LIBRARY LIBRARY OF CONGRESS

November 13, 2023

Twenty-Year-Old OCR Gets A Makeover:

New OCR Pipeline for Chronicling America

National Digital Newspaper Program (NDNP)

Robin Pike
NDNP Coordinator, Library of Congress
rpike@loc.gov

Nathan Yarasavage
NDNP Production Lead, Library of Congress
nyarasavage@loc.gov

LIBRARY OF CONGRESS
SERIAL & GOVERNMENT
PUBLICATIONS DIVISION

National Digital Newspaper Program (NDNP)

Intro to NDNP

What is NDNP?

- Partnership
  - National Endowment for the Humanities (NEH)
  - Library of Congress (LC)
  - State partner organizations
- Develop searchable database of U.S. newspapers
- Funded by the NEH NDNP awards
- Permanently maintained at LC

Information? Ask Star Newsies

CHARLES CHADBOURNE, Star newsboy at the postoffice corner, Third ave. and Union st., is giving Gardner I'Laug the "lowdown" on how to find the city library. Every Star newsie has been appointed a Junior Tourist Guide—note the button on Charley's cap—by the Junior Know Seattle bureau of the Chamber of Commerce.

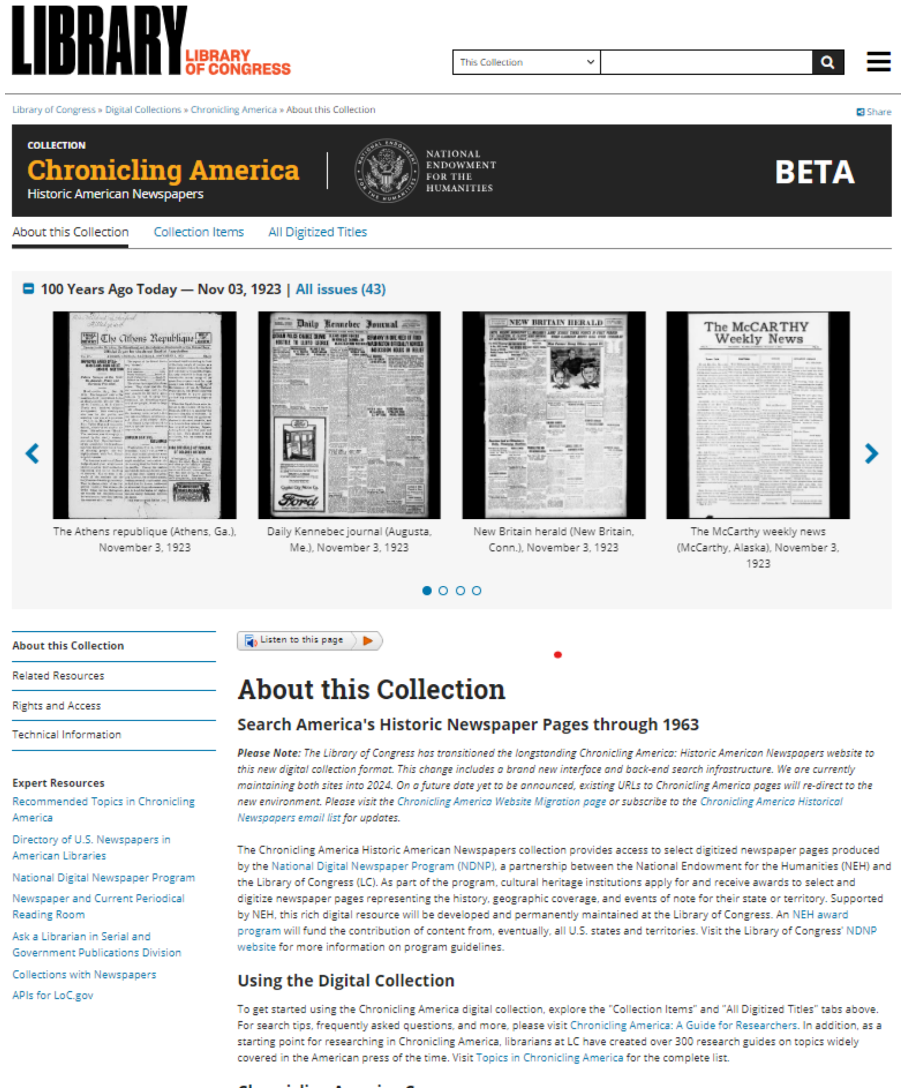
What is Chronicling America?

- Chronicling America
- Free, publicly accessible database of newspapers
  - 1770-1963
  - +21 million pages
  - 3,960 newspaper titles
  - 50 states, DC, PR, VI

SEARCH HISTORICAL NEWSPAPERS 1777-1963
50
STATES
chroniclingamerica.loc.gov
LIBRARY OF CONGRESS · NATIONAL ENDOWMENT FOR THE HUMANITIES
#ChronAm
PEACE!
WAR
JAPAN IN FIGHT
LUSITANIA
PROCLAIMED
MOURNS

Chronicling America Users

- Interface users
  - Students/teachers
  - Genealogists
  - Historians
  - Cultural heritage insts
  - Many other researchers
- Data users
  - Bulk data via API


LIBRARY
LIBRARY OF CONGRESS
This Collection
Library of Congress » Digital Collections » Chronicling America » About this Collection
Share
COLLECTION
Chronicling America
Historic American Newspapers
NATIONAL ENDOWMENT FOR THE HUMANITIES
BETA
About this Collection
Collection Items
All Digitized Titles
Listen to this page
About this Collection
Related Resources
Rights and Access
Technical Information
Expert Resources
Recommended Topics in Chronicling America
Directory of U.S. Newspapers in American Libraries
National Digital Newspaper Program
Newspaper and Current Periodical Reading Room
Ask a Librarian in Serial and Government Publications Division
Collections with Newspapers
APIs for LoC.gov

100 Years Ago Today — Nov 03, 1923 | All issues (43)

The Athens republique (Athens, Ga.), November 3, 1923

Daily Kennebec journal (Augusta, Me.), November 3, 1923

New Britain herald (New Britain, Conn.), November 3, 1923

The McCarthy weekly news (McCarthy, Alaska), November 3, 1923

About this Collection

Search America's Historic Newspaper Pages through 1963

Please Note: *The Library of Congress has transitioned the longstanding Chronicling America: Historic American Newspapers website to this new digital collection format. This change includes a brand new interface and back-end search infrastructure. We are currently maintaining both sites into 2024. On a future date yet to be announced, existing URLs to Chronicling America pages will re-direct to the new environment. Please visit the Chronicling America Website Migration page or subscribe to the Chronicling America Historical Newspapers email list for updates.*

The Chronicling America Historic American Newspapers collection provides access to select digitized newspaper pages produced by the National Digital Newspaper Program (NDNP), a partnership between the National Endowment for the Humanities (NEH) and the Library of Congress (LC). As part of the program, cultural heritage institutions apply for and receive awards to select and digitize newspaper pages representing the history, geographic coverage, and events of note for their state or territory. Supported by NEH, this rich digital resource will be developed and permanently maintained at the Library of Congress. An NEH award program will fund the contribution of content from, eventually, all U.S. states and territories. Visit the Library of Congress' NDNP website for more information on program guidelines.

Using the Digital Collection

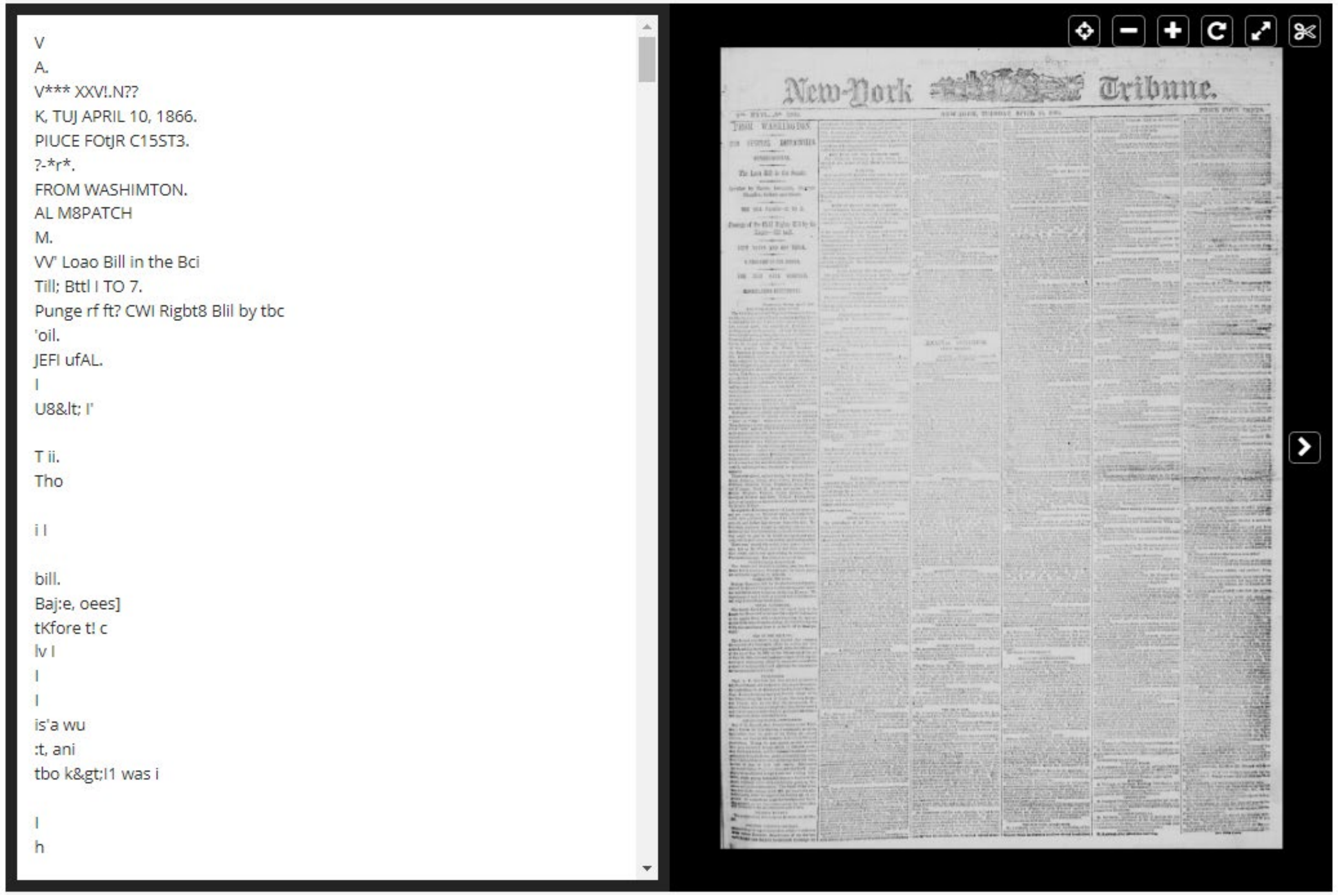
To get started using the Chronicling America digital collection, explore the "Collection Items" and "All Digitized Titles" tabs above. For search tips, frequently asked questions, and more, please visit Chronicling America: A Guide for Researchers. In addition, as a starting point for researching in Chronicling America, librarians at LC have created over 300 research guides on topics widely covered in the American press of the time. Visit Topics in Chronicling America for the complete list.

NDNP Deliverables

- Batch
  - One to multiple titles, multiple reels or print, structured in hierarchy
  - Up to 10,000 pages
- Batch package
  - Images
    - TIFF for preservation – not accessible through website
    - JP2000 for high-res downloads, zooming in/display
    - PDF for quick, full-page downloads
  - METS XML
    - Contextual data about the newspaper page, issue, title, reel
  - ALTO XML file for Optical Character Recognition text (OCR)
    - OCR+page coordinates
    - Enables hit-highlighting

OCR Issues

- OCR Quality
  - Damaged/poorly printed original print
  - Scanning from microfilm, not the original
  - Bad column zoning
  - Tiny text
- OCR quality variability
  - Varies from title to title
  - OCR engines have dramatically improved from 2005-2023



```
V
A.
V*** XXV!.N??
K, TUJ APRIL 10, 1866.
PIUCE FOtJR C15ST3.
?-*r*.
FROM WASHIMTON.
AL M8PATCH
M.
VV' Loao Bill in the Bci
Till; Bttl I TO 7.
Punge rf ft? CWI Rigbt8 Blil by tbc
'oil.
JEFI ufAL.
I
U8&lt; I'

T ii.
Tho

i I

bill.
Baj:e, oees]
tKfore t! c
lv I
I
I
is'a wu
:t, ani
tbo k&gt;I1 was i

I
h
```

National Digital Newspaper Program (NDNP)

NDNP-Open-OCR Pipeline

NDNP-Open-OCR

NDNP-Open-OCR is an **open-source project** developed by the Library of Congress for **re-processing OCR** of NDNP data.

NDNP-Open-OCR

Initial Goals:

- Plan and test OCR reprocessing for a targeted set of pages digitized prior to 2012.
- Incorporate re-processed OCR content into Chronicling America.

Guiding Principles:

- Needs to work at-scale, cost effective, adaptable for NDNP, highly automated

Image Source: https://www.loc.gov/resource/ds.02121/

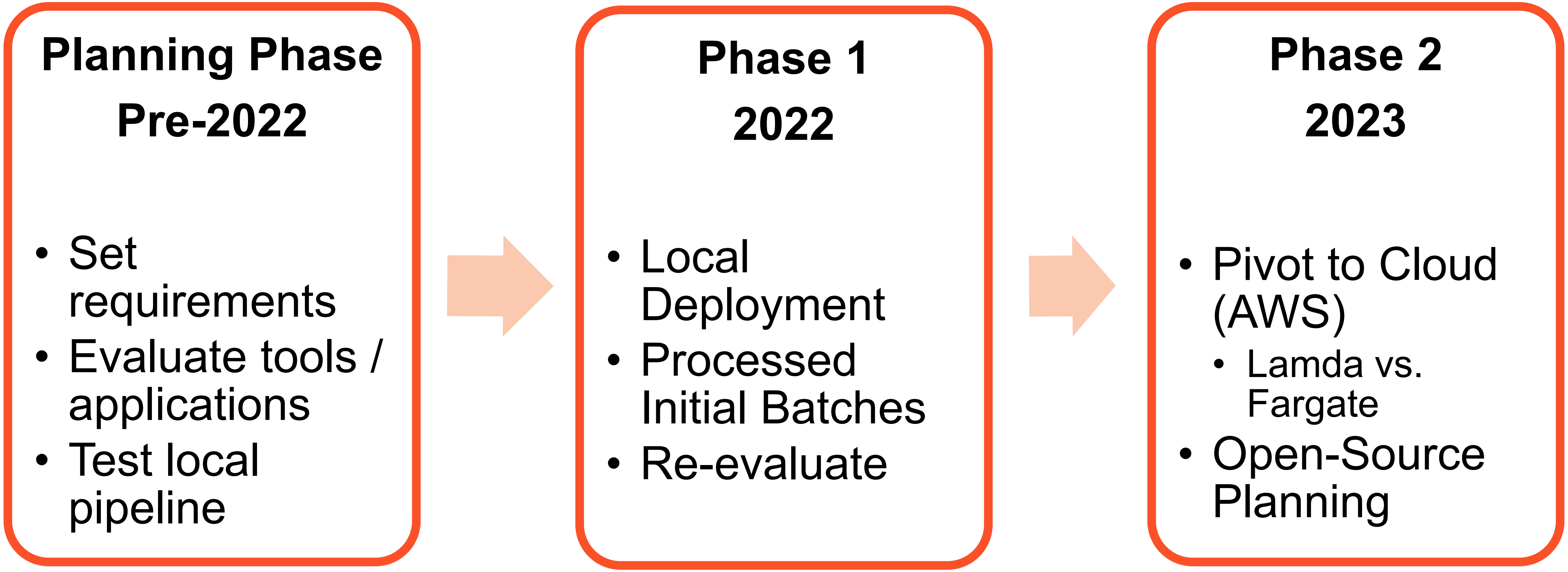
NDNP-Open-OCR

The NDNP-Open-OCR pipeline…

- creates new **ALTO XML** and **PDF** files for NDNP batches,
- can be deployed locally or in common **cloud environments**,
- uses **Tesseract** and custom post-processing steps,
- can be accessed via **command line interface**, and
- has potential to be **adapted** for other data.

NDNP-Open-OCR

Timeline



Planning Phase
Pre-2022

- Set requirements
- Evaluate tools / applications
- Test local pipeline

Phase 1
2022

- Local Deployment
- Processed Initial Batches
- Re-evaluate

Phase 2
2023

- Pivot to Cloud (AWS)
  - Lamda vs. Fargate
- Open-Source Planning

NDNP-Open-OCR Open-Source Tools:

Processing

- **OpenCV** and **Python Pillow Library (PIL)** for **pre-processing of JP2 files**
- **Tesseract** for production of new ALTO OCR and PDF files
- **ExifTool and Ghostscript** for **post-processing PDF files**
  - Preserve RDF metadata
  - Set display and compression settings
- **BeautifulSoup** (Python Library) and **custom Python script** for **post-processing ALTO files**
  - Correct end-of-line hyphenation in OCR

Image Source: https://chroniclingamerica.loc.gov/lccn/sn85066387/1912-03-10/ed-1/seq-17/

NDNP-Open-OCR Open-Source Tools:

Infrastructure

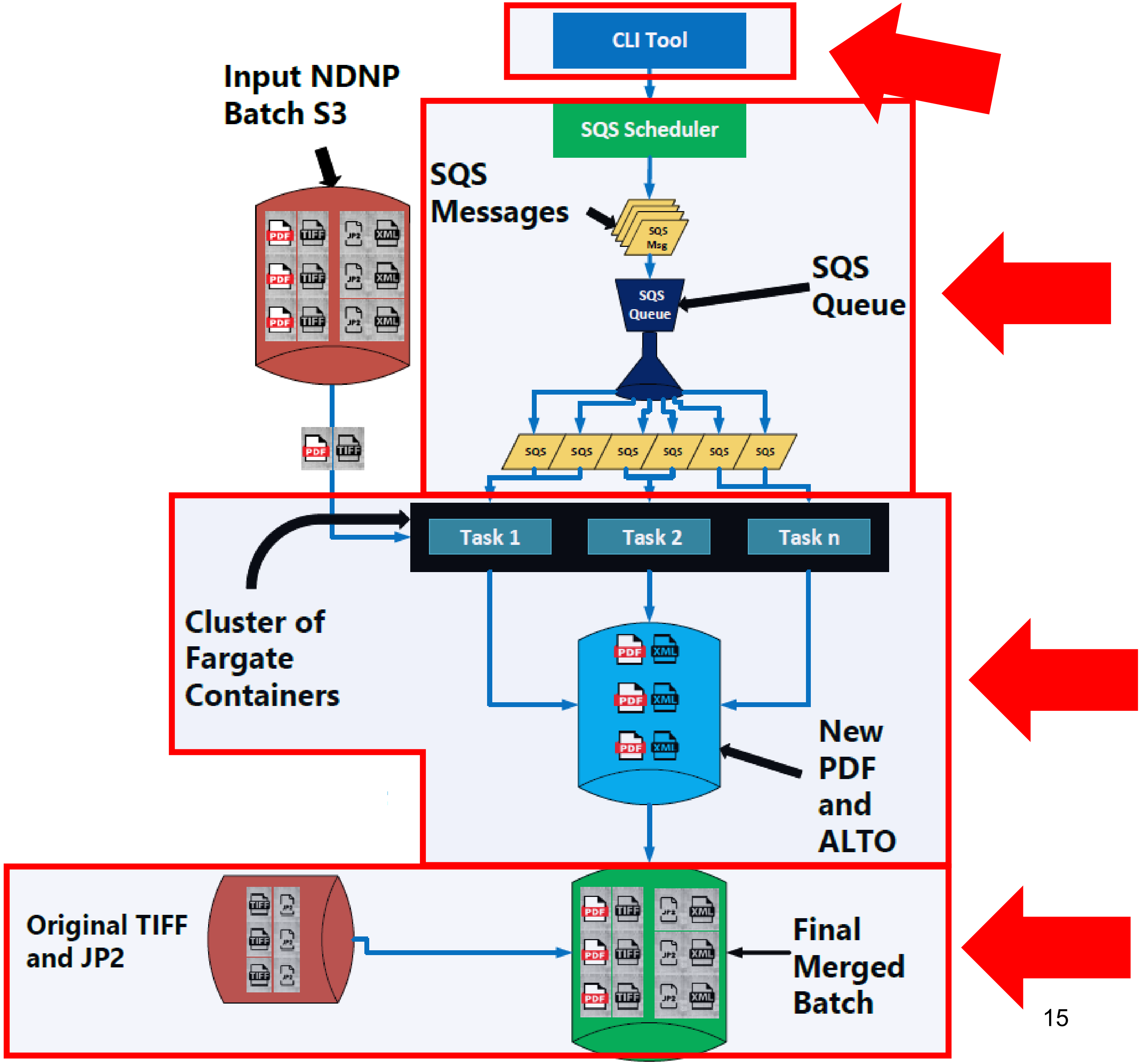
- **Terraform** for IAC (infrastructure as code) / cloud infrastructure **at scale**
- **boto3** (AWS Python Library) for **interfacing with AWS Services**
- **Docker** for **containerizing code** and making easy to run anywhere
- **Flask** (Python backend application) for Fargate tasks and **parallel processing**. Can scale up to as many workers as we want/need.

Image Source: https://chroniclingamerica.loc.gov/lccn/sn85066387/1912-03-10/ed-1/seq-17/

NDNP-Open-OCR AWS Pipeline

- Deployed using Terraform (IAC) tool to Amazon Web Services (AWS) cloud environment.

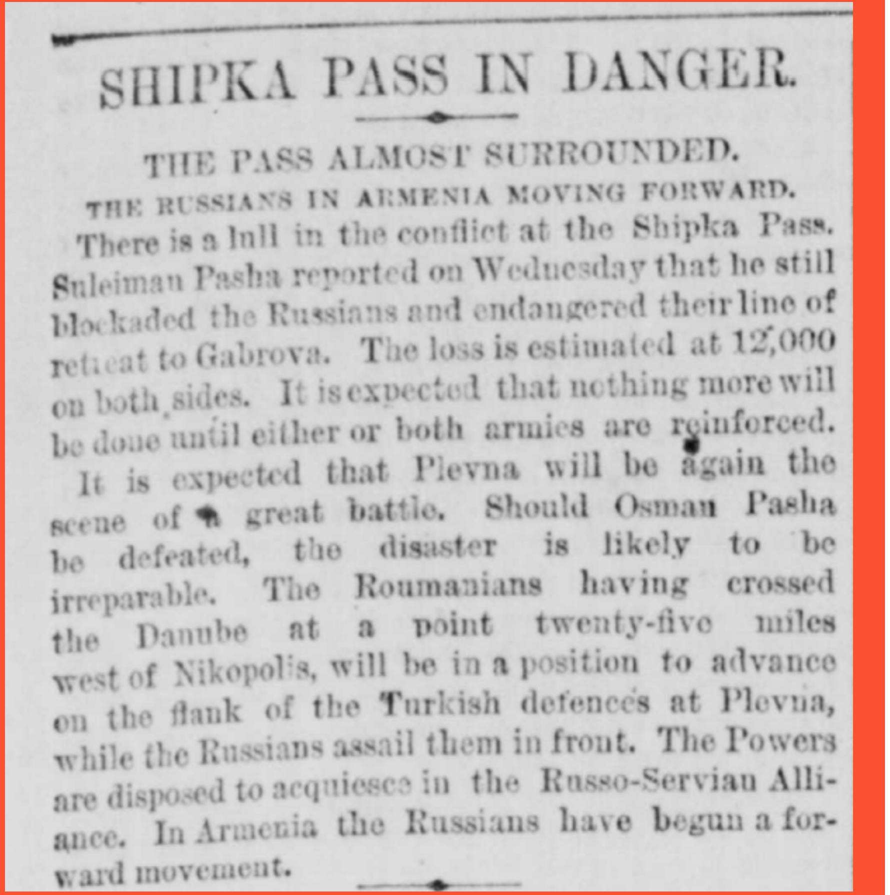
  AWS services used in the pipeline include:
  - Amazon Simple Queue Service (SQS)
  - Serverless services (AWS Fargate)
  - Amazon Simple Storage Service (Amazon S3)



LIBRARY OF CONGRESS

NATIONAL ENDOWMENT FOR THE HUMANITIES

Before Reprocessing



SHIPKA PASS IN DANGER.

THE PASS ALMOST SURROUNDED.

THE RUSSIANS IN ARMENIA MOVING FORWARD.

There is a lull in the conflict at the Shipka Pass. Suleiman Pasha reported on Wednesday that he still blockaded the Russians and endangered their line of retieat to Gabrova. The loss is estimated at 12,000 on both sides. It is expected that nothing more will be done until either or both armies are reinforced.

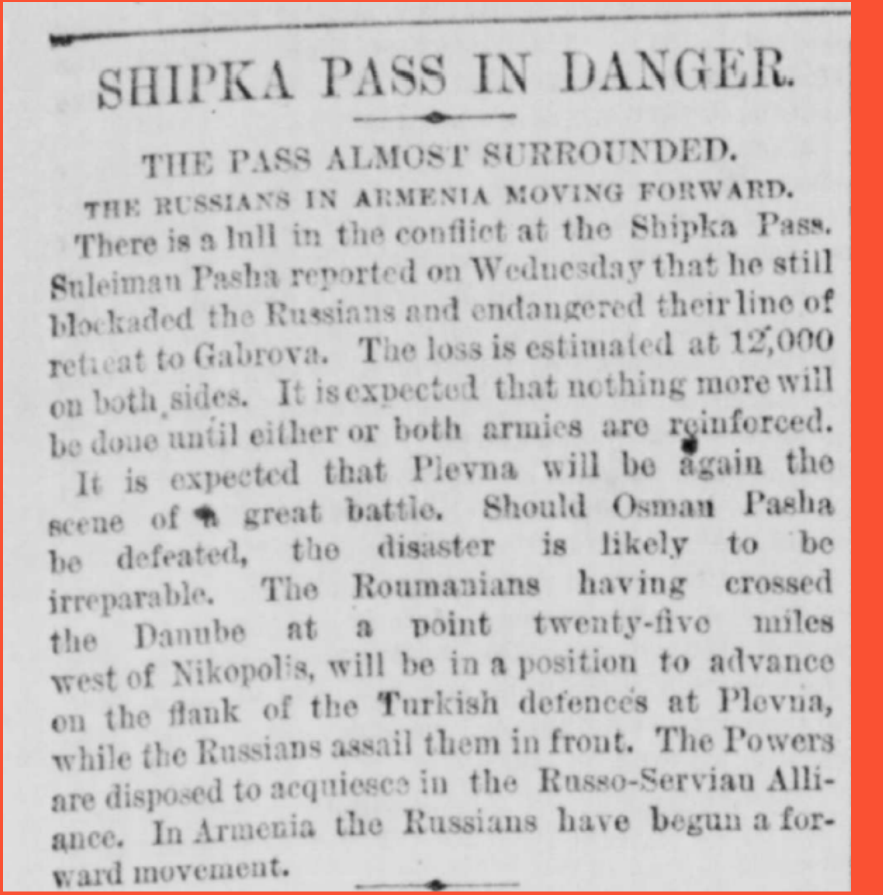
It is expected that Plevna will be again the scene of a great battle. Should Osman Pasha be defeated, the disaster is likely to be irreparable. The Roumanians having crossed the Danube at a point twenty-five miles west of Nikopolis, will be in a position to advance on the flank of the Turkish defences at Plevna, while the Russians assail them in front. The Powers are disposed to acquiesce in the Russo-Servian Alliance. In Armenia the Russians have begun a forward movement.

```
SHIPKA PASS IN DANGEI
[i :? i si ;?!.'. >U? I?; n.
Tin BfSMAWa in ARMENIA v.?vi\?; POSWAK
that in
i L ?mate .1 .it ?-'.
i
; ? -,? ox both 1'e.lifer?
I..1 Ill.lt r.evii.I ?ill In
. - . ? ? i P
;, tbe A-aeetei i.? likely t<>
I.' mum'il.ill? Ii:I4 111?; 1 I??
? nt a I'lillt tVM'LIt.V-MVI- II!
-, will be in a] Ition to adi i
h nf tin* Tur! Iah defence s al Pie?
Tin' I'i.v,
in th" Bi
I
```

LIBRARY OF CONGRESS

NATIONAL ENDOWMENT FOR THE HUMANITIES

After Reprocessing



SHIPKA PASS IN DANGER.

THE PASS ALMOST SURROUNDED.

THE RUSSIANS IN ARMENIA MOVING FORWARD.

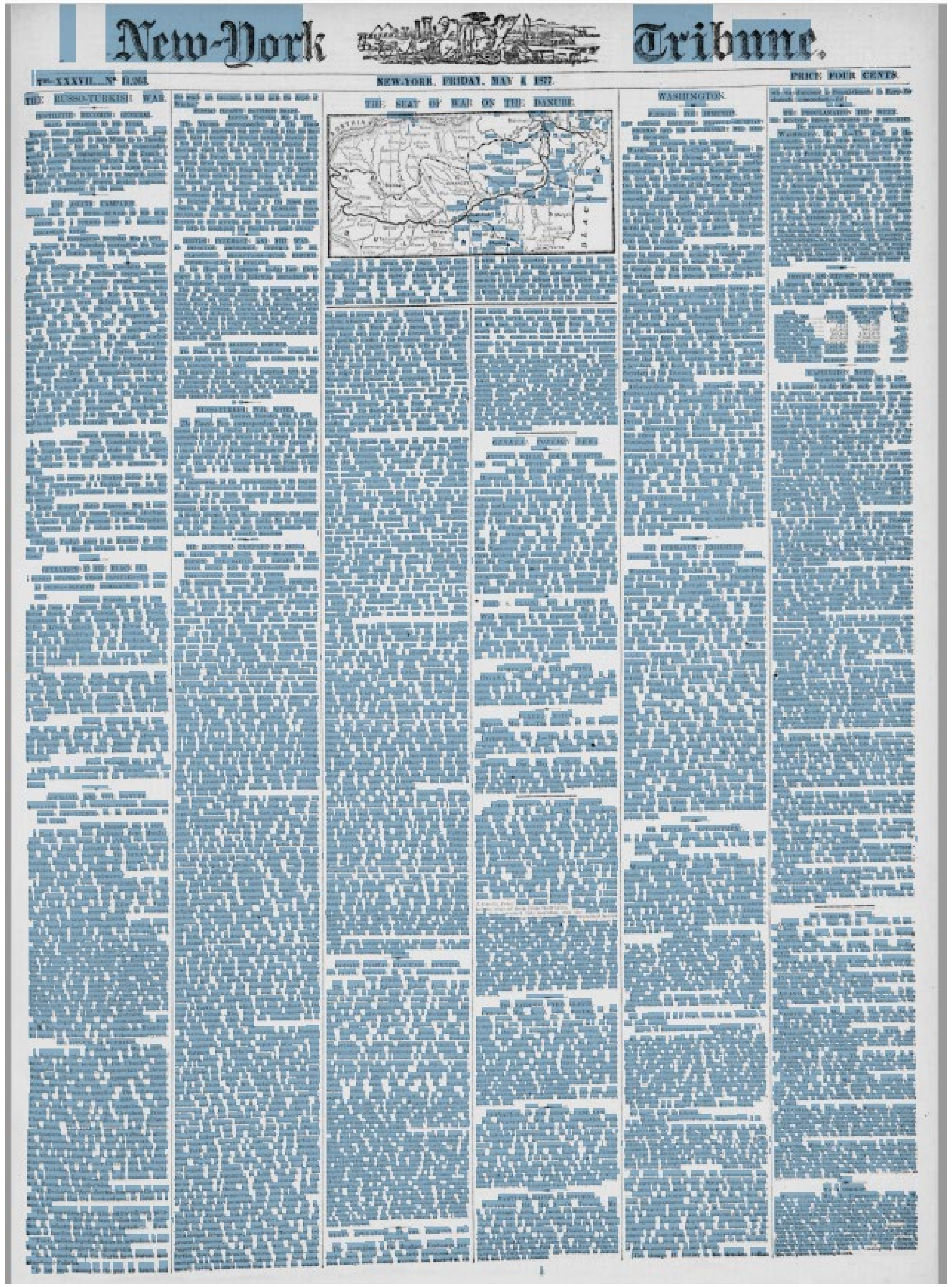
There is a lull in the conflict at the Shipka Pass. Suleiman Pasha reported on Wednesday that he still blockaded the Russians and endangered their line of retreat to Gabrova. The loss is estimated at 12,000 on both sides. It is expected that nothing more will be done until either or both armies are reinforced.

It is expected that Plevna will be again the scene of a great battle. Should Osman Pasha be defeated, the disaster is likely to be irreparable. The Roumanians having crossed the Danube at a point twenty-five miles west of Nikopolis, will be in a position to advance on the flank of the Turkish defences at Plevna, while the Russians assail them in front. The Powers are disposed to acquiesce in the Russo-Servian Alliance. In Armenia the Russians have begun a forward movement.

```
SHIPKA PASS IN DANGER.
THE PASS ALMOST? SURROUNDED.
SSIANS IN ARMENIA MOVING FORWARD.
There is a Inli in the conflict at the Shipka Pass.
Suleiman Pasha reported on Wednesday that he still
plockaded the Russians and endangered thetrline of
reticat to Gabrova. The loss is estimated at 12,000
t isexvected that nothing more will
her or both armies are rginforced.
on both sides. I
be doue until eit
It is expected that Plevna will be again the
scene of ® great battle, Should Osman Pasha
be defeated, the disaster is likely to be
irreparable. The Roumanians having crossed
the Danube at a point twenty-five miles
west of Nikopolis, will be ina position to advance
```

After Reprocessing

- Column-level zoning adherence



New-York Tribune

Vol. XXXVII....Nº 11,363. NEW-YORK, FRIDAY, MAY 4, 1877. PRICE FOUR CENTS.

THE RUSSO-TURKISH WAR.

THE SEAT OF WAR ON THE DANUBE.

WASHINGTON.

LI NEWS BRARY
LIBRARY OF CONGRESS

NATIONAL ENDOWMENT FOR THE HUMANITIES

NDNP-Open-OCR: Releasing as Open-Source Pipeline

- Benefit to Chronicling America researchers/ website users:
  - Improve search results
  - Clean up dirty data for bulk data / text mining users

- Benefit to Library of Congress collections:
  - Smaller Solr index to maintain for LC, future migration and collection maintenance will be easier

- Benefit to NDNP:
  - NDNP awardees and institutions using NDNP standards can use to improve their collections

- Benefit to the Digital Library community:
  - DL community and DH researchers can fork and adapt to local needs

NDNP-Open-OCR

Next Steps

- Finish work on AWS pipeline and Command Line Interface (CLI) Fall 2023
- Export and re-ingest new versions of ~20 batches
- Create “bad OCR” batch nomination process
  - Early batches
  - Low searchability
  - Known missing/duplicate OCR
  - Batches with an unusually high number of “unique words” in Solr
  - Languages with a new OCR engine
- Run ~25 more batches
- Release pipeline, code as Open Source on LC GitHub

Note: NDNP-Open-OCR is still in R&D phase. Details are subject to change.

Thanks!

Robin Pike
NDNP Coordinator, Library of Congress
rpike@loc.gov

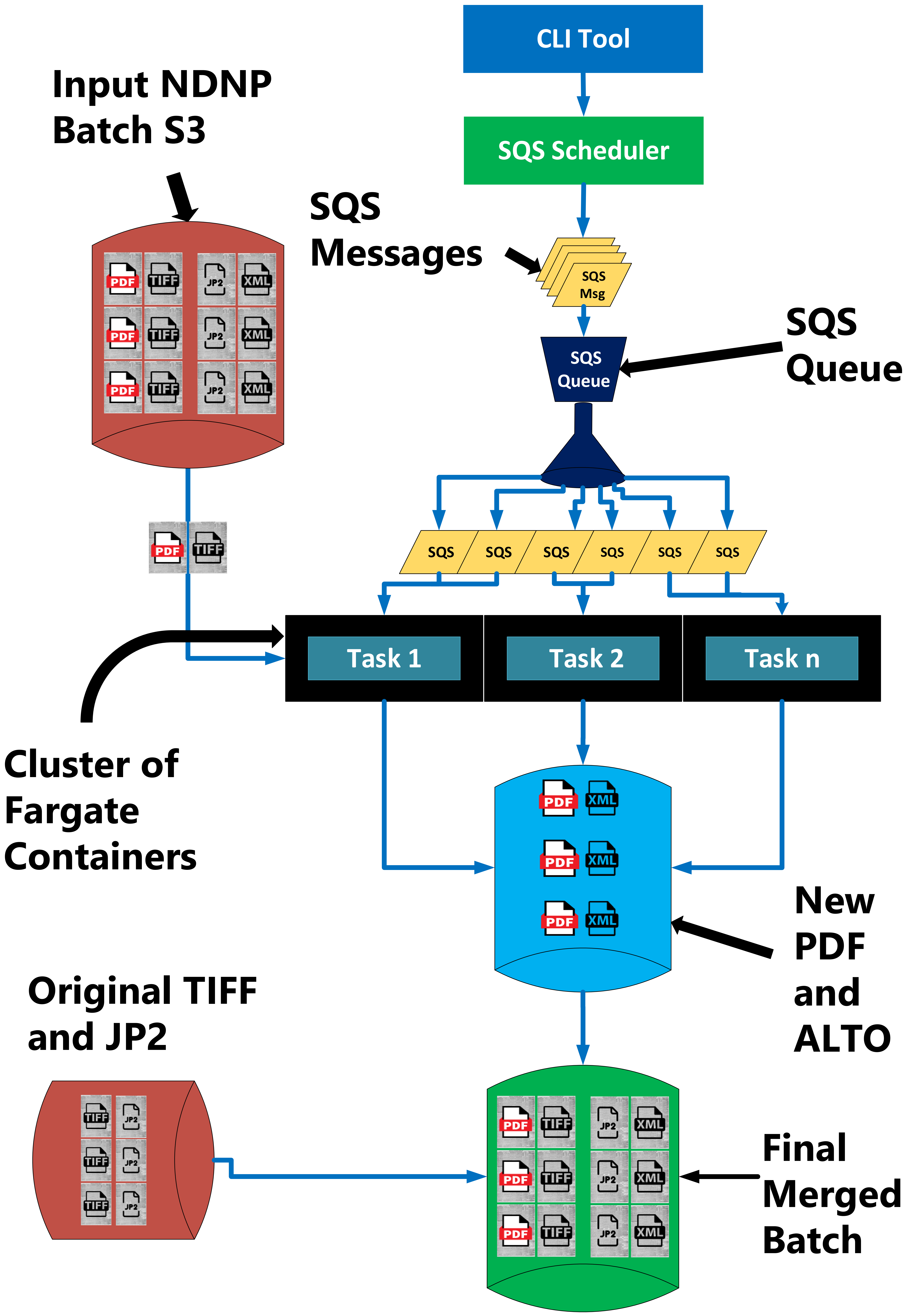
Nathan Yarasavage
NDNP Production Lead, Library of Congress
nyarasavage@loc.gov

Subscribe **for updates**

Questions?

NDNP Open OCR AWS Pipeline

The motivation for the creation of this pipeline is to re-OCR NDNP batch data at-scale in a cost-effective way. The application is deployed via Terraform IAC to Amazon Web Services (AWS) cloud environment.



Workflow

1. The flow begins by triggering a NDNP batch to be processed in the NDNP Open OCR Start step.

2. The SQS Scheduler will read the contents of the NDNP batch data in the INPUT S3 bucket, then creates 1 SQS message for each newspaper page, submitting these to the SQS queue.

3. The Queue then feeds messages to the NDNP Open OCR Fargate tasks for PDF and ALTO file creation.

4. The parallel Fargate tasks read the original TIFF files from the INPUT S3 bucket, and run Tesseract on those to generate new files with greater OCR quality.

5. Both the PDF and ALTO generators will write outputs to a desired location in an output bucket, specified by Terraform IAC code. For each newspaper page in a batch, there will be 1 output PDF and 1 output ALTO file, to be pulled down and merged with local batch data later.

Updated: 11/7/2023
Pipeline subject to change.