| How To Guide – Lead in Drinking Water | |
| --- | --- |
| ` | |
| **February 2023**  **CDC Lead: T.J. Pierce (pwc2@cdc.gov)**  **State Lead: Peter DiPippo (Peter.DiPippo@health.ri.gov)** |  |

# Purpose and Use of this Document

This How-To Guide (HTG) provides a general outline of the steps required for processing state drinking water quality compliance datasets into NCDMs for the National Environmental Public Health Tracking Network (NEPHTN).   This revised guide is a draft for lead in DW only.

The guide is organized into the following sections (e.g., major steps):

1. Staging Table Development
2. XML Dataset Development
3. State Level NCDM Development

# Version History V1.0

|  |  |
| --- | --- |
| Date | Description |
| 12/14/2022 | First draft of a HTG for lead in DW |
| 1/27/2023 | Updated HTG draft for lead in DW |
| 2/1/23 | Third HTG draft for lead in DW |
| 2/3/23 | Final HTG draft for lead in DW |
|  |  |

# How To Guide

| Measures | 1. Annual 90th percentile lead in drinking water |
| --- | --- |
| Data Source(s) | This set of processing steps involves extracting the following types of records for the reporting period from SDWIS:   1. Water system descriptive and location information: the Inventory table; Please make sure that all CWS providing lead in DW data are in the inventory file you submit for the regular DW data submission and 2. Water quality sampling results information: the “Sampling Results” table.   The reporting period is 2018 to 2020. We are asking for 3 years of data to be submitted. |
| Definitions Relevant to Indicator | State water quality Safe Drinking Water Act compliance databases are referred to as “SDWIS” (Safe Drinking Water Information System) since most grantee states use the SDWIS-State database.  States that have a SDWIS-like database can assume that the term “SDWIS” refers to their water quality database(s) as well. |
| HTG Requirements and Cautions | This revised guide covers only lead in DW for a pilot data submission for the Spring 2023 Data Call. |
| Step #1 | ***I.  Staging Table Development***  *See Appendix A for sample Staging Tables.*  **Steps for Staging Table Development**   1. Assemble “Inventory” table – The Inventory staging table has one record for each Community Water System (CWS) for each year that it actively provided service to its retail population, whether systems were active for an entire year or a portion of that year.  Given limitations in SDWIS the hope is that NCDMs will be generated for water systems currently active as well as those that went inactive during the reporting period.  CWS that are currently inactive most often consolidate with other water systems to mitigate ongoing and costly water quality issues.   Recipients, or cooperating data stewards at the state primacy agency, will extract the Inventory table directly from SDWIS or via data flow tailored for Tracking Water NCDMs. The Inventory staging table includes all the elements from the data dictionary: PWS ID Number, Year Associated To, Year Pulled, PWS name, principal city/county served FIPS, number of connections, population served, primary type of source water, latitude, longitude, location derivation code.  Recipients only need to submit the inventory file for the regular drinking water submission. An additional file for lead is not needed. **States who have already submitted inventory data for past years that will be submitted for lead do not need to resubmit the inventory file for those years if the CWS that have been sampled for lead are included in that file. In either case, please make sure that CWS sampled for lead are accounted for and have been submitted in the past or in the current submission.**  *Listed below are processing steps and associated data elements that correspond to the current XML schema:*   1. For each inventory record, calculate geographic coordinates of approximate center of the retail service area.  These coordinates will assist users in identifying water systems; they are not intended for water quality linkage analysis.  Please see Appendix B - Guidance for Estimating Community Water System (CWS) Service Area Representative Point Locations for further reference. 2. For each inventory record in which a representative geographic coordinate has been found, provide a code that describes the derivation technique used.  These codes have been enumerated in the Data Dictionary and can be found in Appendix B. 3. Assemble water quality “Sampling Results” table – The Sampling Results staging table includes one record for each annual 90th percentile value for lead for each CWS reporting.   Water quality sampling data will be extracted directly from SDWIS or received directly from your data steward. The Sampling Results staging table includes the following elements: PWS ID number, year, analyte code, concentration units, concentration, date sampled, aggregation type, summary time period, number of samples.  *Listed below are some notable special cases and processing guidelines that should be adhered to when assembling the Sampling Results staging table:*   1. The analyte code for lead annual 90th percentile is PB90 (lead summary). If individual water sample data is submitted, the code 1030 should be used for that. |
| Step #2 | ***II. XML Dataset Development***  This set of processing steps summarizes and formats the Staging Tables into one XML dataset that conforms to the XML schema.  Please see Appendix E for example XML that have been validated for submission to CDC.   1. The Inventory XML dataset is an annual list of each CWS that was actively delivering water to customers for the years of reporting; it is a direct copy in XML format of the Inventory staging table. 2. Please refer to the draft Data Dictionary for lead in DW only. The current Data Dictionary can be found on SharePoint under Data Submission Documentation.   Listed below are processing steps:   1. Access lead in DW directly from SDWIS or receive annual 90th percentile lead in CWS from your data steward. For non-SDWIS states calculating their own 90th percentile for lead please follow the EPA guidelines ([Lead and Copper Rule: Monitoring and Reporting Guidance for Public Water Systems](https://nepis.epa.gov/Exe/ZyPDF.cgi?Dockey=P100DP2P.txt)) on page 34 and 35. Use the interpolation method to be consistent with SDWIS. |
| Step 3 | ***III. State-Level NCDM Development***  This set of processing steps further summarizes the CWS-level measures (calculated in Step II) into statewide frequencies of water systems and summed population-served by analyte-specific concentration categories.  The data produced in this step are not submitted to CDC.  Listed below are the steps for producing State-level annual NCDMs:   1. Joining the Inventory (Step II.a) and annual CWS-level records from the Water Quality dataset (see Step II.b.1) and grouping by: (1) analyte-specific concentration categories, (2) by lead, and (3) by all years, calculate State-level annual frequencies of CWS and summed population-served. 2. **These steps are for the display of the data. Data from this pilot will not be displayed.** |

# Appendix A – Sample staging Tables

**PWS Inventory**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PWSIDNumber | Year Associated to | Year Pulled | PWSName | PrincipalCountyServedFIPS | PrincipalCityFeatureID | TotalConnections | SystemPopulation | PrimarySourceCode | Latitude | Longitude | LocationDerivationCode |
| NH1234567 | 2010 | 2011 | wild acres development | 33003 | 873526 | 375 | 750 | SWP | 44.1467 | -72.5537 | SA |
| NH0023010 | 2010 | 2011 | green pine mobile home park | 33019 | 873525 | 100 | 200 | GW | 42.9725 | -71.4385 | MFL |

**Sampling Results**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PWSID  Number | Year | SummaryTimePeroid | Date  Sampled | Analyte Code | Concentration | ConcentrationUnits | Aggregation Type | NumSamples |
| NH1234567 | 2022 | 2022 | 2022-12-17 | PB90 | 0.05 | µg/L | 90X | 10 |
| NH1234567 | 2021 | 2021 | 2021-10-9 | PB90 | 5 | µg/L | 90X | 12 |
| NH0012345 | 2022 | 2022 | 2022-10-28 | PB90 | 15 | µg/L | 90X | 16 |
| NH0012345 | 2021 | 2021 | 2021-10-25 | PB90 | 25 | µg/L | 90X | 20 |
|  |  |  |  |  |  |  |  |  |

# Appendix B – guidance for estimating community water system (CWS) service area representative Point locations

This appendix provides a recommended methodology for EPHT State grantees to estimate a central and representative point location for each CWS that is reported to CDC by grantees in annual data calls.  The CDC public portal will display these point locations to orient users while navigating dynamic maps and to facilitate identification of a CWS relative to the approximate retail population in which it serves.  It is expected that these point locations will be displayed in relatively low resolution at scales small enough to depict regional differences, like county-, state-, or national-scale maps.  These locations are not expected to be precise enough or intended to be used in linking health outcome information to water quality measures; they are only intended for diagrammatic purposes.

Grantees are required to report to CDC representative service area locations for each CWS.  Grantees, however, are not required to follow this methodology or to use the recommended data sources.  This document is provided to enumerate the possible derivation methods, given grantees’ collective experience and expertise and when assuming that the described data sources exist at grantee sites.  It is not an exhaustive list of methods or data sources.  Depending on capacity and data sources available at each individual state, a grantee may choose a different methodology or data sources to accomplish the required objective.

As per the Drinking Water XML schema, all point locations shall be reported in North American Datum 1983 decimal degrees coordinates.

General Methodology:  Coordinates for all water system locations are compiled for each available data source.  Data sources are prioritized in order of increasing precision.  The Inventory dataset is updated according to this prioritization, using coordinates that are likely to be less precise, if coordinates of water systems from higher priority data sources are not available.

Complexities not addressed by this document:  The Inventory dataset includes an annual record for each CWS, if the CWS was active for at least a portion of the reported year.  This document, however, does not address annual changes in the water system location.  If grantee water quality and location-related data sources can track annual changes in water system service area locations, then grantees are expected and it is their responsibility to accurately depict these changes in the Inventory dataset.

Ordered below are source datasets containing geographic locations for individual water systems, prioritized from highest to lowest spatial precision.  Each source dataset has a corresponding LocationDerivationCode value, which should be used in the corresponding field of the Inventory dataset.

1. *Service area polygon centroids* (LocationDerivationCode=SA) – Increasingly, States are working to assemble polygonal boundaries that approximate the retail service area of public water systems.  As a GIS layer, the geometric center (or centroid) can be quickly calculated using common GIS tools.  For example, the following article from the ESRI Support website describes how to create and update two fields that describe the centroid of a polygonal layer:  <http://support.esri.com/en/knowledgebase/techarticles/detail/32482>.   For large water systems that have uneven population distribution, grantees may wish to use a population-weighted centroid, by intersecting water system polygons with Census Block centroids and proportionally weighting each centroid by its contribution to the total population within the service area polygon.  Further population-weighted refinements can be accomplished by assuming that people only live within an arbitrary distance (e.g., 500ft) of a street centerline network, and similarly intersecting the resulting buffered segments with service area polygons.
2. *Mean of water system facility locations* (LocationDerivationCode=MFL) – State drinking water primacy agencies typically track geographic locations of important facilities at public water systems.  Some of these facility types are often situated proximate to the retail population.  These include groundwater wellheads, treatment plants, and distribution system sampling stations.  If it is assumed that these types of facilities are close to the retail population, then we can use their mean center as a proxy for the representative system location.  However, because of the sensitivity of some of these facilities, a non-disclosure agreement may be required to release the facility locations to the grantee.  Grantee liaisons to the State primacy agency should communicate and reiterate to the data steward that the use of sensitive coordinates is undertaken in a secure domain.  Any derived locational information that is ultimately made public also completely masks the true geography of the original sensitive facility locations.

Per water system, a centroid or mean location can be calculated from a group of facility locations.  If only one facility location is available and data stewards have strict confidentiality requirements for this location, grantees can reduce precision on the points (e.g., to the nearest hundredth of a decimal degree) or use the GeoMasking tool (See Appendix C) provided by the Geospatial Workgroup and downloadable on the EPHTN SharePoint site to randomly skew the point within an arbitrary distance threshold (e.g., between 200 and 500 meters).  This tool requires the use of a polygonal layer within which the resulting point is constrained; county administrative boundaries trimmed of water features would serve this purpose satisfactorily.

1. *Principal city served* (LocationDerivationCode=PCS) – The Geographic Names Information System (GNIS - <http://geonames.usgs.gov>) place code for each water system’s principal city served is already an element in the Inventory dataset.  Each GNIS entity has a corresponding latitude/longitude in NAD83 decimal degrees that can be used to approximate a water system’s service area.  Alternatively, grantees can use their own State’s place name database or a geocoding service to derive coordinates from the principal city served
2. *Geocoded water system address* (LocationDerivationCode=GSH) – Water quality databases often include contact addresses for each CWS.  Grantees can use in-house geocoding expertise or an external service like BatchGeo.com or Google Fusion Tables (<http://earth.google.com/outreach/tutorial_fusion_yourowndata.html>) to infer coordinate locations from the water system address.  To filter contact addresses that are not proximate to the water system service area, compare the geocoded county of the contact address to the principal county served (which is typically a substring within the PWSID).  Contact addresses for CWS are usually proximate to their service area, but sometimes, and especially for private water systems that serve multiple jurisdictions, a water system contact address is the system’s billing address and can be situated far from the retail service area.
3. *Principal County served* (LocationDerivationCode=PNS) – Using GNIS or commercial data sources, grantees can make use of the centroid of county regions.  In the absence of locations from the previous 4 data sources, this location might be useful for some states that have smaller counties, in which very few water systems serve the population of a single county. In western states this data source will not likely be useful, since these regions can be very large and can be expected to capture too many water systems.
4. *Other* (LocationDerivationCode=O) – Location derived by some method not outlined above, e.g., zip code, etc. Please specify what O is in your metadata file.

# Appendix C – geomasking tool

A PowerPoint demonstration of New York State Department of Health’s GeoMasking Tool is on SharePoint at: <https://partner.cdc.gov/Sites/NCEH/EHHE/tracking/ArchiveWG/Geospatial/GeoMask_Tool_7_11.pdf>

The GeoMasking Tool described in the presentation above is available on SharePoint at:

# Appendix D – Determining appropriate values for non-detects when no detection limit is provided

This guidance is provided to assist grantees in substituting detection limit (DL) values in the Sampling Results Staging Table for observations in which DL values were not provided by their SDWIS data steward.

1. Using sampling observations that already have analyte-specific DLs specified, take the annual median DL for each unique lab and substitute this value for non-detect observations lacking a DL for the same year and same lab.
2. Using sampling observations that already have analyte-specific DLs specified, take the annual median DL and substitute this value for non-detect observations not updated in (1) and lacking a DL for the same year.
3. For the remaining non-detect observations lacking a DL and not updated in (1) or (2), please refer to the Excel file on Share Point, Analyte Detection Methods (<https://partner.cdc.gov/Sites/NCEH/EHHE/tracking/Resources/NCDM/Analyte_detection_methods_01062012.xlsx>) for an appropriate DL.
4. Otherwise, use the following table to determine the detection limit:

|  |  |  |  |
| --- | --- | --- | --- |
| Method ID | Method name | Detection Limit | Reference |
| D3559-03 D, 3559-08 D  Or D3559-15 D | Standard Test Methods for Lead in Water – Atomic Absorption, Graphite Furnace | 5 μg/L | [ASTM International - Standards Worldwide](https://www.astm.org/) |
| 200.5, Rev. 4.2 | Determination of Trace Elements in Drinking Water by Axially Viewed Inductively Coupled Plasma – Atomic Emission Spectrometry | 1.100 μg/L | [Corel Office Document (epa.gov)](https://nepis.epa.gov/Exe/ZyPDF.cgi/P10096US.PDF?Dockey=P10096US.PDF) and [NEMI Method Summary - 200.5](https://www.nemi.gov/methods/method_summary/9798/) |
| 200.8, Rev. 5.4 | Determination of Trace Elements in Waters by Inductively Coupled Plasma - Mass Spectrometry | 0.600 μg/L | [NEMI Method Summary - 200.8](https://www.nemi.gov/methods/method_summary/4665/) |
| 200.9, Rev. 2.2 | Determination of Trace Elements by Stabilized Temperature Graphite Furnace Atomic Absorption | 0.700 μg/L | [NEMI Method Summary - 200.9](https://www.nemi.gov/methods/method_summary/4797/) |
| 3113-B | Metals in Water by Electrothermal Atomic Absorption Spectrometry | 1.000 ug/L | [NEMI Method Summary - 3113 B](https://www.nemi.gov/methods/method_summary/4698/) |
| Palintest/Hach 1001 | Differential Pulse Anodic Stripping Voltammetry | 1 ug/L | [Kemio™ Heavy Metals - Palintest](https://www.palintest.com/products/kemio-heavy-metals/) |

# Appendix E – sample xml validated for submission to cdc

Pasted below are validated examples of the Inventory and Water Quality Levels datasets.  Only one observation is included in the Inventory dataset and the Water Quality Levels dataset.  Please see the Data Dictionary for additional description of allowable values and variable formats.  Please refer to the actual schema files provided on the EPHTN SharePoint site as an authoritative resource for XML syntax, including element naming and sequence.

I.  Inventory dataset:

<?xml version="1.0" encoding="windows-1252" ?>

<PWSInventory xmlns="http://www.ephtn.org/NCDM/ENV/PWSInventory">

  <Header>

    <MCN>4b956457-b578-41c4-b274-f067ba0430da</MCN>

    <JurisdictionCode>CA</JurisdictionCode>

    <ContentGroupIdentifier>PWSINVENTORY</ContentGroupIdentifier>

    <SubmitterInformation>

      <SubmitterEmailAddress>craig.wolff@cdph.ca.gov</SubmitterEmailAddress>

      <SubmitterName>Craig Wolff</SubmitterName>

      <SubmitterTitle>IT/GIS Director</SubmitterTitle>

    </SubmitterInformation>

    <StateFIPSCode>06</StateFIPSCode>

  </Header>

  <Dataset>

    <Row>

      <RowIdentifier>1</RowIdentifier>

      <PWSIDNumber>CA0103040</PWSIDNumber>

      <YearAssociatedTo>2011</YearAssociatedTo>

      <YearPulled>2011</YearPulled>

      <PWSName>NORRIS CANYON PROPERTY OWNERS ASSN.</PWSName>

      <PrincipalCountyServedFIPS>06001</PrincipalCountyServedFIPS>

      <PrincipalCityFeatureID>1658237</PrincipalCityFeatureID>

      <TotalConnections>19</TotalConnections>

      <SystemPopulation>100</SystemPopulation>

      <PrimarySourceCode>GW</PrimarySourceCode>

      <Latitude>

        <LatitudeRange>37.734364</LatitudeRange>

        <!-- comment about LatitudeNS element -->

      </Latitude>

      <Longitude>

        <LongitudeRange>-122.027303</LongitudeRange>

        <!-- comment about LongitudeNS element -->

      </Longitude>

      <LocationDerivationCode>SA</LocationDerivationCode>

    </Row>

  </Dataset>

</PWSInventory>

II. Water Quality Levels dataset (first row is an optional sample-level observation and second row is a summary-level observation)

<?xml version="1.0" encoding="windows-1252" ?>

<WaterQualityLevels xmlns="http://www.ephtn.org/NCDM/ENV/WaterQualityLevels">

  <Header>

    <MCN>f4ba15e2-3933-456f-a7d9-7edbac6fd19e</MCN>

    <JurisdictionCode>CA</JurisdictionCode>

    <ContentGroupIdentifier>WQLPB</ContentGroupIdentifier>

    <SubmitterInformation>

      <SubmitterEmailAddress>craig.wolff@cdph.ca.gov</SubmitterEmailAddress>

      <SubmitterName>Craig Wolff</SubmitterName>

      <SubmitterTitle>IT/GIS Director</SubmitterTitle>

    </SubmitterInformation>

    <StateFIPSCode>06</StateFIPSCode>

  </Header>

  <Dataset>

    <Row>

      <RowIdentifier>1</RowIdentifier>

      <PWSIDNumber>CA0103040</PWSIDNumber>

      <Year>2000</Year>

      <AnalyteCode>PB90 or 1030</AnalyteCode>

      <DateSampled>2000-08-08</DateSampled>

      <ConcentrationUnits>µg/L</ConcentrationUnits>

      <Concentration>4.52</Concentration>

    </Row>

    <Row>

      <RowIdentifier>43571</RowIdentifier>

      <PWSIDNumber>CA0103040</PWSIDNumber>

      <Year>2000</Year>

      <AnalyteCode>PB90</AnalyteCode>

      <DateSampled>2000-08-08</DateSampled>

      <AggregationType>90X</AggregationType>

      <SummaryTimePeriod>2000</SummaryTimePeriod>

      <NumSamples>1</NumSamples>

      <ConcentrationUnits>µg/L</ConcentrationUnits>

      <Concentration>4.52</Concentration>

    </Row>

  </Dataset>

</WaterQualityLevels>

# Appendix F – data quality checks at the gateway, 2022

Starting with the 2015 data call for drinking water data, the following validation checks will occur at the Gateway to shorten and facilitate data validation.

**1. Duplicate records.**  You can check your dataset for duplicate records by looking for more than one record with the following information:

**WQL files**: Year x Aggregation Code (90x) x Summary Time Period (A) x Analyte Code

**PWS Inventory**:  Year Associated To x PWS ID

1. **Concentration Units.** Lead is measured in ug/L. Units are specific to the analyte code; any mismatch will be rejected at the Gateway.

Icon

Description automatically generated with medium confidence CDC’s Environmental Public Health Tracking PRogram

Contact us: [trackingsupport@cdc.gov](mailto:trackingsupport@cdc.gov)

Visit the Tracking Network today: [www.cdc.gov/ephtracking](http://www.cdc.gov/ephtracking)

Follow us on social media:

* Twitter (@CDC\_EPHTracking)
* Facebook (facebook.com/CDCEPHTracking)