Parsing Prickly PDFs 🌵

📍 NICAR 2025 (Minneapolis, MN)

🗓️ March 6, 2025

🗣️ Jeremy Singer-Vine

🌵

Agenda

0. Why this workshop? (5 minutes)
1. What *is* a PDF? (5 minutes)
2. What *is* `pdfplumber` (5 minutes)
3. Let's parse some PDFs! (45 minutes)

Why this workshop?

A ton of data is published only as PDFs (🤯)

NICS Firearm Background Checks

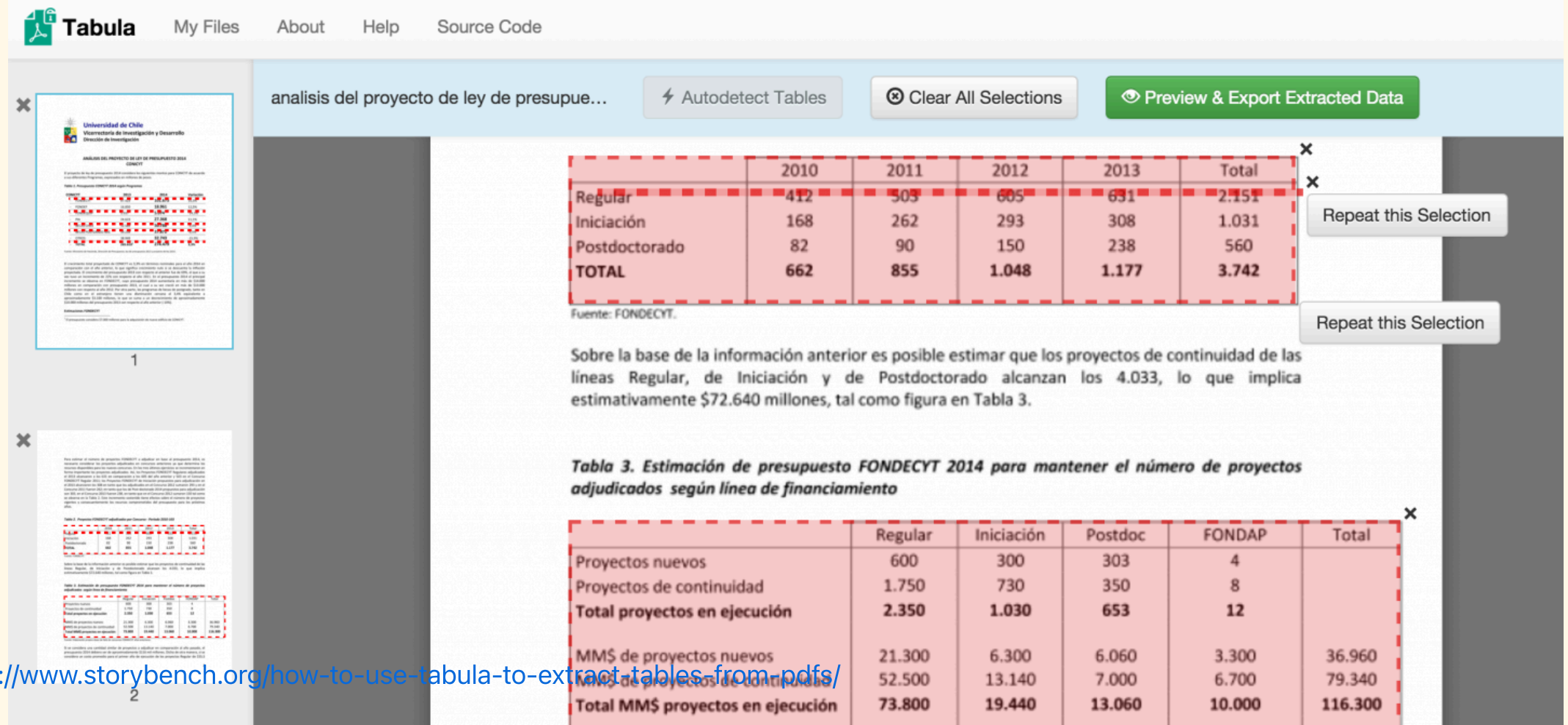
January - 2023

State / Territory	Permit	Permit Recheck	Handgun	Long Gun	*Other	**Multiple	Admin	Pre-Pawn Handgun	Pre-Pawn Long Gun	Pre-Pawn *Other	Redemption Handgun	Redemption Long Gun	Redemption *Other	Returned/Disposition Handgun	Returned/Disposition Long Gun	Returned/Disposition *Other	Rentals Handgun	Rentals Long Gun	Private Sale Handgun	Private Sale Long Gun	Private Sale *Other	Return to Seller - Private Sale Handgun	Return to Seller - Private Sale Long Gun	Return to Seller - Private Sale *Other	Totals
Alabama	13,130	229	20,028	13,674	1,403	1,042	0	12	8	2	2,215	1,004	7	74	0	0	0	0	35	26	11	1	1	0	52,902
Alaska	269	25	2,406	1,779	323	167	0	2	1	0	92	60	2	22	4	0	0	0	0	1	4	0	0	0	5,157
Arizona	8,875	1,389	17,744	8,275	1,763	997	0	9	6	1	1,211	453	6	232	26	2	0	0	9	3	0	1	1	0	41,003
Arkansas	1,944	272	6,484	6,290	485	405	6	13	6	0	979	750	2	0	0	0	0	0	7	2	0	1	2	0	17,648
California	21,900	10,343	35,181	21,735	4,949	0	0	0	0	0	621	329	31	1,444	1,044	184	0	0	7,665	3,009	611	77	36	0	109,159
Colorado	7,464	11	18,373	11,230	2,031	1,668	0	0	0	0	0	0	0	304	54	4	0	0	0	0	0	0	0	0	41,139
Connecticut	8,334	2,849	5,842	1,922	2,514	0	2	0	0	0	0	0	0	0	0	0	0	0	544	169	199	0	0	0	22,375
Delaware	434	0	2,231	1,081	86	61	0	0	0	0	15	10	0	68	0	0	0	0	99	13	1	0	0	0	4,099
District of Columbia	960	1	326	13	0	4	10	0	0	0	0	0	0	2	0	46	0	0	0	0	0	0	0	0	1,362
Florida	19,460	0	58,648	22,950	4,883	2,827	0	10	5	1	3,726	898	7	1,426	117	3	0	0	338	164	53	54	42	0	115,612
Georgia	16,453	0	18,081	9,451	997	785	0	19	4	0	2,011	706	17	72	0	0	0	0	16	4	0	0	1	0	48,617
Guam	0	0	188	79	19	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	292
Hawaii	1,847	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1,848
Idaho	6,159	69	5,234	4,415	649	265	0	1	1	0	286	246	0	91	13	0	0	0	15	15	0	0	0	0	17,459
Illinois	480,752	0	28,370	14,140	1,715	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	524,977
Indiana	372	32,711	22,825	12,560	1,912	1,092	3	6	1	0	781	302	16	43	1	0	0	0	34	28	4	0	1	0	72,692
Iowa	6,740	6,768	2,527	2,790	193	92	11	0	0	0	30	22	3	51	20	3	0	0	3	2	0	0	0	0	19,255
Kansas	1,315	403	6,507	5,577	708	417	0	3	1	3	502	225	5	81	12	3	0	0	6	6	1	1	0	0	15,776
Kentucky	716	301,242	10,744	8,081	628	603	6	7	4	0	1,488	806	4	31	9	0	0	0	35	18	4	2	0	0	324,428
Louisiana	1,879	706	10,755	7,949	870	572	0	8	1	0	910	423	14	0	0	0	0	0	15	12	4	0	0	0	24,118
Maine	522	8	3,614	3,093	352	214	0	0	0	0	59	54	2	18	6	1	0	0	8	4	0	0	0	0	7,955
Mariana Islands	0	0	8	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11
Maryland	21,495	36	12,299	5,849	116	10	0	0	0	0	82	93	1	113	18	2	0	0	0	44	6	0	1	0	40,165
Massachusetts	10,720	0	5,626	2,799	1,506	242	0	1	0	1	9	16	0	2	0	0	0	0	32	23	6	1	0	0	20,984
Michigan	18,535	6,068	21,880	15,056	2,420	915	0	0	1	2	18	195	7	57	3	0	0	0	22	9	2	1	0	0	65,191
Minnesota	11,361	32,523	9,530	8,082	1,016	494	0	0	0	0	145	146	0	77	22	1	0	0	15	8	1	0	1	0	63,422
Mississippi	3,346	0	8,153	6,557	511	373	0	10	8	0	1,366	694	1	23	1	0	0	0	5	2	0	0	0	0	21,050
Missouri	853	87	19,831	14,724	2,133	1,206	2	3	0	1	1,190	546	21	400	48	2	0	0	27	28	4	0	5	0	41,111
Montana	996	0	3,600	3,179	388	253	21	0	2	1	369	357	3	23	9	1	0	0	4	7	2	2	2	0	9,219
Nebraska	4,240	66	227	1,665	31	10	0	0	0	0	2	22	0	26	9	0	0	0	0	3	0	0	0	0	6,301
Nevada	1,265	0	4,875	2,288	386	304	0	0	0	0	259	77	0	4	1	0	0	0	463	162	28	0	0	0	10,112

🌵

There are some great tools for parsing PDF tables

For example, Tabula: 🌐 tabula.technology



	2010	2011	2012	2013	Total
Regular	412	505	605	631	2.151
Iniciación	168	262	293	308	1.031
Postdoctorado	82	90	150	238	560
TOTAL	**662**	**855**	**1.048**	**1.177**	**3.742**

Fuente: FONDECYT.

Sobre la base de la información anterior es posible estimar que los proyectos de continuidad de las líneas Regular, de Iniciación y de Postdoctorado alcanzan los 4.033, lo que implica estimativamente $72.640 millones, tal como figura en Tabla 3.

Tabla 3. Estimación de presupuesto FONDECYT 2014 para mantener el número de proyectos adjudicados según línea de financiamiento

	Regular	Iniciación	Postdoc	FONDAP	Total
Proyectos nuevos	600	300	303	4	
Proyectos de continuidad	1.750	730	350	8	
Total proyectos en ejecución	**2.350**	**1.030**	**653**	**12**	
MM$ de proyectos nuevos	21.300	6.300	6.060	3.300	36.960
MM$ de proyectos de continuidad	52.500	13.140	7.000	6.700	79.340
Total MM$ proyectos en ejecución	**73.800**	**19.440**	**13.060**	**10.000**	**116.300**

https://www.storybench.org/how-to-use-tabula-to-extract-tables-from-pdfs/

But some PDFs are more complicated

Quirky tables

Grand Prix Final 2017 Senior and Junior

PAIRS FREE SKATING JUDGES DETAILS PER SKATER

Rank	Name	Nation	Starting Number	Total Segment Score =	Total Element Score +	Total Program Component Score (factorized)	Deductions -
1	Aljona SAVCHENKO / Bruno MASSOT	GER	6	157.25	80.23	77.02	0.00

#	Executed Elements	Info	Base Value	GOE	J1	J2	J3	J4	J5	J6	J7	J8	J9	Ref	Scores of Panel
1	3Tw3		6.20	2.10	3	3	3	3	3	3	3	3	3		8.30
2	3LzTh		5.50	1.70	2	2	1	3	2	3	3	3	2		7.20
3	3S+2T+2T		7.00	1.40	2	2	2	2	2	2	2	2	1		8.40
4	3T		4.30	1.20	1	2	2	2	1	2	2	2	1		5.50
5	3STh		4.50	0.90	1	1	2	1	0	2	2	2	0		5.40
6	5RLi4		7.50	1.90	2	2	3	3	3	3	3	3	2		9.40
7	3Li4		4.50	1.43	3	2	3	3	3	3	3	3	2		5.93
8	CCoSp4		3.50	1.29	3	2	3	3	2	3	3	2	2		4.79
9	PCoSp4		4.50	1.21	3	2	2	3	2	3	3	2	2		5.71
10	ChSq1		2.00	1.80	3	2	2	3	2	3	3	3	2		3.80
11	BoDs4		4.50	1.80	3	2	2	3	3	3	3	2	2		6.30
12	5ALi4		7.50	2.00	3	2	3	3	3	3	3	3	2		9.50
			61.50												**80.23**
	Program Components			**Factor**											
	Skating Skills			1.60	9.25	9.50	9.50	9.50	9.25	9.50	9.75	9.75	9.25		9.46
	Transitions			1.60	9.50	9.25	9.50	9.75	9.50	9.75	9.50	9.75	9.00		9.54
	Performance			1.60	9.50	9.75	9.75	10.00	10.00	10.00	10.00	9.75	9.25		9.82
	Composition			1.60	9.50	9.50	9.75	9.75	9.50	9.75	9.75	9.75	9.00		9.64
	Interpretation of the Music			1.60	9.75	9.50	9.75	9.50	9.75	9.75	10.00	9.75	9.25		9.68
	Judges Total Program Components Score (factored)														**77.02**
	Deductions:														**0.00**

Rank	Name	Nation	Starting Number	Total Segment Score =	Total Element Score +	Total Program Component Score (factorized)	Deductions -
2	Wenjing SUI / Cong HAN	CHN	4	155.07	79.24	75.83	0.00

#	Executed Elements	Info	Base Value	GOE	J1	J2	J3	J4	J5	J6	J7	J8	J9	Ref	Scores of Panel
1	4Tw2		8.00	1.29	1	1	1	2	1	-2	2	2	1		9.29
2	3T+2T+2T		6.90	1.30	2	2	2	2	1	1	2	2	2		8.20

Recursive tables

FinCEN | Financial Crimes Enforcement Network
United States Department of the Treasury

BSA: [redacted]

BSAR Transcript
BSA: [redacted] and DCN:

Filing Information

Type of Report	Initial Report
Filing Date	10/29/2013
Received Date	10/29/2013
Entry Date	10/30/2013
Submission Method	Electronic discrete filing

Subject Information

Subject 1 of 1 : WCM777 LIMITED

Role	Subject	
Subject Type	Both Purchaser and Payee	
Individual/Organization	Organization	
Last (or Entity) Name	WCM777 LIMITED	
TIN Unknown	Yes	
Form(s) of Identification	Identification Unknown	Yes
Address(es)	Street Address	RM1204 12/F GREENFIELD TOWER OF CONCORDIA PLAZA 1 SCIENCE MUSEUM ROAD
	City	KOWLOON
	ZIP Code Unknown	Yes
	Country	HK
		HK - Enhanced
Corroborative Statement to Filer	No	
Relationship to Reporting Institution(s)	Institution TIN	[redacted]
	Relationship of Subject	Other: Yes CLIENTS CLIENT
Affected Account(s)	Account 1 of 2	
	Account Number	[redacted]

Templated reports

United States Department of Agriculture
Animal and Plant Health Inspection Service

KFRANK
INS-0000826614

Inspection Report

University of California-Berkeley

119 California Hall
Berkeley, CA 94720

Customer ID: **9191**

Certificate: **93-R-0432**

Site: 001

UNIVERSITY OF CALIFORNIA, BERKELEY

Type: ROUTINE INSPECTION

Date: 14-NOV-2022

Little bits of extra text

NICS Firearm Background Checks

January - 2023

State / Territory	Permit	Permit Recheck	Handgun	Long Gun	*Other	**Multiple	Admin	Pre-Pawn Handgun	Pre-Pawn Long Gun	Pre-Pawn *Other	Redemption Handgun	Redemption Long Gun	Redemption *Other	Returned/Disposition Handgun	Returned/Disposition Long Gun	Returned/Disposition *Other	Rentals Handgun	Rentals Long Gun	Private Sale Handgun	Private Sale Long Gun	Private Sale *Other	Return to Seller - Private Sale Handgun	Return to Seller - Private Sale Long Gun	Return to Seller - Private Sale *Other	Totals
Alabama	13,130	229	20,028	13,674	1,403	1,042	0	12	8	2	2,215	1,004	7	74	0	0	0	0	35	26	11	1	1	0	52,902
Alaska	269	25	2,406	1,779	323	167	0	2	1	0	92	60	2	22	4	0	0	0	0	1	4	0	0	0	5,157
Arizona	8,875	1,389	17,744	8,275	1,763	997	0	9	6	1	1,211	453	6	232	26	2	0	0	9	3	0	1	1	0	41,003
Arkansas	1,944	272	6,484	6,290	485	405	6	13	6	0	979	750	2	0	0	0	0	0	7	2	0	1	2	0	17,648
California	21,900	10,343	35,181	21,735	4,949	0	0	0	0	0	621	329	31	1,444	1,044	184	0	0	7,665	3,009	611	77	36	0	109,159
Colorado	7,464	11	18,373	11,230	2,031	1,668	0	0	0	0	0	0	0	304	54	4	0	0	0	0	0	0	0	0	41,139
Connecticut	8,334	2,849	5,842	1,922	2,514	0	2	0	0	0	0	0	0	0	0	0	0	0	544	169	199	0	0	0	22,375
Delaware	434	0	2,231	1,081	86	61	0	0	0	0	15	10	0	68	0	0	0	0	99	13	1	0	0	0	4,099
District of Columbia	960	1	326	13	0	4	10	0	0	0	0	0	0	2	0	46	0	0	0	0	0	0	0	0	1,362
Florida	19,460	0	58,648	22,950	4,883	2,827	0	10	5	1	3,726	898	7	1,426	117	3	0	0	338	164	53	54	42	0	115,612
Georgia	16,453	0	18,081	9,451	997	785	0	19	4	0	2,011	706	17	72	0	0	0	0	16	4	0	0	1	0	48,617
Guam	0	0	188	79	19	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	292
Hawaii	1,847	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1,848
Idaho	6,159	69	5,234	4,415	649	265	0	1	1	0	286	246	0	91	13	0	0	0	15	15	0	0	0	0	17,459
Illinois	480,752	0	28,370	14,140	1,715	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	524,977
Indiana	372	32,711	22,825	12,560	1,912	1,092	3	6	1	0	781	302	16	43	1	0	0	0	34	28	4	0	1	0	72,692
Iowa	6,740	6,768	2,527	2,790	193	92	11	0	0	0	30	22	3	51	20	3	0	0	3	2	0	0	0	0	19,255
Kansas	1,315	403	6,507	5,577	708	417	0	3	1	3	502	225	5	81	12	3	0	0	6	6	1	1	0	0	15,776
Kentucky	716	301,242	10,744	8,081	628	603	6	7	4	0	1,488	806	4	31	9	0	0	0	35	18	4	2	0	0	324,428
Louisiana	1,879	706	10,755	7,949	870	572	0	8	1	0	910	423	14	0	0	0	0	0	15	12	4	0	0	0	24,118
Maine	522	8	3,614	3,093	352	214	0	0	0	0	59	54	2	18	6	1	0	0	8	4	0	0	0	0	7,955
Mariana Islands	0	0	8	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11
Maryland	21,495	36	12,299	5,849	116	10	0	0	0	0	82	93	1	113	18	2	0	0	0	44	6	0	1	0	40,165
Massachusetts	10,720	0	5,626	2,799	1,506	242	0	1	0	1	9	16	0	2	0	0	0	0	32	23	6	1	0	0	20,984
Michigan	18,535	6,068	21,880	15,056	2,420	915	0	0	1	2	18	195	7	57	3	0	0	0	22	9	2	1	0	0	65,191
Minnesota	11,361	32,523	9,530	8,082	1,016	494	0	0	0	0	145	146	0	77	22	1	0	0	15	8	1	0	1	0	63,422
Mississippi	3,346	0	8,153	6,557	511	373	0	10	8	0	1,366	694	1	23	1	0	0	0	5	2	0	0	0	0	21,050
Missouri	853	87	19,831	14,724	2,133	1,206	2	3	0	1	1,190	546	21	400	48	2	0	0	27	28	4	0	5	0	41,111
Montana	996	0	3,600	3,179	388	253	21	0	2	1	369	357	3	23	9	1	0	0	4	7	2	2	2	0	9,219
Nebraska	4,240	66	227	1,665	31	10	0	0	0	0	2	22	0	26	9	0	0	0	0	3	0	0	0	0	6,301
Nevada	1,265	0	4,875	2,288	386	304	0	0	0	0	259	77	0	4	1	0	0	0	463	162	28	0	0	0	10,112

What about AI?

Pros:

- No programming knowledge required
- Can handle messier, less consistent PDFs
- Getting better by the day

Cons:

- Not deterministic
- Can be quite slow
- Can get expensive
- You have no idea what it's really doing

What *is* a PDF?

> The Portable Document Format (PDF) was created by Adobe Systems, introduced at the Windows and OS|2 Conference in January 1993 and remained a proprietary format until it was released as an open standard in 2008. Since then, it is under the control of International Organization for Standardization(ISO) Committee of volunteer industry experts.

https://en.wikipedia.org/wiki/History_of_PDF

PDF Reference

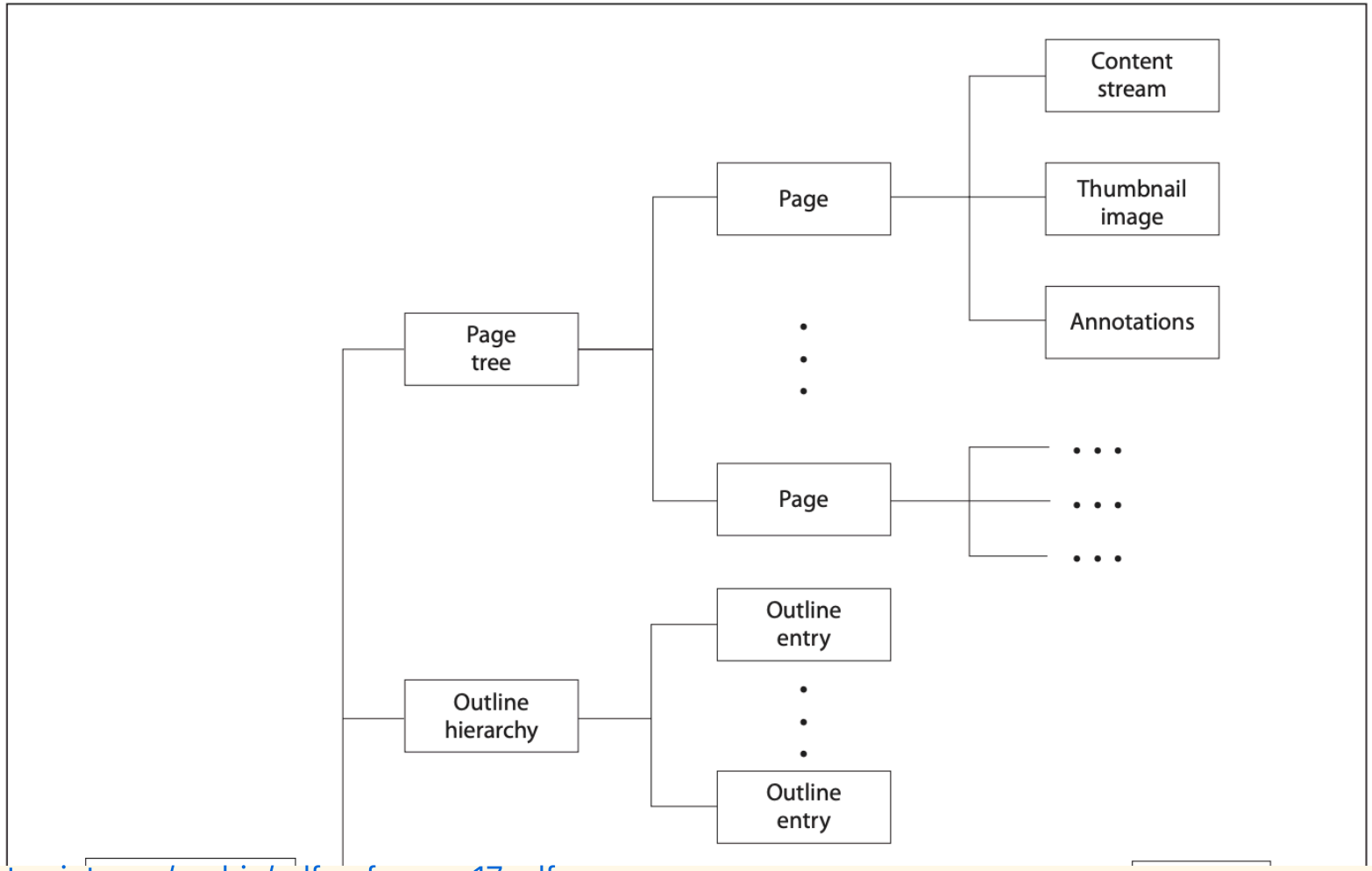
sixth edition

Adobe® Portable Document Format

Version 1.7

November 2006

https://ghostscript.com/~robin/pdf_reference17.pdf



https://ghostscript.com/~robin/pdf_reference17.pdf

TABLE 4.7 Graphics state operators

OPERANDS	OPERATOR	DESCRIPTION
—	**q**	Save the current graphics state on the graphics state stack (see "Graphics State Stack" on page 214).
—	**Q**	Restore the graphics state by removing the most recently saved state from the stack and making it the current state (see "Graphics State Stack" on page 214).
a b c d e f	**cm**	Modify the current transformation matrix (CTM) by concatenating the specified matrix (see Section 4.2.1, "Coordinate Spaces"). Although the operands specify a matrix, they are written as six separate numbers, not as an array.
lineWidth	**w**	Set the line width in the graphics state (see "Line Width" on page 215).

https://ghostscript.com/~robin/pdf_reference17.pdf

HTML is for browsers, PDFs are for … printers

- Browsers are smart … printers are dumb
- HTML is declarative and semantic … PDFs are imperative

```html
<table>
  <tr>
    <th>Name</th>
    <th>Animal</th>
    <th>Age</th>
  </tr>
  <tr>
    <td>Fido</td>
    <td>Dog</td>
    <td>13</td>
  </tr>
  <tr>
    <td>Jojo</td>
    <td>Cat</td>
    <td>7</td>
  </tr>
</table>
```

... versus:

🌵

```
q
    1 0 0 1 50 400 cm
    0.5 G
    0 0 m
    500 0 l
    S
    0 -20 m
    500 -20 l
    S
    0 -40 m
    500 -40 l
    S
    0 -60 m
    500 -60 l
    S
    0 -80 m
    500 -80 l
    S
    0 -100 m
[...]
```

```
W* n^M
BT^M
/F1 9.96 Tf^M
1 0 0 1 303.65 662.26 Tm^M
0 g^M
0 G^M
[(C)-12(A)4(LIFO)-5(R)-12(NIA)4(, )-9(B)4(E)-7(RK)-8(E)4(L)-
9(E)-7(Y)] TJ^M
ET^M
Q^M
q^M
224.33 660.1 288.05 11.52 re^M
W* n^M
BT^M
/F1 9.96 Tf^M
1 0 0 1 422.11 662.26 Tm^M
```

Appearance vs. structure

- Two similar-looking PDFs can have very different underlying structures.
- Every PDF-generating program works slightly differently — and sometimes wrong.

“True” PDFs vs. Image PDFs (🤯)

- **True PDFs** encode each textual and graphical element
- **Image PDFs** are just a stack of images
- Optical character recognition (OCR) can help, but only so much

🌵

So … how do we parse PDFs?

- In this workshop: `pdfplumber`
- It's not the only option, but it's a pretty good one
- … but I'm biased 🙃

🌵

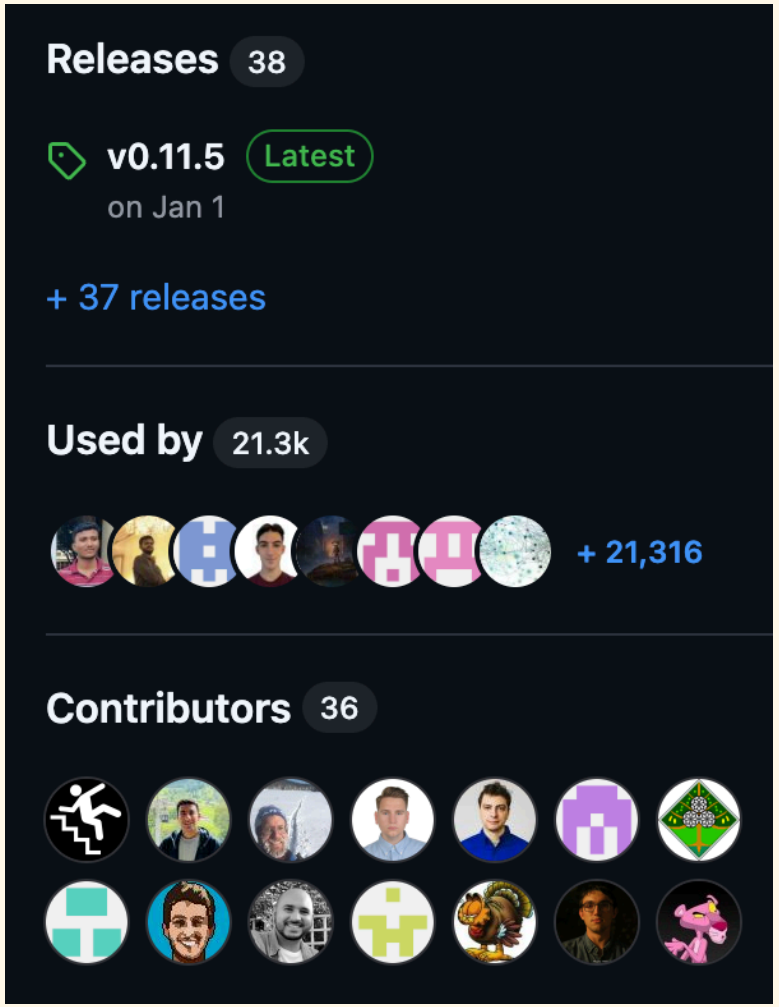
What is `pdfplumber`?

- 🌐 github.com/jsvine/pdfplumber
- Built on top of `pdfminer.six`
- Easier access to individual characters, lines, et cetera
- Convenient methods for data and text extraction

🌵

```
❯ git log b3a7cb8
commit b3a7cb83599863b416c98f08226668d86452116b (tag: v0.0.0)
Author: Jeremy Singer-Vine <jsvine@gmail.com>
Date:   Sun Aug 23 23:12:56 2015 -0400

    Initial commit
```



- State WARN Act layoff notices
- Michigan air pollution violation notices
- Washington State nursing home violations
- Prison-banned book lists
- Figure skating judging results
- Animal Welfare Act inspection reports
- FBI gun background check stats
- TSA traveler complaint counts
- [...]

bit.ly/nicar-2025-pdfplumber

Additional Resources

- Tipsheet from NICAR 2016: **bit.ly/parsing-prickly-pdfs**
- *PDF Explained* (book) by John Whitington: **bit.ly/pdf-explained**
- "Let's write a PDF" (presentation) by Ange Albertini: **bit.ly/lets-write-a-pdf**

bit.ly/pdf-workshop-feedback