Appendix H: Data Management Process

H-GAC's Surface Water Quality Data Management Process & Flow Chart

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. The data may be in the form of Excel spreadsheets, Access tables, scanned field data collection forms, or files downloaded directly from field instrumentation. If data summary checklists have been submitted as electronic files, they are also stored in this folder. Hard copies of data, data summary checklists, calibration records, or other physical data are filed for subsequent data entry by H-GAC staff and for reference during the data review and validation process. In addition, receipt of the data is documented in the "CRP Data Tracking" database, currently found at Q:\CE\Clean Rivers\DATA\Data\CRP Data Tracking.accdb.

   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 may involve manual data entry into an Access data entry form, re-formatting of Excel or Access tables, and other data management tasks as needed. In addition, identifying information such parameter names in the raw data files replaced by TCEQ parameter codes (specific information is found below). Because the measurement performance specifications found in the A7.1 table may vary from one QAPP to another, the working data file must 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 where necessary to create one record containing all observations made at the sample site. Because combination of field and laboratory records is most efficiently performed by joining two Access tables on a common unique field, Excel tables may be imported into an Access 2003 database. In most cases, data are joined on equivalent monitoring station ID, sample date, and end depth values. If not already present in the datasets, the TCEQ monitoring station ID must be added, and the format of the date field must be consistent in the files to be combined.

   a. Note 1: this step is not necessary for City of Houston HHS and Harris County datasets.

   b. Note 2: The electronic data submitted by Eastex Laboratory must be transposed before combination with corresponding field data. This is most efficiently accomplished using SAS PROC TRANSPOSE.

4. The 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.

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

6. The parameter names used by the partner are replaced by the TCEQ parameter code. Precede the code number with an "S" to ensure that the data is read into SAS files as text data.

a. Example: The field or column for dissolved oxygen must be relabeled "S00300" prior to SAS processing.

7. 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.

8. 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.

9. Any parameters that are not included in the A7.1 table for the partner should be removed from the dataset. In most cases, the SAS code will simply omit the parameter from inclusion in the final datasets. It is preferable 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.

10. When a database table(s) or Excel spreadsheet containing all field and laboratory data has been compiled and reformatted as described above, it is saved to the SAS input folder within the "SAS Data Processing" folder (currently at Q:\CE\Clean Rivers\DATA\SAS_Data_Processing) as an Access 2002-2003 database or an Excel 97-2003 file. Note that the version of SAS (9.1.2) in use at H-GAC cannot import or export Office 2007 file types. The input file should be renamed to include a code identifying the partner and the date range of the data.

11. 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.

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

    a. Open and save the SAS program with the same name as the new input file.

    b. Find all references to input and output files within the program, and replace them with the name of the new input file.

    c. Save changes to the program.

    d. Run the program through the step where "Flagged_Records_1" is created.

13. 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.

14. The data manager updates the "CRP Data Tracking" database to include the date of initial SAS processing.

15. The "Flagged_Records_1" table identifies questionable data that must be investigated by the data manager. 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. Total constituents below dissolved constituent
    f. TDS/conductance ratio outside 0.55-0.70
    g. TDS less than total hardness
    h. Nitrate+nitrite concentration is less than nitrite concentration
    i. TDS less than chloride and sulfate;
    j. Inconsistent observed turbidity and water clarity results
    k. Inconsistent water surface and wind intensity results

16. 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. A code is entered into the "Action" column, the "Verification Method" code is entered, and the initials of the responsible party are entered in the "Verified By" column.

    a. "Verification Method" codes currently in use are DR (document review) and PJ (professional judgment).

17. 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 quality assurance officer.

18. All written communications with the staff of partner agencies that are made during the data verification process are printed and retained with the final data package that is retained by H-GAC. Records of telephone conversations are also retained.

19. Before changes are made to each data set, the data manager creates a "Data Summary Report/Sheet" 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.

20. 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/Sheet."

    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/Sheet."

    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.)

21. The SAS program is re-run following action on all flagged records. The flagged records table is read back into the process, and a variety of new tables and files are created. The most important of these 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 prior to generating the second group of tables and files.

22. The data manager queries a subset of data from the "Draft_Data_Matrix" table and reviews it against hard-copy raw data to check for random transcription errors. A sufficient number of records are selected so that when added to the flagged records previously evaluated, at least ten percent of submitted data has been verified against raw data. The query results are printed and retained with the data package as a record of data review.

23. The data manager creates and views a totals query of the "Draft_Data_Matrix" table to identify missing records that have not been addressed in the data summary report.

24. The data manager completes the draft data summary report, and updates the "CRP Data Tracking" database with the date the draft was completed.

25. The summary report is submitted to the quality assurance officer (QAO). The "Draft_Data_Matrix" and draft summary 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.

26. The data manager reviews the written comments, takes the appropriate action, and documents any additional actions on the data summary report. In most cases, the SAS program will be run at least one more time, although a new flagged records table is not routinely created. In the event there has been extensive modification of the input dataset, a new flagged records table may be created. 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). The date of data summary report approval is added to the "CRP Data Tracking" database.

27. 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 emailed to the CRP Project Manager at TCEQ.

    a. The data manager copies the event and result files to the desktop.
    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 SWQMIS user's guide (http://www.tceq.state.tx.us/assets/public/compliance/monops/water/wqm/swqmis_users_guide.pdf , retrieved 8/10/2010).
    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. If this option is selected, document changes to text files appropriately. It may be most convenient to document minor changes to the text files in the "Comments" section of the appropriate record in the "CRP Data Tracking" database.
    e. When no validation errors are found, the upload is completed, and a validator report is created and saved report (with a unique file name) as an html 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 validator report are e-mailed to the CRP Project Manager.
    h. The validator report is saved in the "Data Review and Submission Docs" folder at Q:\CE\Clean Rivers\DATA\Data\Data Review and Submission Docs."

28. The data manager updates the "CRP Data Tracking" database to include the date the files were sent to TCEQ, and add hyperlinks to the data summary and validator reports.

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

30. When the dataset is accepted by TCEQ and loaded into SWQMIS, the data manager updates the "CRP Data Tracking" database to include the acceptance date.

31. All data management activities are documented in an Access database maintained by the Data Manager. The database contains details of receipt, processing, submission, and acceptance by TCEQ, and includes hyperlinks to raw and final datasets, data summary reports, and data validation reports.