International Journal of Population Data Science

Journal Website: www.ijpds.org

U.S. Decennial Census Digitization and Linkage Project

Alexander, T^{1*} and Genadek, K²

¹University of Michigan

²U.S. Census Bureau

The U.S. Census Bureau maintains a large longitudinal research infrastructure that currently includes linked data from the 1940 census, the 2000-2010 censuses, major national surveys going back to 1973, and administrative records dating back to the 1990s. These restricted data are accessible to researchers around the U.S. via the Federal Statistical Research Data Centers (FSRDC) network. The major shortcoming of this infrastructure is that it lacks linkable files from the decennial censuses of 1950 through 1990. Full-count microdata from these censuses are available for research, but datasets from these years do not include respondent names and therefore have not been linked over time. Respondent names for these censuses are available only via the original census returns, which are stored on 258,000 reels of microfilm.

The Decennial Census Digitization and Linkage project (DCDL) is an initiative to recover names from the 1960-1990 censuses and to produce linked restricted microdata files for research use. We describe the results of a pilot project we completed on the 1990 census. For that pilot, we created digital images from census microfilm, hand-keyed "truth data" from those images, supported two teams' attempts to conduct Handwriting Recognition on the images, appended recovered names to already-existing microdata files, and linked the new 1990 census microdata records to previous and subsequent censuses. We describe our processes, the accuracy of the Handwriting Recognition, and the accuracy of the record linkage with the recovered names. We conclude by providing an update on the recently-initiated project to carry out these processes on a production scale for the 1960 through 1990 censuses.

When combined with existing linkages between the censuses of 1940, 2000, 2010, the soon-to-be public 1950 census, and the future 2020 census, DCDL will provide the final component in a massive longitudinal data infrastructure that covers most of the U.S. population since 1940. As a multi-purpose statistical tool, the DCDL will further the U.S. Census Bureau's mission to provide high quality data on the U.S. population and support cutting-edge research in the FSRDC network. The resulting data resource will expand our understanding of population dynamics in the U.S. far beyond what is currently possible, providing transformational opportunities

for research, education, and evidence-building across the social, behavioral, and economic sciences.

*Corresponding Author:

Email Address: jtalex@umich.edu (T Alexander)