# Depositing Data on the NSRR

## Purpose

To describe the process of taking a dataset, stripping it of missing and identifiable data, and publishing the data on [sleepdata.org](https://sleepdata.org).

## Scope

This SOP is intended for use by any clinical research staff who may be involved in the collection, cleaning, or curation of any dataset that is to be hosted by the NSRR.

## Responsibilities

#### Research Assistant

* To generate JSON data dictionary files for each variable in the dataset
* To create an automated process to deidentify and clean any PHI in the dataset
* To identify and verify or note any outliers or implausible data in the dataset

#### Data Manager

* To liaise with other data coordinating centers (DCC) for clarification of undefined variables
* To provide guidance to Research Assistant(s) working on a dataset
* To adjudicate any potential outliers or implausible data that are found in a dataset

#### System Administrator

* To assist with the publication of datasets on [sleepdata.org](https://sleepdata.org)
* To provide guidance, where necessary, in the utilization of tools developed by the NSRR

## Procedure

1. Formation of a GitHub repository

* Research Assistant will initialize a version-controlled GitHub repository specific to the dataset
* GitHub repository should contain folders to house variables, domains, forms, Ruby tests, and data cleaning scripts, along with any other files necessary to maintain version control
  + Examples can be found [here](https://github.com/sleepepi/shhs-data-dictionary)
* Repository version should be initialized to v0.1.0.pre and a CHANGELOG should be created

1. Receipt of the source data

* Data Manager or Research Assistant will receive the source data in an easily accessible format (CSV, SAS, etc.)
* Source data should be imported into a combined SAS dataset, separated by visit (where appropriate), without any formats applied
* Newly created source dataset(s) should be stored on a network location accessible to any relevant research staff

1. Cleaning and deidentification of the source data

* Research Assistant will work with the Data Manager to identify variables that contain PHI or might otherwise be unnecessary to include in the dataset
* Research Assistant will use the newly created source dataset(s) as the basis for the creation of a data cleaning script
  + Script should be written either in SAS or R and hosted publicly with the data dictionary
  + Script should facilitate the importation of source data, scrubbing of missing data, deidentification of PHI, and exportation of finalized, version-controlled datasets

1. Creation of a JSON data dictionary

* Research Assistant will use the “clean” dataset to create a JSON data dictionary in conjunction with the formatting guidelines available on [Spout](https://www.github.com/sleepepi/spout)
* Research Assistant will use Spout to test (spout t) the structure of the data dictionary and the degree to which it covers (spout c) the variables found in the dataset
  + Research Assistant or Data Manager will write additional Ruby tests for Spout, when necessary, to better check the integrity of the data dictionary

1. Publication of the dataset

* Research Assistant or Data Manager will use Spout to deploy (spout d) the completed dataset to a staging server [staging sleepdata.org](https://sleepepi.partners.org/edge/sleepdata)
  + System Administrator will facilitate access to research staff deploying the dataset
* The dataset will be verified by Research Assistant(s) to ensure proper formating of JSON files
* After the dataset has been fully reviewed, Research Assistant or Data Manager will deploy (spout d) the dataset to [sleepdata.org](https://sleepdata.org)
  + System Administrator will facilitate access to research staff deploying the dataset
* System Administrator or Data Manager will notify the NSRR community of the availability of the new dataset via a forum post on Sleepdata.org

## References

1. Spout Style Guide (https://www.github.com/sleepepi/spout)

=============== [Back to README](README.md)