4-22-2022

Digitizing Historical Forest Service Data

Florina Ciaglia
Boise State University

Catherine Olschanowsky
Boise State University

Kelly Hopping
Boise State University

Digitizing Historical Forest Service Data

Floriana Ciaglia, Dr. Catherine Olschanowsky, Dr. Kelly Hopping

BOISE STATE UNIVERSITY
COLLEGE OF ENGINEERING
Department of Computer Science

adapt
DATA FLOW OPTIMIZATIONS LAB

1. Problem Statement

- Ecologists record vegetation data by hand onto physical paper-sheets.
- Historical Forest Data is inaccessible for further analysis and research.

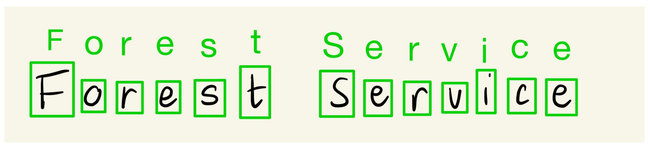
2. Motivation

Vegetation and soil condition data from the Sun Valley, Idaho area has been collected **by hand** and is laying into dusty filing cabinets.

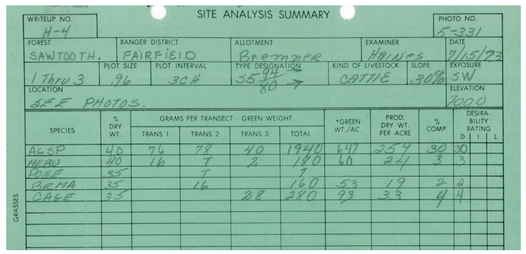
The goal of this project is to **digitize** the data forms to make them available for future scientific research.

3. Optical Character Recognition (OCR)

- Processes image.
- Recognizes ASCII characters in the provided image.
- Extracts the character and saves it into a machine-encoded text.



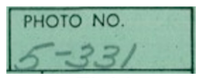
4. Process



Original data format

Ste 1. Identifying sub-fields in the form

- Extract sub-fields from the form using the OpenCV library.



Field

Step 2. Bounding box around single characters

- Crop the image around each single character to feed to the model.

Field → Individual Cells

Step 3. Create CSV file

- We store the RGB value of all the single-character images into a CSV file.
- The CSV file is fed into a pretrained character recognition model.
- The model outputs an ASCII character guess for each image trained on the EMNIST Dataset.

EMNIST 0-9, A-Z and a-z

image courtesy of: https://www.researchgate.net/

5. Results

Our pipeline reaches an accuracy percentage of 47.8% on average, with an increase to 67.6% on the second guesses.

As a comparison, we implemented the Google Cloud Vision API into our code base and ran it on our data reaching an average of 61.2% accuracy score.

Cloud Vision API

image courtesy of: https://cloud.google.com/vision

Forest Service Database

We created a SQL database to store the data yielded by the pipeline and SQLAlchemy to communicate between our Python code base and the SQL database.

SQLAlchemy

image courtesy of: https://hackersandslackers.com/series/mastering-sqlalchemy/

```
-- Inserted values in report table
 r_id | writeup_no | photo_no |  forest  | ranger_district | allotment | examiner |   date   | transect_no
------+------------+----------+----------+-----------------+-----------+----------+----------+-------------
    1 | h-4        | 5-331    | sawtooth | fairfield       | bremner   | haines   | 7/15/73  | / thru 3
    2 | h4         | 5331     | sawtooth | fairfield       | bremner   | h8inrs   | 711517z  | ltbiul3
(2 rows)
```

```
| plot_size | plot_interval | type_designation | livestock | slope | aspect |    location     | elevation |
+-----------+---------------+------------------+-----------+-------+--------+-----------------+-----------+-
| 96        | 36h           | 33 5524          | cattle    | 30%   | sw     | see e photos    | 7000      |
| 9b        | l3ch          | s5yftnx          | cattle    | 7d9   | sw     | seepaotos       | 7oob      |
```

6. Future Development and Challenges

Implement second guess selection to improve guesses of characters when we know the ones we received are wrong. Such as the above ‘examiner’ field.

Implement further checks on the output to improve overall performance. For example, implement a system to parse month, day, and year for the ‘date’ field.

7. Acknowledgements

Boise State’s Research Computing Department. 2017. R2: Dell HPC Intel E5v4 (High Performance Computing Cluster). Boise, ID: Boise State University. DOI: 10.18122/B2S41H. I’d like to thank all the students who contributed to the development of this project: Joshua Soutelo Vieira, Sandra Busch, Chinwendum Njoku, and Isaac Bard.