INSTITUTE of Museum and Library SERVICES

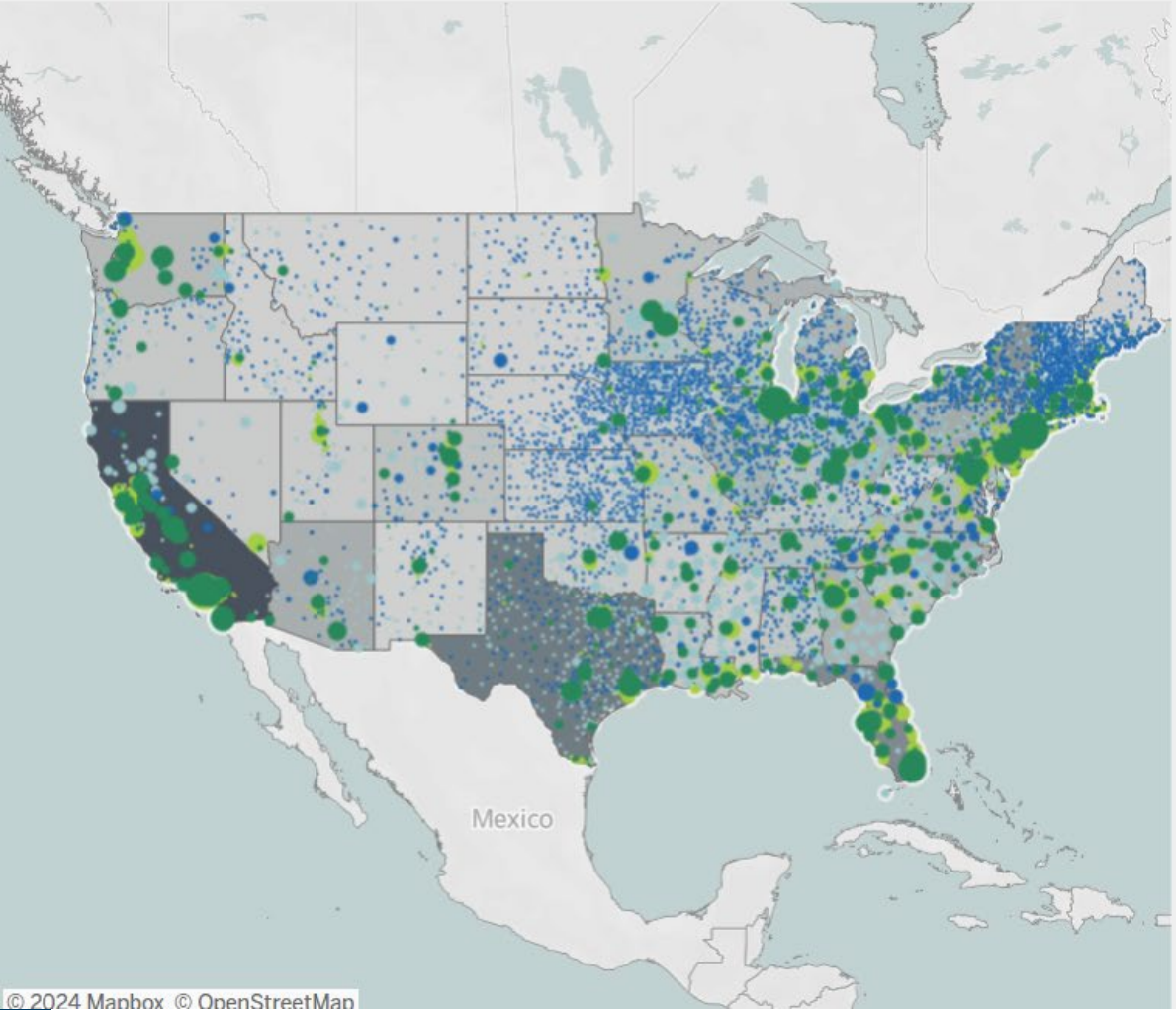
Challenges with Data Linking and Attribution in a Universe Establishment Data Collection

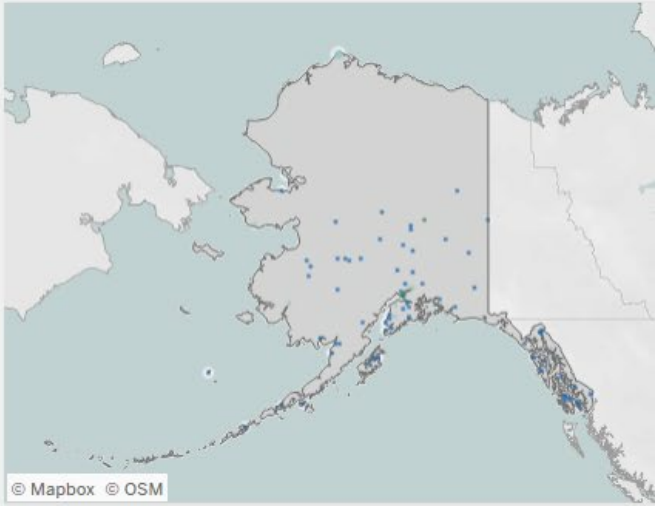
Evan Nielsen, American Institutes for Research
Marisa Pelczar, Institute of Museum and Library Services

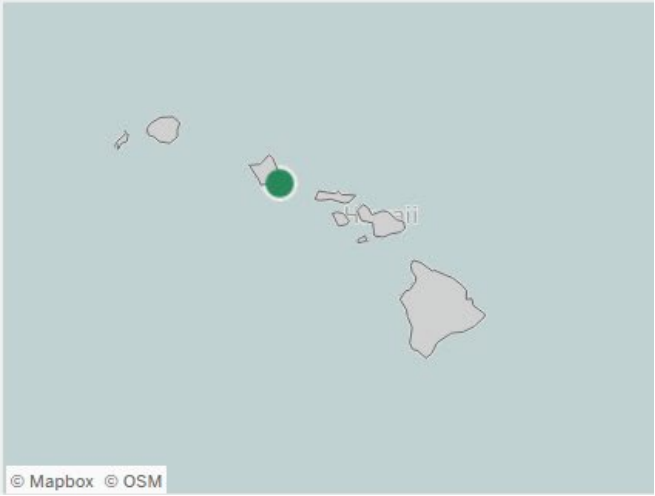
Background: Public Libraries Survey

- Annual census of all public libraries in 50 States, DC, and the outlying territories since 1989 (online data entry portal)
- Data collected at two levels:
  - ~9,200 library systems (“administrative entities” or AEs) and
  - ~17,000 points of service (“outlets”), each associated with one AE
- Response rate >95% of AEs for each year
- Geospatial information appended to both AE and outlet datafiles
  - Latitude/Longitude, Census tract/block IDs, Congressional district, etc.

Public Libraries in the US by Locale







Research Objectives

- Follow national, state, and other subgroup trends (e.g., NCES locale, population size)
- Provide peer comparisons for strategic and operational planning of individual public libraries
- Document the value of public libraries to their communities

Recent Initiatives

- In the past 5 years, IMLS has undertaken a set of initiatives to improve the methodological rigor and utility of the PLS:
  - Operationalizing data elements through stakeholder engagement and cognitive testing in the field
  - Releasing data elements to state respondents several months earlier to help them implement state-level collection

Challenges

- Service area populations are reported by state agencies using different demography sources (e.g., decennial census, annual Census estimates, ACS estimates, state data centers)
  - Affects per capita metrics released by IMLS
- Geographic coordinates of library locations with outdated qualitative jurisdiction indicators
  - Locations exist within multiple geographies
  - Data users would have to make effort to identify and join these geographies

Service Area Population	Number of Public Libraries	Number of Libraries with Branches	Number of Libraries with Book Mobiles	Total Number of Stationary Outlets	Number of Central Libraries	Number of Branch Libraries
1,000,000 or more	37	37	14	1,414	23	1,391
500,000 to 999,999	57	57	29	1,060	40	1,020
250,000 to 499,999	118	116	45	1,220	91	1,129
100,000 to 249,999	367	315	114	1,982	321	1,661
50,000 to 99,999	578	334	109	1,566	551	1,015
25,000 to 49,999	993	301	111	1,738	978	760
10,000 to 24,999	1,731	239	75	2,213	1,723	490
5,000 to 9,999	1,506	99	29	1,665	1,495	170
2,500 to 4,999	1,242	29	6	1,276	1,235	41
1,000 to 2,499	1,462	13	6	1,476	1,461	15
Less than 1,000	930	2	1	933	929	4
NATIONAL	9,021	1,542	539	16,543	8,847	7,696

Research Attempts with Big Caveats

- Two research studies that linked PLS data with other federal data:
  - Lisa Frehill (IMLS) and Melissa Cidade (US Census Bureau): American Housing Survey

https://nces.ed.gov/fcsm/pdf/CSPOS_Frehill_Cidade_2020.05.01_CLEARED.pdf

  - IMLS and AIR: National Household Education Survey 2019
    - Examining correlations between library measures and whether households with children visited library

Research Attempts with Big Caveats

- Both attributed library system (AE-level) data to outlets and then linked outlets to the other data records based on location (i.e., a spatial join).
- Both ended up with findings that had substantial caveats, limiting the utility of those findings.

Inaccuracy Risk

- Linking a household to the closest library outlet may not reflect the outlet that the household visits (or would consider visiting) or even whether the household is located in the legal service area for that library system.

How IMLS Is Mitigating Inaccuracy Risk

- Updating the Geographic Identifiers
  - Revising the GEOCODE category options
    - Separating Places from Minor Civil Divisions
    - Adding multi-place and multi-MCD
    - Separating School Districts to elementary, secondary, unified
  - This change is allowing IMLS to include Census identifiers (GEOIDs) for most library service areas in the FY22 PLS data files

Expected Results from These Efforts

- GEOIDs for library service areas will enable data users to link PLS data to other data sources for the same jurisdictions.

Bias Risk

- Attributing AE-level data to each library outlet in larger library systems (which tend to be in urban and suburban locales) masks neighborhood-level variation between those outlets.

How IMLS Is Mitigating Bias Risk

- Exploring the Collection of Outlet-Level Data
  - Libraries already track many current AE-level data elements at the outlet level so they can aggregate them to the AE level (e.g., library visits).
  - Some states already collect these data elements at the outlet level.
  - IMLS is conducting cognitive labs, engaging in a pilot with a state currently undergoing the change from AE to outlet data collection, planning for respondent research survey

Expected Results from These Efforts

- Outlet-level data will enable data users to analyze PLS data in urban and suburban areas at a neighborhood level, including joining with other small area data sources.

Next Steps

- IMLS will continue to refine the service area geographies for as many AEs as possible, especially multi-area.
- IMLS may formalize collection of state-level shapefiles of library service areas.

Any advice?

- If anyone in the FedCASIC community has had similar experience or has advice, we'd love to hear from you!

Contact:

enielsen@air.org
mpelczar@imls.gov