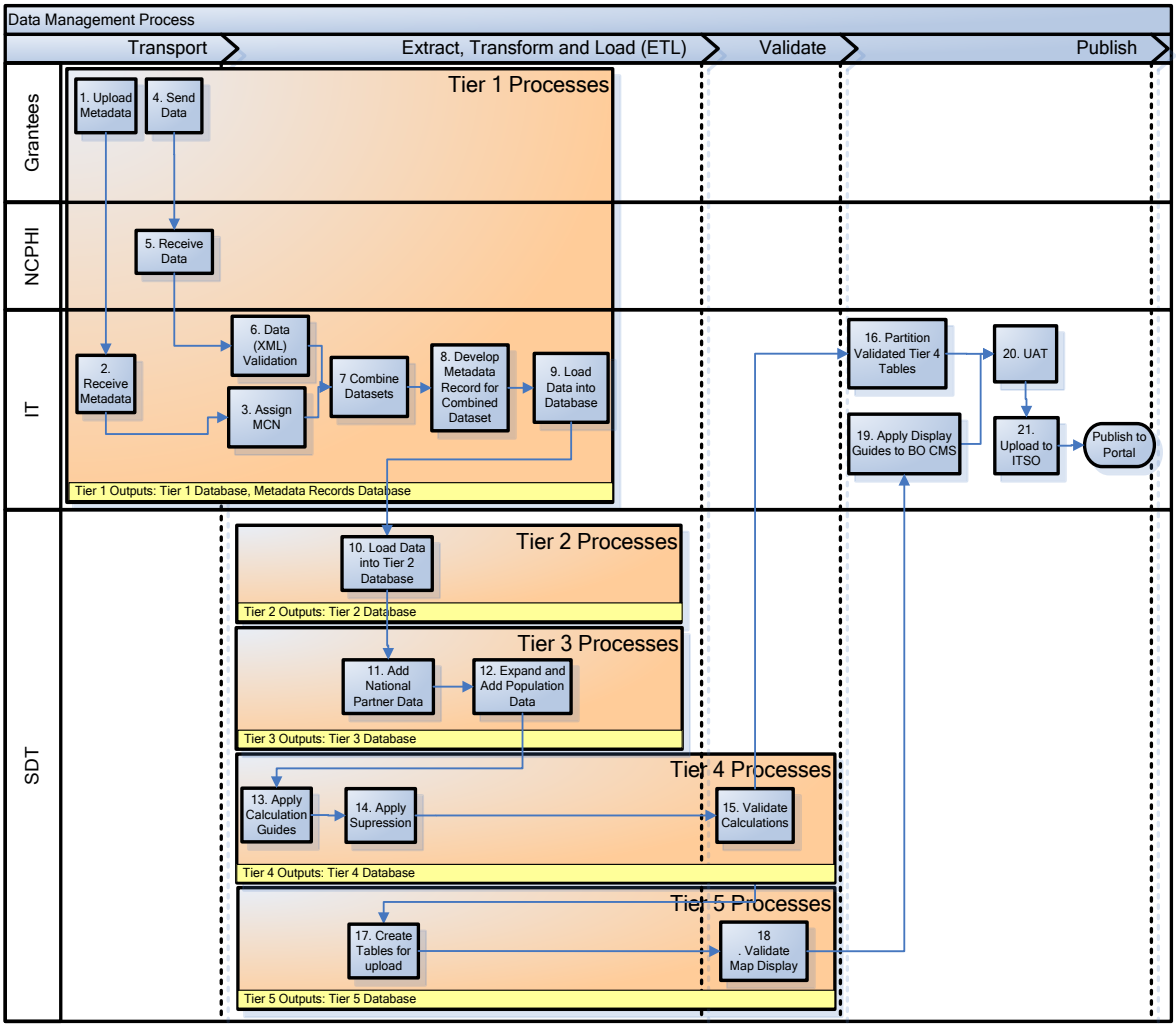
Tracking Branch Data validataion and Management Processes

HIGH LEVEL OVERVIEW

The Data Management processes are categorized by the entities responsible for doing the work (vertical axis) and the phases of work (horizontal axis). The diagram represents the processes at a high-level and does not describe all of the interactions among entities. The numbers in the boxes above are described in more detail in the text below.



DATA MANAGEMENT PROCESSES AND OUTPUTS

TIER 1 PROCESSES

1. UPLOAD METADATA

Grantees can use the Metadata Creation Tool (MCT) to create metadata that complies with the Metadata Standard template.

1.1. Login to MCT
1.2. Create Metadata Record
1.3. Save Metadata Record
1.4. Receive confirmation

EXISTING DOCUMENTATION

- Metadata QA/QC Administrator Checklist (Under Development by MDSG).
- Metadata Content Guidance
    - http://ephtn.sharepointsite.net/snd/MDSG1/Forms/AllItems.aspx?RootFolder=%2fsnd%2fMDSG1%2fMetadata%20Products&FolderCTID=&View=%7bB6798F40%2d4728%2d4D44%2d9B67%2d30F17BB4D2BF%7d
- MCT User Guide
    - Within the MCT - https://ephtsecure.cdc.gov
- List of proposed MCT improvements
    - Current list is within Eventum
- Metadata FAQ
    - Data submission page on share point: http://ephtn.sharepointsite.net/datasubmission/default.aspx
- URL to MD Documents List
    - http://ephtn.sharepointsite.net/snd/MDSG1/Forms/AllItems.aspx?RootFolder=%2fsnd%2fMDSG1%2fMetadata%20Products&FolderCTID=&View=%7bB6798F40%2d4728%2d4D44%2d9B67%2d30F17BB4D2BF%7d

2. RECEIVE METADATA

The Tracking Branch receives metadata that is created using the Metadata Creation Tool (MCT).

2.1. Check for creation of Metadata record
2.2. Review Metadata Record
2.3. Identify errors (if any)

EXISTING DOCUMENTATION

- None

3. ASSIGN METADATA CONTROL NUMBER

Metadata that is uploaded to the MCT is received by the Tracking Branch and assigned a Metadata Control Number (MCN). This MCN is then used to track both the metadata and the corresponding data submission.

3.1. Approve Metadata Record
3.2. Unique MCN assigned

EXISTING DOCUMENTATION

- Transport Guide: http://ephtn.sharepointsite.net/snd/Transport%20Guide/Sending%20NCDMs%20to%20CDC.aspx

- MD MCN FAQ: http://ephtn.sharepointsite.net/datasubmission/default.aspx

4. SEND DATA

Grantees can use either PHIN MS or SDN SFU to send data to CDC. (Note about SAMS). Not that Steps 4.1 - 4.6 may be different if the Data Integrity Validation Engine (DIVE) is used (Documentation TBD).

4.1. Retrieve latest XML schema for submission NCDM
4.2. Create XML file containing required data for submission
  4.2.1. Ensure the XML xsi:schemaLocation value is updated to reflect the location of your stored XML schema
  4.2.2. Assure Science contact email is contained in XML
4.3. Create metadata about submission (using MCT or other appropriate software)
4.4. Submit metadata
4.5. Receive MCN to be inserted into XML file
4.6. Validate XML file, using validation software and schema from step 1
4.7. ZIP the XML file
4.8. Submit XML file via PHIN MS or SDN SFU
  4.8.1. If SFU is used confirmation of receipt will be given via web interface
  4.8.2. If PHIN MS is used confirmation of receipt will be given via PHIN MS client
4.9. Confirm receipt of valid file email to Science contact identified in XML
4.10. If errors identified, correct and resubmit.

EXISTING DOCUMENTATION

- SharePoint Data Submission Site: http://ephtn.sharepointsite.net/datasubmission/default.aspx
- Transport Guide http://ephtn.sharepointsite.net/snd/Transport%20Guide/Sending%20NCDMs%20to%20CDC.aspx

- PHIN MS User Guide
- SDN SFU User Guide

5. RECEIVE DATA

Data are received either via PHIN MS or SDN SFU by the NCPHI PHIN MS Gateway.

5.1. Dataset receipt notice from Gateway

EXISTING DOCUMENTATION

- SharePoint Data Submission Site: http://ephtn.sharepointsite.net/datasubmission/default.aspx
- Transport Guide http://ephtn.sharepointsite.net/snd/Transport%20Guide/Sending%20NCDMs%20to%20CDC.aspx

- PHIN MS Administrator's Guide
- SDN SFU Administrator's Guide

6. DATA VALIDATION

Once the XML data are received by the Tracking Branch, it is run through a series of automated procedures and validated against the XML Schema files and the MCN is validated. Currently the XML Schema only checks the data structure of the submitted data and does not check all the business rules. The DIVE Tool checks for data integrity and Schematron may be used by CDC to further check the data (under development)
If an error is found, the Grantees are notified via email with specific error messages.
6.1. Received data validated against XML Schema
6.2. MCN checked
6.3. Error notice if any sent to Grantee via email

EXISTING DOCUMENTATION

- SharePoint Data Submission Site: http://ephtn.sharepointsite.net/datasubmission/default.aspx

7. COMBINE DATASETS

Datasets from multiple Grantees are combined into a table which contains all the data for a particular dataset or content area.

8. DEVELOP METADATA RECORD FOR COMBINED DATASET

Once datasets from multiple Grantees have been combined, a new metadata record for the combined dataset needs to be created.

EXISTING DOCUMENTATION

- Meta Data MCT Content
  http://ephtn.sharepointsite.net/datasubmission/Tech_Guidance/Metadata_MCTContent_Simple_Field_Guide

- Metadata Content Creation Guide
  http://ephtn.sharepointsite.net/snd/metadata/Reference_Documents/Forms/AllItems.aspx

9. LOAD DATA INTO DATABASE

After initial data validation, the XML data are loaded into a staging SQL Server database. Automated procedures are used to convert the data from XML format into SQL Server. At this point the data are called Tier 1 Data.

TIER 2 PROCESSES

10. LOAD DATA INTO TIER 2 DATABASE

Tier 1 data is copied to Tier 2 thru stored procedures. The Data Management Contractor (DM Team) and the SDT run preliminary checks on the data and look at the results to see if the years, counts etc. look reasonable. If an error is identified, the Grantees are notified by the SDT of the error and asked to clarify or resubmit the data if necessary. Tier 2 data is stored in the same database but in a different table which does not get modified as Grantees send new submissions in. The purpose of the Tier 1 to Tier 2 transformation is to provide the SDT and DM Team with a static copy of the data.

Data are reviewed in the order in which they were submitted. The DM Team uploads summary output to the data submission tracking system (currently residing in Eventum) and also includes any comments or points of data that should be more closely reviewed by the SDT. The SDT does a secondary review and provides a report to the grantees. Because of differences within content areas, the summary reports provided to the grantees differ by content area.

The SDT also updates the data submission tracking system to include and validation summary comments to the grantee and to indicate the current status of the reviewed data. Available statuses are:

1. Not Yet Submitted: File yet to be submitted by grantee.
2. Submitted: File successfully submitted by grantee
3. Submission Failed: File submitted but failed to load
4. In Progress: File in data validation by DM Team
5. Ready for Validation: Data awaiting approval from Science Team
6. Validated: Data passed validation and approved by Science Team
7. Postponed: Data submission is postponed to a date outside the current data call
8. Clarification Requested: Data awaiting additional information from grantee to complete validation process

TIER 3 PROCESSES

11. EXPAND AND ADD POPULATION DATA AS DEFINED BY CALCULATION AND DISPLAY GUIDE

Data expansion consists of ensuring that there is a record for every possible stratification combination for a dataset for each state that provided data for that content area. Expansion is performed by downloading population data from the Census website and summed to the lowest possible stratification level as well as pulling all unique values from the data and cross multiplying to get every possible stratification combination. This process is performed on each set of T3 data to ensure that data are properly displayed in the Public Portal.

Only State and County FIPS Codes will be used in the present iteration of the datasets. Any discrepancies with FIPS codes, such as Grantees/National Partners using obsolete County FIPS Codes, Merging/splitting of County FIPS codes, etc. are resolved prior to the expansion code being applied.

Population data is additionally added in some datasets to the Tier 3 data for use in the calculations of the Crude Rates, Age Adjusted Rates, and Variance and Relative Error calculations. The sum of the population amount by stratification level is included as part of the Tier 4 data.

When new population data becomes available via the Census website, the DM Team downloads it and makes it available to the SDT for their use in validating data.

Currently, all data gets expanded at the Tier 3 stage with the exception of Drinking Water data. Drinking Water data is expanded post-Tier 4 calculation.

The Tracking Branch receives data from multiple national partners. Currently the data are received in different ways depending on the preferences of the National Partner.

Content	Source	Transport Method
Vital Statistics	NCHS	Battelle File Exchange Shared Drive
Cancer	NPCR and SEER	*See Below
Air	EPA	FTP
Lead	CDC Lead Branch	Portable Media
Well Water	USGS	Web Download
Census	Census Bureau	Web Download

* SEER > CDC > DM Team Processes Tier 3 and Tier 4 > Sends Notification to the IT Team

TIER 4 PROCESSES

12. APPLY CALCULATION MEASURE GUIDES

Use the calculation-measure guides to determine which calculations to perform on the data. Stored procedures (which are an implementation of the guides) are then run on the datasets to do the calculations.

Where possible, calculations are done batch-wise to ensure that changes to similar calculations will be propagated across measures for a content area.

13. APPLY SUPRESSION RULES

Suppression rules are applied to the data to preserve confidentiality. Two kinds of suppression rules are applied:

a. Primary Suppression: to ensure that confidential information is not revealed directly by the data. Where necessary, geographic and temporal aggregation is performed to adhere to the data suppression rules identified by the data owners. Where this is not possible, data are removed to preserve confidentiality.
b. Complementary Suppression: to ensure that confidential information cannot be derived by running multiple queries on the data and calculating otherwise suppressed values. Complementary suppression is run once for a given stratification. Suppression determined by that run is applied for all data moving forward, including if the data was updated.

14. VALIDATE CALCULATIONS

The Tier 4 is validated by the SDT and DM Team. Currently data is validated by two independent groups calculating the measures and comparing the results. Measures are calculated independently between the SDT and the DM Team. The results of their validation are compared and discrepancies if any are identified and resolved. Tier 4 data calculations by the DM Team are performed directly in SQL Server. Calculated Tier 4 data are initially run on the DM Team servers. When the Tier 4 process completes its run, a copy of the newly-

created Tier 4 data is placed on the CDC Server for SDT comparison. Once a Tier 4 dataset is deemed valid by the SDT, the DM Team will correspond with the IT Team to coordinate inclusion in the DVP and Public Portal.

Additional stored procedures or views are run to further transform the data so it can be used to generate charts and tables for display on the National Portal. The stored procedures differ based on the datasets but in general they include running additional suppression rules, identifying missing variables and adding data from the National Partner datasets to create query able data. This process applies to some of the Tier 4 tables but not all of them and depending on the dataset, additional or different processes may be used.

- Continue to work on Standardize measure calculation logic such that logic can be validated providing a level of confidence that measures are calculated correctly if using logic that has been tested and is not ever changing. Standardization improves with each data call for the calculation logic
- Work in progress Develop better documentation for stored procedures of logic used.
- The Measure Management System allows the capture of standardized calculation types and what measures they are applied to. New measures can reuse existing calculation types or create new ones in the MMS as needed.
- While the process could be streamlined, the SDT would like to have final approval authority over the datasets.

EXISTING DOCUMENTATION

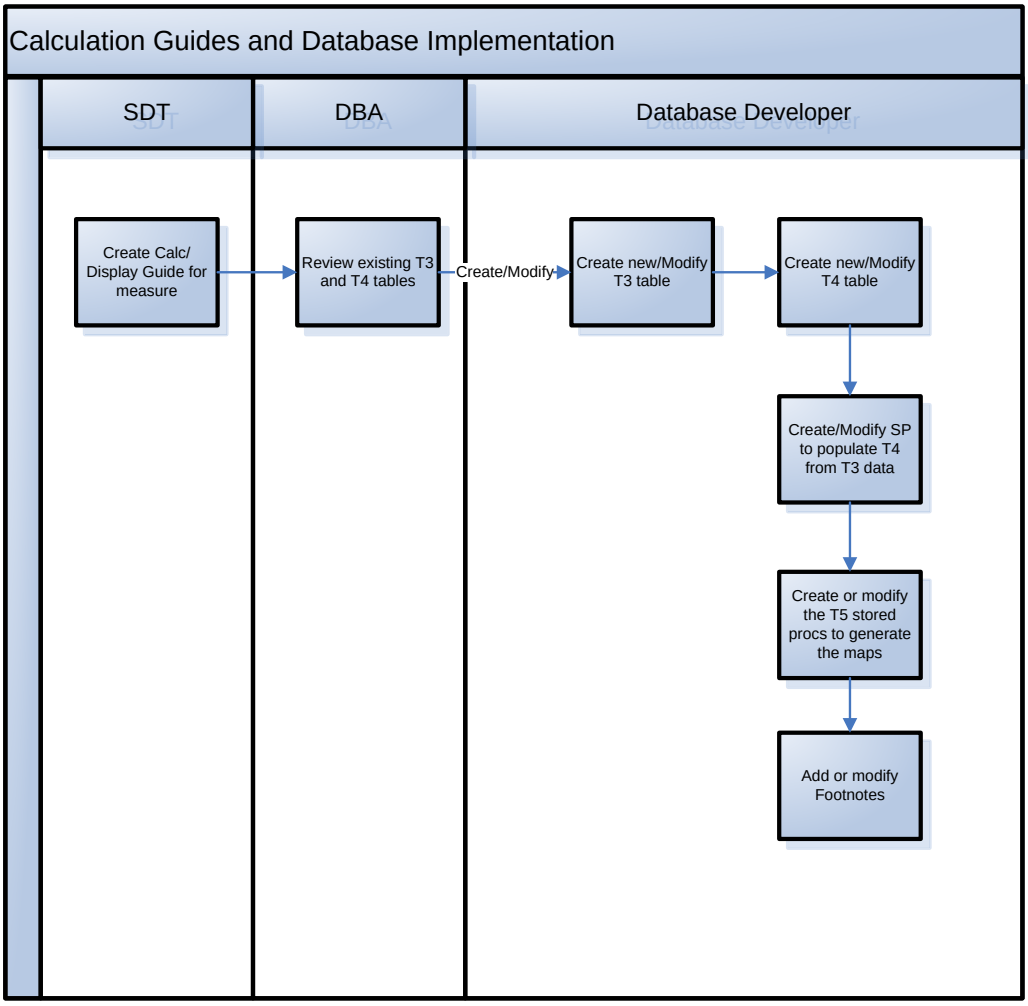
- Calculation and Display Guides
- DVP User Guide

15. PARTITION VALIDATED TIER 4 TABLES

The Tier 4 tables are partitioned by the developer based on the information provided in the Calculation and Display Guides by the SDT. The developer also creates or modifies footnotes based on the new tables that need to be created or modified.

The current Data Management effort is to create a content-area specific Tier 4 table rather than Measure Number-specific Tier 4 data. This minimizes the number of objects in the database and allows for more generic stored procedures for data processing. The calculation guides will be revisited with each data call to see how they relate back to the database tables to see if the portioning of the Tier 4 tables can be further optimized.

The following a generic data flow diagram and may vary by content area:



EXISTING DOCUMENTATION

- Calculation and Display Guides

TIER 5 PROCESSES

16. STORED PROCDECURE UPDATES

The stored procedures that generate TIER 5 are run as a query is made. They need to be updated to handle any new functionality or data depending on the data that is being processed. Otherwise this entire process is handled dynamically in code.

17. VALIDATE MAP DISPLAY

EXISTING DOCUMENTATION

- Calculation and Display Guides
- Many features of map display are controlled via MMS. The MMS can be validated to check certain aspects of each measures map display. Otherwise map displays should be validated on a public portal build.

DATA PUBLISHING PROCESSES

The data publishing processes are the steps involved to display the data on the Public Portal. All calculations are conducted in earlier stages and the following steps show how those calculations stored in the MMS and then displayed on the public portal.

Apply Display Guides to Measure Management System

Display Guides[1] are then applied to the Tier 4 Data. These guides are currently MS Excel files (this will hopefully be moved into the MMS soon) that contain information on how to display the measures and guidance on how to display the measure on tables, charts, and maps. These also inform the creation of the query panel on the National Portal.

Describe how data in the T4 and T5 tables are packaged and sent to the FLEX interface

- Stored Procedures again
- Coreholder (XML schema
    - General info (parts for maps, parts for tables and charts etc.

EXISTING DOCUMENTATION

- Calculation and Display Guides

18. USER ACCEPTANCE TESTING (UAT)

Tracking Branch staff and contractors review the data and display in a secure environment. Issues if any are identified and communicated back to the SDT or the contractors. DVP portal will allow Grantees to view their data and national data for approval before it is published to the National Portal – will improve the QA

- Currently this is a manual and time consuming process and it is very difficult to manually check the results of every combination of variables in the query panel.
- Also reviewing the presentation of the data on the portal – making sure that FLEX is presenting the data as intended in tables, charts, and maps, and that the numbers are correct

EXISTING DOCUMENTATION

- Error reports and support logs (currently in eventum)

19. UPLOAD TO ITSO DATABASES

Once the calculations and displays have been generated, the data tables are uploaded to the ITSO databases for access via a web server. The Tracking Branch does not have direct control over the servers or the database that run the Public Portal. As a result changes in content or features cannot be made very quickly.

- Currently we give ITSO DB database schema structure changes in SQL script
- Once the SQL scripts are run we implement the actual data changes ourselves

EXISTING DOCUMENTATION

- Mid Tier/ITSO Guide