possible.
- Preservation is about establishing communication, trust.

Collaboration with Publishers

- We have the capability to open or hide issues at publisher’s request.
- We can unhide issues when the embargo period expires.

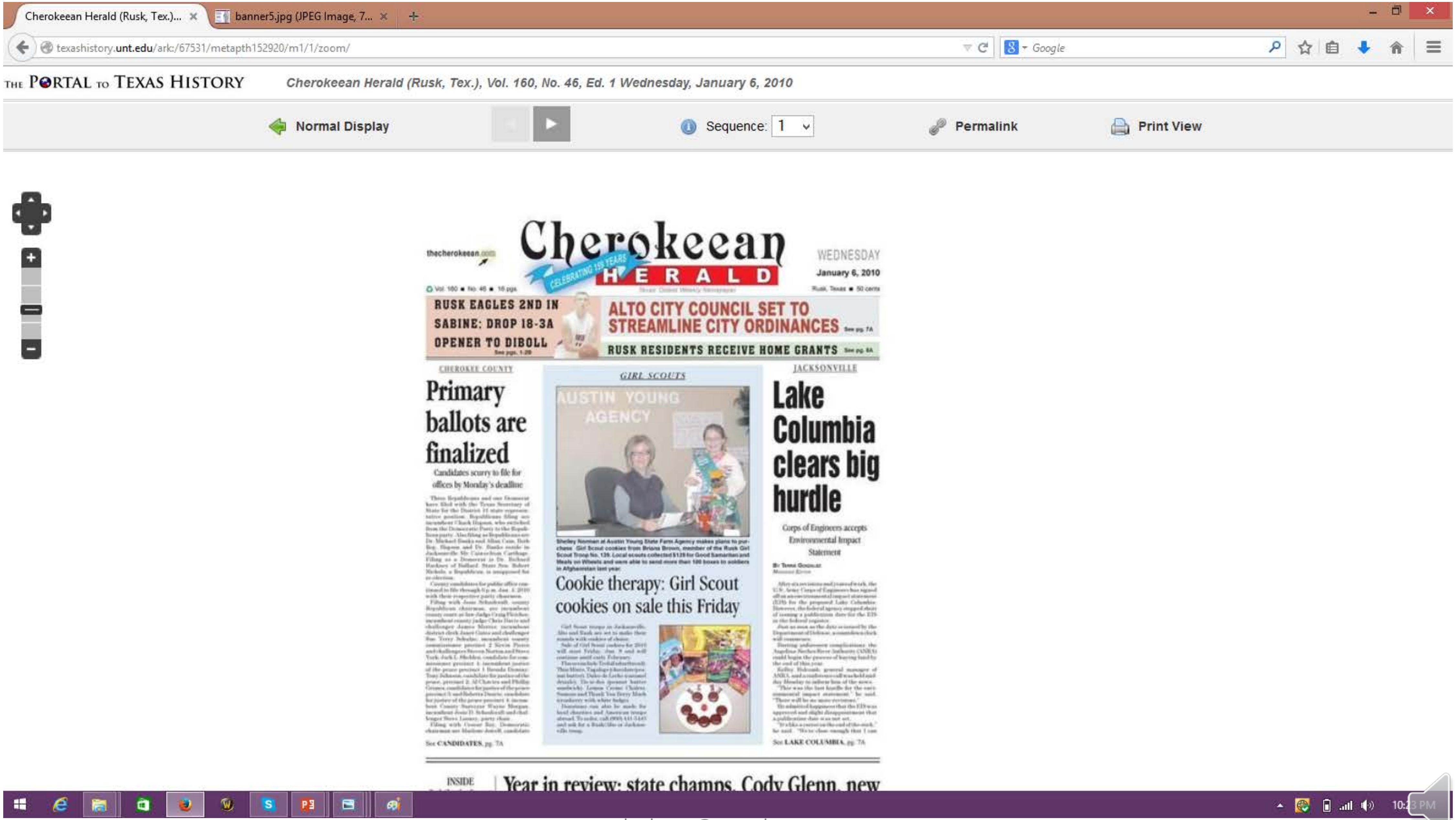


Statistics for Cherokeean Herald

Item Usage | More Data

Totals

Total Number of Items	4,464
Total Number of Visible Items	4,360
Total Number of Hidden Items	104
Total Number of Files	66,167
Total Number of Uses	244,400



Cherokeean Herald (Rusk, Tex.)... ✕ | banner5.jpg (JPEG Image, 7... ✕ | +

texashistory.unt.edu/ark:/67531/metapth152920/m1/1/zoom/

THE PORTAL TO TEXAS HISTORY   *Cherokeean Herald (Rusk, Tex.), Vol. 160, No. 46, Ed. 1 Wednesday, January 6, 2010*

thecherokeean.com

Cherokeean HERALD

CELEBRATING 155 YEARS

WEDNESDAY
January 6, 2010

Vol. 160 ■ No. 46 ■ 16 pgs
Rusk, Texas ■ 50 cents

RUSK EAGLES 2ND IN SABINE: DROP 18-3A OPENER TO DIBOLL See pgs. 1-2B

ALTO CITY COUNCIL SET TO STREAMLINE CITY ORDINANCES See pg. 7A

RUSK RESIDENTS RECEIVE HOME GRANTS See pg. 6A

CHEROKEE COUNTY

Primary ballots are finalized

Candidates scurry to file for offices by Monday's deadline

[illegible]

See CANDIDATES, pg. 7A

GIRL SCOUTS

Shelley Norman at Austin Young State Farm Agency makes plans to purchase Girl Scout cookies from Briana Brown, member of the Rusk Girl Scout Troop No. 139. Local scouts collected $129 for Good Samaritan and Meals on Wheels and were able to send more than 100 boxes to soldiers in Afghanistan last year.

Cookie therapy: Girl Scout cookies on sale this Friday

[illegible]

JACKSONVILLE

Lake Columbia clears big hurdle

Corps of Engineers accepts Environmental Impact Statement

[illegible]

See LAKE COLUMBIA, pg. 7A

INSIDE

Year in review: state champs, Cody Glenn, new

Technology and Standards

File types, software, and metadata

ana.krahmer@unt.edu

Filetypes

- The PDF print master is the preservation copy.
- Save this into JPG format at 400 dpi, from which derivatives are created.

Software

- Adobe Acrobat, batch-scripting capabilities
- Batch renaming application
- Python scripts
- Microservices

A file is a file is a file is a. . .

- After conversion, whether from analog to digital, or from one born-digital filetype to another, processing is very similar.
- Regardless of original format, issues are named according to yyyymmdded: 1901052301= the issue for May 23, 1901, edition 01, of a newspaper.
- Metadata, with minor differences (“physical description” & bagit information), is the same for all newspapers.

Metadata

- Minor bag-info differences: bag-info files (BagIt) for pdfs contain the following information (red text is unique to PDF materials).

  Source-Organization: University of North Texas Libraries
  Organization-Address: P. O. Box 305190, Denton, TX 76203-5190
  Contact-Name: Mark Phillips
  Contact-Phone: 940-565-2415
  Contact-Email: mark.phillips@unt.edu
  External-Description: Newspaper issues of the “NEWSPAPER NAME HERE" published in
  [CITY], Texas. Issues were made available from born-digital
  PDF printmasters. Partner institution is the [partner library here]. Master files were PDF printmasters from which derivative JPGs were created.

Questions?

Email: Ana.Krahmer@unt.edu
Call: 940-565-3367
Visit: http://texashistory.unt.edu/explore/collections/TDNP/