Summary of the SDWIS Prime Data Quality Check Process

# Overview scope of the work

* A process is being developed for large data files from Primacy Agency data systems to be transferred to EPA and loaded onto an EPA system. This data is checked for any deviations from the expected data format or content. Once the file meets the minimal requirements it is run through numerous data quality checks. These checks include those on format, referential integrity, and data quality.

# What steps are completed

* A ‘Go Anywhere’ tool has been set up for transfers of large data files from Primacy Agencies.
* Servers have been set up for the Source, Common, and Prime formats.
* Mapping from Source to Common has been completed for standard SDWIS 3 formats.
* A process has been set up to identify any non-standard data from each Primacy Agency source file.
* A communication system with the Primacy Agencies has been developed to help the Primacy Agencies to rectify any issues with their files.
* Coding the numerous data quality checks has been initiated.
* A process is being developed to identify and prioritize the data quality checks to help Primacy Agency focus their cleanup efforts for the data.

# List of Data Quality Checks to be coded

* North Dakota provided over 300 checks they run on their system. Of those, over 88 were determined to be appropriate for the national data quality checks. All these checks have been coded.
* FedRep includes over 900 data quality checks. 800 have been found to still be relevant to the current data sets. 608 of the checks have been coded.
* New York provided over 100 checks that they run on their system. Of those, over 21 were appropriate for the national data quality checks. All these checks have been coded.
* Idaho provided over 200 checks they run on their system. Of those we have not determined which are appropriate for the national data quality checks yet. None of these checks have been coded.
* CMDP includes some checks run on sample data. None of these checks have been coded yet.

# Data Quality Workgroup

* A Data Quality Workgroup has been initiated to prioritize the data quality checks with Regional and Primacy Agency input.
* Each data quality check will have a priority that will range from ‘must be corrected’ to just ‘noted’ in the data as a potential data quality issue.

# Timing for DQ checks

* All initial data quality checks are expected to be coded by April.
* Preliminary priority for the data quality checks should be complete by end of mid-March.
* Final priority for the data quality checks should be complete by end of March.

# When can IN and Region 7 run data through DQ checks?

* It is expected that in April we can start running IN and Region 7 data through the full process of the data quality checks.
* Based on any issues found from running the data quality checks, corrections or revisions may be made to the checks, to the priorities assigned to the checks or to the process.

# When can the remaining Pilot agencies start having their data run through the DQ checks?

* Once the data quality checks have been initially run on the IN and Region 7 data files, other pilot states can begin submitting files of their data as well as EN Grant recipient states.
* It is expected that in May we can start running other pilot state data through the data quality checks.
* Based on any new issues found from running the other pilot states through the data quality checks, corrections or revisions may be made to the checks, to the priorities assigned to the checks or to the process.
* Any deviations from the standard format or special data fields for a specific Primacy Agency will require additional mapping or formatting of their data (this may require additional time to set up a specific process for this Primacy Agency).
* All non-SDWIS State primacy agencies will need to map their data to SDWIS State 3.3 version to be able to have their data processed through the data quality checks process.

# When can non-Pilot agencies start having their data run through the DQ checks?

* Once the data quality checks have been initially run on the pilot agencies data files, other non- pilot states can begin submitting files of their data.
* It is expected that end of 2019 we can start running other non-pilot state data through the data quality checks.
* Based on any new issues found from running the other non-pilot states corrections or revisions may be made to the checks or to the priorities assigned to the checks.
* Any deviations from the standard format or special data fields for a specific Primacy Agency will require additional mapping or formatting of their data (this may require additional time to set up a specific process for this Primacy Agency).
* All non-SDWIS State primacy agencies will need to map their data to SDWIS State 3.3 version to be able to have their data processed through the data quality checks process.