Proceedings of the Biocomplexity Institute
Technical Report 2022-1774

What is the Curated Data Enterprise? Envisioning the Census Bureau of the Future: A 21st Century Census Curated Data Enterprise

Sallie Keller
Distinguished Professor in Biocomplexity
https://orcid.org/0000-0001-7303-7267
sak9tr@virginia.edu

Kenneth Prewitt
Carnegie Professor of Public Affairs and
Special Advisor to the President,
Columbia University
kp2058@columbia.edu

Social and Decision Analytics Division
Biocomplexity Institute & Initiative
University of Virginia

June 1, 2022

Funding: This report was funded by U.S. Census Bureau Contract Census Agreement No. 01-21-MOU-06 and Alfred P. Sloan Foundation, Grants G-2019-11316 and G-2020-14002.

Citation: Keller S, Prewitt K. What is the Curated Data Enterprise? Envisioning the Census Bureau of the Future: A 21st Century Census Curated Data Enterprise. [Other]. Biocomplexity Institute, University of Virginia; 2022 June. https://doi.org/ https:/10.18130/1gb2-fa71.

UNIVERSITY of VIRGINIA
BIOCOMPLEXITY INSTITUTE

Abstract

The Curated Data Enterprise is both an infrastructure and a continuous evolving ambition. The vision is to empower and enable Census Bureau scientists and their data users to develop new and better measures of America's people, places, and economy. This 21st-century model will exploit multiple data sources across sample surveys, censuses, federal, state, and local administrative data, as well as third-party data, in order to produce more robust, granular, timelier, and comprehensive measures of demographic changes, social trends, and economic activity.

Envisioning the Census Bureau of the Future: A 21st Century Census Curated Data Enterprise

What is the Curated Data Enterprise?
June 2022

Sallie Keller, Distinguished Professor in Biocomplexity, Director of the Social & Decision Analytics Division, Biocomplexity Institute at the University of Virginia, sak9tr@virginia.edu

Ken Prewitt, Carnegie Professor of Public Affairs and Special Advisor to the President, Columbia University, U.S. Census Bureau Director, 1998-2001, kp2058@columbia.edu

A New Census Bureau Data Curation Vision

The Curated Data Enterprise is both an infrastructure and a continuous evolving ambition. The vision is to empower and enable Census Bureau scientists and their data users to develop new and better measures of America's people, places, and economy. This 21st-century model will exploit multiple data sources across sample surveys, censuses, federal, state, and local administrative data, as well as third-party data, in order to produce more robust, granular, timelier, and comprehensive measures of demographic changes, social trends, and economic activity.

A New Census Bureau Data Curation Vision

Across more than two centuries and today's more than 130 surveys and censuses, the Census Bureau has earned the nation's trust for high quality, consistent, and reliable data products. It has achieved this reputation by adhering to strict scientific standards in its operations, including survey data curation standards.

Rigorous, reviewable, and repeatable processes are in place for every survey to provide data accuracy and that total survey error is minimized. These efforts ensure the published data are credible and trusted by the survey community, that geographic, temporal, and demographic comparisons are valid, and that documentation is thorough and clear. Moreover, the Bureau strives to create data products useful to data users and researchers that are relevant, useful, and accessible, adhere to published schedules, and provide documentation of coverage and coherence when estimates are from different sources.

Moving Towards An Enterprise Model of Curation

The Census Bureau is innovating its processes to take advantage of new data sources and data science computing innovations and to adapt to declining survey response rates that are challenging the public and private survey world. The Bureau is also exploring additional means of producing data to address shortfalls that are becoming increasingly inherent with surveys. It has stood up a team representing a cross-section of its demographic, geographic, and economic programs to break down siloed activities and build a new enterprise infrastructure. One component of this infrastructure is the "Frames Program" that will link four key components of the internal architecture, Geospatial, Business, Jobs, and Demographic frames, into shared enterprise resources that will bring together, serve, and support all surveys and best exploit the use of administrative and third-party data. Additionally, this resource will form the foundation for the Curated Data Enterprise, creating a scaffold to hold and link massive amounts of public and private sector data. These innovations represent an evolution beyond the survey-only model that has reached scientific and practical limits in an era of increasing demand for more data, more often, and more urgently. Producing trusted, comprehensive, and reliable estimates in the future will require blending survey methods with other sources of data collection.

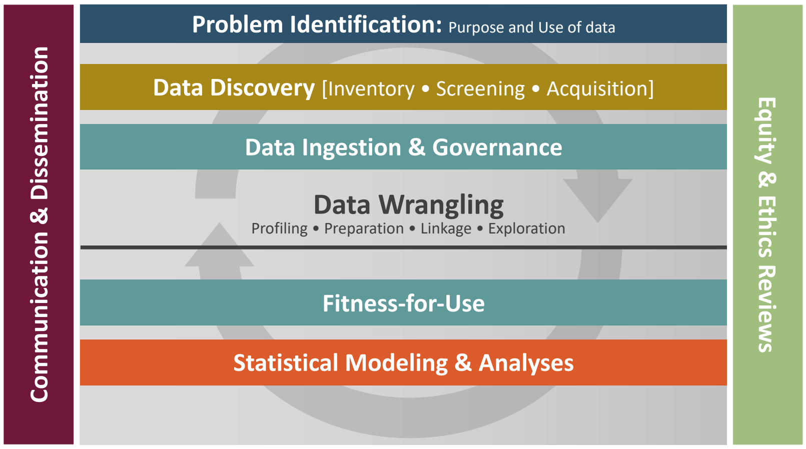
The Curated Data Enterprise transformation requires a new data curation model that greatly expands and enables data discovery and retrieval, maintains data quality across multiple and diverse sources of information, adds value by creating new derived variables, and provides for re-use over time through activities including authentication, archiving, metadata creation, digital preservation, and transformation. Census scientists are engaging with and collaborating with researchers and data users to understand the full potential of this vision.

Consistent with the Bureau's strong tradition for protecting privacy, a key focus of these discussions is to ensure ethical data sharing while adopting procedures to safeguard the confidentiality of respondents' information and at the same time enhance efforts to make published data more findable, accessible, interoperable, trusted, and reusable.

Curated Data Enterprise Framework

The CDE framework provides a rigorous, transparent, and repeatable structure to build the CDE in the context of purpose and use. For a given purpose, the curation steps in Figure 1 include defining the purpose and use that motivates the curation (specific problem to be addressed); discovery of potential data sources relevant to the purpose and use (inventory, screening, and acquisition); data ingestion and governance; data wrangling to understand and prepare data for evaluation (data profiling, data preparation and linkage, and data exploration); fitness-for-use assessment in relation to the purpose; statistical modeling and analyses to extract metrics from or prepare the data products; communication and dissemination of results with stakeholders; and an ongoing ethics and equity review throughout the entire process.

Figure 1. Curated Data Enterprise Framework


Communication & Dissemination
Problem Identification: Purpose and Use of data
Data Discovery [Inventory • Screening • Acquisition]
Data Ingestion & Governance
Data Wrangling
Profiling • Preparation • Linkage • Exploration
Fitness-for-Use
Statistical Modeling & Analyses
Equity & Ethics Reviews

About the University of Virginia's Social and Decision Analytics Division

The **Social and Decision Analytics Division (SDAD)** is a leading Division in the Biocomplexity Institute at the University of Virginia. The Biocomplexity Institute is at the forefront of a scientific evolution, applying a deeply contextual approach to answering some of the most pressing challenges to human health and well-being within our changing environment. SDAD was created in the fall of 2013 to extend the Biocomplexity Institute's capabilities in social informatics, policy analytics, and program evaluation. The researchers at SDAD form a multidisciplinary team, with expertise in statistics, policy and program evaluation, economics, political science, psychology, computational social science, and data governance and information architecture. SDAD's mission is to embrace today's data revolution, developing evidence-based research and quantitative methods to inform policy decision-making and evaluation.