Extracting Field Signboard Text with Google Vision

Matthew D. Harris - AECOM
DataONE / Prov-a-Thon August 31, 2017
Bernalillo, NM

Archaeological Signboard

easy to read example

Archaeological Signboard

not so easy to read example

Goals

- Identify text in image repository

- Clean and validate identified text

- Assign text as image metadata

- Searchable database of images, text, and exif data

```
### plugin your credentials
options("googleAuthR.client_id" = "...")
options("googleAuthR.client_secret" = "...")
## use the fantastic Google Auth R package
### define scope!
options("googleAuthR.scopes.selected" = c("https://www.googleapis.com/auth/cloud-platform"))
googleAuthR::gar_auth()


photo_loc  <- paste0(getwd(), "/photos/boards/")
image_name <- 'P6190096.jpg'
image_url <- paste0(photo_loc, image_name)
file.exists(image_url)
image_text = getGoogleVisionResponse(image_url,
                                      feature = 'TEXT_DETECTION')

print(image_text)
```

Call to GoogleVision API
requires GoogleCloud account and API keys

```
> print(image_text)
  locale                                                              description                      vertices
1     en WOODLANDS\nHAMILTON DISTRICT\nMONITORING\nAREA 2\nTRENCH 1\nPATH\n6/20/16 AECOM\n 598, 855, 855, 598, 628, 628, 889, 889
2   <NA>                                                                WOODLANDS 656, 795, 794, 655, 628, 634, 650, 644
3   <NA>                                                                 HAMILTON 599, 713, 712, 598, 652, 657, 683, 678
4   <NA>                                                                 DISTRICT 729, 837, 836, 728, 658, 663, 689, 684
5   <NA>                                                               MONITORING 654, 802, 801, 653, 684, 690, 712, 706
6   <NA>                                                                     AREA 672, 740, 739, 671, 715, 717, 735, 733
7   <NA>                                                                        2 762, 774, 773, 761, 718, 718, 736, 736
8   <NA>                                                                   TRENCH 635, 745, 744, 634, 741, 745, 770, 766
9   <NA>                                                                        1 771, 791, 790, 770, 746, 747, 772, 771
10  <NA>                                                                     PATH 663, 738, 737, 662, 779, 783, 811, 807
11  <NA>                                                                  6/20/16 599, 675, 674, 598, 855, 857, 883, 881
12  <NA>                                                                    AECOM 728, 855, 854, 727, 859, 863, 889, 885
```

GoogleVision Output

text and bounding box coordinates

Identified Text

Easy to read example

Identified Text

Not so easy to read example

```
status ← "success"
GV_tags ← image_text[1,"description"] %>%
  str_replace_all("\n", " ") %>%
  str_trim()
tryCatch(
  exifr::exifr(image_url,
        exiftoolargs=paste0("-XMP:Subject=", "'",
                            GV_tags, "'",
                            " -overwrite_original")),
  error=function(e) "NULL"
)
```

Writing Image Metadata

Code example using exifr package

```
> results
         filename        status                                                                              tags
1      1_test.jpg  no features                                                                              <NA>
2    IMG_0210.jpg      success                                RACE ST CONNECTOR MONITORING FEATURE 3 12/21 16
3    IMG_0223.jpg      success                               RACE ST CONNECTOR MONITORING FEATURE 2 12/ 21 16
4    IMG_0250.jpg      success                                      RACE ST CONNECTOR MONITOFR RE ACCOM
5    IMG_0629.jpg      success                                       RACE ST CONNECTOR MONITOR URE 30
6    IMG_0647.jpg      success                                    RACE ST CONNECTOR MONITORIN FEATURE
7    IMG_0766.jpg      success     WOODLAN H AMILTON DISTRICT MONITORI ING WATER LINE EAST PROFILE BRICK STRUCTURE AECOM
8    IMG_0797.jpg      success WOODLANDS HAMILTON DISTRICT MONITORING WATER LINE TRENCH BRICK STRUCTURE PLANVIEW 11/3/16 AECOM
9    IMG_2409.jpg      success                          RACE STREET CONNECTOR MONITORING FEA 6 1/241 AECOM
10   IMG_2428.jpg      success                                           RACE STR CTOR ONITORING AE
11   P6130019.jpg      success                  WOODLAND HAMILTON AREA 3 EU 1 FEATURE 1 OPENING PLANVIEW 613/16
12   P6140048.jpg      success     WOODLANDS HAMILTON DISTRICT MONITORING AREA 3 EU 1 FEATURE 1 SOUTH PROFILE 7 615/16 AECOM
13   P6160055.jpg      success              WOODLANDS HAMILTON DISTRICT MONITORING AREA TRENCH 2 PLOW SCARS AECOM
14   P6190086.jpg      success            WOODLANDS HAMILTON DISTRIOT MONITORING AREA 2 TRENCH 1 PATH 6/20/1 AECOM
15   P6190096.jpg      success            WOODLANDS HAMILTON DISTRICT MONITORING AREA 2 TRENCH 1 PATH 6/20/16 AECOM
16   P8040145.jpg      success           JAPANESE GARDEN UNIT 1 FEATURE 2 WEST PROFILE 1.95-2,65 FT BD 8/3/15 AECOM
17   P8040167.jpg      success                 APANESE GARDEN STPK12 EU 12.6 2-25 FT BD 8/4/15 AECOM
```
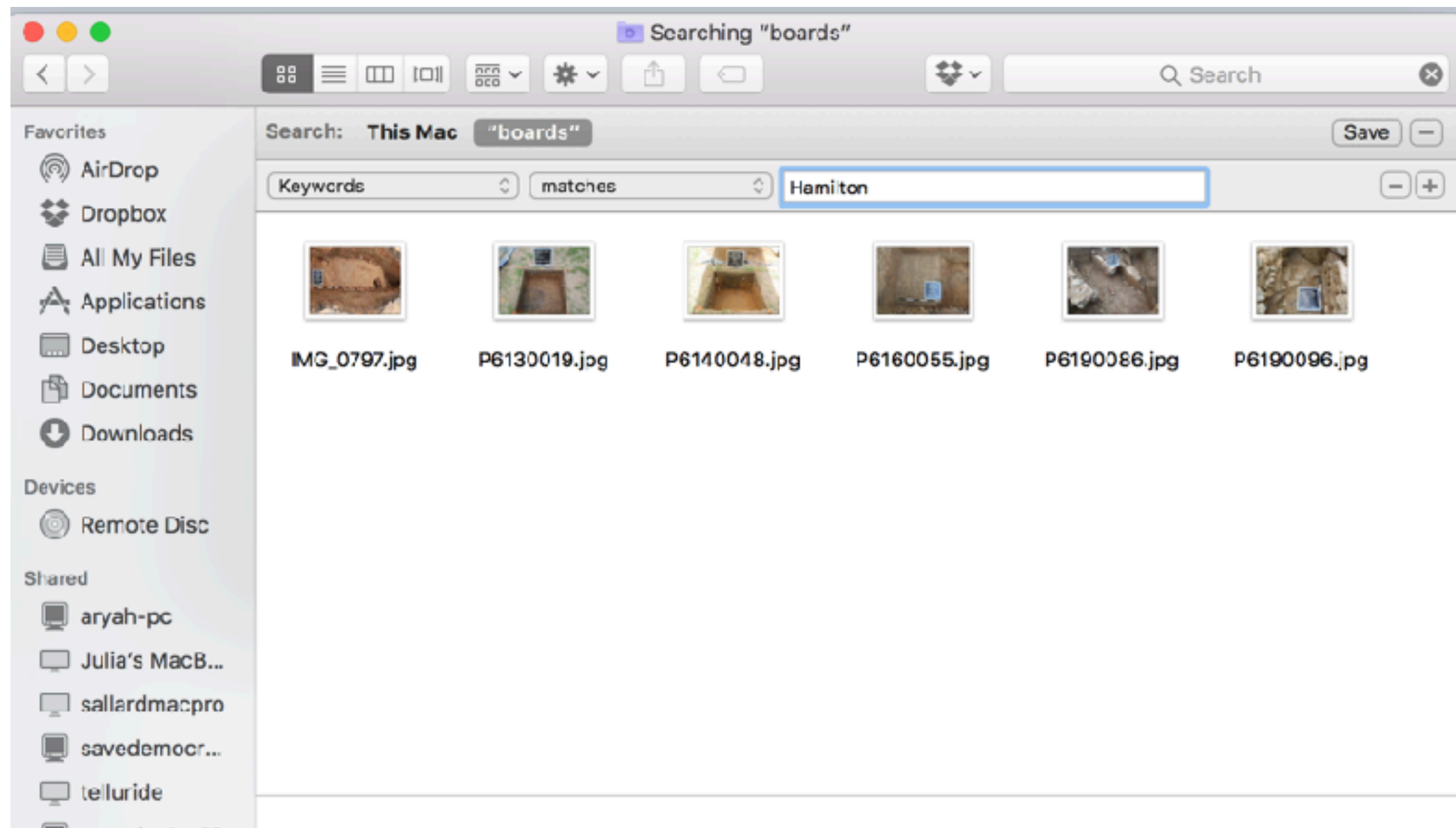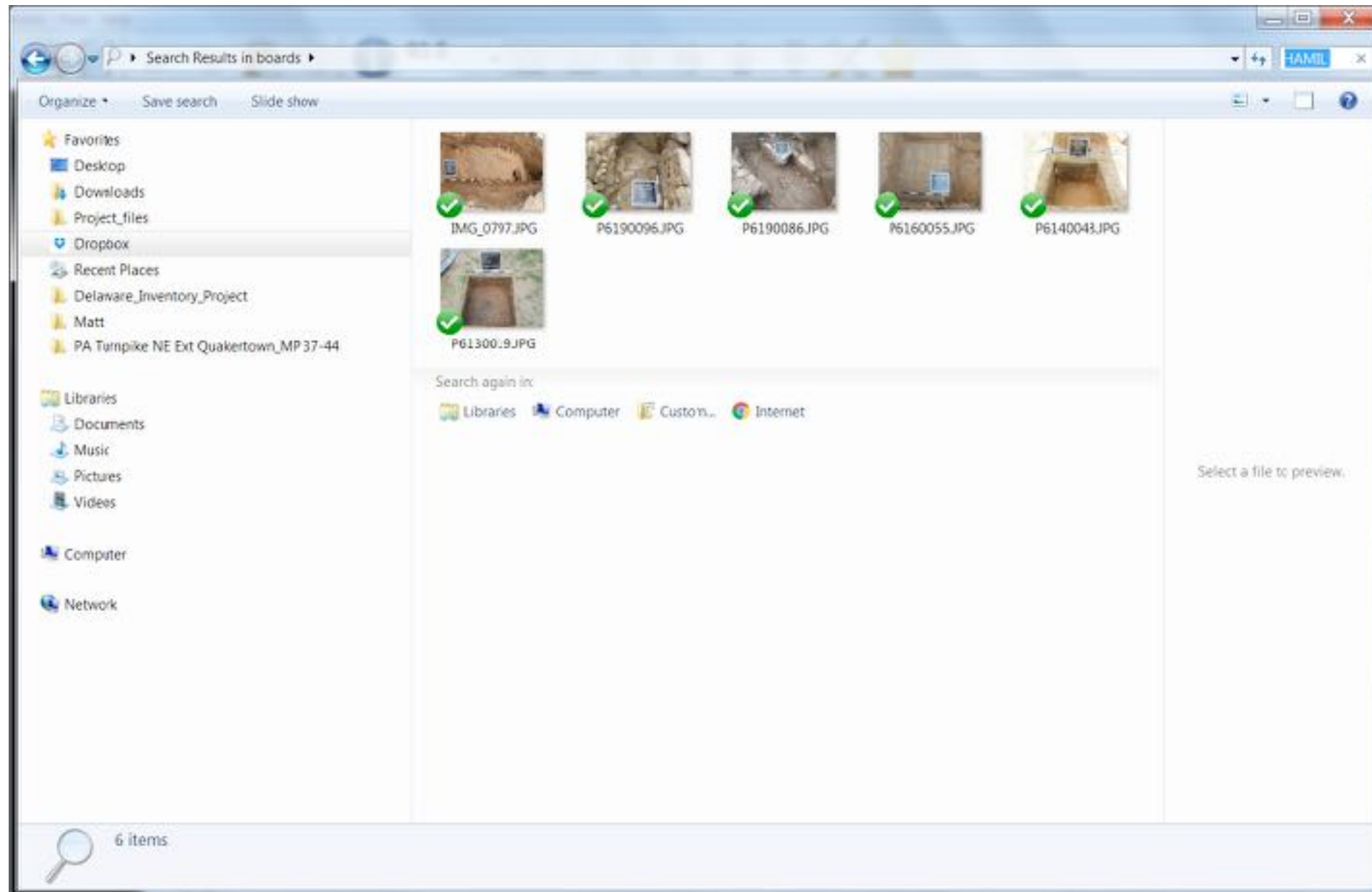
Documentation of Tags

Datatable of tags per image, add all of there data & metadata from here

Search by Tags

Example of OS X search by image tag derived from GoogleVision

Search by Tags

Example of Windows search by image tag derived from GoogleVision