Plan Overview

A Data Management Plan created using DMP Tool

DMP ID: https://doi.org/10.48321/D1D58CCB2E

Title: EAGER: PBI: Designing Public Use Microdata Business Areas (PUMBAs) for Innovation

Creator: Matt Williams - **ORCID:** 0000-0001-8894-1240

Affiliation: RTI International (rti.org)

Principal Investigator: Jennifer Ozawa, Christine Task, Matthew Williams

Contributor: Rob Chew

Funder: National Science Foundation (nsf.gov)

Funding opportunity number: NSF 23-1

Template: NSF-SBE: Social, Behavioral, Economic Sciences

Project abstract:

The team proposes to develop and validate a process for generating new geographic subdivisions especially well-suited for synthetic data release of business microdata. This would be analogous to the Public Use Microdata Areas (PUMA) developed by the Census Bureau for the American Community Survey (ACS). The goal of the proposed work is to aggregate geographic areas (e.g., census tracts or zip code tabulation areas) into contiguous regions of similar size (number of business) and homogeneity (similar types of business).

Start date: 09-01-2024

End date: 08-31-2026

Last modified: 07-08-2024

Copyright information:

EAGER: PBI: Designing Public Use Microdata Business Areas (PUMBAs) for Innovation

Roles and responsibilities

The DMP should outline the rights and obligations of all parties as to their roles and responsibilities in the management and retention of research data. It should also consider changes to roles and responsibilities that will occur should a principal investigator or co-PI leave the institution or project. Any costs should be explained in the Budget Justification pages.

The PI and co-PI's will be responsible for ensuring all team members adhere to the data management plan. If a PI or co-PI leaves their institution or the project, a replacement PI or co-PI will be appointment. Williams and Chew at RTI will manage the geospatial files and methods code outside of the Federal Statistical Research Data Center (FSRDC). Christine Task at Knexus will provide additional quality assessments and checks. All PI's will work with Census Bureau and FSRDC representatives to conform to requirements for working within the FSRDC, including archiving within the FSRDC for reproducibility.

Expected data

The DMP should describe the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project. It should then describe the expected types of data to be retained.

- Several geographic data files which cluster counties or zip codes into regions.
    - These will be shared along with the code used to generate them.
    - Some of the inputs will be based on licensed data from ESRI, but the outputs will simply be county or zip code IDs.
- Data from the Annual Business Survey (ABS) from the Census Bureau and National Center for Science and Engineering Statistics
    - Evaluation metrics for these regions based on ABS microdata
    - Synthetic microdata files based on ABS and geographic regions
    - Summary statistics of business characteristics for these custom geographies will be generated from statistical models.
    - These data files will be accessible through the FSRDC. We will work with Census and NCSES to evaluate if these files will be made available to the public.

Period of data retention

SBE is committed to timely and rapid data distribution. However, it recognizes that types of data can vary widely and that acceptable norms also vary by scientific discipline. It is strongly committed, however, to the underlying principle of timely access, and applicants should address how this will be met in their DMP statement.

We will share data and artifacts as soon as possible. For example, when we present findings at conference or submit for publication, we will strive to have the available at the time of submission. We will adhere to requirements and best practices for sharing our data products within the FSRDC environment.

Data format and dissemination

The DMP should describe data formats, media, and dissemination approaches that will be used to make data and metadata available to others. Policies for public access and sharing should be described, including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements. Research centers and major partnerships with

industry or other user communities must also address how data are to be shared and managed with partners, center members, and other major stakeholders.

- We will use open formats for data storage, for example geojson and csv.
- Any data generated within the FRSDC will be reviewed for disclosure by the Census and NCSES.

Data storage and preservation of access

The DMP should describe physical and cyber resources and facilities that will be used for the effective preservation and storage of research data. These can include third party facilities and repositories.

- For any data that is able to be exported from the FSRDC, we plan to use a repository such as the Harvard Dataverse to host our data.
- We plan to share our code through Github, likely associating any updates or modifications we've made to existing code with the original repositories.
- For data and code within the FSRDC, we will comply with requirements for storage and archiving and provided as detailed information as possible for other research teams to replicate results.

Additional possible data management requirements

More stringent data management requirements may be specified in particular NSF solicitations or result from local policies and best practices at the PI’s home institution. Additional requirements will be specified in the program solicitation and award conditions. Principal Investigators to be supported by such programs must discuss how they will meet these additional requirements in their Data Management Plans.

N/A

Planned Research Outputs

Dataset - "Public Use Microdata Business Areas Definitions"

This dataset will contain one or more clustered geographic areas, where the aggregation is based on similar characteristics such as business industry and density. Different definitions will be based on alternative clustering algorithms and changing the tuning parameters (such as minimum size).

Software - "Code for Generating PUMBAs"

We will apply and extend existing Python and R code for clustering micro-areas.

Dataset - "Synthetic Microdata for PUMBAs"

If synthetic microdata is deemed by the Census and NCSES as safe to release, we will share the microdata externally.

Dataset - "Small Area Estimates for ABS PUMBAs"

This dataset will contain summary statistics for each micro area based on Annual Business Survey. These estimates will be produced from statistical models. If cleared by the Census Bureau and NCSES for release, these will be shared publicly.

Software - "Code for Evaluating PUMBAs"

The code for evaluating the externally defined PUMBAs. If approved for release from the FSRDC by the Census and NCSES, will share publicly.

Software - "Code for Generating SAE"

The code for generating the small area estimates for PUMBAs. Any modifications to existing libraries in R will be shared publicly.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Public Use Microdata Business Areas Definitions	Dataset	2025-09-22	Open	Harvard Dataverse		Creative Commons Attribution Non Commercial 4.0 International	ISO 19115	No	No
Code for Generating PUMBAs	Software	2025-09-22	Open	GitHub		Creative Commons Attribution Non Commercial 4.0 International	None specified	No	No
Synthetic Microdata for PUMBAs	Dataset	2026-09-22	Restricted	Harvard Dataverse Federal Statistical Research Data Center		None specified	None specified	No	Yes
Small Area Estimates for ABS PUMBAs	Dataset	2026-09-22	Open	Harvard Dataverse		Creative Commons Attribution Non Commercial 4.0 International	ISO 19115	No	No
Code for Evaluating PUMBAs	Software	2026-09-22	Open	GitHub		Creative Commons Attribution 4.0 International	None specified	No	No
Code for Generating SAE	Software	2026-09-22	Open	GitHub		Creative Commons Attribution 4.0 International	None specified	No	No