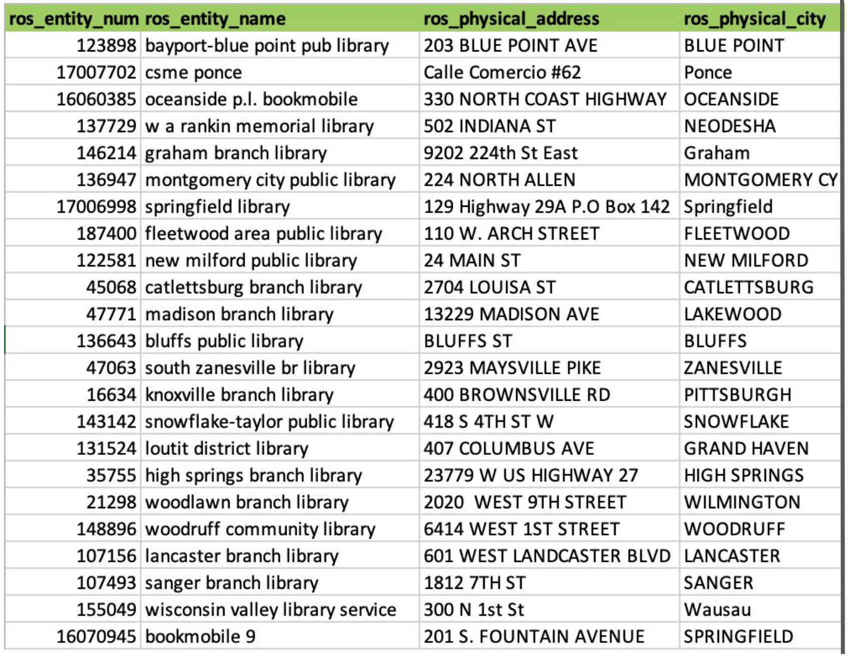
Data Integration Lecture

Spring 2021 Week 4

LIS 546: Data Curation II

Public Library Data Example

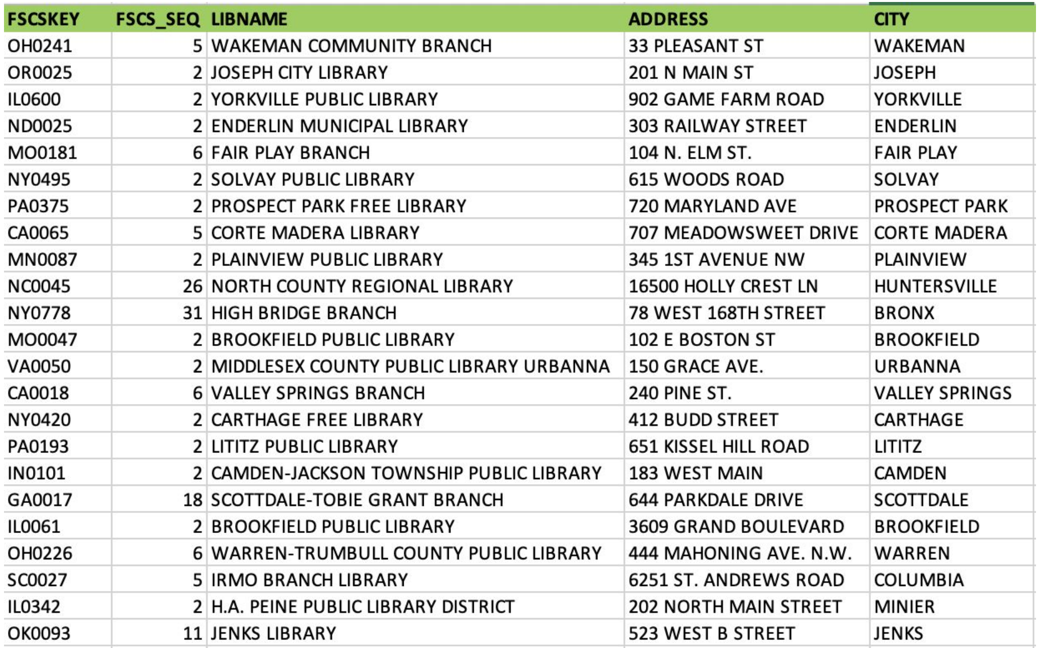
FCC E-rate Program: Libraries (>230,000 rows)



ros_entity_num	ros_entity_name	ros_physical_address	ros_physical_city
123898	bayport-blue point pub library	203 BLUE POINT AVE	BLUE POINT
17007702	csme ponce	Calle Comercio #62	Ponce
16060385	oceanside p.l. bookmobile	330 NORTH COAST HIGHWAY	OCEANSIDE
137729	w a rankin memorial library	502 INDIANA ST	NEODESHA
146214	graham branch library	9202 224th St East	Graham
136947	montgomery city public library	224 NORTH ALLEN	MONTGOMERY CY
17006998	springfield library	129 Highway 29A P.O Box 142	Springfield
187400	fleetwood area public library	110 W. ARCH STREET	FLEETWOOD
122581	new milford public library	24 MAIN ST	NEW MILFORD
45068	catlettsburg branch library	2704 LOUISA ST	CATLETTSBURG
47771	madison branch library	13229 MADISON AVE	LAKEWOOD
136643	bluffs public library	BLUFFS ST	BLUFFS
47063	south zanesville br library	2923 MAYSVILLE PIKE	ZANESVILLE
16634	knoxville branch library	400 BROWNSVILLE RD	PITTSBURGH
143142	snowflake-taylor public library	418 S 4TH ST W	SNOWFLAKE
131524	loutit district library	407 COLUMBUS AVE	GRAND HAVEN
35755	high springs branch library	23779 W US HIGHWAY 27	HIGH SPRINGS
21298	woodlawn branch library	2020 WEST 9TH STREET	WILMINGTON
148896	woodruff community library	6414 WEST 1ST STREET	WOODRUFF
107156	lancaster branch library	601 WEST LANDCASTER BLVD	LANCASTER
107493	sanger branch library	1812 7TH ST	SANGER
155049	wisconsin valley library service	300 N 1st St	Wausau
16070945	bookmobile 9	201 S. FOUNTAIN AVENUE	SPRINGFIELD

https://opendata.usac.org/E-rate/E-rate-Recipient-Details-And-Commitments/avi8-svp9

IMLS Public Library Survey (>17,000 rows)

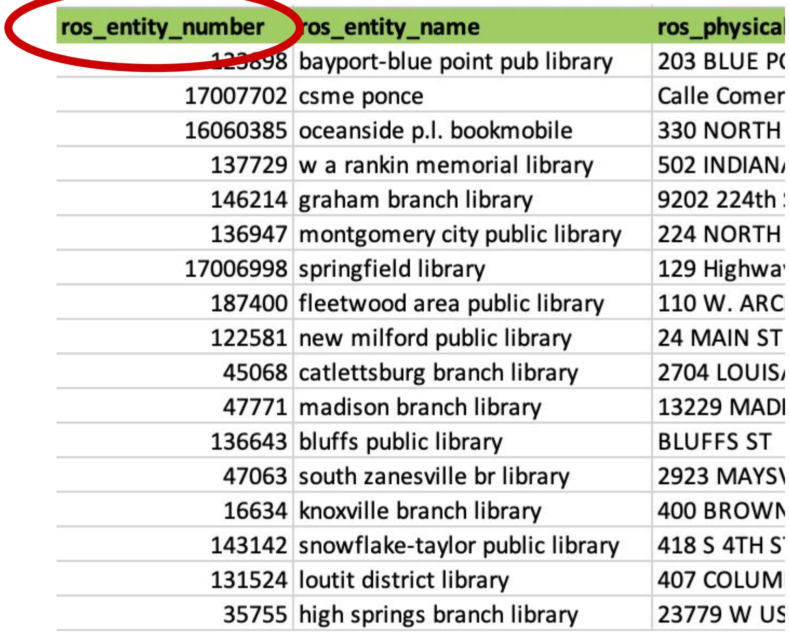


FSCSKEY	FSCS_SEQ	LIBNAME	ADDRESS	CITY
OH0241	5	WAKEMAN COMMUNITY BRANCH	33 PLEASANT ST	WAKEMAN
OR0025	2	JOSEPH CITY LIBRARY	201 N MAIN ST	JOSEPH
IL0600	2	YORKVILLE PUBLIC LIBRARY	902 GAME FARM ROAD	YORKVILLE
ND0025	2	ENDERLIN MUNICIPAL LIBRARY	303 RAILWAY STREET	ENDERLIN
MO0181	6	FAIR PLAY BRANCH	104 N. ELM ST.	FAIR PLAY
NY0495	2	SOLVAY PUBLIC LIBRARY	615 WOODS ROAD	SOLVAY
PA0375	2	PROSPECT PARK FREE LIBRARY	720 MARYLAND AVE	PROSPECT PARK
CA0065	5	CORTE MADERA LIBRARY	707 MEADOWSWEET DRIVE	CORTE MADERA
MN0087	2	PLAINVIEW PUBLIC LIBRARY	345 1ST AVENUE NW	PLAINVIEW
NC0045	26	NORTH COUNTY REGIONAL LIBRARY	16500 HOLLY CREST LN	HUNTERSVILLE
NY0778	31	HIGH BRIDGE BRANCH	78 WEST 168TH STREET	BRONX
MO0047	2	BROOKFIELD PUBLIC LIBRARY	102 E BOSTON ST	BROOKFIELD
VA0050	2	MIDDLESEX COUNTY PUBLIC LIBRARY URBANNA	150 GRACE AVE.	URBANNA
CA0018	6	VALLEY SPRINGS BRANCH	240 PINE ST.	VALLEY SPRINGS
NY0420	2	CARTHAGE FREE LIBRARY	412 BUDD STREET	CARTHAGE
PA0193	2	LITITZ PUBLIC LIBRARY	651 KISSEL HILL ROAD	LITITZ
IN0101	2	CAMDEN-JACKSON TOWNSHIP PUBLIC LIBRARY	183 WEST MAIN	CAMDEN
GA0017	18	SCOTTDALE-TOBIE GRANT BRANCH	644 PARKDALE DRIVE	SCOTTDALE
IL0061	2	BROOKFIELD PUBLIC LIBRARY	3609 GRAND BOULEVARD	BROOKFIELD
OH0226	6	WARREN-TRUMBULL COUNTY PUBLIC LIBRARY	444 MAHONING AVE. N.W.	WARREN
SC0027	5	IRMO BRANCH LIBRARY	6251 ST. ANDREWS ROAD	COLUMBIA
IL0342	2	H.A. PEINE PUBLIC LIBRARY DISTRICT	202 NORTH MAIN STREET	MINIER
OK0093	11	JENKS LIBRARY	523 WEST B STREET	JENKS

https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey

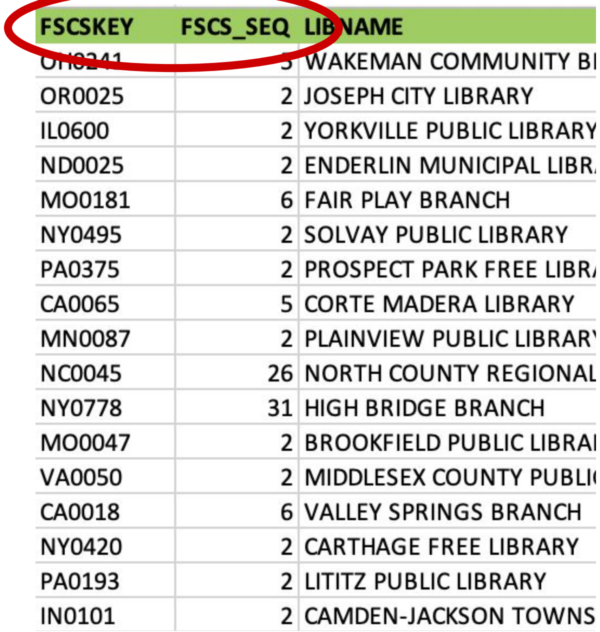
ID Variables (‘Primary Key’ in SQL)

FCC E-rate Program: Libraries



ros_entity_number	ros_entity_name	ros_physical
123898	bayport-blue point pub library	203 BLUE PC
17007702	csme ponce	Calle Comer
16060385	oceanside p.l. bookmobile	330 NORTH
137729	w a rankin memorial library	502 INDIAN
146214	graham branch library	9202 224th
136947	montgomery city public library	224 NORTH
17006998	springfield library	129 Highwa
187400	fleetwood area public library	110 W. ARC
122581	new milford public library	24 MAIN ST
45068	catlettsburg branch library	2704 LOUIS
47771	madison branch library	13229 MADI
136643	bluffs public library	BLUFFS ST
47063	south zanesville br library	2923 MAYSV
16634	knoxville branch library	400 BROWN
143142	snowflake-taylor public library	418 S 4TH ST
131524	loutit district library	407 COLUM
35755	high springs branch library	23779 W US

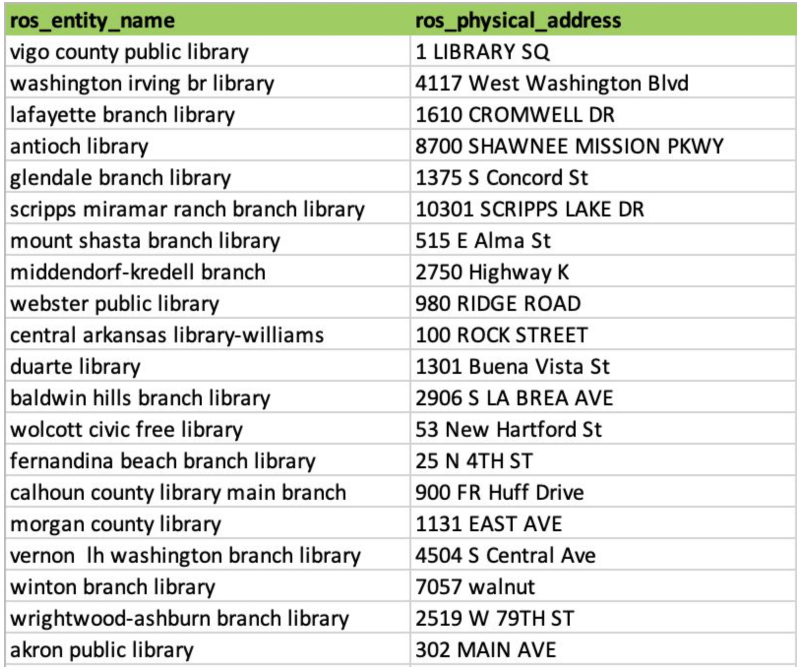
IMLS Public Library Survey



FSCSKEY	FSCS_SEQ	LIBNAME
OH0241	5	WAKEMAN COMMUNITY BI
OR0025	2	JOSEPH CITY LIBRARY
IL0600	2	YORKVILLE PUBLIC LIBRARY
ND0025	2	ENDERLIN MUNICIPAL LIBR
MO0181	6	FAIR PLAY BRANCH
NY0495	2	SOLVAY PUBLIC LIBRARY
PA0375	2	PROSPECT PARK FREE LIBR
CA0065	5	CORTE MADERA LIBRARY
MN0087	2	PLAINVIEW PUBLIC LIBRAR
NC0045	26	NORTH COUNTY REGIONAL
NY0778	31	HIGH BRIDGE BRANCH
MO0047	2	BROOKFIELD PUBLIC LIBRA
VA0050	2	MIDDLESEX COUNTY PUBLI
CA0018	6	VALLEY SPRINGS BRANCH
NY0420	2	CARTHAGE FREE LIBRARY
PA0193	2	LITITZ PUBLIC LIBRARY
IN0101	2	CAMDEN-JACKSON TOWNS

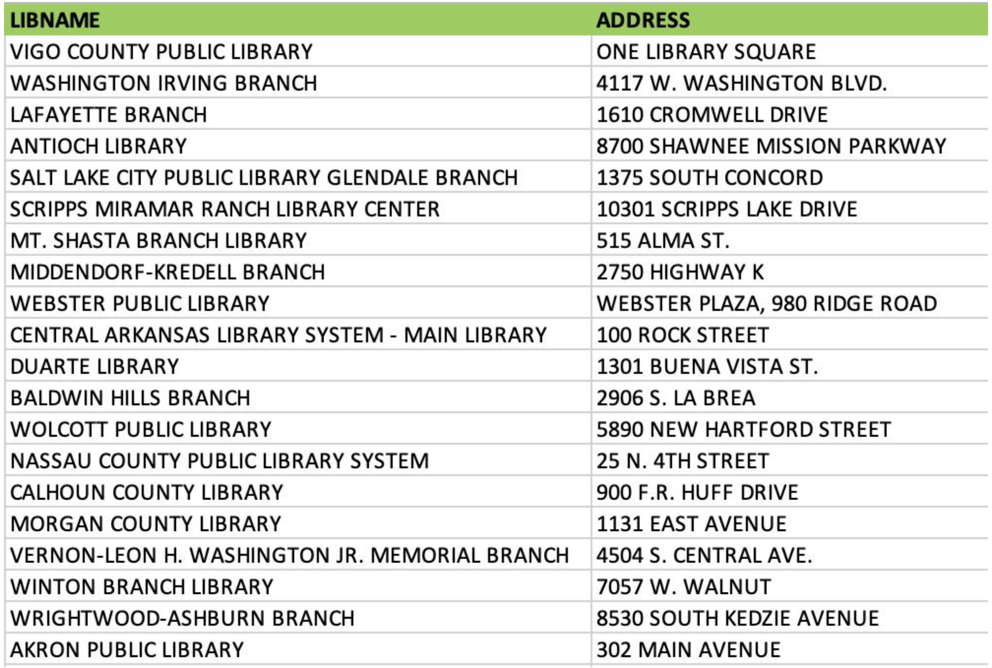
Our Key Variables: Name and Address

FCC E-rate Program: Libraries



ros_entity_name	ros_physical_address
vigo county public library	1 LIBRARY SQ
washington irving br library	4117 West Washington Blvd
lafayette branch library	1610 CROMWELL DR
antioch library	8700 SHAWNEE MISSION PKWY
glendale branch library	1375 S Concord St
scripps miramar ranch branch library	10301 SCRIPPS LAKE DR
mount shasta branch library	515 E Alma St
middendorf-kredell branch	2750 Highway K
webster public library	980 RIDGE ROAD
central arkansas library-williams	100 ROCK STREET
duarte library	1301 Buena Vista St
baldwin hills branch library	2906 S LA BREA AVE
wolcott civic free library	53 New Hartford St
fernandina beach branch library	25 N 4TH ST
calhoun county library main branch	900 FR Huff Drive
morgan county library	1131 EAST AVE
vernon lh washington branch library	4504 S Central Ave
winton branch library	7057 walnut
wrightwood-ashburn branch library	2519 W 79TH ST
akron public library	302 MAIN AVE

IMLS Public Library Survey

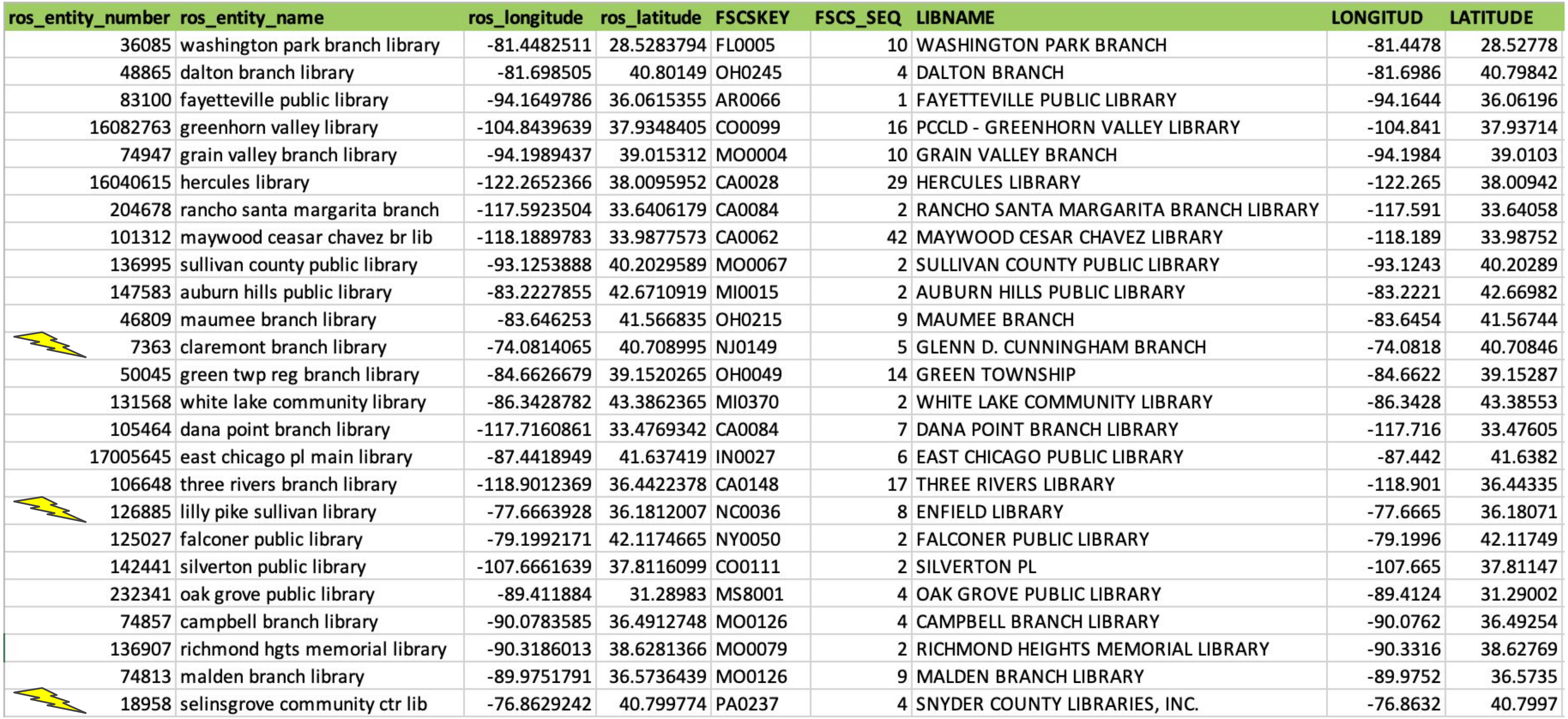


LIBNAME	ADDRESS
VIGO COUNTY PUBLIC LIBRARY	ONE LIBRARY SQUARE
WASHINGTON IRVING BRANCH	4117 W. WASHINGTON BLVD.
LAFAYETTE BRANCH	1610 CROMWELL DRIVE
ANTIOCH LIBRARY	8700 SHAWNEE MISSION PARKWAY
SALT LAKE CITY PUBLIC LIBRARY GLENDALE BRANCH	1375 SOUTH CONCORD
SCRIPPS MIRAMAR RANCH LIBRARY CENTER	10301 SCRIPPS LAKE DRIVE
MT. SHASTA BRANCH LIBRARY	515 ALMA ST.
MIDDENDORF-KREDELL BRANCH	2750 HIGHWAY K
WEBSTER PUBLIC LIBRARY	WEBSTER PLAZA, 980 RIDGE ROAD
CENTRAL ARKANSAS LIBRARY SYSTEM - MAIN LIBRARY	100 ROCK STREET
DUARTE LIBRARY	1301 BUENA VISTA ST.
BALDWIN HILLS BRANCH	2906 S. LA BREA
WOLCOTT PUBLIC LIBRARY	5890 NEW HARTFORD STREET
NASSAU COUNTY PUBLIC LIBRARY SYSTEM	25 N. 4TH STREET
CALHOUN COUNTY LIBRARY	900 F.R. HUFF DRIVE
MORGAN COUNTY LIBRARY	1131 EAST AVENUE
VERNON-LEON H. WASHINGTON JR. MEMORIAL BRANCH	4504 S. CENTRAL AVE.
WINTON BRANCH LIBRARY	7057 W. WALNUT
WRIGHTWOOD-ASHBURN BRANCH	8530 SOUTH KEDZIE AVENUE
AKRON PUBLIC LIBRARY	302 MAIN AVENUE

Custom Algorithm 😣 (Emoji: persevering face)

1. Clean strings (lowercase, remove symbols and extra spaces, abbreviations)
2. Perform geospatial matching on latitude and longitude (no further apart than 0.4 miles) using the R fuzzy join package - if more than one match, choose closest in distance
3. Calculate string “distances” between the library name, a substring of the library name, and the address using the R stringdist package - if the distances meet a certain threshold, make them a match

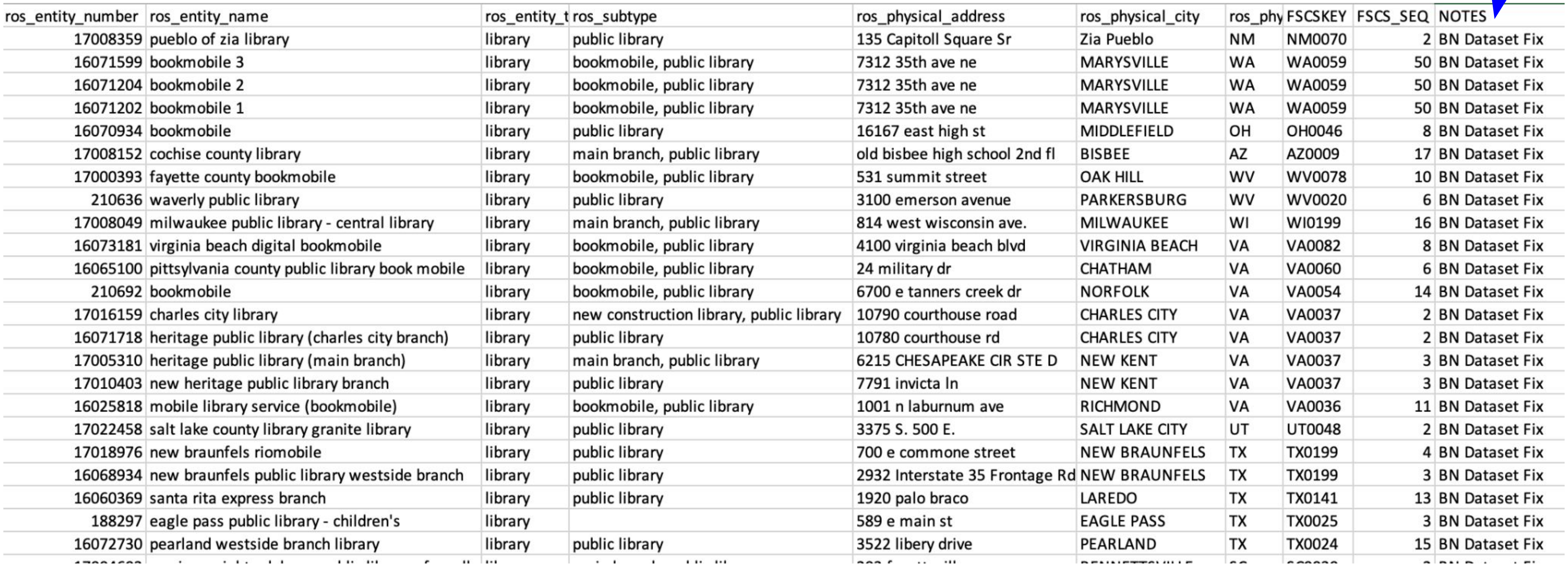
Merged!



ros_entity_number	ros_entity_name	ros_longitude	ros_latitude	FSCSKEY	FSCS_SEQ	LIBNAME	LONGITUD	LATITUDE
36085	washington park branch library	-81.4482511	28.5283794	FL0005	10	WASHINGTON PARK BRANCH	-81.4478	28.52778
48865	dalton branch library	-81.698505	40.80149	OH0245	4	DALTON BRANCH	-81.6986	40.79842
83100	fayetteville public library	-94.1649786	36.0615355	AR0066	1	FAYETTEVILLE PUBLIC LIBRARY	-94.1644	36.06196
16082763	greenhorn valley library	-104.8439639	37.9348405	CO0099	16	PCCLD - GREENHORN VALLEY LIBRARY	-104.841	37.93714
74947	grain valley branch library	-94.1989437	39.015312	MO0004	10	GRAIN VALLEY BRANCH	-94.1984	39.0103
16040615	hercules library	-122.2652366	38.0095952	CA0028	29	HERCULES LIBRARY	-122.265	38.00942
204678	rancho santa margarita branch	-117.5923504	33.6406179	CA0084	2	RANCHO SANTA MARGARITA BRANCH LIBRARY	-117.591	33.64058
101312	maywood ceasar chavez br lib	-118.1889783	33.9877573	CA0062	42	MAYWOOD CESAR CHAVEZ LIBRARY	-118.189	33.98752
136995	sullivan county public library	-93.1253888	40.2029589	MO0067	2	SULLIVAN COUNTY PUBLIC LIBRARY	-93.1243	40.20289
147583	auburn hills public library	-83.2227855	42.6710919	MI0015	2	AUBURN HILLS PUBLIC LIBRARY	-83.2221	42.66982
46809	maumee branch library	-83.646253	41.566835	OH0215	9	MAUMEE BRANCH	-83.6454	41.56744
⚡ 7363	claremont branch library	-74.0814065	40.708995	NJ0149	5	GLENN D. CUNNINGHAM BRANCH	-74.0818	40.70846
50045	green twp reg branch library	-84.6626679	39.1520265	OH0049	14	GREEN TOWNSHIP	-84.6622	39.15287
131568	white lake community library	-86.3428782	43.3862365	MI0370	2	WHITE LAKE COMMUNITY LIBRARY	-86.3428	43.38553
105464	dana point branch library	-117.7160861	33.4769342	CA0084	7	DANA POINT BRANCH LIBRARY	-117.716	33.47605
17005645	east chicago pl main library	-87.4418949	41.637419	IN0027	6	EAST CHICAGO PUBLIC LIBRARY	-87.442	41.6382
106648	three rivers branch library	-118.9012369	36.4422378	CA0148	17	THREE RIVERS LIBRARY	-118.901	36.44335
⚡ 126885	lilly pike sullivan library	-77.6663928	36.1812007	NC0036	8	ENFIELD LIBRARY	-77.6665	36.18071
125027	falconer public library	-79.1992171	42.1174665	NY0050	2	FALCONER PUBLIC LIBRARY	-79.1996	42.11749
142441	silverton public library	-107.6661639	37.8116099	CO0111	2	SILVERTON PL	-107.665	37.81147
232341	oak grove public library	-89.411884	31.28983	MS8001	4	OAK GROVE PUBLIC LIBRARY	-89.4124	31.29002
74857	campbell branch library	-90.0783585	36.4912748	MO0126	4	CAMPBELL BRANCH LIBRARY	-90.0762	36.49254
136907	richmond hgts memorial library	-90.3186013	38.6281366	MO0079	2	RICHMOND HEIGHTS MEMORIAL LIBRARY	-90.3316	38.62769
74813	malden branch library	-89.9751791	36.5736439	MO0126	9	MALDEN BRANCH LIBRARY	-89.9752	36.5735
⚡ 18958	selinsgrove community ctr lib	-76.8629242	40.799774	PA0237	4	SNYDER COUNTY LIBRARIES, INC.	-76.8632	40.7997

Perfection!

Reality: Fun with Data Entry by Hand



ros_entity_number	ros_entity_name	ros_entity_t	ros_subtype	ros_physical_address	ros_physical_city	ros_phy	FSCSKEY	FSCS_SEQ	NOTES
17008359	pueblo of zia library	library	public library	135 Capitoll Square Sr	Zia Pueblo	NM	NM0070	2	BN Dataset Fix
16071599	bookmobile 3	library	bookmobile, public library	7312 35th ave ne	MARYSVILLE	WA	WA0059	50	BN Dataset Fix
16071204	bookmobile 2	library	bookmobile, public library	7312 35th ave ne	MARYSVILLE	WA	WA0059	50	BN Dataset Fix
16071202	bookmobile 1	library	bookmobile, public library	7312 35th ave ne	MARYSVILLE	WA	WA0059	50	BN Dataset Fix
16070934	bookmobile	library	public library	16167 east high st	MIDDLEFIELD	OH	OH0046	8	BN Dataset Fix
17008152	cochise county library	library	main branch, public library	old bisbee high school 2nd fl	BISBEE	AZ	AZ0009	17	BN Dataset Fix
17000393	fayette county bookmobile	library	bookmobile, public library	531 summit street	OAK HILL	WV	WV0078	10	BN Dataset Fix
210636	waverly public library	library	public library	3100 emerson avenue	PARKERSBURG	WV	WV0020	6	BN Dataset Fix
17008049	milwaukee public library - central library	library	main branch, public library	814 west wisconsin ave.	MILWAUKEE	WI	WI0199	16	BN Dataset Fix
16073181	virginia beach digital bookmobile	library	bookmobile, public library	4100 virginia beach blvd	VIRGINIA BEACH	VA	VA0082	8	BN Dataset Fix
16065100	pittsylvania county public library book mobile	library	bookmobile, public library	24 military dr	CHATHAM	VA	VA0060	6	BN Dataset Fix
210692	bookmobile	library	bookmobile, public library	6700 e tanners creek dr	NORFOLK	VA	VA0054	14	BN Dataset Fix
17016159	charles city library	library	new construction library, public library	10790 courthouse road	CHARLES CITY	VA	VA0037	2	BN Dataset Fix
16071718	heritage public library (charles city branch)	library	public library	10780 courthouse rd	CHARLES CITY	VA	VA0037	2	BN Dataset Fix
17005310	heritage public library (main branch)	library	main branch, public library	6215 CHESAPEAKE CIR STE D	NEW KENT	VA	VA0037	3	BN Dataset Fix
17010403	new heritage public library branch	library	public library	7791 invicta ln	NEW KENT	VA	VA0037	3	BN Dataset Fix
16025818	mobile library service (bookmobile)	library	bookmobile, public library	1001 n laburnum ave	RICHMOND	VA	VA0036	11	BN Dataset Fix
17022458	salt lake county library granite library	library	public library	3375 S. 500 E.	SALT LAKE CITY	UT	UT0048	2	BN Dataset Fix
17018976	new braunfels riomobile	library	public library	700 e commone street	NEW BRAUNFELS	TX	TX0199	4	BN Dataset Fix
16068934	new braunfels public library westside branch	library	public library	2932 Interstate 35 Frontage Rd	NEW BRAUNFELS	TX	TX0199	3	BN Dataset Fix
16060369	santa rita express branch	library	public library	1920 palo braco	LAREDO	TX	TX0141	13	BN Dataset Fix
188297	eagle pass public library - children's	library		589 e main st	EAGLE PASS	TX	TX0025	3	BN Dataset Fix
16072730	pearland westside branch library	library	public library	3522 libery drive	PEARLAND	TX	TX0024	15	BN Dataset Fix

Image from: https://www.grunge.com/162741/the-reason-why-cats-are-afraid-of-cucumbers/