Understand and Adapt the 2020 Administrative Record Census for a Demographic Frame

Dominic Frecentese

Supervised by Jennifer Ortman and Michael Ratcliffe

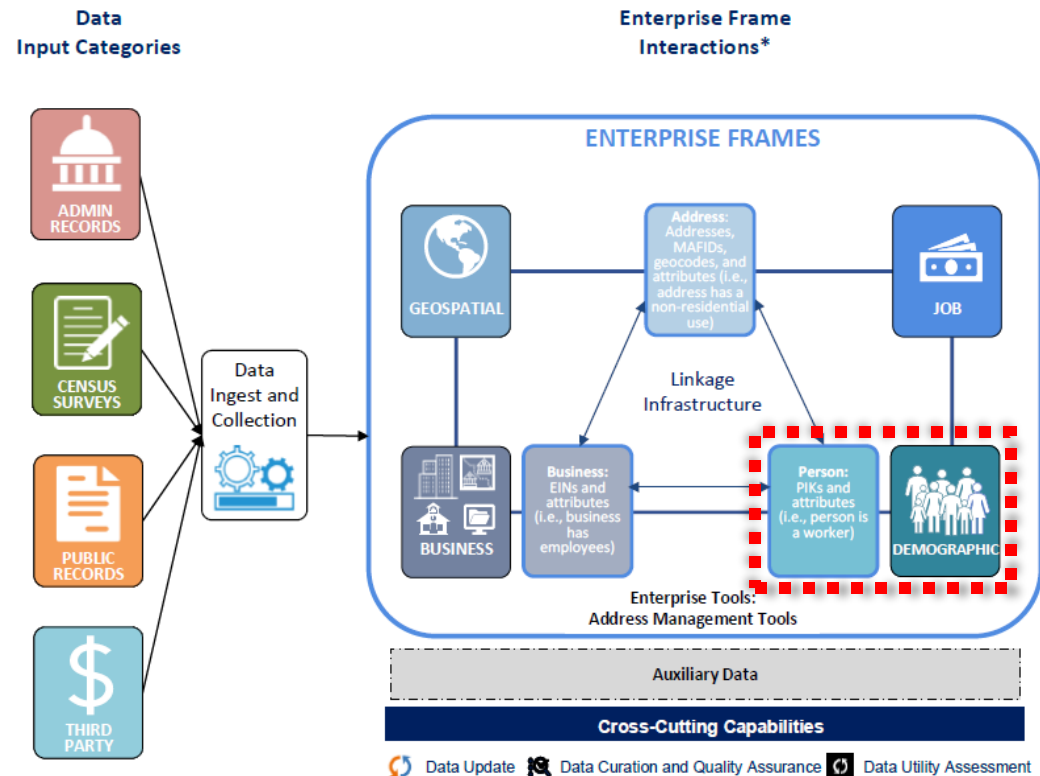Demographic Directorate - Frames Program

What is a "Demographic Frame"?

The Demographic Frame is part of a larger Frames Program intended to make data more **accessible** and **consistent** across the Census Bureau
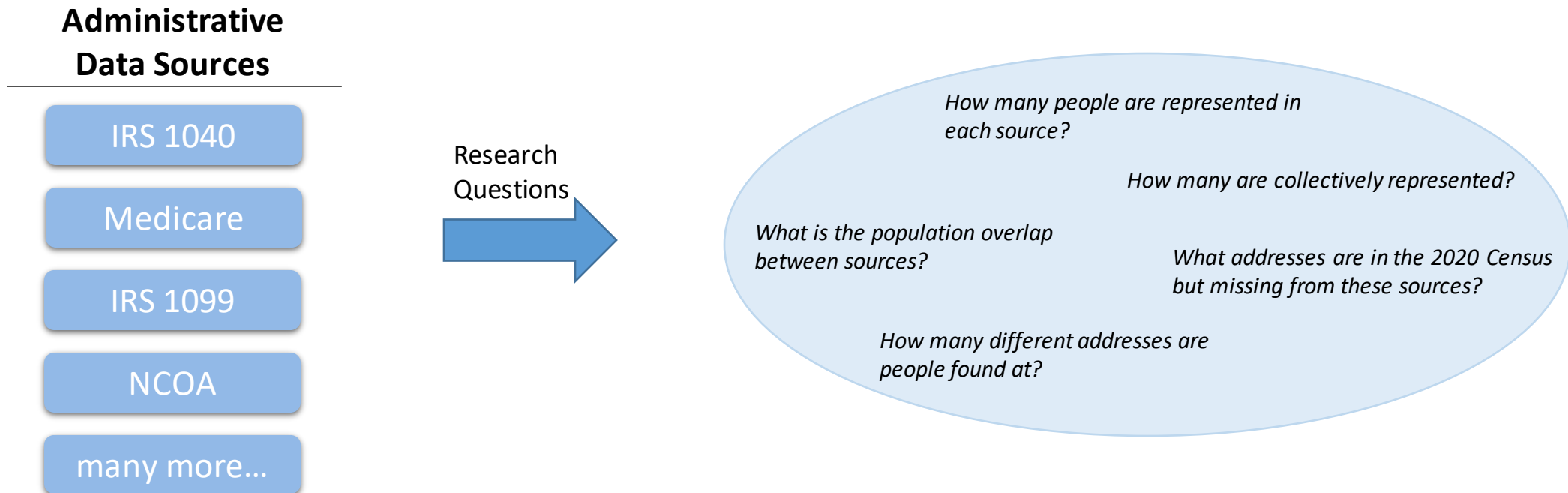
Why build it?

There are currently duplicative efforts within the Bureau to collect, process and store demographic data. The Demographic Frame will reduce redundant work, and make it easier to access this information

What will the Demographic Frame look like?

A comprehensive person-level dataset containing demographic, social, and economic characteristics
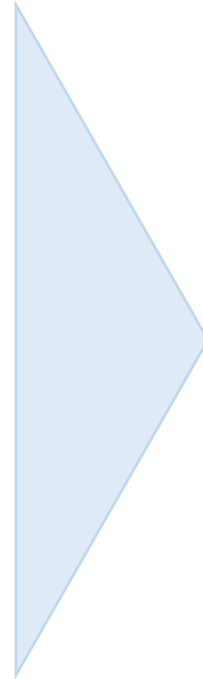
Issue: Before building a Demographic Frame we need to better understand the data that the frame will be built upon

Administrative Data Sources

- IRS 1040
- Medicare
- IRS 1099
- NCOA
- many more…

Research Questions →

How many people are represented in each source?

How many are collectively represented?

What is the population overlap between sources?

What addresses are in the 2020 Census but missing from these sources?

How many different addresses are people found at?

United States® Census Bureau

Objective: Inform Demographic Frame design through an improved understanding of administrative data sources

Specific Goals

- Prioritize data sources for the demographic frame

- Assess the feasibility of implementing a Title-13 only frame (e.g. no IRS data)

- Measure the population coverage lost from the exclusion of 2020 Census-use only data sources
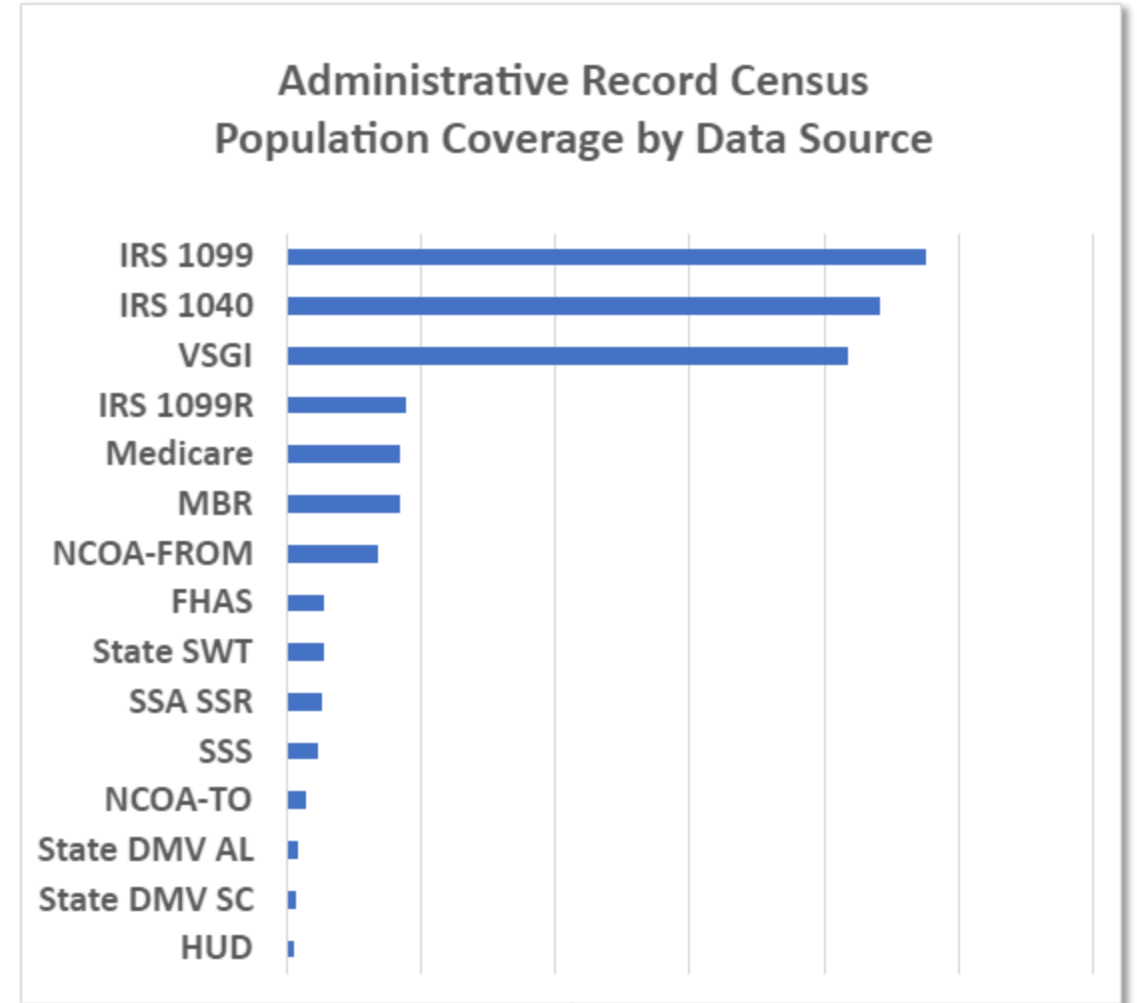
- Understand lapses in population coverage

Approach

- Calculate population coverage of each data source

- Use statistical measure of similarity (Overlap Coefficient) to quantify data source population similarity

- Compare to the 2020 Census Unedited File (CUF) to identify missing people and addresses

Notable findings

- The majority of person-address combinations are found within just 1 administrative data source

- IRS data is extremely important for a Demographic Frame - removing IRS data sources would result in a ~25% reduction in the Administrative Record population

- Three administrative data sources cover >90% of the Administrative Record Population

- 2020 Census-use only data sources gave visibility to ~5% additional persons not found in other administrative record sources; we will no longer have these people in a Demo Frame

Administrative Record Census
Population Coverage by Data Source

United States® Census Bureau

Conclusion: What we've learned and where to go from here

Learning

- Data Quality checks were helpful for identifying anomalous persons / addresses / characteristics

- AdRec Census and Demographic Frame datasets are large (~1 Terabyte) and challenging to analyze with the resources of the IRE

- Many of the problems that need to be solved on the Demographic Frame are related to data design, data processing, and data infrastructure

Recommendation

- Automate and periodically run these checks for the Demographic Frame

- Consider migrating to other storage/compute environments (e.g. Relational Database Management System)

- Consider hiring a dedicated Data Engineer on the Demographic Frame or Frames Program overall

United States® Census Bureau