# SOP4CWD Project: Data Sharing Instructions and Guidelines

## Cornell Wildlife Health Lab
## April 2020

This document provides state agencies with information about preparing and transferring data to Cornell for use in the **SOP4CWD** project.

### Handling of Sensitive Information by State Agencies

The entire **SOP4CWD** collaboration will work together to ensure sensitive human data is not shared with the public. Data such as names, telephone numbers, and physical addresses of humans that played a role in the chain of command of a sample should be removed by state agencies prior to data transmission to Cornell. Although Cornell will conduct a final check to cleave such information out of the datasets, state agencies should generally avoid including any such information in the files uploaded to the Open Science Framework (OSF).

Beyond sensitive information, all agency personnel should be aware of data security concerns when transferring, storing, and using data provided by state agencies. Accordingly, all agency personnel are expected to adhere to best practices described in the **SOP4CWD** Data Use Agreement (DUA).

### Data Transmission by State Agencies

Agencies should only share datasets by uploading them directly into the technological infrastructure of OSF. Specifically, Cornell has constructed a dedicated **SOP4CWD** Project folder in OSF that contains a network of state-specific subfolders ("Components"). Email and other methods of file sharing have not been approved for use in the DUA, because they cannot provide the level of security and sensitivity requested by collaborators. OSF was selected for data transmission, because it provides secure data transfer and storage and has flexible user access control to meet the needs of all collaborators in the **SOP4CWD** Project.

State agency personnel will be able to access only their state-specific OSF Component and should upload their data [and data descriptions; see below] to their state's OSF Component. For example, data provided by the New York Department of Conservation should be uploaded to the OSF under "*SOP4CWD/State Surveillance Programs & Data/New York CWD Surveillance Program Documentation and Data*". Similarly, data provided by the Tennessee Wildlife Resources Agency should be uploaded to the OSF under "*SOP4CWD/State Surveillance Programs & Data/Tennessee CWD Surveillance Program Documentation and Data*". State agency personnel will not be able to access data uploaded to the OSF Components for other states.

### Data Access for State Agencies

Access to data in each state's OSF Component will be limited to a small group of research personnel, even within the greater **SOP4CWD** project. Each agency must specify by name the individuals that should have access to their OSF Component. Please provide to Cornell (1) a primary data management

contact who will facilitate data transmission from the agency to Cornell, and (2) a list of agency contacts by whom the data may be autonomously accessed. Cornell will allocate permissions of each state's OSF Component only to individuals named in (1), (2), and Cornell Research Personnel responsible for standardizing the raw data (as agreed upon in the DUA).

## Generalized Dataset Process

Individuals that interact with the data at every step of the project play a critical role in helping to transition the data from its origin at several independent agencies into the single cohesive format necessary for region-scale modeling. As the state agency, you are the first link in the data chain. Your role in this chain is to document and organize your data prior to its transmission to your state's OSF Component. Once the data arrives in OSF, Cornell will process each state's data in such a way that completes the transformation into the single standard format necessary for analysis. Cornell's processing of your data will only include activities that are directly warranted for region-wide **SOP4CWD** analysis, such as broad-scale data proofing (e.g. searching for obviously incorrect data, such as a deer that is 100 years old, or a latitude/longitude that falls in the Atlantic Ocean), data cleaning (e.g. correcting obvious errors, such as simple typos), data standardizing (e.g. converting identical variables into the same units of measurements), categorical binning (e.g. grouping of data across a scale of the least common denominator), and data formatting (e.g. ensuring each variable has the correct technical/software characteristics).

## The Three Primary Elements of Each Contributed Dataset

The entire transformative chain of data from each independent source into the pooled region-wide dataset cannot be completed unless a great deal of descriptive information ("metadata") accompanies each dataset. State agencies that share data with **SOP4CWD** must therefore prepare the data itself, plus a host of specific documentation that will aid Cornell in the pooling of your data into the region-wide set. Such metadata must include a **data dictionary** and a **written summary of the dataset**.

The **data dictionary** must specify the convention that your agency uses/used for both categorical and numerical variables. For categorical variables, you must include their definitions, regardless if the terminology seems obvious. For example, if your dataset uses the term "juvenile", you must also specify the ages that constitute the juvenile class, plus the method that your biologists used to make that determination. Again, even if it seems obvious, the dictionary must also specify the units of measurement and the significant digits for all numerical variables. These definitions will help Cornell know which data can be directly pooled into the regional data, which data must be standardized prior to pooling, and which data must always be left unpooled. Examples of dictionaries are in the example data packets (described below).

The **written summary of the dataset** should generally summarize the who, what, when, where, and why of the dataset, and include any key caveats that may give rise to statistical complications down the road. This written summary can summarize data collection and curation processes or have appended to it the written agency protocols used up to the point when the data is transferred into the OSF. This summary will help Cornell ascertain which data can be directly pooled into the regional data, which data must be standardized prior to pooling, and which data must always be left unpooled. Examples of such summaries are in the example data packet (described below).

The **data dictionary** and the **written summary** will follow the same privacy and sharing conventions as the datasets they describe. Data handling protocols for **SOP4CWD** are described in the DUA.

## Primary Organization of Contributions

In addition to the data and its documentation, agency attention to the organization of their data prior to its transfer into OSF will aid the overall **SOP4CWD** project.

## Preferred Files and Content

Your primary data management contact should strive to upload into OSF:

| File | Contents | File Format |
|---|---|---|
| **Datasets** | The data file(s) | Preferred formats: TXT, CSV<br>Other formats: XLS, XLSX, ACCDB |
| **Each dataset should have an accompanying:** | | |
| **Data Dictionary** | Descriptions of the categories and measurements included in the dataset. For each field: a description, allowed categories or accepted values, measurement descriptions (e.g. units and significant digits) | Preferred formats: TXT, CSV<br>Other formats: XLS, XLSX, DOC, DOCX |
| **Written Summary** | A written description of the dataset that can be easily understood by someone who was not involved in the data generating process. The Summary should generally include (a) what the data describes, (b) the sampling scheme used to collect the data, (c) why the data was collected, (d) where the data was collected, (e) when the data was collected, (f) a *generic* who [volunteers/professionals] did the collecting, (g) any known data limitations or caveats, (h) whether the data has been manipulated by agency professionals prior to transfer, and (i) any other relevant information for using the dataset appropriately. The summary may also include the protocols used to collect the data, such as surveillance program documentation, staff data collection training protocols, etc. | Preferred formats: TXT, CSV<br>Other formats: DOC, DOCX, PDF |

## Preferred File Names

File names for datasets should be sufficiently unique and descriptive to identify the content. File names for datasets should include:

- State abbreviation
- Agency initials
- Dataset contents (avoid uncommon abbreviations used just to shorten file names)
- Date range for which the dataset is relevant
- Date that the dataset was transmitted to OSF

For example, a great file name is *NY_DEC_CWD_Sample_Test_Data_2019_exported_20200325_mock.*

Each dataset should be accompanied by its **data dictionary** and **written summary**, and all three file names should reflect the association. In the above example, the **data dictionary** would be named *NY_DEC_CWD_Sample_Test_Data_2019_exported_20200325_ data_dictionary_mock,* while the **written summary** would be named *NY_DEC_CWD_Sample_Test_Data_2019_exported_20200325_written_summary_mock.*

### Versioning

New data will be collected over the course of the **SOP4CWD** project. Accordingly, agencies will periodically transmit new data into their OSF Component, where the information will be pooled with previous data to update predictions from the **SOP4CWD** models.

Previous copies of the dataset [plus their **data dictionaries** and **written summaries]** should *not* be deleted from the OSF, as previous project analyses and deliverables relied on their use. Instead, your primary data management contact should upload the latest version of your data into the same OSF Component folder, differentiating among the previous and recent versions through the date:

Older version: *NY_DEC_CWD_Sample_Test_Data_2019_exported_**20200325****
Latest version: *NY_DEC_CWD_Sample_Test_Data_2019_exported_**20210217****
* The date is **bolded** to emphasize the difference between the two versions.

Corresponding **data dictionaries** and **written summaries** for these recent data files should also reflect the newest version. Although it may appear redundant, it is crucial to upload an updated **data dictionary** and an updated **written summary** with each new version. Such documentation tells the processors and modelers whether the information in the newest version should be handled differently than the previous data. If no changes occurred in data collection or handling at the agency level, then the previous **data dictionary** and **written summary** may be copied and renamed to reflect their association with the newest version of the data. The **written summary** for the new version of the data should explain what makes the new version different from the previous version. For example, the new version contains the same data as in the previous version with the addition of 129 new deer; or the new version contains data on 129 new deer that can be added to the previous data.

## Data Priorities

We have described in detail the three primary components of each shared dataset plus their preferred organization and naming conventions. It is understood, however, that agency-specific circumstances can

limit the preparation that an agency can do with their data prior to its transmission into OSF. In such circumstances, the data and its documentation are more important than its organization and naming convention. Cornell can aid in organizing and naming your data, but Cornell cannot produce nor describe your data. All efforts on the part of states to deliver well documented and organized data are appreciated. Rachel Abbott (rca74@cornell.edu) is Cornell's point of contact for this process and can address your concerns or answer questions regarding your efforts up to and including data transfer to Cornell.

## Desired Datasets

### I. CWD Testing Data

Description: A spreadsheet that contains CWD-related test results (positive *and* negative) that were collected any time and at any location within the state. This includes data from all sources of mortality, including hunting season sampling programs, meat processing submissions, road kills, clinical suspect cases, targeted removal efforts, or any other source of mortality.

Temporal and Spatial Scope: Statewide data for the 2018-2019 and 2019-2020 hunting seasons are priority for **SOP4CWD**. However, statewide data from other sources of submission, as well as all data from previous years is also highly desirable.

Data Structure: The spreadsheet rows should constitute individual deer (i.e. one and only one individual deer per row), while the spreadsheet columns should constitute the variables that describe that individual.

Important Column Variables: This will depend on the data available in your agency, but it should generally contain the highest resolution data that describes who, what, where, when, and why:

- Who: The unique identifier (ID). Any ID that differentiates individual deer across space and time.
- What: The highest resolution demographic information available at the time of death (such as age/sex).
- Where: The highest spatial resolution available at the time of death (such as latitude/longitude). If lat/long is unavailable, go up one level to the next scale (e.g. nearest town or city). If that is also unavailable, keep going up: township, county, range, or even state. If possible, also indicate the source of the locational data (GPS, hunter reported, etc.)
- When: The highest temporal resolution available at the time of death (such as DD/MM/YYYY). If day is unavailable, go up to the next scale: month, season, or even year.
- Why: A category of why the animal was tested (hunter harvest, roadkill, exhibited clinical signs, etc.)

Note: While a single spreadsheet for all deer is preferred, some states use separate spreadsheets to report the demographic data (e.g. age/sex) and the test results of the same individual. Separate spreadsheets may be uploaded into OSF provided the records are linkable through each individual's unique ID.

Excluded Variables: Omit any fields containing personal identifying information of humans, such as hunter names, phone numbers, hunter ID numbers, and hunter addresses, or any names, phone numbers, or addresses of individuals who participated in the chain of command of the samples. Any

other sensitive information about a member of the public, a caller, law enforcement officer, sharp-shooter, agency representative, agency contractor, or other agency affiliate, should also be scrubbed from the data prior to transmission into OSF.

Examples:
*NY_DEC_CWD_Sample_Test_Data_2019_exported_20200325_mock*
*NY_DEC_CWD_Sample_Test_Data_2019_exported_20200325_data_dictionary_mock*
*NY_DEC_CWD_Sample_Test_Data_2019_exported_20200325_written_summary_mock*

## II. Population Demographic Data

Description: A spreadsheet describing the white-tailed deer population in the highest spatial and demographic detail possible. Important demographic information includes population estimates, annual age/sex survival rates and fertility, and any additional demographic information that has been summarized by your biometricians, each reported at the county (or Game Management Unit (GMU)) scale. Data at multiple scales, if existing, should be provided, as post hoc binning can occur at Cornell as needed.

Temporal and Spatial Scope: Statewide data for 2018-present are crucial. Statewide data from previous years is also highly desirable.

Data Structure: The spreadsheet rows should constitute discrete population segments (i.e. age/sex segments), preferably at the county (or smaller area) scale, while the spreadsheet columns should constitute the variables that describe the abundances and shared vital rates of that population segment during that time.

Important Column Variables: The list of variables that you include will depend on the demographic data that is available in your state. Annual time series counts (estimations) of all age/sex categories by county (or smaller area) is most desirable, but annual time series counts of the reproductive deer segment by county (or smaller area) will do. Should county-wise estimates not exist, then go up one level and include annual state-wise estimates by age/sex. Annual age/sex vital rates (survival rates, fertility rates, as calculated by your biometricians) in each county (or smaller area) are also highly desirable.

Note: While a single spreadsheet is preferred, some states will have several data files that describe all aspects of the demographic properties of their age/sex segments at various scales. Separate spreadsheets are OK to upload into the OSF provided they are fully described through their own data dictionaries and written summaries.

Excluded fields: Omit any field containing identifying information, such as names, phone numbers, and addresses of individuals who participated anywhere in the chain of command of the data. Any other sensitive information about a member of the public, a caller, law enforcement officer, agency representative, agency consultant, or any other agency affiliate, should also be excluded.

Examples:
*NY_DEC_Demographic_Data_2019_exported_20200325_mock*
*NY_DEC_Demographic_Data_2019_exported_20200325_data_dictionary_mock*
*NY_DEC_Demographic_Data_2019_exported_20200325_written_summary_mock*

## III. Population Management and Hunting Data

Description: A spreadsheet that describes the activities used by the agency to manage the deer population across the state. Data should include information on direct control measures (e.g. agency sharpshooting) and indirect control measures (e.g. hunting). These data will function as the baseline framework from which the **SOP4CWD** models will calculate and optimize state-specific management alternatives.

Temporal and Spatial Scope: Statewide data for 2018-present are the most important. Statewide data from previous years is also highly desirable.

Data Structure: The spreadsheet rows should constitute all possible combinations of county and control measures, while the spreadsheet columns should describe activities within each of those management combinations.

Important Column Variables: The list of variables that you include will depend on the data that is available in your state, but they should generally include the what, when, where:

- Hunting season variables, including *hunting season name (ex. early bowhunting, crossbow, rifle), hunting GMUs, opening and closing dates of each season, total tags given (by age/sex segment), and total tags filled (by date, age/sex segment).* Please provide the highest resolution data available.
- Targeted agency population control variables, including *demographic, spatial, and temporal culling targets* and data describing the *successes of those efforts.* Please provide the highest resolution data available.

Note: While a single spreadsheet is preferred, some states will have several data files that describe all aspects of their management and at various scales. Separate spreadsheets may be uploaded into OSF provided they are fully described through their own data dictionaries and written summaries.

Excluded Variables: Omit any fields containing personal identifying information, such as hunter names, phone numbers, and hunter addresses, or sharp-shooter names, phone numbers, or addresses. Any other sensitive information about a member of the public, caller, law enforcement officer, agency representative, or agency affiliate should also be excluded.

Examples:
*NY_DEC_Harvest_Data_2019_exported_20200325_mock*
*NY_DEC_Harvest_Data_2019_exported_20200325_data_dictionary_mock*
*NY_DEC_Harvest_Data_2019_exported_20200325_written_summary_mock*

## IV. CWD Introduction Risk Hazard Data

Description: A spreadsheet that includes locations that, through human activity or business, constitute "riskier" locations for CWD introduction, infection, or spread, such as commercial meat processing or taxidermy facilities. This data augments disease predictions through the incorporation of the anthropogenic sources of disease amplification vectors.

Temporal and Spatial Scope: Statewide data for 2018-present are the most important. Statewide data from previous years would also be highly desirable.

Data Structure: The spreadsheet rows should constitute facilities (i.e. one facility per row), and the spreadsheet columns should constitute the variables that describe that facility.

Important Column Variables: The list of variables that you include will depend on the data that is available in your state, but they should generally include the what, how, when, and where:

- Taxidermist variables, including *location, years of operation, and hygienic practices*
- Wild Game Processor variables, including *location, years of operation, and hygienic practices*
- Captive Cervid Facility variables, including *location, years of operation, and hygienic practices*
- Carcass Disposal variables, including *location, years of operation, and hygienic practices*

Note: While a single spreadsheet is preferred, some states will have several data files that describe the facilities within their states. Other states may not have this information at all. Please provide as much information as available, and know that separate spreadsheets may be uploaded into OSF provided they are fully described through their own data dictionaries and written summaries.

Excluded Variables: Omit any information that conveys personal identification, such as hunter names, phone numbers, hunter permit numbers, or addresses, taxidermist names, permit numbers, EINs, phone numbers, or addresses, processor names, permit numbers, EINs, phone numbers of addresses, or names, phone numbers, permit numbers, EINs or addresses of any other individual or business that participated in the chain of command of the samples or in the disposal of carcasses. Any other sensitive information about a member of the public, a caller, law enforcement officer, agency representative, or agency affiliate, should also be excluded.

Examples:
*NY_DEC_Hazard_Data_2019_exported_20200325_mock*
*NY_DEC_Hazard_Data_2019_exported_20200325_data_dictionary_mock*
*NY_DEC_Hazard_Data_2019_exported_20200325_written_summary_mock*

## V. CWD Surveillance Activity Data

Description: A spreadsheet that describes current agency-wise surveillance activity. This data will provide the framework by which state-specific surveillance recommendation options will be optimized.

Temporal and Spatial Scope: Statewide data for 2018-present are the most important. Statewide data from previous years would also be highly desirable.

Data Structure: The spreadsheet rows should constitute combinations of counties and sampling types, while the spreadsheet columns should constitute the variables that describe the cost and benefits reported in each combination.

Important Column Variables: The list of variables that you include will depend on the data that is available in your state, but they should generally include the what, when, how, where, and why:

- Agency surveillance information that includes *agency priorities (e.g. knowledge of spread, minimizing cost, enhancing public outreach), informational goal (e.g. knowledge of prevalence, knowledge of presence or absence, knowledge of changes in prevalence), type of surveillance (e.g. passive, targeted), locations of previous emphasis, location-specific staffing limitations or*

*considerations, historical/current policies around disease management epidemic and endemic areas, etc.*

Note: While a single spreadsheet is preferred, some states will have several data files that describe the historical and current surveillance activities within their states. Separate spreadsheets may be uploaded into OSF provided they are fully described through their own data dictionaries and written summaries.

Excluded Variables: Any identifying information about a member of the public, caller, law enforcement officer, agency representative, or agency affiliate should be excluded.

Examples:
*NY_DEC_Surveillance_Data_2019_exported_20200325_mock*
*NY_DEC_Surveillance_Data_2019_exported_20200325_data_dictionary_mock*
*NY_DEC_Surveillance_Data_2019_exported_20200325_written_summary_mock*