# Data Preparation Guide for Vesta

## 1. Geographic and Temporal Setup

Before working with your data file (the data prior to importing it into Vesta), always begin by preparing the **geographic** and, if applicable, **temporal** components, if your data file contains these.

Vesta is designed to analyze data across space and time, but a time identifier is not required to import data into the software. If your data file does not include a time field, you can simply select, “Always Valid” during import to indicate that your data have no temporal components.

Your data file may include:

### GEOGRAPHIC IDENTIFIER:

A **geographic identifier** (e.g., state FIPS, county FIPS, ZIP code, census tract ID) matches areal or point locations to collected data. This is particularly important for merging data files in Vesta, but we’ll get into that later.

### TIME/TEMPORAL IDENTIFIER:

A **time identifier** uniquely represents each time. Your time information can appear as a single-time observation (e.g., YYYY) or as a time interval with start and end times.

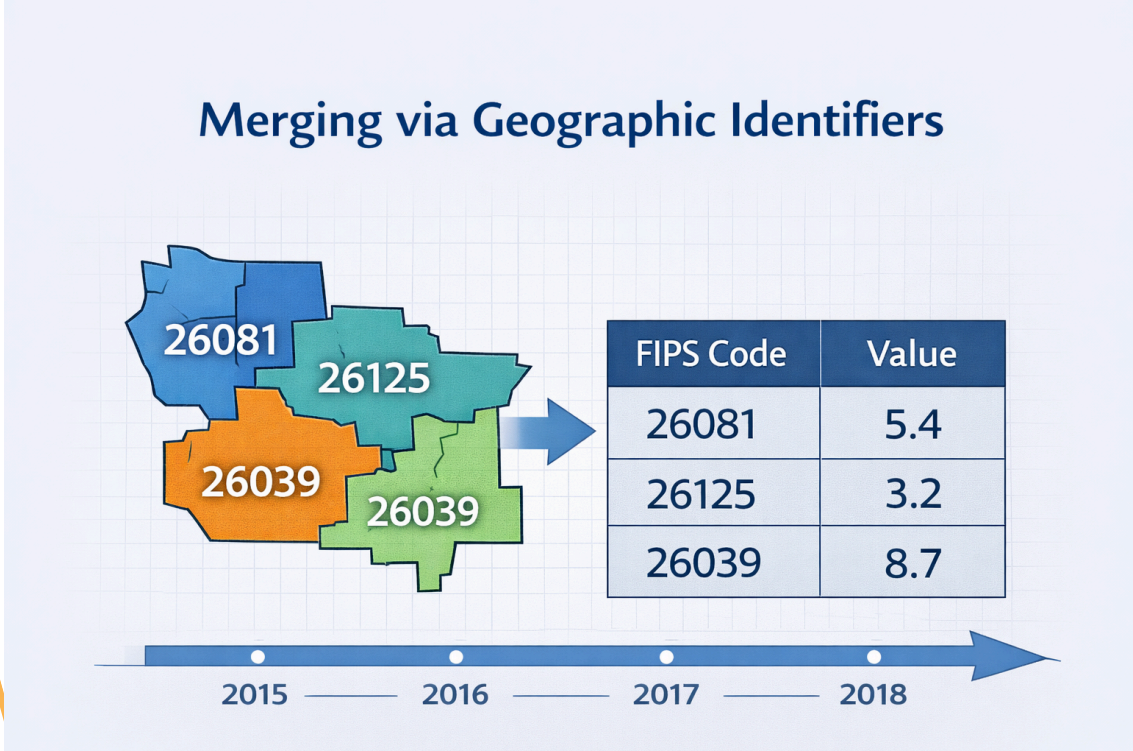
# 1.1 Geographic Identifiers

Having a functional shapefile is key to ensuring you data can be properly imported and used in Vesta. Obtain shapefiles from reliable sources such as the U.S. Census Bureau's TIGER/Line Files.

Each geographic polygon must be linked to a **unique identifier** (e.g., FIPS code). Identifiers can be stored as either **numeric** or **string** variables in a data file. What matters is that the values are identical across data files for merging. These identifiers are what Vesta uses to merge your attribute data with spatial boundaries from a shapefile.

**Attribute data** are the tabular data values associated with geographic features, such as counts, rates, or characteristics, that are linked to spatial boundaries using the geographic identifier, as described above.

Some attribute data are captured as **point data** with sampling coordinates. These may already exist in a shapefile or a spreadsheet. If these exist in Excel files, these need to be expressed as latitude and longitude (WGS 1984 format).



# 1.2 Time or Temporal Identifiers

Temporal information is optional in Vesta, but when present, it allows the software to animate and analyze changes over time. If your data file does **not** contain a time variable, simply import as **Always Valid**. If your data file spans multiple times (e.g., years, dates), you must include a **time variable** so Vesta can determine when the data changes.

If your data file does contain temporally changing information, Vesta recognizes the following time variable formats:

### SINGLE-TIME COLUMN

Create **one column** with one of the following formats:

1. YYYY
2. MMDDYYYY
3. YYYYMMDD
4. MM/DD/YYYY (US-date)

### TIME-INTERVALS, TWO COLUMNS

If your data represents a duration, include **two columns**:

1. Start time
2. End time

Follow these guidelines to ensure your file imports and aligns correctly with other sources. Here is an example file with a time-variable using YYYY below:

| Location_ID | Year | Latitude | Longitude | Rate (%) |
|---|---|---|---|---|
| AA | 2015 | 42.2808 | 83.743 | 7 |
| AA | 2016 | 42.2808 | 83.743 | 9 |
| AA | 2017 | 42.2808 | 83.743 | 12 |
| AA | 2018 | 42.2808 | 83.743 | 18 |

Temporal data files consist of repeated measurements collected at defined intervals over time. For example, census-based data are collected periodically and are understood to represent population characteristics for a limited time window (e.g., several years), after which updated measurements are required to reflect current conditions.

# 2. Data Preparation Procedure

## 2.1 Importing Raw Data

Begin by importing your **raw data file** into your preferred analysis environment (e.g., R, Python, SPSS, SAS). Always keep the original file unchanged – work on a clean copy instead. This ensures you can return to the unaltered source if errors occur.

Different formats require special care:

- **Excel (.xlsx)**: May be prone to datatype misclassification (e.g., numbers imported as text)
- **Text or CSV files**: Pay attention to delimiters and column headers to ensure data is read correctly

While importing, record notes on the following:

- Whether each variable is imported with the correct datatype (numeric, categorical, string, etc.)
- Any issues with encoding (such as UTF-8 vs ANSI), delimiters, or column names that need fixing

Below is an example of a raw dataset that has been imported from the CDC:

| # | sitecode | sitename | year | age | sex | race4 | race7 | sexid | sexid2 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Arkansas | Arkansas [AR] | 2021 | 12 years old or younger | N/A | N/A | N/A | Questioning | Other/Questioning |
| 2 | Arkansas | Arkansas [AR] | 2021 | 12 years old or younger | N/A | N/A | N/A | Other | Other/Questioning |
| 3 | Arkansas | Arkansas [AR] | 2021 | 13 years old | N/A | N/A | N/A | Heterosexual | Heterosexual |
| 4 | Arkansas | Arkansas [AR] | 2021 | 14 years old | N/A | N/A | N/A | Gay or Lesbian | Sexual Minor |
| 5 | Arkansas | Arkansas [AR] | 2021 | 14 years old | N/A | Hispanic/Latino | Hispanic/Latino | Other | Other/Questioning |
| 6 | Arkansas | Arkansas [AR] | 2021 | 15 years old | N/A | N/A | N/A | Gay or Lesbian | Sexual Minor |
| 7 | Arkansas | Arkansas [AR] | 2021 | 15 years old | N/A | White | White | Other | Other/Questioning |
| 8 | Arkansas | Arkansas [AR] | 2021 | 15 years old | N/A | N/A | N/A | Heterosexual | Heterosexual |
| 9 | Arkansas | Arkansas [AR] | 2021 | 16 years old | N/A | Hispanic/Latino | Hispanic/Latino | Heterosexual | Heterosexual |
| 10 | Arkansas | Arkansas [AR] | 2021 | 16 years old | N/A | N/A | N/A | Gay or Lesbian | Sexual Minor |
| 11 | Arkansas | Arkansas [AR] | 2021 | 16 years old | N/A | White | White | Bisexual | Sexual Minor |
| 12 | Arkansas | Arkansas [AR] | 2021 | 16 years old | N/A | N/A | N/A | Heterosexual | Heterosexual |
| 13 | Arkansas | Arkansas [AR] | 2021 | 12 years old or younger | Male | Black or African American | Black or African American | Heterosexual | Heterosexual |
| 14 | Arkansas | Arkansas [AR] | 2021 | 12 years old or younger | Male | Hispanic/Latino | Hispanic/Latino | Heterosexual | Heterosexual |
| 15 | Arkansas | Arkansas [AR] | 2021 | 12 years old or younger | Male | N/A | N/A | Bisexual | Sexual Minor |
| 16 | Arkansas | Arkansas [AR] | 2021 | 13 years old | Male | Hispanic/Latino | Hispanic/Latino | Heterosexual | Heterosexual |

## 2.2 Handling Missing Data

Missing data should be replaced with a **placeholder value** that will never occur naturally in your dataset. For example, using '-99' for continuous variables if all actual values fall between 0 and 100.

For **categorical variables**, use a **consistent symbol** such as " " (empty string), "-" or "NA". Choose one format and apply it uniformly across the dataset so Vesta can recognize missing entries correctly.

The goal is to make missing values clear, consistent, and easy for Vesta to handle during aggregation and mapping.

In the example below, missing values have been replaced with -99:

| # | sitecode | sitename | year | age | sex | race4 | race7 | sexid | sexid2 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Arkansas | Arkansas [AR] | 2021 | 12 years old or younger | -99 | -99 | -99 | Questioning | Other/Questioning |
| 2 | Arkansas | Arkansas [AR] | 2021 | 12 years old or younger | -99 | -99 | -99 | Other | Other/Questioning |
| 3 | Arkansas | Arkansas [AR] | 2021 | 13 years old | -99 | -99 | -99 | Heterosexual | Heterosexual |
| 4 | Arkansas | Arkansas [AR] | 2021 | 14 years old | -99 | -99 | -99 | Gay or Lesbian | Sexual Minor |
| 5 | Arkansas | Arkansas [AR] | 2021 | 14 years old | -99 | Hispanic/Latino | Hispanic/Latino | Other | Other/Questioning |
| 6 | Arkansas | Arkansas [AR] | 2021 | 15 years old | -99 | -99 | -99 | Gay or Lesbian | Sexual Minor |
| 7 | Arkansas | Arkansas [AR] | 2021 | 15 years old | -99 | White | White | Other | Other/Questioning |
| 8 | Arkansas | Arkansas [AR] | 2021 | 15 years old | -99 | -99 | -99 | Heterosexual | Heterosexual |
| 9 | Arkansas | Arkansas [AR] | 2021 | 16 years old | -99 | Hispanic/Latino | Hispanic/Latino | Heterosexual | Heterosexual |
| 10 | Arkansas | Arkansas [AR] | 2021 | 16 years old | -99 | -99 | -99 | Gay or Lesbian | Sexual Minor |
| 11 | Arkansas | Arkansas [AR] | 2021 | 16 years old | -99 | White | White | Bisexual | Sexual Minor |
| 12 | Arkansas | Arkansas [AR] | 2021 | 16 years old | -99 | -99 | -99 | Heterosexual | Heterosexual |
| 13 | Arkansas | Arkansas [AR] | 2021 | 12 years old or younger | Male | Black or African American | Black or African American | Heterosexual | Heterosexual |
| 14 | Arkansas | Arkansas [AR] | 2021 | 12 years old or younger | Male | Hispanic/Latino | Hispanic/Latino | Heterosexual | Heterosexual |
| 15 | Arkansas | Arkansas [AR] | 2021 | 12 years old or younger | Male | -99 | -99 | Bisexual | Sexual Minor |
| 16 | Arkansas | Arkansas [AR] | 2021 | 13 years old | Male | Hispanic/Latino | Hispanic/Latino | Heterosexual | Heterosexual |

## 2.3 Checking and Correcting Datatypes

After importing your data, verify that each variable has the appropriate datatype. Datatypes determine how software interprets and processes your data.

| VARIABLE | VESTA DATATYPE OPTIONS |
|---|---|
| Continuous | numeric (integer or float) |
| Categorical | string or factor, numeric (integer) – sometimes integers |
| Geographic | string or integer, depending on how your shapefile stores them. For example, ZIP codes may be numeric, while state names are strings. |

## 2.4 Subset Large Datasets

Many users work with external and very large datasets – sometimes millions of records. Before you begin aggregating or merging, **create a subset** that includes only the variable and observations relevant to your analysis.

Start by filtering key variables such as **year**, **population group**, or **geographic area**. Subsetting allows you to:

- Verify variable names and spellings (e.g., consistent state names)
- Identify and correct missing geographic entries
- Check data types before scaling up

If you plan to link external datasets, subset each one to the variables you actually need – this keeps files small and easier to merge.

## 2.5 Aggregating Data by Geographic Units

GIS software – and Vesta in particular – operate at the geographic unit level. If your dataset includes multiple observations per area (for example, individuals within each state or county), you must **aggregate** the data so that each geographic unit has a single record. Of note, if your data contains point data, then this step is not necessary.

Common aggregation methods:

- Proportion (%) – for categorical responses (e.g., percentage of youth reporting depression)
- Mean or Median – for continuous measures (e.g., average income or BMI)

For instance, when analyzing youth mental health outcomes by state, aggregate individual responses into a single state-level measure.

You can also aggregate by **merging datasets** that share the same geographic identifier (e.g., state FIPS code). Select your aggregation approach based on your **research question**:

- If studying compositional variables (e.g., proportion of students with suicidal ideation), use proportion
- If focusing on average levels (e.g., mean income by county), use means or medians

Find example of aggregated data below:

| | sitecode | AgeTotalN | %12YearsOrOlder | %13YearsOld | %14YearsOld | %15YearsOld | %16YearsOld | %17YearsOld | %18YearsOldOrOlder | %MissingAge | SexTotalN | %MissingSex | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Arizona | 1133 | 0.4 | 0.4 | 29.0 | 26.7 | 23.7 | 15.3 | 4.5 | -99.0 | 1133 | 0.6 | |
| 2 | Arkansas | 1481 | 0.7 | 0.5 | 8.6 | 26.1 | 29.0 | 21.4 | 13.7 | 0.1 | 1481 | 0.8 | |
| 3 | Colorado | 769 | 0.4 | 0.3 | 20.3 | 26.4 | 24.7 | 20.5 | 6.9 | 0.5 | 769 | 1.4 | |
| 4 | Florida | 4380 | 0.3 | 0.5 | 14.9 | 28.7 | 24.8 | 20.5 | 10.1 | 0.1 | 4380 | 0.5 | |
| 5 | Hawaii | 9508 | 0.2 | 1.2 | 18.1 | 27.7 | 23.5 | 22.9 | 6.4 | 0.1 | 9508 | 0.5 | |
| 6 | Illinois | 5577 | 0.1 | 0.2 | 16.3 | 28.2 | 25.6 | 21.7 | 7.8 | 0.1 | 5577 | 0.7 | |
| 7 | Indiana | 1899 | 0.3 | 0.2 | 11.7 | 34.5 | 26.8 | 16.5 | 9.7 | 0.3 | 1899 | 1.3 | |
| 8 | Iowa | 1326 | 0.2 | 0.5 | 21.3 | 26.6 | 27.4 | 19.3 | 4.5 | 0.2 | 1326 | 0.5 | |
| 9 | Kansas | 1383 | 0.4 | 0.3 | 18.2 | 34.6 | 22.9 | 17.9 | 5.4 | 0.3 | 1383 | 1.3 | |
| 10 | Kentucky | 3938 | 0.2 | 0.2 | 15.8 | 27.4 | 28.6 | 20.6 | 7.1 | 0.1 | 3938 | 0.7 | |
| 11 | Maryland | 70474 | 0.4 | 0.4 | 21.8 | 27.1 | 24.9 | 20.5 | 4.7 | 0.1 | 70474 | 0.8 | |
| 12 | Massachusetts | 2945 | 0.2 | 0.1 | 11.9 | 27.1 | 26.0 | 23.3 | 11.3 | 0.1 | 2945 | 0.5 | |
| 13 | Michigan | 5638 | 0.4 | 0.8 | 19.8 | 28.7 | 25.4 | 19.6 | 5.2 | 0.2 | 5638 | 0.8 | |
| 14 | Mississippi | 2197 | 1.0 | 0.3 | 18.6 | 26.9 | 24.8 | 20.4 | 7.8 | 0.3 | 2197 | 1.1 | |
| 15 | Missouri | 1490 | 0.1 | 0.1 | 17.0 | 35.0 | 24.4 | 17.1 | 6.2 | 0.1 | 1490 | 0.7 | |
| 16 | Nebraska | 634 | 0.2 | 0.3 | 14.5 | 27.0 | 26.3 | 25.2 | 6.5 | -99.0 | 634 | 0.3 | |
| 17 | Nevada | 2989 | 0.4 | 0.5 | 21.4 | 27.0 | 26.0 | 17.9 | 6.7 | 0.1 | 2989 | 1.3 | |
| 18 | New Hampshire | 25001 | 0.2 | 0.1 | 18.3 | 27.2 | 24.5 | 22.4 | 7.1 | 0.2 | 25001 | 1.6 | |
| 19 | New Jersey | 1646 | 0.1 | 0.4 | 19.1 | 23.9 | 25.6 | 24.1 | 6.7 | -99.0 | 1646 | 0.5 | |

Showing 1 to 18 of 30 entries, 31 total columns

# 3. Data Import

Once you have completed the steps outlined above to prepare your data file, you are ready to import it into Vesta!

For guidance on working with your data on importing your data into Vesta, please refer to the Vesta documentation available at the link below:

https://biomedware.com/vesta/docs/DataUse/DataImport/

**If you have any suggestions to improve this guide or further questions about preparing your data for importing into Vesta, please reach out to support@biomedware.com**