Appendix H: Data Management Process

H-GAC's Surface Water Quality Data Management Process

1. When the data manager receives field and laboratory data from individual local partners, all electronic files are saved in the partner's 'Raw Data' folder. Electronic files may include tabular data, data summary checklists, calibration records, and laboratory quality assurance documents that are referenced during the data review and validation process. The data files may be in the form of Excel spreadsheets, Access tables, scanned field data collection forms, or files downloaded directly from field instrumentation. Transfer of data is through e-mail or file sharing services, such as ShareFile and OneDrive. When a data set or individual files are received through any method, an e-mail confirmation is sent to the submitter. The e-mails are stored within an Outlook folder, which has a retention policy of 7 years.

   No modifications or corrections are made to files in the raw data folders.

2. Raw data files are then copied to the partner's "Working Data" folder. All modifications to the data prior to SAS processing are performed on the files in the "Working Data" folder. Compilation of the submitted data, where necessary, is performed by the H-GAC data manager. This typically involves combining and re-formatting spreadsheets or database tables, as well as other data management tasks. Field/variable names are changed to standardized formats, parameter names in the raw data files are replaced by TCEQ parameter codes, and data types are changed as required. Most of these tasks are performed after the data has been imported into the SAS environment for processing. In rare cases (e.g. to correct a data entry error or add data that was not entered prior to submission) H-GAC staff may enter data manually into the working file or add SAS code to make the change. Because the measurement performance specifications found in the A7.1 table may vary from one QAPP to another, the working data file does not include data collected under two different QAPPs. The file may, however, contain information from more than one month within the fiscal year covered by an individual QAPP.

3. Field and laboratory data for specific sample sites (monitoring stations) are combined during SAS processing.

4. During SAS processing, all fields (columns) in the compiled dataset are renamed and reformatted to comply with SWQM data management guidelines. Consult the most recent version of the "Data Management Reference Guide for Surface Water Quality Monitoring" for further information.

   a. The fields containing sample site, sample date, sample time, and sample depth are renamed STATION_ID, ENDDATE, ENDTIME, and ENDDEPTH respectively.

   b. The parameter names used by the partner are replaced by the TCEQ parameter code, preceded by an "S" to ensure that the data is read by SAS procedures as text data.

   c. Example: The field or column for dissolved oxygen is renamed "S00300".

5. The units of measurement as reported by the partner may not comply with SWQM guidelines. In most cases the SAS code will make the conversion to the correct units. If it is discovered that the code for conversion has not been written or is incorrect, or if the partner does not report the results consistently, manual conversion of the units may be necessary. In many cases, the SAS code will

flag any records reported in the wrong units for other reasons (below or above screening values, for example), and the correction can be made using SAS.

6. If the SAS code does not include an algorithm for reformatting dates and times, the data manager ensures that these data are formatted as mm/dd/yyyy and hh:mm respectively prior to import.

7. The partner may submit data for parameters that are not included in the A7.1. In most cases, the SAS code will simply omit the parameter from inclusion in the final datasets. It is better to modify the SAS code if unwanted parameters appear in the final dataset.

   Note: While references appear in this document to modification of the SAS code, these are for expository purposes only. The code should only be modified by a person who is very familiar with SAS programming in general, and the CRP processing code in particular.

8. When a database table(s) or Excel spreadsheets containing all field and laboratory data have been compiled and reformatted (if needed) as described above, they are saved to the SAS input folder within the "SAS Data Processing" folder (currently at "G:\CE\Databases\Clean_Rivers_Program\SAS_Data_Processing\Input") as an Access database or an Excel file. The input file should be renamed to include a code identifying the partner and the date range of the data.

9. As part of SAS processing, tables containing laboratory – specific quantitation limits, TCEQ minimum and maximum screening values, and site name / monitoring station ID correspondences are imported for comparison to the partner data. At the beginning of the period under which a specific QAPP is applicable, the data manager ensures that the tables containing this information correspond (where applicable) to the A7.1 tables. The data manager updates these tables at other times as needed.

10. The data manager modifies the SAS program used for the partner's most recent dataset for processing of the current data as follows.

    a. The most recent SAS program for the partner is saved with a name identifying the partner and date range of the data.

    b. All references to input and output files within the program are replaced with a name identifying the partner and date range of the data, and the program is saved.

11. The SAS program creates a new Access database in the "Access" folder within the "SAS Data Processing" folder. The database should have the same name as the input file.

    a. The database contains at least two tables: The "Input_Data_Matrix" that contains all data in the input file, and the "Flagged_Records_1" table.

12. The "Flagged_Records_1" table identifies questionable data that must be investigated by the data manager and quality assurance officer (QAO). The table is generated from comparisons against screening levels to identify outliers, quantitation limit tables to identify improperly reported data, and a variety of other comparisons. The program includes algorithms to identify the following:

a. Reported values beyond TCEQ screening limits (outliers);
b. Values reported as negative numbers;
c. Illegal values (e.g., results for qualitative parameters that are not in the range of allowed values);
d. Reported orthophosphate that exceeds the reported total phosphate;
e. Nitrate+nitrite concentration is less than nitrite concentration;
f. Inconsistent/irregular observed turbidity and water clarity results;
g. Inconsistent/irregular water surface and wind intensity results; and
h. Other algorithms are added to the QA protocol as needed.

13. The data manager is responsible for reviewing each flagged record against available raw data, data submittal checklists from the partner agency, instrument calibration records, and so forth, and where necessary obtaining additional information from the partner agency in order to determine the appropriate action to be taken. The flagged records table contains a variety of fields for documenting the disposition of the problem. In summary, a flagged record is accepted (on the basis of verification by the data manager), replaced with a corrected value, or deleted. Any corrections, deletions, or additions to the data set are noted so that they can be included in the Data Summary Report submitted to TCEQ.

14. At present, there is a subset of data quality problems that cannot be identified or corrected using the flagged records table. It may be necessary to make changes to the input file to correct some errors and inconsistencies identified during subsequent review by the data manager or QAO.

15. E-mail communications with the staff of partner agencies that are made during the data verification process are retained with the final data package that is stored within H-GAC's file storage system.

    a. An example of a singular data set's folder is "G:\CE\Databases\Clean_Rivers_Program\Current Text and Validator Files\2023 Datasets\Aug 2023 Deliverables\HG Jan - Feb 2023"

    b. Each data set submitted to TCEQ has a file folder where the final pipe-delimited text files, Data Summary Report, Validator Report, and the temporary SWMQIS Validator Report link are stored. These files are submitted to TCEQ. In addition to final data set files, this folder contains any communication or files where data has been reviewed or validated by H-GAC staff or staff of local partners.

16. Before changes are made to each data set, the data manager creates a "Data Summary Report" for that specific data set. The Data Summary Report is created from the most recent Data Summary Report for that partner agency and saved with the name of the current data set. All changes to the data and/or action taken on the data set are documented in this report. In addition, summary narratives discussing missing data, outliers that were verified and accepted, explanations of variations in reporting the data, failure to meet A7.1 LOQs, and so forth are also included. Pertinent information from the data submittal checklist submitted by the partner agency is also included in the final report. This report is submitted to TCEQ with each data set.

17. The data submittal checklist submitted by the partner agency is reviewed for the following, at minimum:

a. If the quality control information included in the report indicates that data has been reported that did not meet the measurement performance specifications of the A7.1 tables, it will be removed from the dataset. The removal will be noted on the "Data Summary Report".

b. If the quality control information included in the report indicates that data has been reported that did not meet method-specific quality control criteria, the impact on data usability will be evaluated. Data may be removed from the dataset if legal defensibility is questionable. The removal will be noted on the "Data Summary Report".

c. The post-calibration error limits in the partner agency's data submittal checklist shall be checked against requirements, as well as raw calibration records if available.

d. Reports of missing data, and the reasons that the data is missing (QC failure, spilled sample, could not sample site, etc.).

18. The SAS program may be re-run following action on flagged records where revision(s) to the input files were necessary. New tables and files are created and over-ride previously created SAS outputs. The most important of these outputs are the "Draft_Data_Matrix" and the pipe-delimited text files that are submitted directly to TCEQ.

    a. The portion of the SAS code that assigns TAG ID numbers is edited during the SAS program execution phase.

19. Once the SAS program is finalized for a data set, the data manager reviews the pipe-delimited text files. Each event and its relative results are reviewed for completeness, transcription errors, reasonableness, and conformity with the QAPP's A7.1 table. Thus, all data submitted to TCEQ has been reviewed by the data manager. The completed review document (Excel spreadsheet) is saved in the data set's file folder (See comment 15a).

20. The data manager's review file(s) and Data Summary Report is submitted to the QAO. The "Draft_Data_Matrix" and draft Data Summary Report are reviewed by the QAO , who identifies all values that, in the QAO's judgment, are unreasonable, are unverified outliers, or are otherwise questionable. Written comments and concerns are returned to the data manager for further investigation and correction of the dataset (where warranted). Newly identified discrepancies are investigated and documented on the Data Summary Report.

21. The data manager reviews the written comments, takes the appropriate action, and documents any additional actions on the Data Summary Report. If action is taken, the change is most commonly performed in the pipe-delimited text files and saved, over-riding the previously created text files. However, if the change(s) are significant, the SAS program may be re-run for that data set. Any changes to the text files or original input files for SAS program re-runs are documented in the Data Summary Report.

22. The written comments from the quality assurance officer, with annotations by the data manager, are retained with the data package as a record of data review and modification (where applicable).

23. The text files created by the SAS program and the final Data Summary Report are then submitted to TCEQ by the data manager. The data is first submitted to the SWQMIS (database) validation

algorithm to obtain a Validation Report; the files are then e-mailed to the CRP Project Manager at TCEQ. E-mails related to the submission of data are also stored in the data set's file folder.

a. The data manager stores the event and result files within the data set's folder.
b. Each file is edited to remove the header line (field names).
c. The data manager logs into the SWQMIS system, and submits the files and data summary report as described in the SWQM Data Management Reference Guide (https://www.tceq.texas.gov/waterquality/data-management/dmrg_index.html , published in June 2019) or the most current version of the same.
d. If the system identifies validation errors, upload is canceled, and the validation errors are investigated and corrected. In some cases, this may involve editing the text files only.
e. When no validation errors are found, the upload is completed, and a Validator Report is created and saved (with a unique file name) as a PDF file.
f. The data manager reviews the Validator Report to identify remaining discrepancies between the dataset, Data Summary Report, and A7.1 table requirements that may have been missed. The appropriate actions, to include resubmission of the data to obtain a revised Validator Report, are performed.
g. The text files, Data Summary Report, and the Validator Report are e-mailed to the CRP Project Manager.
h. The Validator Report is saved within the data set's folder. All files related to the dataset are saved in one folder. A subfolder contains the review process where any correspondence with the partner or QAO are documented as well (See comment 15a).

24. The data manager updates the "CRP Dataset Status" tracking document to include the date the files were sent to TCEQ.

    a. The "CRP Dataset Status" tracking document is currently located at the following pathway: "G:\CE\Databases\Clean_Rivers_Program\CRP Data Management\CRP Dataset Status_Updated_DDMONTHYEAR".

25. If the CRP Project Manager identifies further problems with the dataset, the appropriate action is taken and revised data sets or data correction requests (where appropriate) are submitted. E-mail communications with the CRP project manager are retained on file with the data package to serve as a record of validation and modification of the dataset.

26. When the data set is accepted by TCEQ and loaded into SWQMIS, the data manager updates the "CRP Dataset Status" tracking document to include the acceptance date.

27. Data management activities are documented in the Excel tracking document ("CRP Dataset Status") maintained by the data manager. The tracking document contains each data set submitted to TCEQ, its status, date of submission, and date of acceptance to SWQMIS. All data set files, and any correspondence related to the data set are saved within a single file folder for that data set. These folders are organized by fiscal year, data deliverable date, by local partner, and finally by the date period of the data set.