

Using U.S. Census Bureau Documentation to Revisit and Rebuild 20th Century Data Complexities

David A. Bleckley, University of Michigan; John Sullivan, U.S. Census Bureau; Douglas Rosin, University of Michigan

Data and documents are produced within a specific historical context, for a designated purpose, and within an available technological capacity. The interplay of this context, purpose, and capacity informs the decisions producers make regarding what information gets recorded via what media for how long. When future users access, analyze, and interpret data and documentation, their ability to decode, understand, and apply these collections in a new time, for a novel purpose, and with emerging technologies depends upon the producer's decisions. Today's user relies on the past producer's documentation to extend the utility of the data.

Data producers may curate their data, metadata, and documentation in a forward-looking manner—anticipating the technological advances to come that will overcome the limitations in utility today. However, some data and documentation are produced for a point-in-time use without curation for future use. Depending on how well-curated the data are, today's user may need to use documentation accompanying a dataset to learn necessary background information, identify metadata elements, and determine how the data may be applied to the current task at-hand. In this paper, we present the various methods we used to digitize historical physical data collections and to leverage documentation in these processes.

Our work to translate physical copies of five data collections into machine-readable datasets is part of the larger Decennial Census Digitization and Linkage project (DCDL). They each were produced with defined purposes in mind, but we are using them in manners ranging from slightly different from the original purpose to completely novel. The 1960 and 1970 Enumeration District Listings (EDLs) are collections of hand-typed books used by the U.S. Census Bureau's Age Search service to locate microfilm reels of decennial census records based on Enumeration District (ED). We are using these as quality checks for our microfilm scanning, to ensure we have the correct EDs on each reel. The 1970 Block Header Record (BHR) is a paper printout of a previously computerized dataset, associating ranges of street addresses to 1970 EDs, census tracts, and blocks. The BHR is being used to help link newly-digitized name data with existing 1970 Census microdata. Similarly, the 1980 Address File allows us to associate an exact street address and census geography with decennial questionnaires scanned from microfilm. Our team is using those relationships to better link name data to the 1980 Census microdata and, in turn, to the composite reference files that make up the backbone of Census linkage infrastructure. The 1990 Group Quarter (GQ) records are microfiche copies of the person-level 1990 Census records for individuals living in group quarters like dormitories, prisons, nursing homes, etc. These will allow us to assign names to 1990 Census GQ microdata and, ultimately, link these records to other census data.

We begin with a brief description of the methods used to digitize these five datasets from physical collections to machine-actionable datasets. We discuss various computer vision-based preprocessing steps taken to prepare the scanned images of these collections. We present two

implementations of optical character recognition (OCR) used to translate images into machine-readable text. We also touch upon our data parsing and quality assurance steps.

The second section of the paper focuses on our use of documentation and data to inform data linkage and application. Specifically, we discuss our use of various documentation files produced during the original production of the data, notably, the decennial census “Procedural Histories.” We describe efforts to trace the original data production process and present examples of how our intended use of these recovered datasets is enabled and constrained by historical documentation. For example, by connecting documents and data, we can follow the drafting of boundaries of a census geographic unit (e.g., an enumeration district) onto a map, the enumeration of structures and address coding to breaker sheet images on microfilm to microdata files, and the decisions, revisions and metadata made along the way.

We describe diverse applications of documentation, ranging from informing novel use of existing datasets to transforming documents into datasets themselves. We make extensive use of U.S. Census Bureau documentation to facilitate linkage between datasets and to guide our use of datasets for the purposes of DCDL. Documents help us better understand the relationships between datasets, but our work also required us to discover the relationships between data and the documentation itself, ensuring that we know which files can be used together. Revisions, decisions and errors that are, at some level, inherent to all data production pose major challenges even when we think certain files can be used together. Through our digitization efforts, we translated documentation into data by transforming the EDLs from an index for single datapoints into a crosswalk for entire datasets at scale.

Documentation is a vital element in the research ecosystem. It clarifies and communicates scientific work, provides vital context to datasets, and creates foundations upon which future endeavors may be built. Data produced today for one intention may serve myriad purposes in the future, and documentation allows future analysts to better understand the potential opportunities and limitations of a dataset. Presenting our experience using documentation of past data collection, tabulation, and storage, this paper seeks to provide insights into the incorporation of documentation in the research process and the production of documentation today for tomorrow’s researchers.