



# Introduction to Optical Character Recognition (OCR) Software

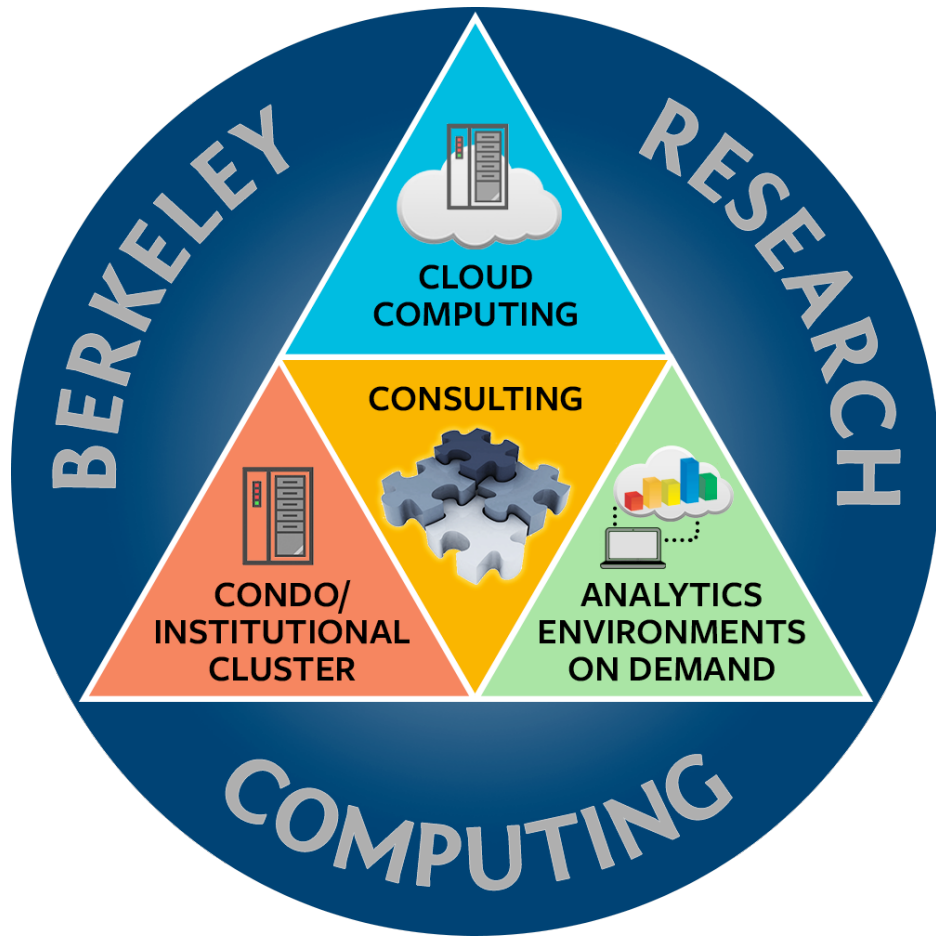
D-Lab workshop, February 28, 2017

Quinn Dombrowski, DH Coordinator, Research IT, [quinnd@berkeley.edu](mailto:quinnd@berkeley.edu)

Research IT

Advancing Research@Berkeley

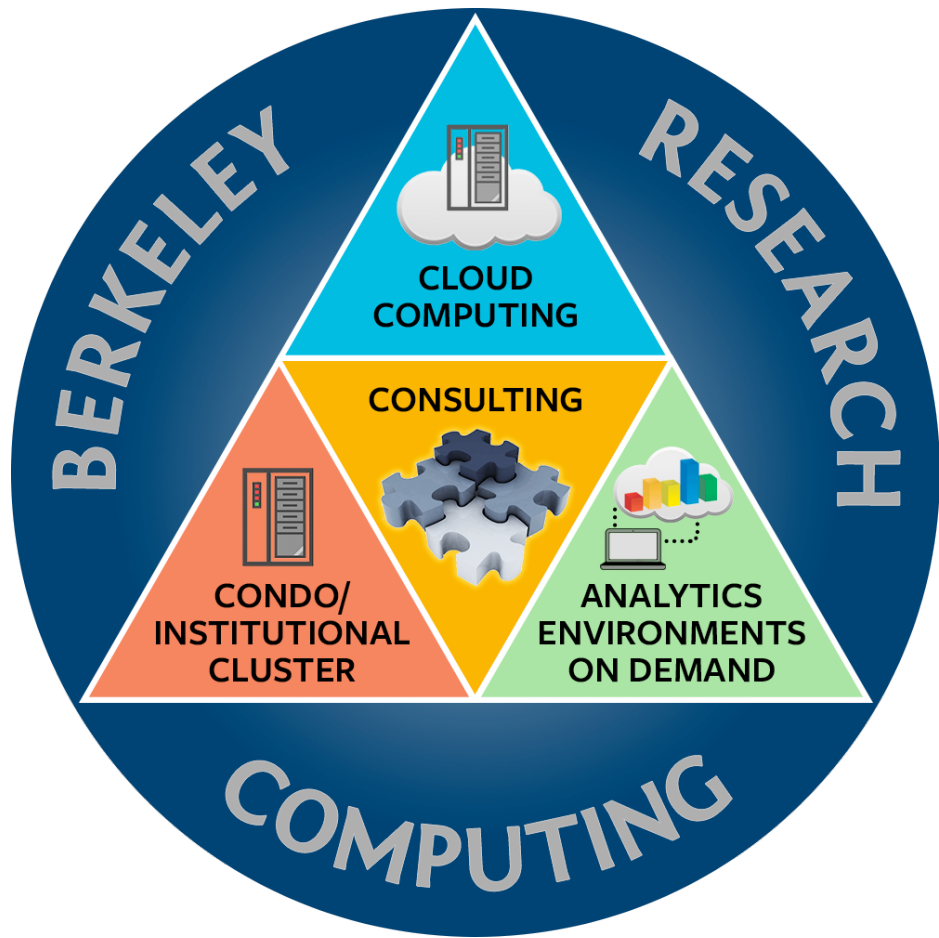
# Berkeley Research Computing



Research IT

Advancing Research@Berkeley

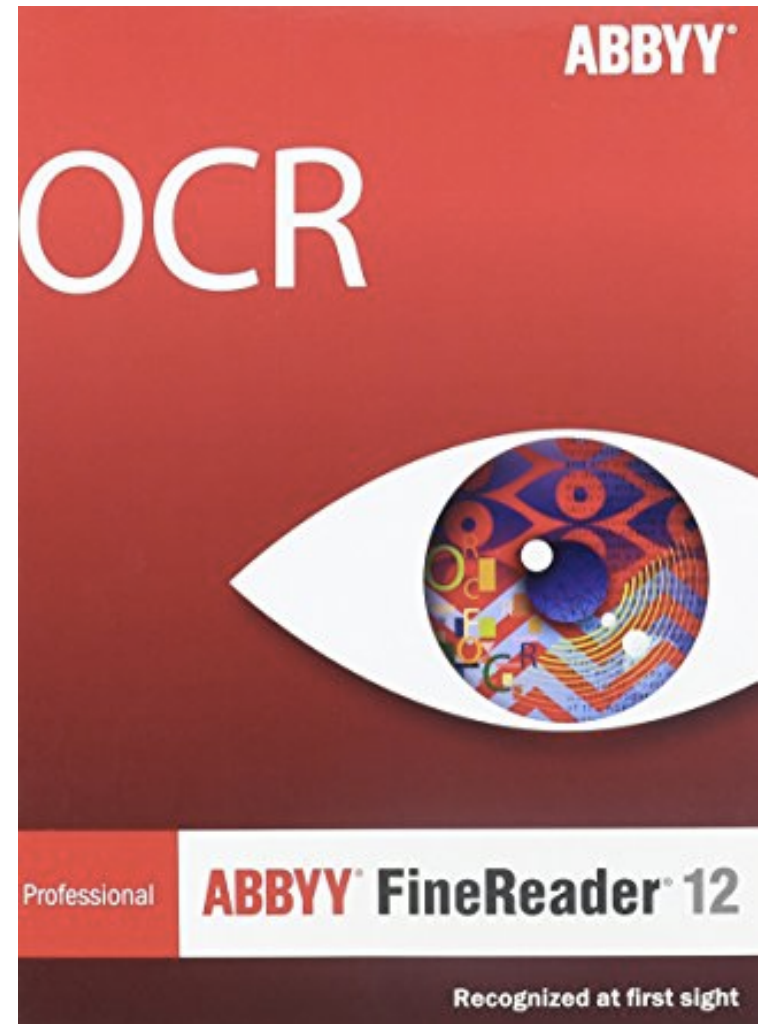
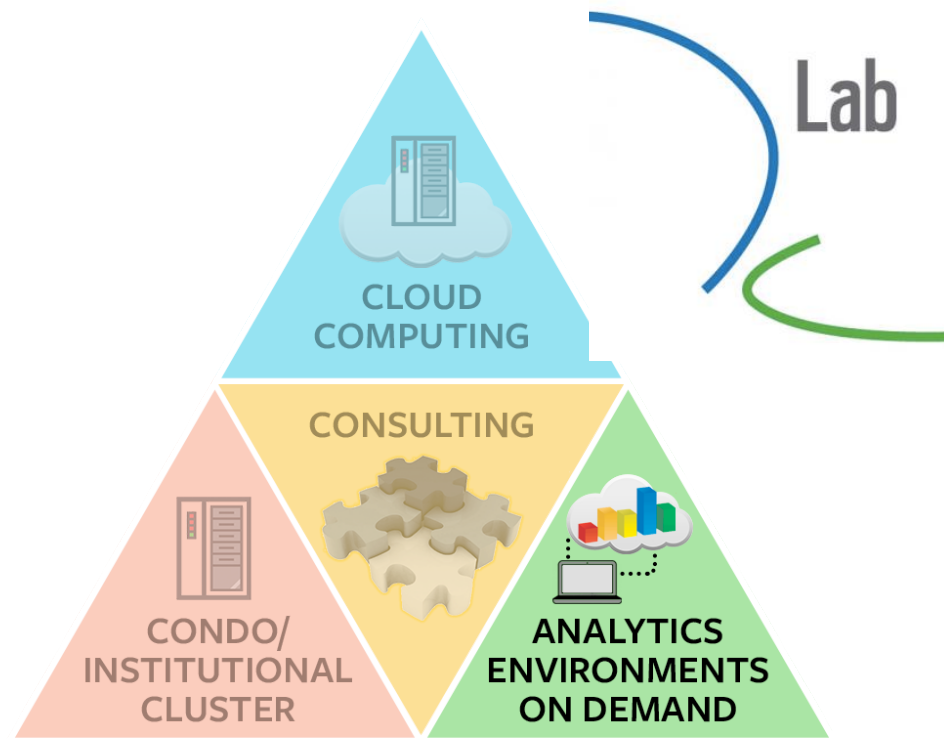
# Berkeley Research Computing + DH



Research IT

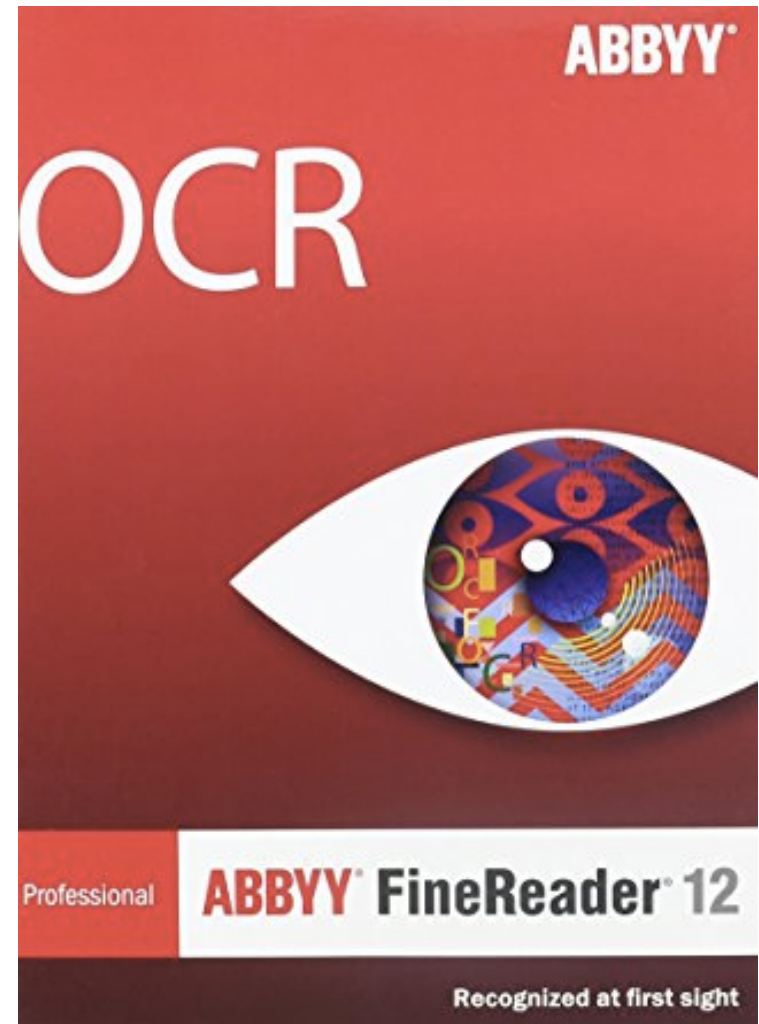
Advancing Research@Berkeley

# Optical Character Recognition (OCR)



# Optical Character Recognition (OCR)

- Complex layout
- Non-English (includes Chinese, Japanese, Korean, Cyrillic, Greek, Arabic, Hebrew)
- Multilingual
- **Precision**



# When precision matters

Untitled document [1] - ABBY FineReader 10 Professional Edition

File Edit View Document Page Areas Tools Help

New Task Open Scan Read Save Editable copy

Pages Document Language English

Image Edit Image Analyze Text Picture

Text Check Spelling Next Error Previous Error

**Nasdaq & AMEX**

Stocks in bold rose or fell 5% or more

Track your investments with our continuously updated stocks. Visit us on the web at [money.usatoday.com](http://money.usatoday.com)

32-week High				32-week Low			
Low	Stock	Last Change	High	Low	Stock	Last Change	High
9.19	ABX Air n	7.52	-0.10	45.71	Biomet	36.71	-0.42
33.25	ACMoore	13.58	-1.57	2.76	Biomira	1.46	+0.03
31.38	ADA-ES	20.96	+3.16	9.07	BioScrip	8.05	0.34
27.14	ADC Tel rs	23.21	+0.13	8.50	BirchMt gn	50.0	4.57
30.40	ADECP	27.32	+0.73	18.21	BioTech T	204.66	-0.84
16.45	AFC Ent s	15.40	-0.14	52.73	BluCoat	41.29	+0.70
8.37	ASE Tel	7.76	+0.40	44.35	BlueNile	40.30	-1.10
19.25	ASML Hld	21.24	+0.46	26.45	BobEvn	22.99	...
27.38	ASV Inc s	26.76	+0.14	15.94	Bookham	5.94	+0.06
19.82	ATI Tech	17.89	+0.68	11.80	Borland	6.68	+0.14
33.62	ATMI Inc	29.95	+1.29	31.90	BostPrv	31.18	-0.07
39.20	ATP O&G	38.40	-0.39	18.42	BttmInt	11.53	+0.20
4.24	AVI Bio	3.62	-0.22	14.68	BrigExp	12.10	-0.23
				46.72	BrightHrz s	38.90	-0.80

52-week High Low Stock Last Change

-- A --				52-week High	52-week Low	Stock	Last Change	
9.19	6.89	ABX Air n	7.52	-0.10	45.71	Biomet	36.71	-0.42
33.25	12.40	ACMoore	13.58	-1.57	2.76	Biomira	1.46	+0.03
31.38	13.51	ADA-ES	20.96	+3.16	9.07	BioScrip	8.05	0.34
27.14	12.88	ADC Tel rs	23.21	+0.13	8.50	BirchMt gn	50.0	4.57
30.40	16.70	ADECP	27.32	+0.73	18.21	BioTech T	204.66	-0.84
16.45	10.47	AFC Ent s	15.40	-0.14	52.73	BluCoat	41.29	+0.70
8.37	4.50	ASE Tel	7.76	+0.40	44.35	BlueNile	40.30	-1.10
19.25	12.75	ASML Hld	21.24	+0.46	26.45	BobEvn	22.99	...
27.38	16.39	ASV Inc s	26.76	+0.14	15.94	Bookham	5.94	+0.06
19.82	10.47	ATI Tech	17.89	+0.68	11.80	Borland	6.68	+0.14
33.62	20.53	ATMI Inc	29.95	+1.29	31.90	BostPrv	31.18	-0.07
39.20	16.76	ATP O&G	38.40	-0.39	18.42	BttmInt	11.53	+0.20
4.24	1.99	AVI Bio	3.62	-0.22	14.68	BrigExp	12.10	-0.23
					46.72	BrightHrz s	38.90	-0.80

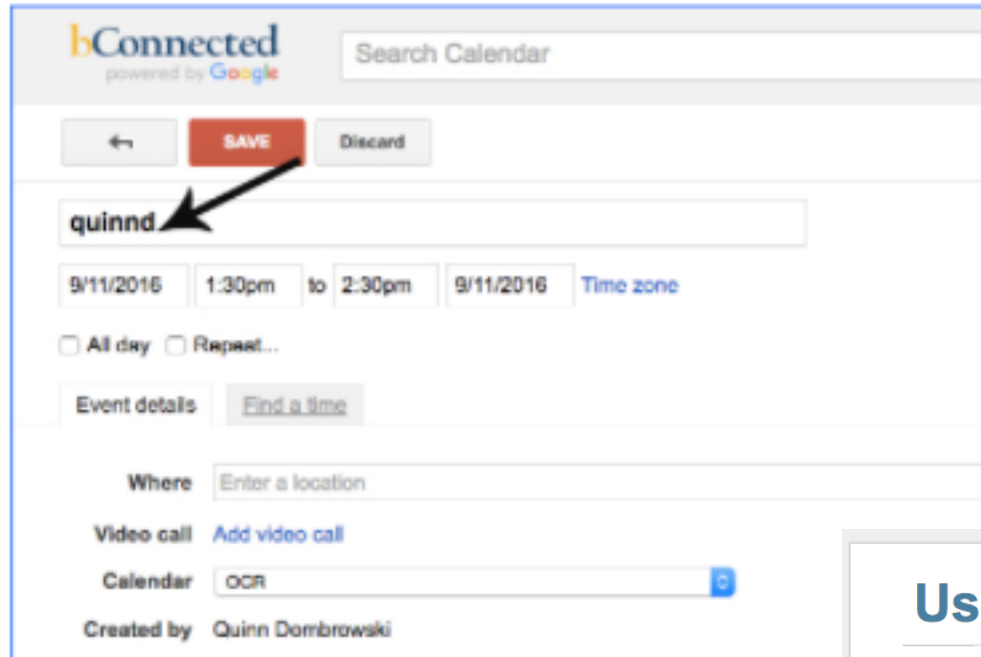
52-week High Low Stock Last Change

Low	Stock	Last Change	52-week High	52-week Low	Stock	Last Change
45.7	32.5	Blomet	36.7	-0.42		
2.7	1.20	Biomira	1.46	+		
6	5.13	BioScrip	8.05	0.34		
9.07	50.6	BirchMt gn	50.0	4.57		
68.81	5	BioTech T	204.66	-0.84		
212.	131.	BluCoat	41.29	+0.70		
25	03	BlueNile	40.30	-1.10		
8.5	1.4	BobEvn	22.99	...		
0	0	Bookham	5.94	+0.06		
18.2	10	Borland	6.68	+0.14		
1	73	BostPrv	31.18	-0.07		
82.7	13.8	BttmInt	11.53	+0.20		
3	6	BrigExp	12.10	-0.23		
44.3	24.1	BrightHrz s	38.90	-0.80		
5	5					
26.4	19					

78% 81%

Click here to view zoomed image (Ctrl+F5).

# Google Calendar + Docs + Forms



Sign up for Research IT's experimental OCR virtual research desktop

\* Required

Your name \*

Your answer

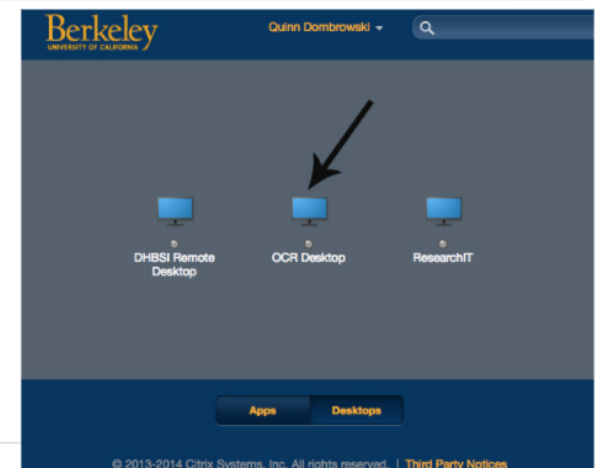
Email address \*

Your answer

## Using FineReader via Citrix

Double-click on **OCR Desktop** to connect to a remote Microsoft Windows computer that has **ABBYY FineReader** installed.

*(You will probably only have access to one icon).*



Research IT  
Advancing Research@Berkeley

# OCR at scale



Postdoc Adam Anderson

+





# OCR at scale



An OCR Engine that was developed at HP Labs between 1985 and 1995... and now at Google.

Project Home | [Downloads](#) | [Wiki](#) | [Issues](#) | [Source](#)

Summary | [Updates](#) | [People](#)

## Project Information

★ Starred by 1864 users

[Activity](#) High  
[Project feeds](#)

Code license  
[Apache License 2.0](#)

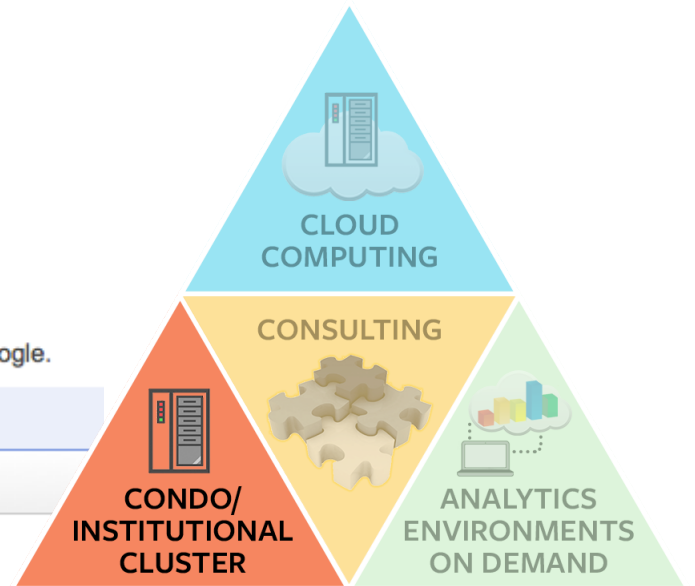
Labels  
OCR, Utility, CPlusPlus, Google

## Background

The Tesseract OCR engine was one of the top 3 engines in the 1990s but it is probably one of the most accurate open source OCR engines. Image input is managed by the [Leptonica Image Processing Library](#).

## Important Download Information:

The language data files are separate from the engine.



# OCR at scale

## Convert all pdf files in working directory to images

```
In [21]:  
  
def runGhostscript(pdfFile):  
    print("filename: ", pdfFile)  
    if filename.endswith(".pdf") :  
        name, extension = os.path.splitext(pdfFile)  
  
        # Ghostscript executable is inside the container.  
        # TEMPLATE: gs -dBATCHE -dNOPAUSE -dQUIET -sDEVICE=pdfwrite  
        GHOSTSCRIPTCMD = 'gs -dBATCHE -dNOPAUSE -dQUIET -sDEVICE=pdfwrite  
        gcmd = GHOSTSCRIPTCMD.format(tesseractScratchDataDirectory, pdfFile)  
  
        #  
        # convert pdf to png  
        #  
        print("singularity cmd: ", SINGULARITYCMD)  
        print("gs cmd: ", gcmd)  
        #result = subprocess.call(GHOSTSCRIPTCMD)  
        result = !SINGULARITYCMD $gcmd  
        print("gs result: ", result)
```

```
In [22]:  
  
from multiprocessing import Pool  
  
pdffileList = []  
for filename in os.listdir(scratchDataDirectory):  
    print("filename: ", filename)  
    if filename.endswith(".pdf") :  
        pdffileList.append(filename)  
  
        name, extension = os.path.splitext(filename)  
        pdfnamelist.append(name)  
  
print("pdffileList: ", pdffileList)  
print("filenameList: ", pdfnamelist )  
  
#  
# multiprocessing the pdf to png work work  
#  
pool0 = Pool(20)  
pool0.map(runGhostscript, pdffileList)  
pool0.close()  
pool0.join()
```

```
filename: OAAS_1_-_OA_Bib_2003.pdf  
filename: OACC_1976.pdf  
pdffileList: ['OAAS_1_-_OA_Bib_2003.pdf', 'OACC_1976.pdf']
```

## Run tesseract on all image files in the working directory

```
def runTesseract(imagefile):  
    print("imagefile : ", imagefile)  
  
    # template: tesseract --tessdata-dir /opt/tessdata /scratch/germanocr_Page_01.png germanocr  
    TCMD = 'tesseract --tessdata-dir /opt/tessdata {}{} {}{} -l eng'  
    #  
    # ocr the png  
    #  
    basename, ext = os.path.splitext(imagefile)  
    tcmd = TCMD.format(tesseractScratchDataDirectory, imagefile, tesseractScratchDataDirectory)  
    print("tesseract cmd: ", tcmd)  
    #print("singularity cmd: ", SINGULARITYCMD)  
  
    result = !SINGULARITYCMD $tcmd  
    print("tesseract result: ", result)  
    TCMD = 'tesseract --tessdata-dir /opt/tessdata {}{} {}{} -l eng'
```

```
from multiprocessing import Pool  
  
imageList = []  
for imagename in os.listdir(scratchDataDirectory):  
    if imagename.endswith(".png"):  
        imageList.append(imagename)  
  
#  
# multiprocessing the ocr work  
#  
pool = Pool()  
pool.map(runTesseract, imageList)  
pool.close()  
pool.join()
```

```
imagefile : OACC_1976-175.png  
imagefile : OACC_1976-112.png  
imagefile : OAAS_1_-_OA_Bib_2003-94.png  
imagefile : OAAS_1_-_OA_Bib_2003-78.png  
imagefile : OACC_1976-68.png  
imagefile : OAAS_1_-_OA_Bib_2003-67.png  
imagefile : OAAS_1_-_OA_Bib_2003-85.png  
imagefile : OACC_1976-4.png  
imagefile : OACC_1976-129.png  
imagefile : OACC_1976-35.png  
tesseract cmd: tesseract --tessdata-dir /opt/tessdata /scratch/OACC_1976-175.png /scratch/OACC_1976-175 -l eng  
imagefile : OAAS_1_-_OA_Bib_2003-9.png
```

# Tesseract with complex layouts

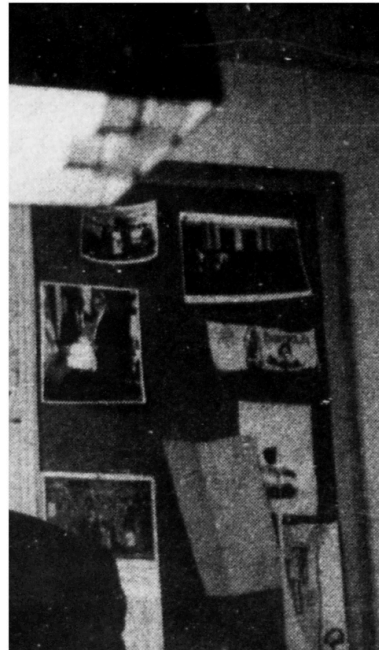
ified that rabbits could make up no more than five percent of her sales volume.

Then last summer, Crymes continues, "I was called to the office. [The Rouse leasing agent] said to me, 'We are tired of bears. We want you to get rid of all your bears by October 1.' I said, 'Excuse me, but this is July 28. That's a little sudden.' They said, 'Well, we think bears are passe, and your sales have been down.' I have to report [sales figures to management] every month, and sales were down, but they were down for a lot of people in Harborplace. It was not just me, although at the time I didn't really know it."

Crymes states that she had always been a good tenant and had been active in the Harborplace Merchants Association, and that as a result Rouse officials "said, 'We would like you to stay, but we'd like you to come up with a new idea, like a store selling sweatshirts or sweaters.'" Management offered other suggestions

ified that rabbits could make up no more than five percent of her sales volume.

Then last summer, Crymes continues, "I was called to the office. [The Rouse leasing agent] said to me, 'We are tired of bears. We want you to get rid of all your bears by October 1.' I said, 'Excuse me, but this is July 28. That's a little sudden.' They said, 'Well, we think bears are passe, and your sales have been down.' I have to report [sales figures to management] every month, and sales were down, but they were down for a lot of people in Harborplace. It was not just me, although at the time I didn't really know it."



## Tut-Tut, It's Putt-Putt

### THE BROKERAGE BUILDS A MINI-GOLF COURSE

IT WAS INEVITABLE THAT THE ALMOST sublime campiness of miniature golf—the flamingos, the windmills, the lighthouses—would one day metamorphose into high art. Yesterday's tacky is today's tacky in a socio-economic context—a real or hyper-real reflection of the sensibility of the time. Ergo art. Right? Not necessarily bearing this in mind, the people at the Brokerage nonetheless decided to leap on the miniature golf bandwagon (who can blame them?) and open Baltimore's first indoor, year-round miniature golf course. To stir some excitement, they dreamt up a design-and-construct-your-own-golf-hole competition.

a; mi  
" U;  
,\_ wt «'0'  
la'mfamwfi-ww"  
l.  
4!;  
i.  
'1  
W1», 1,  
3  
i  
.3.  
'

Research IT

Advancing Research@Berkeley

# Early modern OCR



<http://emop.tamu.edu/>

*Arcadia* LIB. I.  
friendship between riuals, and beaurie taught the beholders chastitie? He was going on with his praises, but *Stephan* bad him stay, and looke: and so they both perceiued a thing which floated drawing nearer and nearer to the banke; but rather by the fauourable working of the Sea, then by any selfe industrie: They doubted a while what it should be, till it was cast vpon euē hard before them: at which time they fully saw that it was a man. Vherupon running for pitie sake vnto him, they found his hands (as it should appeare, constant friends to his life than his memorie) fast griping vpon the edge of a square small coffer, which lay all vnder his breast: els in himselfe no shew of life, so as the boord seemed to be but a beere to carrie him aland to his  
10 Sepulcher. So drew they vp a yong man of so goodly shape, and well pleasing fauour, that one would thinke death had in him a louely countenance; and, that though he were naked, nakednesse was to him an apparell. That sight increasēd their compassion; and their compassion called vp their care; so that lifting his feete about his head, making a great deale of late water come out of his mouth, they layd him vpon  
15 some of their garments, and fell to rub and chafe him, til they brought him to recover both breath the seruant; and warmth the companion of liuing. At length opening his eyes, he gaue a great groan; a doleful note, but a pleasant dittie: for by that

ill) h mgm ne m  
adii:iiot bigg in meGXWa'-  
owne a monfter' kind more " z  
ti be aeth n  
hischrefl':f ety ;fh lm lil

twal m abi? id an  
' which if muchm dw- ldo not deceiue meas -  
i -'Worth)f to bc a fan [X]?1ngl:'?a gallater offendpr T his ay  
p fgx:!?tjhëcaufe!tt

may hc enartfo;ontofaysh?mn becaulëitw- tibet-uerfo.-  
Readëz:tthell atjymmldlahm?sesstrlhefflllgës your good  
- melaminst l lind-hamamde mt, but ap e at. Apd-  
lea solon saphetterrtuli, thanlasln a, a e  
r ftmp gia a'm fsathars'eymu willen-onti- ue tolouethe  
V and mostë

Research IT

Advancing Research@Berkeley



# Thanks. Questions?

For more information visit

[research-it.berkeley.edu/ocr](https://research-it.berkeley.edu/ocr)

Email:

[quinnd@berkeley.edu](mailto:quinnd@berkeley.edu)

**Research IT**

Advancing Research@Berkeley