DATA FROM PDFS: LET'S GET IT!

Beyond the Numbers Conference

November 10, 2022

Christine Murray

Bates College
Lewiston, Maine

OUTLINE

1. The Objective
2. Adobe Acrobat
3. Tabula
4. Camelot/Excalibur
5. Amazon Textract
6. Summing Up

THE OBJECTIVE

THE PROBLEM

PDFs are great for humans to read. But not for machines.

When researchers want to analyze the data trapped in a PDF, what tools can we recommend?

CRITERIA

What we're looking for in a PDF-to-tabular-data tool:

1. Easy to use.
2. Accurate output.
3. Free to low cost.
4. Works with different kinds of PDFs.
5. Provides ability to automate.

Extra credit: built-in OCR, output in open file formats, recognition of non-Roman characters, and good documentation.

OUR PDF SAMPLES

Tools were tested with a sample of PDFs representing real world situations:

1. Born digital
    - Simple table structure.
    - Complex table structure.
2. Scanned
    - With OCR
    - Without OCR (image)
    - Bad scan

BORN DIGITAL DOCUMENTS

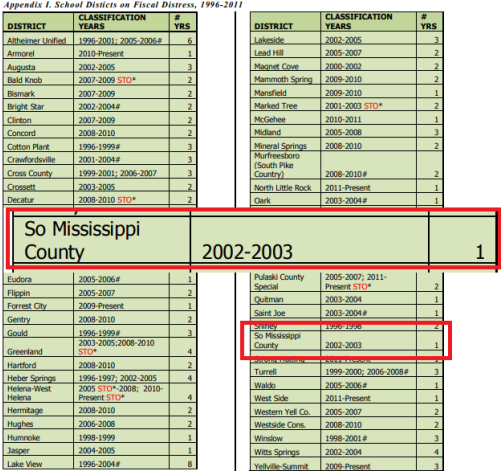
http://muskie.usm.maine.edu/Publications/WorkingWaterfronts.pdf

Table 1
PORT AND HARBOR FACILITIES ON THE MAINE COAST

Municipality	Boat Launch or Ramp	Commercial Land/ Buildings	Float System	Incomplete	Marina or Yacht Club	National Park	Other	State or Municipal Park	Wharf or Pier	Total
Addison	2									2
Bangor			1						2	3
Bar Harbor	2	2			1				7	12
Bath	1	3			2				4	10
Belfast									6	6
Biddeford	3		1		2					6
Blue Hill	3								3	6
Boothbay	2	1	2						6	11
Boothbay Harbor	1				3				24	28
Bremen									3	3
Brewer									2	2
Bristol	1	3							7	11
Brooksville					1				1	2

BORN DIGITAL DOCUMENTS

https://scholarworks.uark.edu/oepbrief/57/



Appendix I. School Districts on Fiscal Distress, 1996-2011

DISTRICT	CLASSIFICATION YEARS	# YRS
Altheimer Unified	1996-2001; 2005-2006#	6
Armorel	2010-Present	1
Augusta	2002-2005	3
Bald Knob	2007-2009 STO*	2
Bismark	2007-2009	2
Bright Star	2002-2004#	2
Clinton	2007-2009	2
Concord	2008-2010	2
Cotton Plant	1996-1999#	3
Crawfordsville	2001-2004#	3
Cross County	1999-2001; 2006-2007	3
Crossett	2003-2005	2
Decatur	2008-2010 STO*	2
Eudora	2005-2006#	1
Flippin	2005-2007	2
Forrest City	2009-Present	1
Gentry	2008-2010	2
Gould	1996-1999#	3
Greenland	2003-2005;2008-2010 STO*	4
Hartford	2008-2010	2
Heber Springs	1996-1997; 2002-2005	4
Helena-West Helena	2005 STO*-2008; 2010-Present STO*	4
Hermitage	2008-2010	2
Hughes	2006-2008	2
Humnoke	1998-1999	1
Jasper	2004-2005	1
Lake View	1996-2004#	8

DISTRICT	CLASSIFICATION YEARS	# YRS
Lakeside	2002-2005	3
Lead Hill	2005-2007	2
Magnet Cove	2000-2002	2
Mammoth Spring	2009-2010	2
Mansfield	2009-2010	1
Marked Tree	2001-2003 STO*	2
McGehee	2010-2011	1
Midland	2005-2008	3
Mineral Springs	2008-2010	2
Murfreesboro (South Pike Country)	2008-2010#	2
North Little Rock	2011-Present	1
Oark	2003-2004#	1
Pulaski County Special	2005-2007; 2011-Present STO*	2
Quitman	2003-2004	1
Saint Joe	2003-2004#	1
Shirley	1996-1998	2
So Mississippi County	2002-2003	1
Turrell	1999-2000; 2006-2008#	3
Waldo	2005-2006#	1
West Side	2011-Present	1
Western Yell Co.	2005-2007	2
Westside Cons.	2008-2010	2
Winslow	1998-2001#	3
Witts Springs	2002-2004	4
Yellville-Summit	2009-Present	3

So Mississippi County 2002-2003 1

SCANNED, WITH OR WITHOUT OCR

https://catalog.hathitrust.org/Record/101795524

receipts averaged $75 million.

Table 3. Estimated quantity of potatoes sold, season average price received by farmers and value, Maine, 1959 to 1974[1]

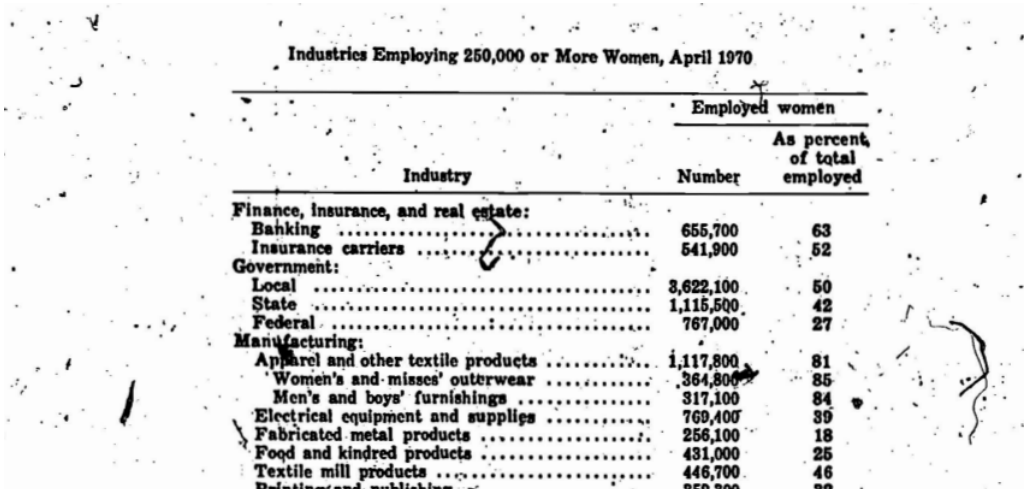
Year	Quantity sold[2]	Price per Cwt.	Sales value
	1000 cwt.	dollars	1000 dollars
1959	30,482	2.32	70,718
1960	30,254	1.36	41,145
1961	32,814	1.13	37,080
1962	34,754	1.22	42,400
1963	33,478	1.93	64,613
1964	34,599	3.83	132,514
1965	31,582	2.36	74,534
1966	33,810	1.70	57,477
1967	32,513	1.36	44,218
1968	31,652	1.85	58,556
1969	30,040	2.20	66,088
1970	30,810	1.98	61,004
1971	32,403	1.70	55,085
1972	28,872	4.10	118,375
1973	25,260	7.25	183,135
1974	29,582	2.90	85,788

[1] *Potatoes and Sweet Potatoes*, CRB-SRS, U.S. Department of Agriculture, Stat. Bul. 409, July 1967, for crops for 1959-1964. Annual issues of *Potatoes and Sweet Potatoes*, Pot. (6) Aug. for 1965 and years that follow.

[2] Consists of potatoes sold for all purposes including food, seed, processing, and livestock feed.

SCANNED, BADLY

https://eric.ed.gov/?id=ED068010



Industries Employing 250,000 or More Women, April 1970

Industry	Employed women	
	Number	As percent of total employed
Finance, insurance, and real estate:		
Banking	655,700	63
Insurance carriers	541,900	52
Government:		
Local	3,622,100	50
State	1,115,500	42
Federal	767,000	27
Manufacturing:		
Apparel and other textile products	1,117,800	81
Women's and misses' outerwear	364,800	85
Men's and boys' furnishings	317,100	84
Electrical equipment and supplies	769,400	39
Fabricated metal products	256,100	18
Food and kindred products	431,000	25
Textile mill products	446,700	46

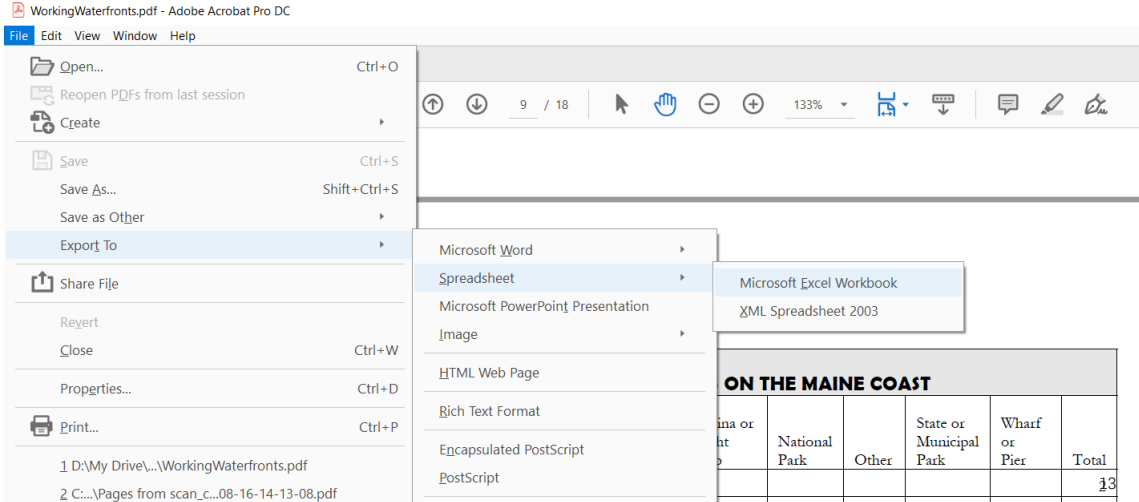
WHAT I LOOKED AT

1. Adobe Acrobat
2. Tabula
3. Camelot & Excalibur
4. Amazon Textract

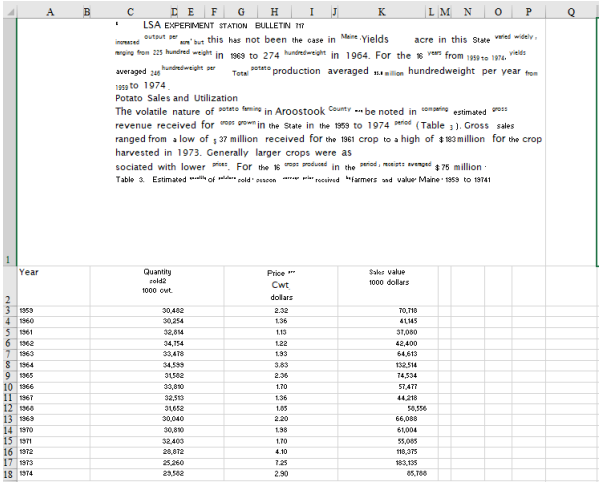
Other tools include ABBYY FineReader, PDF Table Extractor, pdfminer.six, pdfplumber, SLICEmyPDF, and PDFTables

ADOBE ACROBAT

ADOBE ACROBAT



ADOBE ACROBAT



LSA EXPERIMENT STATION BULLETIN 717

increased output per acre, but this has not been the case in Maine. Yields per acre in this State varied widely, ranging from 225 hundredweight in 1963 to 274 hundredweight in 1964. For the 16 years from 1959 to 1974, yields averaged 248 hundredweight per acre. Total potato production averaged 11.4 million hundredweight per year from 1959 to 1974.

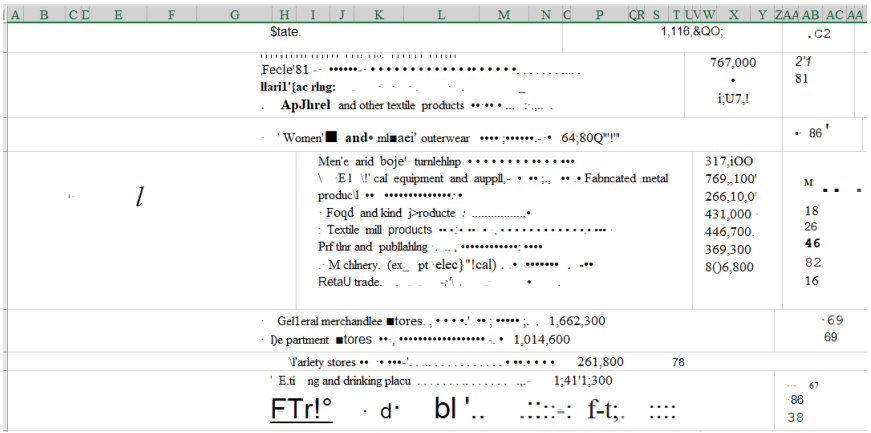
Potato Sales and Utilization

The volatile nature of potato farming in Aroostook County may be noted in comparing estimated gross revenue received for crops grown in the State in the 1959 to 1974 period (Table 3). Gross sales ranged from a low of $ 37 million received for the 1961 crop to a high of $ 183 million for the crop harvested in 1973. Generally larger crops were associated with lower prices. For the 16 crops produced in the period, receipts averaged $ 75 million.

Table 3. Estimated quantity of potatoes sold, season average price received by farmers and value, Maine, 1959 to 1974[1]

Year	Quantity sold[2] 1000 cwt.	Price per Cwt dollars	Sales value 1000 dollars
1959	30,462	2.32	70,716
1960	30,254	1.36	41,145
1961	32,614	1.13	37,080
1962	34,754	1.22	42,400
1963	33,478	1.93	64,613
1964	34,589	3.83	132,514
1965	31,562	2.36	74,534
1966	33,810	1.70	57,477
1967	32,513	1.36	44,218
1968	31,652	1.65	52,556
1969	30,040	2.20	66,088
1970	30,810	1.98	61,004
1971	32,403	1.70	55,085
1972	28,872	4.10	118,375
1973	25,260	7.25	183,135
1974	23,562	2.90	65,786

ADOBE ACROBAT: IT TRIED



ADOBE ACROBAT: THE VERDICT

Pros

- Easy to use for the point-and-clickers.
- Pretty accurate, especially for born digital PDFs but also clear scans.
- Built-in OCR.
- Options for non-Roman characters.
- You may already have it installed.

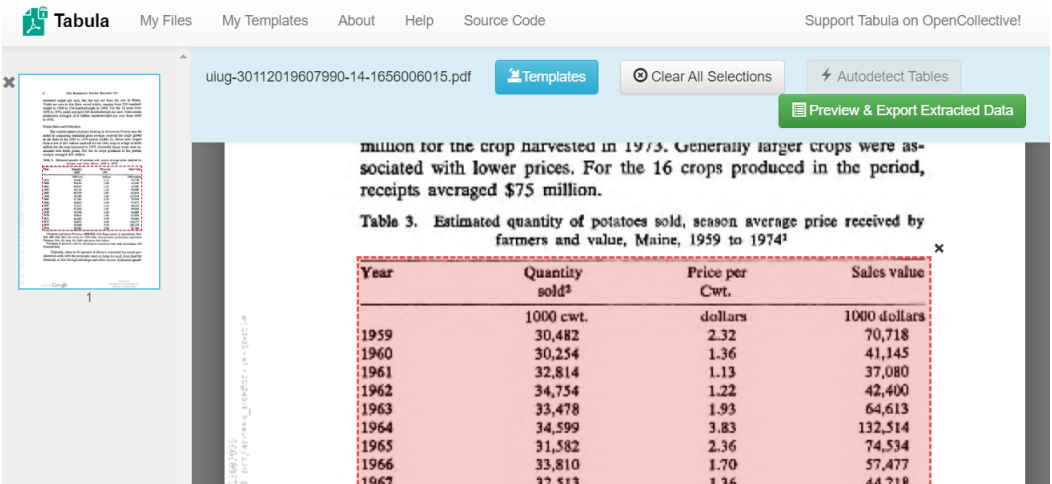
Cons

- Proprietary, requiring a license.
- No CSV output option.
- Exports non-table content along with the table.
- No option to automate.

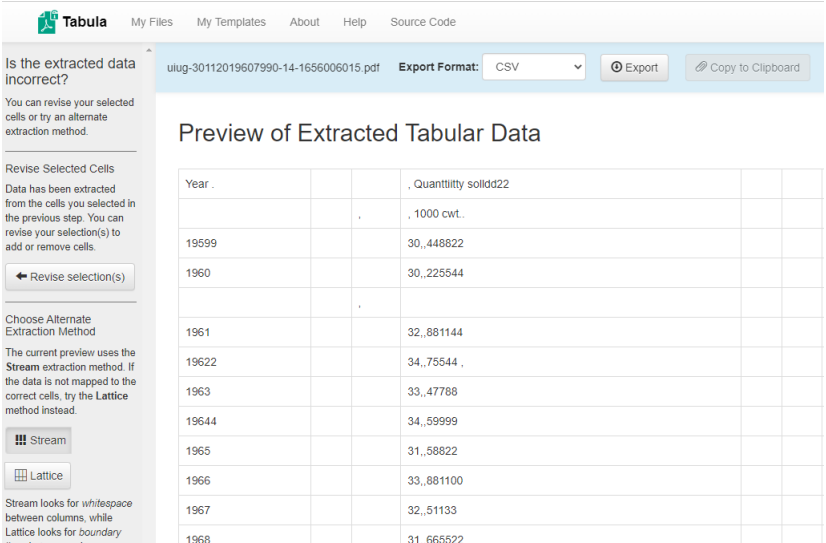
TABULA

TABULA

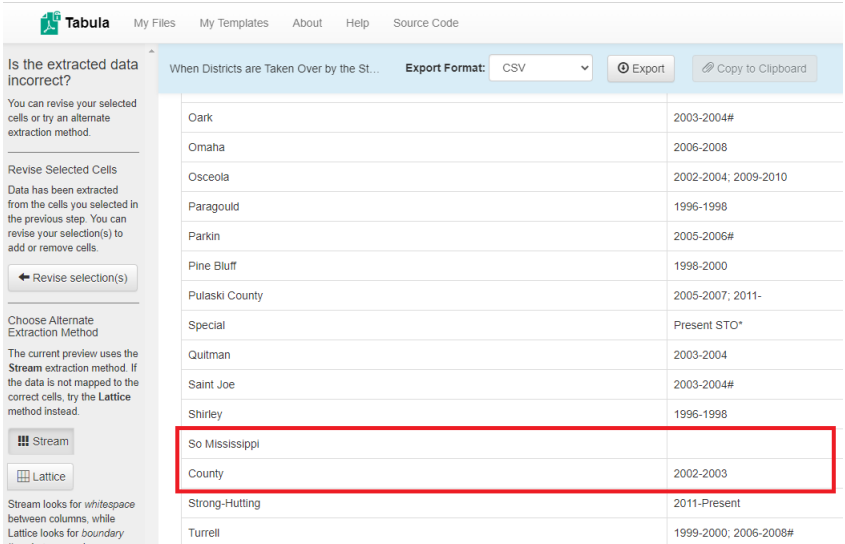
https://tabula.technology/ Tabula is a Java program that runs in your browser.



TABULA: STREAM OR LATTICE



TABULA: CELL PROBLEMS



TABULA: THE VERDICT

Pros

- Free, open source.
- Easy to install and use.
- Pretty accurate, especially for PDFs with simple table layout.
- Also has options for Python or R.

Cons

- No OCR, so that must be done separately.
- Had some issues with complex table layout.
- Minimal documentation.
- Not in active development.

CAMELOT/EXCALIBUR

CAMELOT

Camelot is a Python library. **https://camelot-py.readthedocs.io/en/master/**

In [45]:
```
tables = camelot.read_pdf('working-waterfronts.pdf', pages='1,2' , strip_text='\n')
tables
df = tables[0].df[2:]
colnames = tables[0].df.values[1]
df.columns = colnames
df
```

Out[45]:

	Municipality	Boat Launch or Ramp	Commercial Land/ Buildings	Float System	Incomplete	Marina or Yacht Club	National Park	Other	State or Municip Pa
2	Addison	2							
3	Bangor			1					
4	Bar Harbor	2	2			1			
5	Bath	1	3			2			
6	Belfast								
7	Biddeford	3		1		2			
8	Blue Hill	3							
9	Boothbay	2	1	2					

CAMELOT

In [6]:
```
tables[0].df[1:]
```

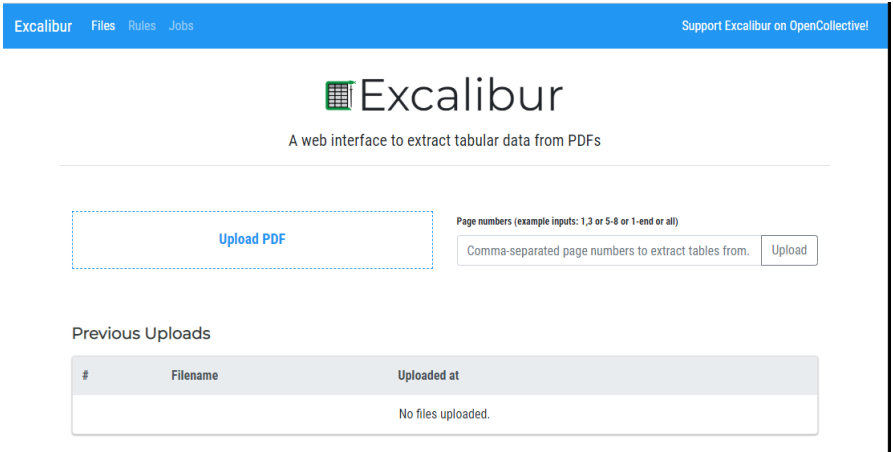
6	Bismark	2007-2009	2	Mansfield	2009-2010	1
7	Bright Star	2002-2004#	2	Marked Tree	2001-2003 STO*	2
8	Clinton	2007-2009	2	McGehee	2010-2011	1
9	Concord	2008-2010	2	Midland	2005-2008	3
10	Cotton Plant	1996-1999#	3	Mineral Springs	2008-2010	2
11	Crawfordsville	2001-2004#	3	Murfreesboro (South Pike Country)	2008-2010#	2
12	Cross County	1999-2001; 2006-2007	3			
13	Crossett	2003-2005	2	North Little Rock	2011-Present	1
14	Decatur	2008-2010 STO*	2	Oark	2003-2004#	1
15	Delaplaine	1996-1997#	1	Omaha	2006-2008	2
16	Dermott	2011-Present	1	Osceola	2002-2004; 2009-2010	3
17	Dierks	2005-2007	2	Paragould	1996-1998	2
18	Dollarway	2005-2007	2	Parkin	2005-2006#	1

In [7]:
```
print(tables[0].parsing_report)
```

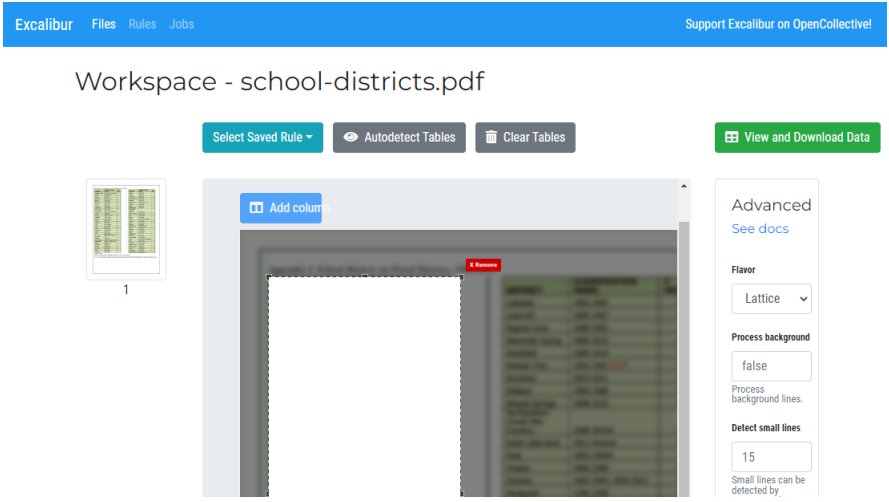
```
{'accuracy': 99.01, 'whitespace': 50.25, 'order': 1, 'page': 1}
```

EXCALIBUR

Excalibur is an interface for Camelot that runs in your browser, and functions similarly to Tabula. **https://excalibur-py.readthedocs.io/en/master/**



EXCALIBUR


Excalibur
Files
Rules
Jobs
Support Excalibur on OpenCollective!
Workspace - school-districts.pdf
Select Saved Rule
Autodetect Tables
Clear Tables
View and Download Data
1
Add colum
X Remove
Advanced
See docs
Flavor
Lattice
Process background
false
Process background lines.
Detect small lines
15
Small lines can be detected by

CAMELOT/EXCALIBUR: THE VERDICT

Pros

- Free, open source.
- Handled born digital and scanned documents well.
- Many options for tweaking table detection and cell parsing.
- Perfect for automating.
- Exports as CSV, JSON, Excel, HTML files or a SQLite database.
- Good documentation.

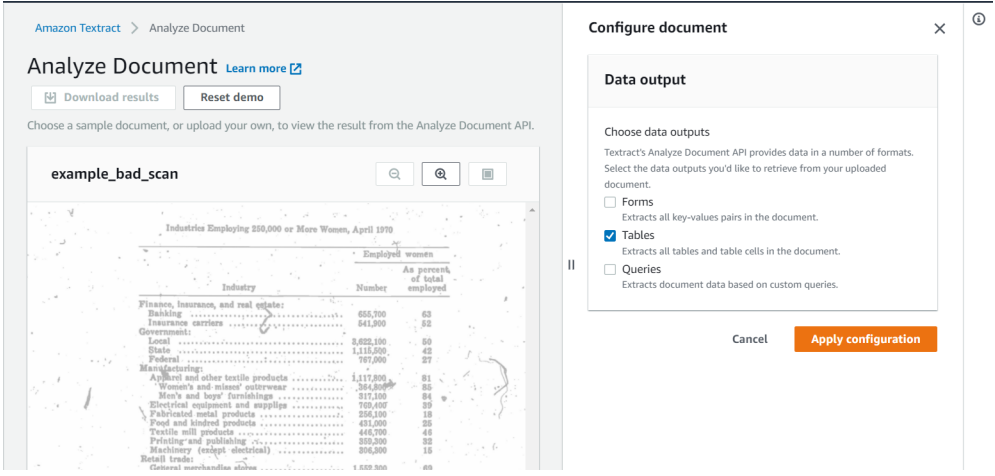
Cons

- Not easy to install, particularly Excalibur.
- Auto-detection of tables and table types seems limited. You must already know the layout of the file and page.
- No OCR, so that must be done separately.

AMAZON TEXTRACT

AMAZON TEXTRACT

https://aws.amazon.com/textract/

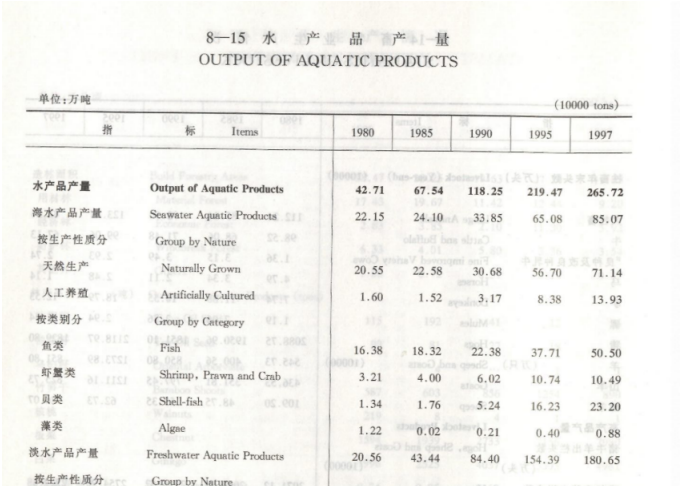


AMAZON TEXTRACT

Mariufacturing:		
Apparel and other textile products	1,117,800	81
Women's and misses' outerwear	364,806	85
Men's and boys' furnishings	317,100	84
Electrical equipment and supplies	769,400	39
Fabricated metal products	256,100	18
Foqd and kindred products	431,000	25

AMAZON TEXTRACT: WHAT ABOUT NON-ROMAN?

From the Jiangsu Statistical Yearbook of 1998:



8—15 水 产 品 产 量

OUTPUT OF AQUATIC PRODUCTS

单位：万吨 (10000 tons)

指标 Items		1980	1985	1990	1995	1997
水产品产量	**Output of Aquatic Products**	**42.71**	**67.54**	**118.25**	**219.47**	**265.72**
海水产品产量	Seawater Aquatic Products	22.15	24.10	33.85	65.08	85.07
按生产性质分	Group by Nature					
天然生产	Naturally Grown	20.55	22.58	30.68	56.70	71.14
人工养殖	Artificially Cultured	1.60	1.52	3.17	8.38	13.93
按类别分	Group by Category					
鱼类	Fish	16.38	18.32	22.38	37.71	50.50
虾蟹类	Shrimp, Prawn and Crab	3.21	4.00	6.02	10.74	10.49
贝类	Shell-fish	1.34	1.76	5.24	16.23	23.20
藻类	Algae	1.22	0.02	0.21	0.40	0.88
淡水产品产量	Freshwater Aquatic Products	20.56	43.44	84.40	154.39	180.65
按生产性质分	Group by Nature					

AMAZON TEXTRACT: WHAT ABOUT NON-ROMAN?

'#	'0001	'DE Items	'1980	'1985	'200 1990	'1995	'1997
'*****	'Output	'of Aquatic Products	'000019 4	'67.54	'118.25	'219.47	'265.72
'	'Seawater	'Aquatic Products	'22.15	'19.67 24.	'33.85	'65.08	'85.07
'	'Group	'by Nature	'6.33	'	'	'	'
'	'Naturally	'08.1 Grown	'20.55	'22.58	'30.68	'56.70	'71.14
'	'Artificially	'Cultured	'1.60	'1.52	'3.17	'8.38	'13.93
'	'Group	'by Category	'115	'192	'asluM41	'	'-
'143	'Fish	'	'16.38 000	'18.32	'22.38	'37.71	'50.50
'	'Shrimp,	'Prawn and Crab	'3.21	'4.00	'6.02	'10.74	'- 10.49
'	'Shell-fish	'	'1.34	'603 1.76	'5.24	'16.23	'23.20
'	'Algae	'	'1.22	'0.02	'0.21	'0.40	'0.88
'*******	'Freshwater	'Aquatic Products	'20.56	'43.44	'84.40	'154.39	'180.65
'	'Group	'by Nature ITOS	'	'	'	'	'
'	'Naturally	'Grown	'10.02	'10.72	'16.42.	'25.47	'30.25
'	'Artificially	'Cultured	'10.54 uns	'32.72	'67.98	'128.92	'150.40
'	'Group	'by Category	'	'biss	'gesile	'	'
'	'Fish	'	'17.95	'41.00	'79.58	'139.34	'161.17

AMAZON TEXTRACT: THE VERDICT

Pros

- Very accurate.
- Handles the weirdest PDFs.
- Demo option is free and processes up to 11 pages.
- API can be used to automate in scripts.
- Good documentation.

Cons

- Requires an AWS account.
- There is a charge beyond a certain number of documents.
- Somewhat intimidating.
- Files do not remain on your computer as they do in other options.
- No non-Roman capabilities.
- Amazon.

SUMMING UP

MY PICKS

- Adobe Acrobat for the impatient, for small jobs, and for non-Roman or not-yet-OCR'd documents (and if you already have it).
- Tabula for the impatient who don't want to pay.
- Camelot for the patient, cost-conscious power user.
- Textract for the power user with weird documents, who doesn't mind paying a little.

FURTHER READING

Anurag Bejju.
Document Intelligence: The art of PDF information extraction, January 2022.

Andreiwid Sheffer Corrêa and Pär-Ola Zander.
Unleashing tabular content to open data: A survey on pdf table extraction methods and tools.
In Proceedings of the 18th Annual International Conference on Digital Government Research, pages 54–63, 2017.

Nosheen Fayyaz, Shah Khusro, and Shakir Ullah.
Accessibility of Tables in PDF Documents: Issues, Challenges, and Future Directions.
Information Technology and Libraries, 40(3), September 2021.

Shah Khusro, Asima Latif, and Irfan Ullah.
On methods and tools of table detection, extraction and annotation in PDF documents.
Journal of Information Science, 41(1):41–57, February 2015.

Gustav Rosén.
Analysis of Tabula: A PDF-Table extraction tool, 2019.

Questions?