

View email in your browser

Share this email with a friend

## New Search Feature: Optical Character Recognition (OCR)

With more than 92 million pages of digitized records available to search in the National Archives Catalog, we are always working on ways to improve search results to better help you find what you're looking for.

That's why we're excited to share a new feature in the Catalog: Optical Character Recognition, or OCR.



PV2 Ruben Gutierrez a general mechanic, 4/21/1997, National Archives Identifier 6504747

#### What is OCR?

OCR converts images that contain typed, handwritten, or printed text into text that can be read and searched by a computer.

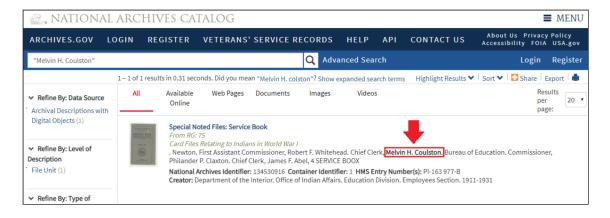
Previously, records in the Catalog were only searchable based on the titles, descriptions, and other fields entered by archivists, or by tags and transcriptions entered by citizen archivists. Now, with OCR capability, text from some images in the Catalog can be extracted, making that text searchable and more likely to come up in your search results.

Currently, the Catalog's new OCR engine is applied to records in either JPG or PDF format added to the Catalog since June 2019. NARA is exploring how to retroactively process records from before that point, but right now this feature applies to millions of pages!

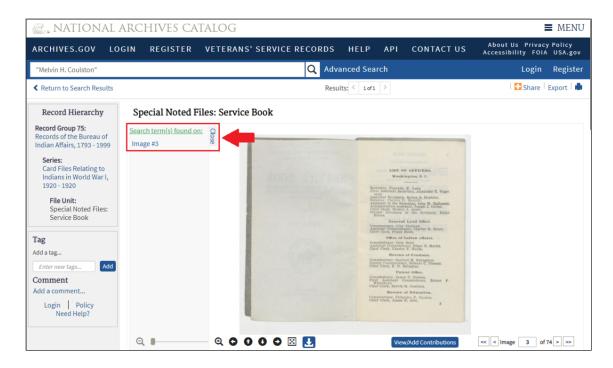
Here's what you can expect and how it works:

A search for <u>"Melvin H. Coulston"</u> returns this Bureau of Indian Affairs record with OCR data. The search term is bolded.

SEARCH TIP - If you are searching for a name or phrase, surround it in quotation marks to do an exact phrase search.



To explore the results further, click on the blue description title. On the description page that follows, you can see the pages where your search term is found. They are listed below the description title and to the left of the image viewer:



Likewise, you may also see the page thumbnails highlighted beneath the image viewer that contain your search term. Clicking on any of the pages in the list or a highlighted page thumbnail will take you to that page.



#### Try it out!

You can test the capability yourself by running one of the following searches and clicking the first result returned for each:

- "Dornier Aviation Plants"
- "Did you bail out? No"
- "Sea Devil Torpedo Attack"

We still have work to do! Right now, we are investigating options to re-process items for OCR that were in the Catalog prior to June 2019. Additionally, records that are only available in PDF format currently do not provide the page jumping or highlighting capability.

OCR is not perfect! While this technology helps to make records more searchable, we still find human-entered transcriptions to be more accurate than OCR, so we still need your help as citizen archivists to transcribe records in the Catalog, and help decipher that tricky handwriting!

#### In case you were interested...

**Technical specs:** NARA's new OCR engine is powered by the open source <u>Tesseract</u> software. As records are added to NARA's <u>Amazon Web Services (AWS) S3 cloud storage</u>, it is run through image processing powered by a series of <u>AWS Lambda</u> functions.

# News from the Innovation Hub: 500,000 pages scanned by Citizen Archivists!

Citizen Scanners in the National Archives Innovation Hub have officially scanned 500,000 pages of historical records!

Following scanning, each digitized page is then added to the National Archives Catalog for anyone to view and download. Currently, nearly 490,000 pages of records scanned in the Hub are available in the National Archives Catalog. You can <u>view all records scanned by citizen contributors</u> in the Catalog so far.

Here is the 500,000th page scanned in the Innovation Hub:

*
0 9-447.
Entroise - george
har ong in 1.303,511 Department of the Interior,
axtonio Orrace
od 27 from his
Washington, D. C. april 16, 1907
Sex: To slid thin Birean in preventing my one falsely personating you, or otherwise committing fraud
in your name, or on account of your service, you are required to masser fully the questions emmarated below.
You will please return this circular under cover of the todaesd suveleps which requires no postage.  Very respectfully.
antonis Dardell
200 Gen strice
Joen Haven Com
The state of the s
1. When were you tount some. Taken from behing in chillhood
2. Where were you been? Answer. Chinese
1. When did you called Answer, 12th Getates 1862.
4. Where did you exist? Answer, City of New Heaven learn
A. When had you lived before you califord? Aroner. Charles
5. What was your occupation address at culletiment? Answer Server Abdress to
8. White were you discharged? Answer, 25 th Junion, 1883
2. Where were you discharged? Answer. Lite of New Years
10. Where have you lived since discharge? Give dates, as nearly as possible, of any clumges of residence.
Chimmen to have to New Heron land
1862
11. What is your present occupation? Asserts. Treats.  12. What is your height? Asserts. J. fort. 2/2. inches. Your weight? / 2/2.
The color of your ayer? Lake ask. The color of your hair? Merch Kank Your complexion?
Are there may permanent marks or sears on your person? If so, distribe them,
18. What is your full moun? Please write it on the line below, in ink, in the manner is which you are
excusioned to sign it, in the process of two witnesses who can write,
45 NO O
U. 27 8 Antini Mardell .
0 0 1
William & Showfree Day April 20th 100
(2 Min D. Thompson
, sand the sand sand sand sand sand sand sand sand

This is a form that Antonio Dardell filled out to apply for an increase in his Civil War veteran's pension. Antonio Dardell is an interesting soldier, being one of the small number of Chinese American soldiers who fought in the American Civil War.

Dardell had likely been adopted as a young boy by an American whaling ship captain, and brought to Connecticut. He enlisted in the 27th Connecticut Infantry Regiment at age 19. After the war he worked as a tinsmith, and in 1880 he became a naturalized citizen, showing his honorable discharge certificate in lieu of a Declaration of Intent. More information can be found in this "The Blue, the Gray, and the Chinese" blog post.

This pension file was scanned as part of a two-week project by two citizen scanners who were interested in digitizing records relating to Asian American history. In addition to Antonio Dardell's pension, the citizen scanners also worked on his Compiled Military Service Record, and the CMSRs and pensions of other Chinese American soldiers.

Many thanks to everyone who's worked in or adjacent to the Hub now or in the past! Here's to the next 500,000 pages.

Now that these pages are scanned and added to our Catalog, help transcribe them to make them more searchable! Our latest <u>citizen archivist mission</u> features records scanned by citizen contributors in our Innovation Hub.

### **Get started transcribing!**

Interested in Citizen Scanning? Learn more about the <u>National Archives Innovation Hub</u> in Washington, DC.

Questions or comments? Email us at <a href="mailto:catalog@nara.gov">catalog@nara.gov</a>.









Privacy policy

Subscribe or Unsubscribe

<u>Powered by Mailchimp</u>