OpenIndex OCR

Using open source technology to enhance digital archives

Meet the teAm

W. Carrington Powell III

I was born at an early age in Southern Oregon and have slowly moved north to find colder weather. I have a minor in Computer Information Technology and a passion for historical preservation.

Sponsor Organization

Lisa Oberg

Interim Director / History of Science and Medicine Curator, UW Special Collections

UW Special Collections

Collects and preserves rare books, manuscripts, records, photographs, moving images, and architectural drawings

The Issue at Hand

- Because of the nature of the materials, researchers are required to physically visit the collection to view materials
- This constraint necessitates that patrons must already have some idea what is in the collection, so they can request the right boxes and view the material in the allotted appointment time
- Without a proper index, much of the material is as good as invisible
- The digital world is changing this, opening up the archive to researchers from all around the world

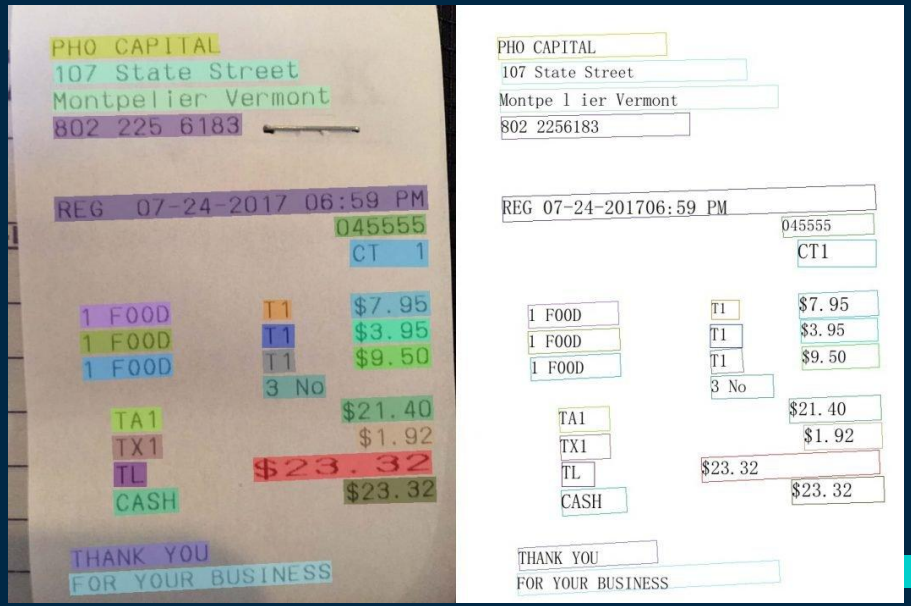
A Gap to be Bridged

1.2 million cards!

Catalog of materials from 1850-1996

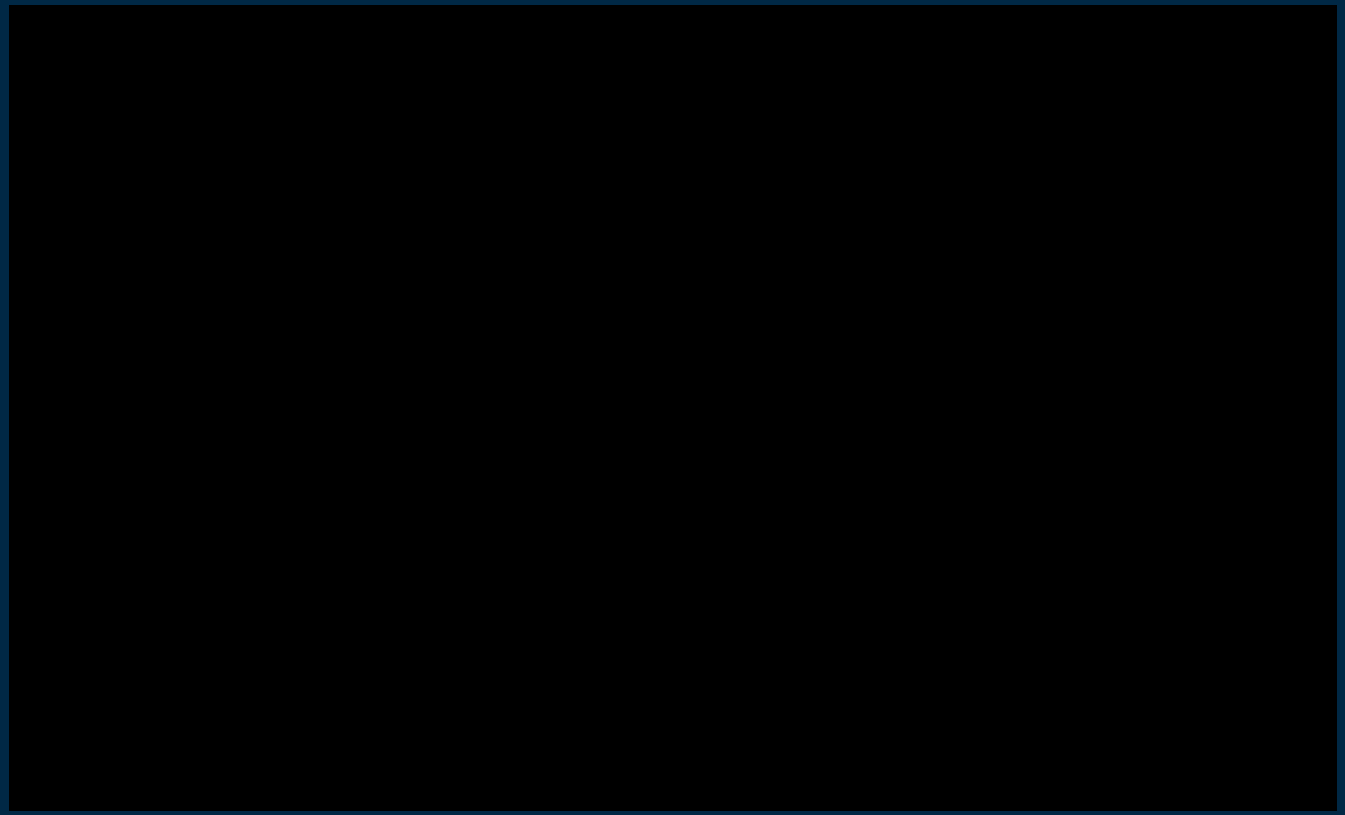
What is OCR?

- When scanning images (like this receipt) we have text information but no way to actually use it
- With Optical Character Recognition, a computer can recognize the text within the image and overlay a layer of machine readable text on top



An Open Source Solution

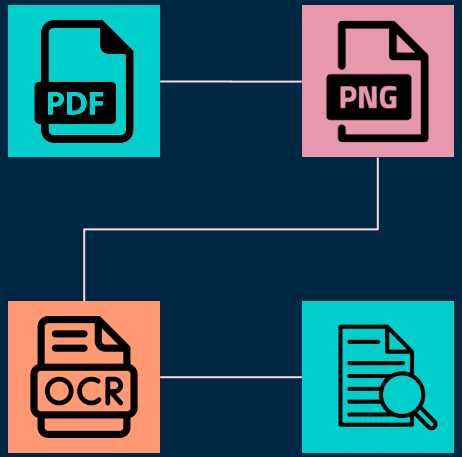
Demonstration and output



The Approach

Input

A user uploads a PDF file of any size



Preparation

The file is first transformed into an array of images

Processing

Tesseract reads the images and creates a layer of text

Output

The images are recombined into a new PDF, ready to download!

Next steps

- While the tool is finished, the example files have not yet been updated
- I am working on transforming all files and getting them updated on the Klondike Nugget webpage
- This can be a tool in a larger kit for archivists to digitize their collections