# Census Bureau’s Business Frame Prototype: Linking Additional Data Sources to the Business Register

Author: Jessica Wellwood and Lori Zehr

To be presented at Wiesbaden Group on Business Registers, October 2023, The Netherlands

Note: Any opinions and conclusions expressed herein are those of the author and do not reflect the views of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data (Disclosure Review Board (DRB) approval number:  CBDRB-FY23-ESMD002-031).

The authors would like to thank the Business Frame Research and Development Staff (Erica Marquette, Michael Booker, Christine Cai, Jessica Pearson, and Jeremy Stash) for their contribution in the technical development of the Business Frame Prototype and for their expert review of this report.

# Abstract

The U.S Census Bureau is undergoing a significant transformation process to modernize and revolutionize how we do business to ensure we continue to achieve our mission: To serve as the nation’s leading provider of quality data about its people and economy.

Response rates are continuing to decline, jeopardizing not only data quality, but also the ability to publish granular statistics. Yet, in direct contrast, data users need more data, faster, and at lower geographical levels. To achieve this, we must increase our operational efficiencies by leveraging our existing data more effectively while also incorporating new data sources.

To facilitate this transformation, a data infrastructure must exist that allows Census Bureau staff to access and link various data sources. Sound methodologies for linking datasets must be developed and integrated into this infrastructure. **We are constructing a Business Frame prototype that allows the Frame to coexist with the Business Register with the Business Register continuing to provide full coverage of the business populations**. This prototype will also include a collection of auxiliary data from multiple sources integrated with the master list of units via establishing new linkages. This will provide an enhanced statistical infrastructure that will strengthen research capabilities, promote innovation, simplify processes and systems, and promote managing data as a strategic asset. Existing statistical programs can leverage this robust foundation as they look to meet the demand for more frequent, timely, relevant, and granular data products.

This paper will discuss the **various data sources currently being evaluated for the Business Frame prototype, the methods for linking multiple data sources, the newly defined data architecture guiding the construction of the prototype, and the challenges associated with this project.**

# Background

The Census Bureau’s Economic Directorate’s mission is to use transparent, scientific methods to provide broad and impartial information on the changing U.S. economy for use in public and private sector decision making.[[1]](#footnote-1) To fulfill this mission, the directorate has developed a multitude of statistical programs designed to measure and profile U.S. businesses and government organizations. These programs include the Economic Census, Census of Governments, and over 60 surveys taken monthly, quarterly, annually, or periodically including twelve monthly or quarterly principal economic indicators, extensive compilations of administrative records, and many research and technical studies. Among other things, Federal agencies, policy makers, economists, and private industries use Census Bureau economic data for market analysis, economic forecasting, Gross Domestic Product (GDP) calculations, monetary policies, input-output (I/O) tables, and evaluating the nation’s economic health.

Traditionally, the Economic Directorate’s business model has relied on complex statistical methods to define the target population, select a representative sample, and initiate large-scale collection operations to gather relevant data directly from the sampled population. Administrative data is then used to validate and edit reported data and to impute for non-response cases. In addition, analysts conduct comprehensive reviews of both the micro and macro data ensuring data errors are identified and corrected. This strategy has produced high quality statistical products viewed as the gold standard, trusted sources for economic statistics.

However, this business model is no longer sustainable. Data users want more data, and they want it faster. Yet, programs are faced with declining response rates, budget reductions, and privacy concerns. These opposing forces require the transformation of economic programs and the adoption of a new business model. Programs must find the balance in providing the right data at the right time and at manageable cost without sacrificing quality.

One solution is to leverage existing data more effectively. Data is the Census Bureau’s strategic edge. In addition to the multitude of survey programs, efforts have increased dramatically to ingest a variety of third-party data sources. However, due to a de-centralized approach to survey programs and data management, awareness and access to various data sources is uncoordinated and unintegrated. At the core of the Census Bureau’s transformation lies the creation of a Business Frame that will exist as a renewed data infrastructure that centralizes and links existing data together, in meaningful ways.

The Business Frame will be a rich collection of harmonized business data from multiple sources, integrated together and stored in a central location. It will provide an enhanced statistical infrastructure that will facilitate managing data as a strategic asset, **strengthen research capabilities, promote innovation, and simplify processes**. Existing statistical programs can leverage this robust foundation as they look to meet the demand for more timely, relevant, and granular data products. New data products can be developed by utilizing the new infrastructure.

To achieve this vision, we have constructed an initial prototype of the Business Frame. This prototype is a single relational database that integrates four different data sources within the Business Frame prototype, allowing those sources to be linked to the Business Register. This paper will discuss the **various data sources currently being evaluated for the prototype, the methods for linking multiple data sources, the newly defined data architecture guiding the construction of the prototype, and the challenges associated with this project.**

# Data Sources

The Business Frame prototype establishes linkages for four data sources previously not integrated with the Business Register (BR). Each data source was chosen because it demonstrated a specific dimension of utility and linking methodology. Below are specifics on each data source and why it was selected.

## Business Register

The Business Register (BR) is a current and comprehensive database of business establishments located in the United States and its territories maintained for statistical program use[[2]](#footnote-2). The BR includes business location, industry classification, and other operating data such as employment, payroll, and receipts. The BR is integral to the Business Frame prototype because it serves as the source of the sampling frame for the Economic Census and other economic surveys.

The BR includes businesses operating in the government sector if they pay federal taxes. The duplicative nature of government sector information in the BR originates from the following institutional practices. The BR classifies establishments based on the North American Industry Classification System (NAICS). However, due to differences in reporting structures for government establishments versus business establishments, the Census Bureau does not consider NAICS to be an adequate classification system for government programs. That is, the purpose of NAICS is to classify an establishment’s primary business activity whereas the government programs at the Census Bureau measure all activity performed by a government entity. To address this, the Census Bureau created and maintains the Governments Master Address File (GMAF) separately from the BR.

## The Governments Master Address File

The Governments Master Address File (GMAF) includes a list of state and local governments in the United States. Local governments include counties, cities, townships, school districts and special districts (such as water, fire, and library districts). A multitude of data is collected for each government through various economic survey programs. Data collected include general information on each government, such as activities and functions performed, government type and operating status, as well as financial information regarding revenues, expenditures, payroll, reemployment, pensions, and enrollment information specific to school districts[[3]](#footnote-3).

Maintaining two separate registers for businesses and governments introduces some duplication and redundancy because there are units in industries that are potentially surveyed in both government and industry programs. To illustrate this point, we identified potential overlapping records in the BR and the GMAF. There are about 122,000 active government entities recorded in the GMAF. About 94% the active government units in the GMAF have the potential to overlap with records in the BR.

Linking the BR and the GMAF will benefit both registers. This has the potential to improve both frames in the following ways:

* The number of unclassified or misclassified records can be reduced by leveraging classification variables on both registers.
* Sharing or enhancing imputation methods will improve the quality of data.
* Improving the quality of contact information for businesses on the Business Register could improve the Census Bureau’s ability to initiate contact with a respondent which could in turn increase response rates.
* Lastly, if we leverage already validated information about a business, this could reduce respondent burden.

## Longitudinal Business Database

The Longitudinal Business Database (LBD) is a derivative of the Business Register[[4]](#footnote-4). Whereas the Business Register represents a single reference period, the LBD links establishments records over time, using annual snapshots from the Business Register dating back to 1976. The LBD contains annual employment, payroll, and industry classification for all workplace establishments across all industries. Because the BR is a snapshot at a single point in time and the LBD is longitudinal in nature, one must leverage the LBD to determine age and size of establishments and enterprises in the BR. The primary purpose of the LBD is to measure business dynamics (i.e., openings and closings (births and deaths) and other structural changes of businesses). The LBD also constructs measurements of size and age of establishments and enterprises with economic activity.

The LBD was selected for inclusion in the Business Frame prototype because it contains longitudinal linkages; linking the LBD to the BR and storing in a central location will promote awareness and encourage use of the LBD while adding a longitudinal dimension to the BR. For example, after linking the LBD to the BR, programs could incorporate firm age into sampling parameters. Further, firm age as a measure of business dynamism (i.e., entry of new firms and creation of employment) is related to economic growth. Linking this information to the BR could result in constructing new indicators measuring the health of the economy. Firm age is also important because firms in different age categories respond differently to the introduction of new technology, trade shocks, and economic upswings or downturns in general. Age could be as important as size in stratifying firm samples and targeting survey questions. Linking the LBD to the BR will facilitate measurement of these phenomena.

## Point-of-sale data provided by Circana

Circana is a private sector market research company that collects retailer point-of-sale data[[5]](#footnote-5). The Census Bureau purchases data from Circana covering a small number of retailers comprising thousands of establishments. This data is received monthly after Circana has completed its internal processing, editing, and analysis of the data. The data set contains sales revenue aggregated by product code for each establishment from which Circana receives data. This data not only provides timely sales information for select retailers in the Business Register at the establishment level, but it also introduces a new level of granularity by providing product details that are mapped to North American Product Classification Codes (NAPCS). Connecting the point-of-sale data to the BR enriches the BR data by providing organizational linkages; that is, the Circana data validate establishment to parent company links. Further, linking Circana to the BR adds the timeliest establishment-level data as well as relevant data to the BR that could possibly be used to assist with determining a company’s operational status. This data was selected for the Business Frame prototype to demonstrate the ability to link third-party data to the BR from a reference period more current than the BR. Further, increasing access and awareness of this timely and granular level of detail offers improvement opportunities for programs, as well as the development of new measurements. For example, the Census Bureau has already used third-party data like Circana to construct experimental monthly state retail sales data[[6]](#footnote-6).

## Non-employer Statistics by Demographics (NES-D)

NES-D is a data product created using administrative records from the BR and data collected from the Decennial Census, the American Community Survey, and other demographic administrative data sources. The NES-D program assigns demographic characteristics for each establishment based on the demographic characteristics of the business owners, adjusting for when multiple owners exist for a single business. Characteristics produced are sex, ethnicity, race, and veteran status. Additionally, data for business owner age, place of birth, and citizenship status are calculated. By linking NES-D to the BR, a demographic dimension is created for businesses for use in additional data products. In addition, this data provides the foundation to link the Demographic and Business Frames together[[7]](#footnote-7).

# Methodology

The team employed three methodologies when linking data to the Business Register in support of the Business Frame Prototype: common identifier, probabilistic matching, and analyst review. The three methodologies are described below.

## Common Identifier

A common identifier is a shared identifier between two or more data sources. This is the preferred method when linking data together. It requires few resources and yields highly accurate results. The BR contains many different identifiers for each record which proved beneficial when linking to various data sources. Of the four initial data sources linked to the BR for the Business Frame prototype, three were able to use a common identifier. However, different common identifiers were used across the three sources. The rest of this section illustrates the depth and variety of identifiers on the BR.

Currently the BR has two types of units: **administrative units** used for tax administration, and **statistical units** used for survey programs.

### *Administrative Units*

The administrative units on the BR are the EIN and the PIK. An **EIN** (employer identification number) is a unique nine-digit number assigned to businesses operating in the US by the Internal Revenue Service (IRS) for tax administration. However, the identifier is used for other administrative purposes, including State Unemployment Insurance systems. Because of its broad use, the EIN provides a link between the BR and many other data sources. An EIN can provide information for one or more establishments. An enterprise could also have one or more EINs associated with it. When an EIN is known for an establishment or establishments, we link those establishments to their corresponding EIN, and when applicable, link EINs to their parent enterprises. Similarly, the **PIK** (protected identification key) is a unique nine-digit anonymous identifier as unique as a social security number. In the context of the BR, the PIK serves as a tax identifier for businesses operating as sole proprietors.

While it is labor intensive maintaining so many identifier structures and their associated linkages, it does make linking data sources possible for the Business Frame prototype. When identifiers are shared between the data sources, we can assume the link is accurate with reasonable confidence.

Currently, all linkages described above are utilized to link data to the BR for the Business Frame prototype. LBD data is linked to the BR using the BR ID. NES-D data is linked using EIN and PIK. Circana data is first mapped to an enterprise identifier via analyst review; it is then linked to the BR using the combination of store number and enterprise identifier.[[8]](#footnote-8)

### *Statistical Units*

An **establishment** is a single location where business or other activities are conducted. An establishment is the basic unit on the BR. All establishments are assigned a BR ID with a unit type of establishment[[9]](#footnote-9). This ID is used for internal processing. All other units on the BR are built from establishments.

An **Enterprise** is a collection of establishments under common ownership or control. An example of an enterprise is when multiple grocery stores are owned by a single parent company[[10]](#footnote-10). The parent company is an enterprise, and each grocery store location is a child establishment. Enterprises are also assigned a BR ID; however, their unit type is enterprise. Enterprises are also assigned an ‘**Alpha** ID’, a unique 6-digit code. Establishment IDs are linked to their parent enterprise. Many enterprises also maintain a ‘**store number**’ for establishments under their ownership. In these cases, the store number is an ID created by the enterprise, or parent company. Due to the importance of store number in a company’s internal tracking, Census Bureau programs ask businesses to report any store numbers they use. We store this in the BR and provide it on survey forms when corresponding with businesses.

## Probabilistic Matching

To link the BR and the GMAF, we plan to use probabilistic matching. We have conducted some testing using an algorithm called MAMBA (Multiple Algorithm Matching for Better Analytics)[[11]](#footnote-11). The MAMBA matching methodology uses machine learning to conduct pair-wise matching for 13 string comparators measuring the similarity between business names and addresses and assigns a probability score to each comparison. The scores are then averaged to construct a single composite score indicating the likelihood that the records are a match. Implementation of this algorithm is complex and resource intensive. However, we believe the investment in MAMBA will yield more matches with higher quality.

## Linking Architecture

A key component of the Business Frame is to develop a data architecture that is agile in nature, flexible in structure, and available on a need-to-know basis. Anticipating that more data sources will be incorporated into the Business Frame prototype in the coming years, determining the best organizing principle is key. The model needs to be expandable, adaptable, and sustainable. After reviewing the data architectures for the four data sources we ingested into the prototype, we decided to organize around the administrative and statistical units. Each unit type has a primary table containing information about the unit. Separate tables are then created containing the data items for those units and related back to the primary unit table. Further, another table is then created to relate one unit to another unit. As a result, several tables exist simply to relate one unit type to another, serving as a series of crosswalks between data sources.

For example, the data architecture for the administrative unit EIN is illustrated in Figure 1. The grey box is a table containing the administrative unit defined as EIN. In this example, the EIN unit table contains one row for each EIN in the BR. Key information about the unit is contained at the column level. The information includes, but is not limited to, name, operating status, NAICS code, and legal form of organization (LFO). Tables grouping similar data items from the BR (e.g., payroll and revenue tables) are connected to the EIN unit table using EIN as a primary key.

Figure 1. Example data architecture for an administrative unit



In addition to the tables storing data items for EIN units, a series of tables exist to house the relationship of an EIN to a different unit type. For instance, as noted in the ‘common identifier’ section, an EIN is a unit created for tax administration purposes. Connected to each EIN is at least one establishment whose tax information is reported under that EIN. EINs may also be connected to an enterprise or parent company. All establishments and enterprises are assigned an Employer Unit ID. To capture this complexity, a table is created containing all EIN-to-employer unit relationships. One record exists for each unit combination between EINs and employer units. One EIN may be listed multiple times in this table if multiple establishments are reporting tax information under a single EIN, or if the EIN has a parent company. See Table 1 for an example. The primary key on this table is the combination of the EIN and Employer Unit ID. Users can query a particular Employer Unit in this table, then join it with any of the EIN data tables using the EIN ID, resulting in both EIN data and employer unit data for a business.

***Table 1. EIN to Employer Unit Table (Data are fictional and for illustrative purposes.)***

|  |  |  |
| --- | --- | --- |
| EIN | Employer Unit | Employer Type |
| 123456789 | 200123456 | Establishment |
| 123456789 | 200987654 | Establishment |
| 123456789 | 200554433 | Enterprise |

The connection between an EIN and an employer unit is a critical linkage because it connects the administrative unit to the statistical unit. Connecting these two ID structures provides the foundation to link other data sources to the BR. In Figure 2 below, grey boxes represent units and blue boxes represent data sources. All data sources are linked to either a statistical or administrative unit using a common ID, except for the GMAF data. The NES-D data is linked to administrative units using EIN or PIK. LBD and BR data derived from administrative tax data are connected to statistical units using employer ID. Circana data is also connected to the employer ID but does not use the employer ID as the common identifier. Instead, it uses a combination of Alpha (6-digit ID assigned to enterprises) and store number (an ID assigned to individual establishments within an enterprise). All establishments linked to an enterprise have an enterprise Alpha ID associated with them in the BR. To connect Circana, all records containing the specific Alpha ID are mapped to the records in Circana with the same Alpha. Store number is then used to connect individual establishments within each Alpha grouping.

***Figure 2. Business Frame Conceptual Data Architecture***



This data architecture is expandable. As new data sources are identified, they will first be assessed for a common identifier to connect them to an existing unit on the Business Frame. If the unit does not currently exist in the BR, then probabilistic matching methodology will be applied to create a crosswalk between the new unit and an established BR unit. This is how the GMAF and BR units were linked. A new Government Unit table was created and matched to employer units on name and address via probabilistic matching. Data from the GMAF was then stored in a table connected to the Government Unit Table. Users can link data from any of the other data sources in the Business Frame using the Government to Employer Unit crosswalk. Expansion is flexible, limitless, and consistent with promoting data integrity and usability among users. The Business Frame will enhance and improve our existing data infrastructure through establishing new linkages.

# Challenges and Next Steps

Connecting data and storing those connections in a central location allows programs to leverage existing data more effectivity. As response rates decline and the demand for data products increases, fully utilizing existing data is a cost-effective solution to this challenge. However, many challenges are inherent when linking data from multiple sources. Both conceptual and physical differences exist across the data sources. Examples of physical differences between source data include different file types (e.g, CSV, SAS, JSON), field formats (e.g., numeric, integer, date), data structures (wide vs long structures), different naming conventions applied to field names, and varying degrees of metadata. These differences in file types and formats create challenges with ingesting and interpreting the data. We developed a complex transformation process and thorough testing procedures to ensure files are correctly harmonized without compromising the integrity of the data during the load process. Likewise, conceptual differences between the data sources threaten the integrity and interpretation of the linked data. For example, the definition of revenue is not the same between the Circana data and the tax records. Yet both sources contain a revenue variable. Users may not fully understand the differences in definitions when developing a blended data product leveraging revenue values for a portion of the measured population and tax data for the other.

Differences in how units are defined pose unique linking challenges. A record may appear to be the same between the GMAF and the BR, but the two differ significantly. For example, both the GMAF and BR may contain a record with the same name and address; the record may share the fictitious name *School Board ABC*. In the BR, that unit represents the high school, middle school, and three elementary schools. However, in this example, *School Board ABC* represents the central office of the school district in the GMAF. When comparing these two records, the name and address match exactly, yet the data items differ significantly because the BR and GMAF unit definitions differ as described in the aforementioned example. This is an example of how sharing data across programs with unit definitional differences introduces error. The challenge itself lies in uncovering that a difference exists in the first place.

Each of the Business Frame prototype data sources are governed by different laws and regulations. Information in the BR, the LBD, and internal processing of the NES-D is confidential and protected by Title 13 and Title 26. Title 13 protects data collected by the U.S. Census Bureau and Title 26 protects Federal Tax Information[[12]](#footnote-12),[[13]](#footnote-13). Data provided by Circana are protected by Title 13 and a Nondisclosure agreement with Circana. The GMAF, however, is publicly available. Title 13 and Title 26 requirements must be considered when developing the Business Frame prototype architecture and must be adhered to when managing access to the Business Frame prototype.

The final and arguably most difficult challenge is maintaining the data integrity of both the source data and the subsequent linkage of data sources. The accuracy, completeness, and quality of linked data can be difficult to assess. For instance, many differences across data sources are unknown. Until the inconsistencies between the data sources comprising the Business Frame prototype are known, users risk interpreting data incorrectly.

After establishing the Business Frame prototype, the next steps are to evaluate the linkages established in the Business Frame prototype. The evaluation results will be documented and distributed to internal users. As appropriate, we will apply disclosure avoidance techniques to the evaluation results and disseminate to external audiences. Additionally, we will develop user guide documentation for users of the Business Frame prototype, and eventual Business Frame, to aid users in selecting the appropriate fields and linkages within the Business Frame prototype.

# Conclusion

The Census Bureau must transform to continue to meet the changing demand of data users. Linking multiple data sources at the U.S. Census Bureau represents a necessary 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.[[14]](#footnote-14) The Business Frame is at the core of this transformation of shifting from a survey-centric to a data-centric ecosystem. It is providing a collection of rich, harmonized business data from multiple sources, integrated, and stored in a central location. It will provide an enhanced statistical infrastructure that will strengthen research capabilities, promote innovation, simplify processes, and promote managing data as a strategic asset. Existing programs producing economic statistics will leverage this robust foundation as they look to meet the demand for more timely, relevant, and granular data products.

1. Learn more about the Economic Directorate here: <https://www.census.gov/about/leadership/economic-programs.html> [↑](#footnote-ref-1)
2. See an overview of the Business Register here: <https://www2.census.gov/ces/wp/2016/CES-WP-16-17.pdf> [↑](#footnote-ref-2)
3. See an overview of the Government Master Address File here: <https://www.census.gov/programs-surveys/gus/technical-documentation/methodology/information-collection.html> [↑](#footnote-ref-3)
4. Read about the Longitudinal Business Data base here: <https://www.census.gov/programs-surveys/ces/data/restricted-use-data/longitudinal-business-database.html> [↑](#footnote-ref-4)
5. Learn how circana data is used here: <https://www.nber.org/system/files/chapters/c14270/c14270.pdf> [↑](#footnote-ref-5)
6. Learn more about the experimental monthly state retail sales data here: <https://www.census.gov/library/visualizations/interactive/monthly-state-retail-sales.html> [↑](#footnote-ref-6)
7. Learn more about the Frames Program here: <https://www.census.gov/newsroom/blogs/research-matters/2022/10/technology-transformation.html> [↑](#footnote-ref-7)
8. Alpha IDs are explained in the next section [↑](#footnote-ref-8)
9. There are several unit types, defined as: **Single-unit (SU)**, an enterprise (company or organization) with one establishment; **Multi-unit (MU)**, an establishment owned or controlled by a company or organization with more than one establishment; **Submaster (SBM)**, EIN submaster (representing all establishments linked to a specific EIN for multi-unit enterprises); **Enterprise (ENT),** Multi-unit enterprise ID; and **Single-unit EIN Nonemployer (SUN), a**n enterprise (company or organization) with one establishment but no paid employees. [↑](#footnote-ref-9)
10. An enterprise can be a single-unit or multi-unit. A single-unit enterprise is a company or organization with one establishment. A multi-unit enterprise is a company or organization with more than one establishment. For multi-unit enterprises, unit type = ‘ENT’ for the parent ID; ‘MU’ for each child establishment; and ‘SBM’ for an EIN-level ID.) [↑](#footnote-ref-10)
11. The MAMBA methodology is described in the following working paper: <https://www.census.gov/library/working-papers/2018/adrm/ces-wp-18-46.html> [↑](#footnote-ref-11)
12. <https://www.census.gov/history/www/reference/privacy_confidentiality/title_13_us_code.html> [↑](#footnote-ref-12)
13. <https://www.census.gov/history/www/reference/privacy_confidentiality/title_26_us_code_1.html> [↑](#footnote-ref-13)
14. 2023 National Academies of Science Consensus Report titled “Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources” [↑](#footnote-ref-14)