| How-To Guide – Drinking Water | |
| --- | --- |
| Provided by CDC’s Environmental Public Health Tracking Program | |
| 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).

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

# How-To Guide

| Measures | 1. Annual mean contaminant concentration values by CWS. 2. Percent of CWS by mean contaminant concentrations by year. 3. Percent of CWS by maximum contaminant concentrations by year. 4. Percent of people served by CWS by mean contaminant concentrations by year. 5. Percent of people served by CWS by maximum contaminant concentrations by year. 6. Percent of CWS by mean contaminant concentrations by quarter for atrazine, nitrates, HAA5 and TTHM only. 7. Percent of people served by CWS by mean contaminant concentrations by quarter for atrazine, nitrates, HAA5 and TTHM only. 8. Number of people receiving water from CWS. 9. Number of CWS. |
| --- | --- |
| 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; and 2. Water quality sampling results information: the “Sampling Results” table.   The reporting period is 1999 to the most current year having complete water quality information within SDWIS (e.g., 2021 for the 2022 data call). |
| 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 | The Guide covers ten analytes (Arsenic, Nitrates, Disinfection By-products: Haloacetic Acids and Total Trihalomethanes, Atrazine, DEHP, PCE, TCE, Radium, and Uranium) and locational information requirements for water systems. All ten analytes are treated generically in this HTG unless analyte-specific details are otherwise noted. |
| 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.  *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 compliance sample for each of the ten tracked analytes that are attributable to each CWS in the Inventory table.   Water quality sampling data will be extracted directly from SDWIS or via data flow tailored for Tracking Water NCDMs by the Exchange Network.  The Sampling Results staging table includes the following elements: PWS ID number, year, analyte code, concentration units, concentration, date sampled, aggregation type, number of sampling locations, summary time period, number of samples, number of non-detects.  *Listed below are some notable special cases and processing guidelines that should be adhered to when assembling the Sampling Results staging table:*   1. All samples with results below the detection limit must have a non-zero and positive detection limit value provided.  If a detection limit is not available from the source data, grantees are expected to estimate the detection limit from available data and/or provide a standard detection limit number.  Guidance for determining detection limits when one is not provided is available in Appendix D. 2. Compliance sampling done by drinking water wholesalers, that have interties with and sold to the CWS having a retail population, should be included in the Sampling Results staging table, if SDWIS captures this information accurately and completely.  Each importing CWS should be attributed with wholesalers’ applicable sampling results data. 3. For Uranium samples that are reported in pCi/L, convert pCi to µg using the following conversion: 0.67 pCi/µg  (or pCi/L x 1.49 = µg/L)   **Note**: The Drinking Water Team is tasked to develop SAS and SQL indicator calculation packages that take the two Staging Tables as input and accomplishes Steps II and III below.  These SAS and SQL programs will be developed by and shared among grantees (please see Share Point). |
| Step #2 | ***II. XML Dataset Development***  This set of processing steps summarizes and formats the Staging Tables into two XML datasets that conform 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. The Water Quality XML dataset includes summary-level measures of water contaminant concentrations.  The Water Quality dataset has a finite number of columns, and incorporates a hierarchical data structure, such that including additional analytes in future data calls will not require additional XML schema changes.  The reader is referred to the Data Dictionary as reference in further understanding this structure.  The current Data Dictionary can be found on SharePoint under Data Submission Documentation.   Listed below are processing steps that are affected by changes to the new XML schema:  1. Calculate annual mean and maximum concentration values, count of unique samples, count of unique sampling stations, count of samples that resulted in non-detect, and date of last sample.  Append these annual measures to dataset and code these records as “summary-level” data with an “annual” time period type.  For samples coded as non-detect a concentration of half the detection limit is used before summarizing.  2. In the same manner as Step I above, calculate CWS-level quarterly mean concentration values, count of unique samples, count of unique sampling stations, count of samples that resulted in non-detect, and date of last sample for **Nitrate, Atrazine, TTHM, and HAA5 only.**  **Special note for Steps II.b.1 & II.b.2 Above**:  For Nitrate, Arsenic, Atrazine, DEHP, TCE, PCE, Uranium and Radium, annual [and quarterly for Nitrate and Atrazine only] average concentration values are derived from first averaging by sampling station, then averaging by CWS.  For disinfection-by-products (TTHM and HAA5) annual and quarterly average concentration values are derived from first averaging by day, then by CWS.   Maximums for all 10 analytes are derived by taking the annual maximum for each CWS. |
| 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.  Standard concentration categories have not been specified by CDC for each analyte.  However, previous SAS indicator packages have suggested and used benchmarks relative to the regulatory Maximum Contaminant Level (MCL) for each analyte.  In practice, grantees and CDC have been afforded the option of choosing which analyte-specific concentration categories work best for them.  Listed below are the steps for producing State-level annual and quarterly 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 all analytes, and (3) by all years, calculate State-level annual frequencies of CWS and summed population-served.   In the same manner as Step III.a above, and using quarterly CWS-level records from the Water Quality dataset (see Step II.b.2), calculate state-level quarterly frequencies of CWS and summed population-served for Nitrate, Atrazine, TTHM, and HAA5 only. |

# 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**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| State FIPS code | PWSID  Number | Year | Date  Sampled | Analyte Code | Concentration | ConcentrationUnits | Aggregation Type | NumSampling Locations |
| 33 | NH1234567 | 2001 | 11/28/2001 | 1040 | 0 | mg/L | Mean |  |
|  | NH1234567 | 2002 | 10/9/2002 | 1005 | 5 | µg/L | Maximum |  |
|  | NH0012345 | 2003 | 10/28/2003 | 2950 | 15 | µg/L |  |  |
|  | NH0012345 | 2004 | 10/25/2004 | 2456 | 25 | µg/L |  |  |
|  |  |  |  |  |  |  |  |  |
| SummaryTime  Period | Num Samples | NumNon  Detects |  |  |  |  |  |  |

# 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](https://partner.cdc.gov/Sites/NCEH/EHHE/tracking/Resources/NCDM/Analyte_detection_methods_01062012.xlsx" \t "_blank)) for an appropriate DL.

# 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 only two observations are included in the Water Quality Levels dataset (one for sample- and one summary-level data).  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>WQL</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>4006</AnalyteCode>

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

      <ConcentrationUnits>ug/L</ConcentrationUnits>

      <Concentration>4.5075</Concentration>

    </Row>

    <Row>

      <RowIdentifier>43571</RowIdentifier>

      <PWSIDNumber>CA0103040</PWSIDNumber>

      <Year>2000</Year>

      <AnalyteCode>4006</AnalyteCode>

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

      <AggregationType>MX</AggregationType>

      <NumSamplingLocations>1</NumSamplingLocations>

      <SummaryTimePeriod>2000</SummaryTimePeriod>

      <NumSamples>1</NumSamples>

      <NumNonDetects>0</NumNonDetects>

      <ConcentrationUnits>ug/L</ConcentrationUnits>

      <Concentration>4.5075</Concentration>

    </Row>

  </Dataset>

</WaterQualityLevels>

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

Starting with the 2015 data call for community 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 (MX, M) x Summary Time Period (A, Q) x Analyte Code

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

**2. Concentration Units.** Eight of the ten analytes are measured in ug/L; nitrates use mg/L and radium uses pCi/L.  Units will be specific to the analyte code; any mismatch will be rejected at the Gateway.

**3. Quarterly means greater than annual maximums.** This error often occurs in submission of purchased water data. Please check your files to make sure the concentrations for purchased water are used in the calculation of the mean and maximum value for that CWS.

**4. Number of non-detects greater than the number of samples.** The number of non-detects can only be less than or equal to the number of samples.

**5.  We will only accept quarterly data for atrazine (2050), nitrate (1040), HAA5 (2456) and TTHM (2950).** Submission of quarterly data for analytes other than the 4 listed above will cause your data file to be rejected at the Gateway.

**6. We request that you only submit mean (“X”) quarterly data** for atrazine (2050), nitrate (1040), HAA5 (2456) and TTHM (2950); please do not submit quarterly maximum data.

**7. Too high to be a non-detection. *New in 2022***

**8. Annual mean and annual maximum data must be included in the WQL file(s).** Every analyte should have both annual mean and annual maximum data submitted in the WQL file.

***New in 2022.***

# Appendix G – lowest detection limits for analytes

|  |  |  |  |
| --- | --- | --- | --- |
| **Analyte Code** | **Analyte Description** | **Lowest Detection Level (LDL)** | **Half LDL** |
| 1005 | Arsenic | 0.02 | 0.01 |
| 2050 | Atrazine | 0.003 | 0.002 |
| 2456 | HAA5 | 0.5 | 0.25 |
| 2950 | TTHM | 0.01 | 0.005 |
| 2039 | DEHP | 0.46 | 0.23 |
| 1040 | Nitrate | 0.002 | 0.001 |
| 2987 | PCE | 0.01 | 0.005 |
| 2984 | TCE | 0.002 | 0.001 |
| 4010 | Combined Radium 226 & 228 | 0.03 | 0.02 |
| 4006 | Uranium | 0.001 | 0.0005 |

The lowest detection level and half the lowest detection level will be used during data validation to identify analyte concentrations that may be invalid. Please use these values to check your data submission for concentrations that may be artificially low due to a units conversion issue.

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)