De-identification Policy for Data Ingested by the Operating System

1. De-identification Purpose and Objective

Smart Columbus is a demonstration-based program to improve access and mobility in Columbus, Ohio. Each Smart Columbus project will collect and use data – sometimes of known participants – that can assist in project development, deployment, operation, or performance measurement. This personally identifiable data can help a project succeed; however, the release of such data can irreparably harm the person whose identity is revealed. To facilitate project success while protecting participants' privacy and identities, Smart Columbus requires data providers to de-identify personal and sensitive data before data is sent for Operating System ingestion.

This policy documents the process to de-identify such information so that it cannot be used to identify individuals or that reasonable basis does not exist to believe that the information could be used to identify individuals.

Please email questions about de-identification to Smart Columbus Operating System staff at smartcolumbusos@columbus.gov.

2. Who Should Read this Policy

This policy applies to all data that the Operating System will ingest; therefore, any individual, group or company sending data to the Operating System is responsible for reviewing this document and de-identifying data accordingly before submitting it.

3. Terminology and Definitions

- **Data Management Plan (DMP).** A Smart Columbus Program governing document for data management. The DMP is available online at smart.columbus.gov
- **Data Privacy Plan (DPP).** A Smart Columbus Program governing document for data privacy. The DPP is available online at smart.columbus.gov
- **Data Provider.** Any individual, department, or entity approved to provide information – after the Smart Columbus team vets and evaluates it via the data ingestion workflow process – to the Operating System
- **De-identification (also "Sanitization").** A process that seeks to minimize the possibility of associating personal information, to include identifying information, with a data subject NIST SP 800-188[1]
- **De-identified Information.** Information that is not individually identifiable and unprotected by the federal privacy and security regulations. Smart Columbus considers information to be "de-identified" when the program has no reasonable basis to believe that the information can be used to identify an individual

[1] *Garfinkle, Simson L. National Institute of Standards and Technology. U.S. Department of Commerce.* De-Identifying Government Datasets*. NIST Special Publication 800-188 (2nd Draft). December 2016. Available at: https://csrc.nist.gov/CSRC/media/Publications/sp/800-188/draft/documents/sp800_188_draft2.pdf.*

- **Direct Identifiers.** Information that can be used to identify an individual such as name, postal address, Social Security number (SSN), and date of birth
- **Filtering/Scrubbing.** A process of de-identification that filters and discards the anonymous ID assigned during de-identification, thus removing the means to reverse-engineer the original ID
- **Health Insurance Portability and Accountability Act of 1996 (HIPAA).** U.S. legislation that establishes data-privacy and security provisions for safeguarding medical information
- **Non-Personally Identifiable Information (Non-PII).** Any data that is not Personally Identifiable Information (PII). Non-PII is encrypted data and data reasonably de-identified of PII and Sensitive PII (SPII). Publicly-available PII is Non-PII for the purposes of this policy
- **Payment Card Information (PCI).** PII that includes information about credit cards, account numbers, or financial transactions associated with an individual, and that is subject to the Payment Card Industry Data Security Standard (PCI DSS)
- **Payment Card Industry Data Security Standard (PCI DSS).** An information security standard for organizations that manage cardholder data held by branded credit card companies
- **Personally Identifiable Information (PII).** Information that can be used to distinguish or trace an individual's identity, such as their name, Social Security number (SSN), biometric records, location data, etc., alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, and mother's maiden name. The definition of PII is not anchored to any single category of information or technology. Rather, it requires a case-by-case assessment of the specific risk that an individual can be identified by examining the context of use and combination of data elements. Non-PII may become PII when additional information is made publicly available. This applies to any medium and any source that, when combined with other available information, could be used to identify an individual
- **Protected Health Information (PHI).** Any PII related to health status, provision of health care, or payment for health care that a health-care provider (or its business associates) creates or collects, and that can be linked to a specific individual
- **Publicly Available PII.** Non-PII, for the purposes of this policy
- **Re-Identification.** The act of combining anonymous and/or de-identified data sources to identify individuals
- **Sensitive PII (SPII).** A subset of PII that if lost, compromised, or disclosed without authorization, could result in substantial harm, embarrassment, inconvenience, or unfairness to an individual. SPII requires stricter handling guidelines because compromised data has increased risk to an individual. The following PII is always (de facto) sensitive, with or without any associated personal information:
  - SSN
  - Passport number
  - Driver's license number
  - Vehicle Identification Number (VIN)
  - Biometrics, such as finger or iris print
  - Financial account number such as credit card, bank account number, or Common Payment System Identification
  - Health information, including medical history, mental or physical condition, or medical treatment or diagnosis
  - Medicare status

- Alien Registration Number.

In addition to de facto SPII, some PII may be deemed sensitive based on context. Some PII becomes SPII when paired with another of the following identifiers:

- Citizenship or immigration status
- Ethnic, religious, or sexual orientation or lifestyle information
- Last four digits of SSN
- Date of birth
- Criminal history
- Mother's birth name

- **Smart Columbus Participants.** People involved in one or more of the Smart Columbus portfolio projects

4. Principles and Standards for Handling PII

Adherence to the following principles and standards is expected by data providers and the Smart Columbus team when it comes to the de-identification of data:

- Smart Columbus and data providers will protect Smart Columbus participant's PII, SPII, PCI, and PHI through the de-identification methods discussed in this document
- Any information that does not contain direct identifiers or a means to re-associate identifiers may be used or disclosed without authorization
- Each data provider is expected to apply a filtering/scrubbing process to all data being shared with the Operating System
- Data shall not be filtered/scrubbed to the point where it no longer has value to the users. If a dataset or data point loses its value through de-identification, it shall be deleted
- Data provided to the Operating System that is de-identified without means of re-identification is exempt from the HIPAA privacy rule or the Smart Columbus DPP; however, de-identified data that becomes re-identified is again protected based on its status as PII, SPII, PHI, or PCI, and HIPAA, PCI DSS, or Smart Columbus DPP regulations may apply
- Each data provider is responsible for determining whether information the Operating System requests or data provided to the Operating System through a contractual relationship contains PII, PHI, SPII, or PCI. Information to which these terms apply is subject to all applicable regulations, and data providers must de-identify such data before providing it to the Operating System

5. Types of Data to be De-identified

PII has several subcategories (SPII, PHI, PCI and GIS) and this document has broken out data types into two categories to assist in the de-identification process.

1. Data containing PII, SPII, PHI, or PCI
2. Data containing GIS information

Each dataset should be reviewed in light of both categories of de-identification methods identified in Chapter **6**.

6. Approved Methods for De-identification

6.1. Process for De-identifying Data Containing PII, SPII, PHI, or PCI

1. Remove or de-identify the following 20 data elements for all individuals:
    a. Names
    b. All elements of dates (except years) for dates directly related to an individual including birth date, admission dates, discharge dates, date of death, encounter dates, surgery dates, all ages over 89 years, and all elements indicative of such ages
    c. Telephone numbers
    d. Fax numbers
    e. Email addresses
    f. SSNs
    g. Medical record numbers
    h. Health plan beneficiary numbers
    i. Account numbers
    j. Credit card numbers
    k. Checking account numbers
    l. Certificate/license numbers
    m. Trip identifiers including origin, destination, and critical intervals
    n. Vehicle identifiers and serial numbers including license plate numbers
    o. Device identifiers and serial numbers
    p. Universal Resource Locators (URLs)
    q. Internet Protocol (IP) address numbers
    r. Biometric identifiers including finger and voice prints
    s. Full-face photographic images and any comparable images
    t. Any other unique identifying number, characteristic, or code that may identify an individual.
2. Check free text fields for identifiable elements. Eliminate all unstructured free text fields containing PII that cannot be specifically filtered and removed

6.2. Process for De-identifying GIS Information

De-identifying GIS information is critical to minimizing PII recovery. GIS data can be used to identify an individual's PII such as travel behavior and SPII including but not limited to locations of residence, employment and visited medical facilities.

Two approved methods to de-identify GIS information are available for the Smart Columbus program: the SharedStreets method and the method developed by the USDOT ITS-JPO's Wyoming Connected Vehicle Project (see **Sections 6.2.1** and **6.2.2** below).

Smart Columbus may approve other GIS de-identification methods on a case-by-case basis. Please contact the Smart Columbus team via email at smartcolumbusos@columbus.gov to discuss an alternative method.

The Data Curator must approve de-identification methodology for all data containing PHI.

6.2.1. SHAREDSTREETS MOBILITY METRICS

This de-identification methodology is available at github.com/sharedstreets.

6.2.2. USDOT ITS-JPO'S WYOMING CONNECTED VEHICLE PROJECT

This de-identification methodology is available at github.com/usdot-its-jpo-data-portal/privacy-protection-application. See the CVDI-User-Manual.

7. Updates to this De-identification Policy

The Smart Columbus team periodically will review this policy. Smart Columbus reserves the right to update the policy at any time. The Smart Columbus team will post all modifications to this policy on the program website smart.columbus.gov.