# CenSoc-Numident Sibships User Guide*

## Description

These files allow researchers to identify sibships in CenSoc-Numident data.

The CenSoc-Numident dataset links Social Security Numident records (the Berkeley Unified Numident Mortality Database) with the 1940 Census. Sibships (sibling groups) in the CenSoc-Numident are located by finding individuals with shared parent names as recorded in Social Security Numident records. We use two methods to match siblings in Numident records:

1) The **exact** match method identifies siblings only with exactly identical parent names (after names have undergone cleaning and standardization). This is the most stringent match method.
2) The **flexible** match method permits parents' names to be slightly different, within a threshold defined by Jaro–Winkler string distance, in addition to exact matches. This allows siblings to be matched even in cases of minor misspellings, mistranscriptions, or spelling variations in parents' names among sibships (e.g., mother's maiden name recorded as "Brannum" and "Branum"). This increases the number of siblings found, but has higher potential to falsely match unrelated individuals. Most (but not all) individuals and sibling connections identified using the exact match method are also identified using this method.

Sibling groups are identified among individuals in the CenSoc-Numident who died at age 65+ in the years 1988-2005. They may be of any gender composition and contain from 2 to 8 siblings each. An overview of the resulting sibships created by each method is presented below:

| Match method | Number of individuals | Number of sibships | Mean sibship size |
|---|---|---|---|
| Exact match | 908,472 | 426,952 | 2.13 |
| Flexble match | 1,185,055 | 551,641 | 2.15 |

We note that sibships are established using only information from Social Security Numident data. As such, no information from the 1940 Census is used to match siblings, and false linkages between Numident data and the 1940 Census may create erroneous sibships. For detailed documentation of sibling identification methodologies and characteristics of sibships in Social Security Numident data, please see the paper: *Methods for Identifying Siblings in Administrative Mortality Data* available online at https://censoc.berkeley.edu/documentation/.

## Usage

Each sibling identifier dataset consist of two columns: a unique individual identifier `HISTID`, and an identifier for each sibling group. The HISTID of one sibling within each sibship is used as the group identifier.

```
# Show first 5 rows the exact match
head(numident_sibs_exact, 5)
```

```
##                                    HISTID                   sib_group_id_exact
##                                    <char>                               <char>
## 1: 005FECAD-F17C-472E-8216-43A12B6C5F4E 00827BE8-8C60-4197-AF3B-1F435B7B57D4
## 2: 00827BE8-8C60-4197-AF3B-1F435B7B57D4 00827BE8-8C60-4197-AF3B-1F435B7B57D4
## 3: 00BF0DC9-9C00-46C5-AC51-79E421DE88C0 00CD53F6-7B60-43AB-908F-BE9DEDF59A7C
## 4: 00CD53F6-7B60-43AB-908F-BE9DEDF59A7C 00CD53F6-7B60-43AB-908F-BE9DEDF59A7C
## 5: 008DD3B4-0BFE-424A-94DD-722A568963EB 00CE2BAB-A4EA-4FC0-BA9B-DCC4685F7A3B
```

*Last updated: 05 November, 2024

These files must be used in conjunction with the CenSoc-Numident dataset. Users will need to merge the datasets using the unique identifier `HISTID`, as in the example code below:

```
# Read CenSoc-Numident
numident <- fread("censoc_numident_v3.csv")
# Read CenSoc-Numident siblings IDs
numident_sib_id <- fread("censoc_numident_siblings_exact_match_v1.csv")
# Attach numident variables to siblings, keeping only people in sibships
numident_sibs <- inner_join(numident, numident_sib_id, by = "HISTID")
```