Gannett Training

Data Scraping 101: Web Pages | .PDFs

Presented by

Mike Reilley
mikereilley1@gmail.com | @journtoolbox | journaliststoolbox.org

GANNETT

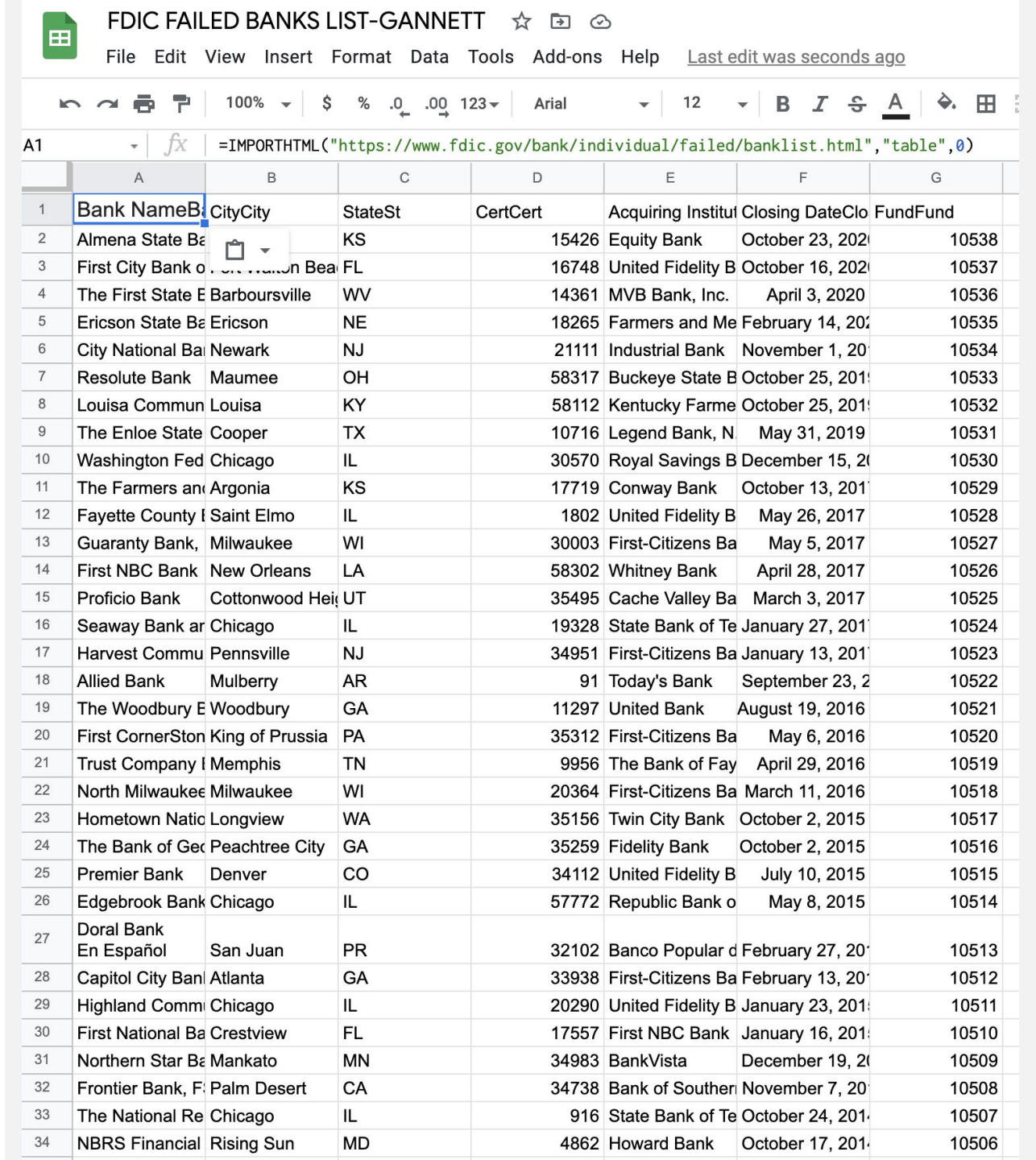
Problem: I requested public records but the PIO sent them as a .PDF or web page link

You need the data as a spreadsheet so you can sort and filter the data to find trends, data points, do math calculations, etc.

You can scrape web page tables with a basic formula in Google Sheets. Just add the web address

You also can scrape web page tables with a Google Chrome browser plug-in

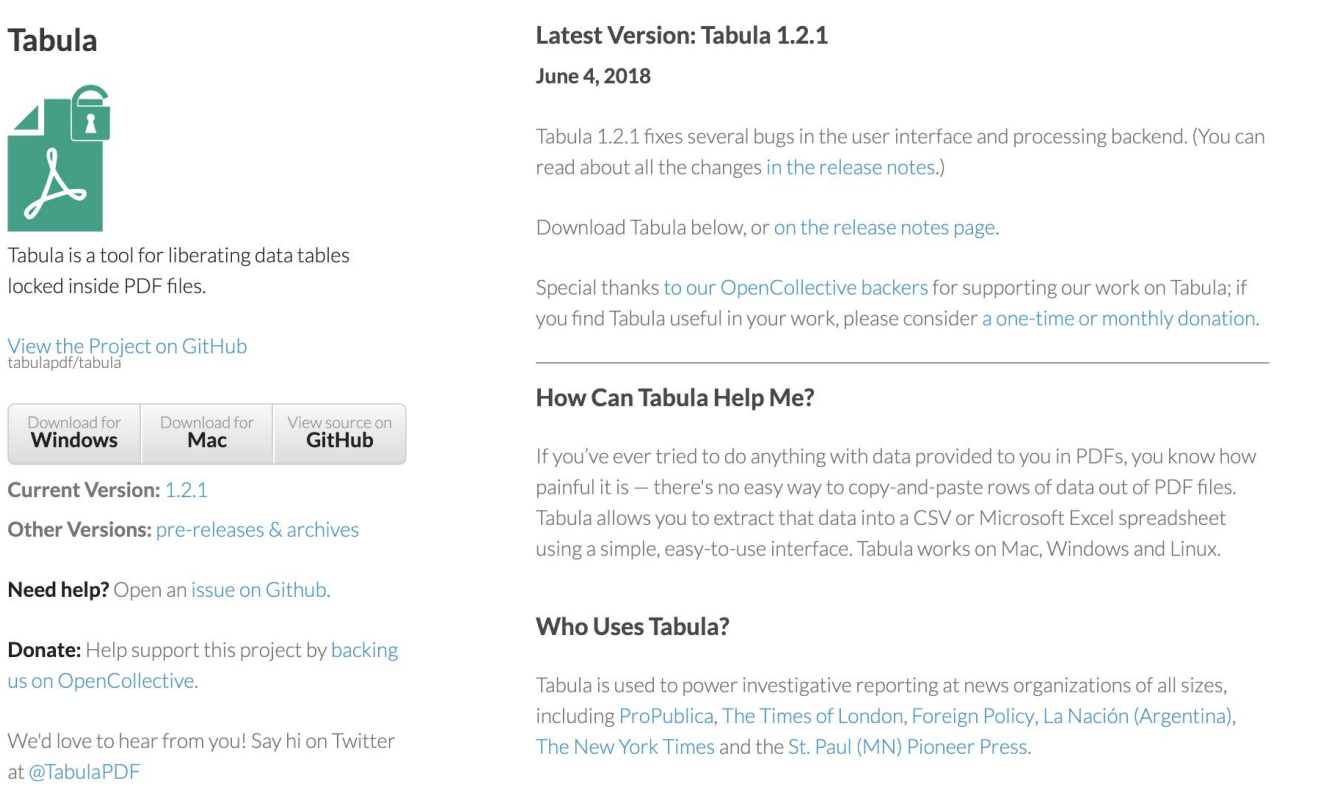
Scrape .PDFs with a free tool called Tabula that you download to your desktop



	A	B	C	D	E	F	G
1	Bank NameBa	CityCity	StateSt	CertCert	Acquiring Institut	Closing DateClo	FundFund
2	Almena State Ba		KS	15426	Equity Bank	October 23, 202	10538
3	First City Bank o	n Bea	FL	16748	United Fidelity B	October 16, 202	10537
4	The First State E	Barboursville	WV	14361	MVB Bank, Inc.	April 3, 2020	10536
5	Ericson State Ba	Ericson	NE	18265	Farmers and Me	February 14, 202	10535
6	City National Ba	Newark	NJ	21111	Industrial Bank	November 1, 20	10534
7	Resolute Bank	Maumee	OH	58317	Buckeye State B	October 25, 201	10533
8	Louisa Commun	Louisa	KY	58112	Kentucky Farme	October 25, 201	10532
9	The Enloe State	Cooper	TX	10716	Legend Bank, N	May 31, 2019	10531
10	Washington Fed	Chicago	IL	30570	Royal Savings B	December 15, 2	10530
11	The Farmers an	Argonia	KS	17719	Conway Bank	October 13, 201	10529
12	Fayette County I	Saint Elmo	IL	1802	United Fidelity B	May 26, 2017	10528
13	Guaranty Bank,	Milwaukee	WI	30003	First-Citizens Ba	May 5, 2017	10527
14	First NBC Bank	New Orleans	LA	58302	Whitney Bank	April 28, 2017	10526
15	Proficio Bank	Cottonwood Hei	UT	35495	Cache Valley Ba	March 3, 2017	10525
16	Seaway Bank ar	Chicago	IL	19328	State Bank of Te	January 27, 201	10524
17	Harvest Commu	Pennsville	NJ	34951	First-Citizens Ba	January 13, 201	10523
18	Allied Bank	Mulberry	AR	91	Today's Bank	September 23, 2	10522
19	The Woodbury E	Woodbury	GA	11297	United Bank	August 19, 2016	10521
20	First CornerSton	King of Prussia	PA	35312	First-Citizens Ba	May 6, 2016	10520
21	Trust Company I	Memphis	TN	9956	The Bank of Fay	April 29, 2016	10519
22	North Milwaukee	Milwaukee	WI	20364	First-Citizens Ba	March 11, 2016	10518
23	Hometown Natio	Longview	WA	35156	Twin City Bank	October 2, 2015	10517
24	The Bank of Geo	Peachtree City	GA	35259	Fidelity Bank	October 2, 2015	10516
25	Premier Bank	Denver	CO	34112	United Fidelity B	July 10, 2015	10515
26	Edgebrook Bank	Chicago	IL	57772	Republic Bank o	May 8, 2015	10514
27	Doral Bank En Español	San Juan	PR	32102	Banco Popular d	February 27, 201	10513
28	Capitol City Banl	Atlanta	GA	33938	First-Citizens Ba	February 13, 201	10512
29	Highland Commu	Chicago	IL	20290	United Fidelity B	January 23, 201	10511
30	First National Ba	Crestview	FL	17557	First NBC Bank	January 16, 201	10510
31	Northern Star Ba	Mankato	MN	34983	BankVista	December 19, 2	10509
32	Frontier Bank, F	Palm Desert	CA	34738	Bank of Southern	November 7, 201	10508
33	The National Re	Chicago	IL	916	State Bank of Te	October 24, 201	10507
34	NBRS Financial	Rising Sun	MD	4862	Howard Bank	October 17, 201	10506

Data Scraping 101

- Once we scrape the data, we'll show you some basic filtering tools
- How to scrape a native .PDF in Tabula vs. a scanned document
- Discuss which tables cannot be scraped and what to do when when you need that data
- Work with these exercises as we watch the training video: http://bit.ly/gannettdata



Tabula

Tabula is a tool for liberating data tables locked inside PDF files.

View the Project on GitHub
tabulapdf/tabula

Download for Windows | Download for Mac | View source on GitHub

Current Version: 1.2.1

Other Versions: pre-releases & archives

Need help? Open an issue on Github.

Donate: Help support this project by backing us on OpenCollective.

We'd love to hear from you! Say hi on Twitter at @TabulaPDF

Latest Version: Tabula 1.2.1

June 4, 2018

Tabula 1.2.1 fixes several bugs in the user interface and processing backend. (You can read about all the changes in the release notes.)

Download Tabula below, or on the release notes page.

Special thanks to our OpenCollective backers for supporting our work on Tabula; if you find Tabula useful in your work, please consider a one-time or monthly donation.

How Can Tabula Help Me?

If you've ever tried to do anything with data provided to you in PDFs, you know how painful it is — there's no easy way to copy-and-paste rows of data out of PDF files. Tabula allows you to extract that data into a CSV or Microsoft Excel spreadsheet using a simple, easy-to-use interface. Tabula works on Mac, Windows and Linux.

Who Uses Tabula?

Tabula is used to power investigative reporting at news organizations of all sizes, including ProPublica, The Times of London, Foreign Policy, La Nación (Argentina), The New York Times and the St. Paul (MN) Pioneer Press.