

SOCRATES: Guidelines for data formatting

1 Introduction and scope

The web repository socialcontactdata.org curates and disseminates data from social contact surveys conducted worldwide. By providing a common framework for data organisation and documentation, the repository aims to facilitate data sharing, comparability across studies, and reuse in modelling and public health research. Datasets hosted on the repository are intended for use with the [SOCRATES tools](#) and the [socialmixr](#) R package, both of which rely on standardised data structures and variable definitions to enable consistent data access, validation, and downstream analysis. To support these goals, contributed datasets are required to follow a set of general formatting rules.

2 Identifiers and naming conventions

2.1 Study identifier (NAMETAG)

Each social contact survey included in the repository is assigned a unique NAMETAG for identification purposes. The NAMETAG has the following structure:

$$\text{NAMETAG} = \text{NAME_YEAR_COUNTRY}$$

where **NAME** corresponds to the surname of the first author of the earliest publication presenting the dataset, **YEAR** to the year of that publication, and **COUNTRY** to the English name of the country in which the survey data were collected. If data were collected in multiple countries, the dataset must be split by country and assigned one NAMETAG per country.

Example: For a social contact survey conducted in Belgium and published by Doe et al. in 2012, the corresponding NAMETAG is:

$$\text{DOE_2012_BELGIUM.}$$

All data files must be provided as plain text comma-separated values (**.csv**) files, using a comma (,) as field delimiter and must not contain row names or implicit index columns.

2.2 Record and entity identifiers

The standardised social contact data data structure (as utilized by [socialmixr](#) and the [SOCRATES tools](#)) relies on a set of identifiers that operate at different levels of the data. Each identifier has a unique and well-defined role and must be used consistently across files. **In this standardised**

format, these identifiers are implemented as column names and are used to link records across data tables.

- **part_id**: identifies a single unique identifier for the respondent's entry. This ID must be unique across the entire submission.
- **sday_id**: identifies the calendar date associated with a given **part_id**, in YYYYMMDD format.
- **cont_id**: identifies a single contact event.
- **hh_id**: identifies a household.
- **panel_id**: identifies the same underlying individual across multiple survey rounds or waves (optional).

Note: all variable names defined in this guideline are case-sensitive and must match their definitions exactly. For example, **part_id** must not be renamed to **Part_ID**, **participant_id**, or **partID**. Deviations will prevent correct parsing by the [SOCRATES](#) and [socialmixr](#) compatible tools.

2.3 Missing values

All missing or not-collected values should be encoded as **NA**. Empty cells, whitespace, placeholder values (e.g. -99, 999), or empty strings are not recommended. The missing value encoding should be used consistently across all files and variables, regardless of data type.

3 General formatting rules

Each survey should be organised as a set of relational datasets, with one file per data entity. To ensure consistency across submissions, the following files are required:

File	Description	Primary key ¹	Foreign key(s) ²
Participant file	Participant-level information	part_id	hh_id
Contact file	Contact-level social interactions	contact_id	sday_id, part_id
Household file	Household-level information	hh_id	–
Survey day file	Survey day characteristics	sday_id	part_id
Time-use file (optional)	Time-use information per survey day	part_id	sday_id
Dictionary file	Variable definitions and coding	–	–

Participant-, contact-, and household-level information must each be provided in two separate files. For each data entity, contributors should distinguish between:

- a **common file**, containing variables that are typically available across most social contact surveys and are required for integration into the **socialmixr R package** and the **SOCRATES platform**;

¹A primary key uniquely identifies records within a dataset.

²A foreign key references a primary key in another dataset to enable linkage across files.

- an **extra file**, containing survey-specific variables that are not part of the common data structure.

The common files are used for harmonisation and standard analyses, whereas the extra files preserve additional information specific to the original survey. The **common** and **extra** files for a given data entity can be merged using the primary key. The extra file must contain only variables that are not included in the corresponding common file; variables defined in the common file must not be duplicated. The name of each file must start with the corresponding NAMETAG.

4 File specifications

The sections below specify the required variables for each **common** data file, grouped by data entity. All variables listed for a **common** file must be present as columns in the corresponding dataset. If a variable was not collected in a given survey, the column must still be included and filled with NA values. Only the **common** files are subject to this requirement (the structure of the common file must not be modified); extra files may contain an arbitrary set of additional variables.

1. Participant common file (file name: NAMETAG_participant_common.csv)

The participant common file contains core demographic information for each survey participant, with exactly one record provided per participant. Longitudinal linkage across survey rounds is handled via `panel_id` (see Section 5).

Column	Type	Coding	Notes
<code>part_id</code>	string / integer	unique identifier	Primary key; unique within the submission.
<code>hh_id</code>	string / integer	household identifier	Foreign key to the household common file; use NA if not collected.
<code>part_age</code>	integer	years	Exact age in years (numeric). Use NA if not collected.
<code>part_age_est_min</code>	integer	years	Lower bound of estimated age interval (numeric); otherwise NA.
<code>part_age_est_max</code>	integer	years	Upper bound of estimated age interval (numeric); otherwise NA.
<code>part_gender</code>	string	F, M, 0	0 denotes a non-binary or other gender category as defined by the original survey; use NA if not collected.

Age encoding rules:

Participant age must be provided using one of the following approaches:

- If exact age in completed years is known, it should be recorded in `part_age` and both `part_age_est_min` and `part_age_est_max` must be set to NA.
- If age is reported as a range or estimated interval, the lower and upper bounds must be recorded in `part_age_est_min` and `part_age_est_max`, respectively, and `part_age` must be set to NA.

Providing age information redundantly across these fields is not recommended.

2. **Contacts common file** (file name: `NAMETAG_contact_common.csv`)

The contacts common file contains contact-level information. Each record corresponds to one contact reported by one participant on one survey day. Multiple location indicators may be TRUE for a single contact if applicable. If studies collect additional or alternative location categories beyond those defined in the common file, these must be stored in the contacts extra file. The information to be included and the required order are the following:

Column	Type	Coding	Notes
<code>part_id</code>	string / integer	participant identifier	Foreign key to the participant file (<code>part_id</code>).
<code>cnt_id</code>	string / integer	unique identifier	Primary key. Must be unique within the survey submission.
<code>cnt_gender</code>	string	F, M, 0	Gender of the contacted person; use NA if not collected.
<code>cnt_age_exact</code>	integer	years	Exact age of the contacted person; otherwise NA.
<code>cnt_age_est_min</code>	integer	years	Lower bound of estimated age interval; otherwise NA.
<code>cnt_age_est_max</code>	integer	years	Upper bound of estimated age interval; otherwise NA.
<code>cnt_age_multi</code>	string / integer	category code	Age-class code if reported in categories*
<code>cnt_home</code>	boolean	TRUE/FALSE	Contact occurred at the participant's home; use NA if not collected.
<code>cnt_work</code>	boolean	TRUE/FALSE	Contact occurred at the participant's workplace; use NA if not collected.
<code>cnt_school</code>	boolean	TRUE/FALSE	Contact occurred at school; use NA if not collected.
<code>cnt_transport</code>	boolean	TRUE/FALSE	Contact occurred while using transport; use NA if not collected.
<code>cnt_leisure</code>	boolean	TRUE/FALSE	Contact occurred during a leisure activity; use NA if not collected.
<code>cnt_others</code>	boolean	TRUE/FALSE	Contact occurred at any other location; use NA if not collected.
<code>frequency_multi</code>	string / integer	category code	Frequency code if multiple frequency options are used*.
<code>duration_exact</code>	numeric	minutes (or specify)	Exact duration if reported numerically; otherwise NA.
<code>duration_multi</code>	string / integer	category code	Duration-class code if reported in categories*.
<code>phys_contact</code>	boolean	TRUE/FALSE	Indicator for physical contact; use NA if not collected.

(*) If categorical codes are used, the dictionary file must include the complete mapping from codes to labels.

3. Household common file (file name: `NAMETAG_hh_common.csv`)

The household common file contains core household-level information. Exactly one record

must be provided per household.

Column	Type	Coding	Notes
hh_id	string / integer	unique identifier	Primary key. Must be unique within the survey submission.
country	string	ISO alpha-2 code	Country in which the survey data were collected, encoded using ISO 3166-1 alpha-2 (e.g. BE for Belgium, and FR for France).
hh_size	integer	count	Household size; use NA if not collected.

4. Survey day file (file name: NAMETAG_sday.csv)

The survey day file contains metadata describing the calendar date associated with each `part_id`.

Column	Type	Coding	Notes
part_id	string / integer	participant identifier	Primary key to the participant file (<code>part_id</code>).
sday_id	string	YYYYMMDD	Foreign key. Survey date identifier.
day	integer	1–31	Calendar day corresponding to <code>sday_id</code> .
month	integer	1–12	Calendar month corresponding to <code>sday_id</code> .
year	integer	4-digit year	Calendar year corresponding to <code>sday_id</code> .
dayofweek	integer	0–6	Day of week (0=Sunday, 1=Monday, ..., 6=Saturday).

5. Dictionary file (file name: NAMETAG_dictionary.xlsx)

The dictionary file documents all variables included in the submitted datasets. It must include, for each variable:

- (a) the variable name
- (b) a short description
- (c) the data type (e.g. integer, floating point, string)
- (d) the allowed values and their meaning (for categorical variables).

The dictionary must be organised by file. If a spreadsheet format is used (e.g. `.xlsx`), it should contain one sheet per data file.

6. Extra files (file name: NAMETAG_xxx_extra.csv)

The **extra** files contain survey-specific variables that are not included in the corresponding **common** file. For each data entity, the common and extra files must share the same primary key, which ensures a one-to-one match between records across data tables.

Commonly missed step: mandatory empty columns

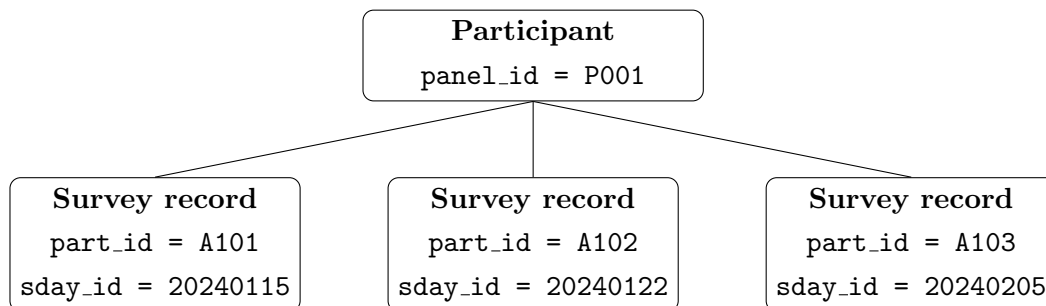
All variables listed in Section 4 must be present as columns, even if no data were collected. In such cases, columns must be included and filled with NA. Deleting unused columns will break compatibility with the socialmixr import standard.

5 Longitudinal and panel studies

In longitudinal or panel studies, participants may be observed on multiple survey occasions; in such cases, time-varying information must be represented using separate `part_id` values, linked via `panel_id` where applicable. Specifically,

- `part_id` must uniquely identify each participant–survey instance (i.e. each record) and must remain unique across the entire submission.
- `panel_id` should be used to identify the same underlying individual across multiple survey rounds or waves.

Records corresponding to the same individual across different survey rounds must share the same `panel_id` but have distinct `part_id` values. The `panel_id` is not included in the participant common file and, when available, must be provided in the participant extra file and linked using `part_id`.



*Note: **panel_id** identifies the participant, not the survey wave.
Survey occasions are identified by **part_id** and **sday_id**.*

Figure 1: Schematic representation of participant identifiers in longitudinal studies.

6 Zenodo submission and metadata requirements

SOCRATES-formatted datasets are disseminated via [Zenodo](#) to support long-term archiving, versioning, and reuse. To ensure datasets hosted on [socialcontactdata.org](#) are discoverable and interoperable, contributors must adhere to the following Zenodo metadata requirements.

Versioning

Each dataset release must specify a clear version number using a numeric scheme (e.g. v1.0,

v1.1). New versions must be created for any change affecting data content, structure, or documentation.

Community and keywords

All submissions must be associated with the *Social Contact Data* Zenodo community. Contributors are encouraged to include keywords such as `social contacts`, `contact data`, and `contact matrix` to facilitate discovery.

Note that each dataset corresponds to a single survey implementation. If multiple surveys are conducted within the same country and year, they must be submitted as separate datasets, each with its own Zenodo record and unique identifier. The information provided in the Zenodo (title, description, dates, and geographic scope) must be consistent with the contents of the submitted data files.

7 Submission workflow

To make a dataset available through socialcontactdata.org, contributors must complete the following steps:

1. Format the dataset according to the SOCRATES data formatting guidelines described in this document.
2. Upload the formatted dataset and accompanying documentation to [Zenodo](https://zenodo.org).
3. Contact the socialcontactdata.org team to request that the Zenodo record be linked and indexed on the platform.

8 Contact and support

Questions regarding data formatting or submission should be directed to:

[socialcontactdata@gmail.com]