Table of Contents

List of Figures

List of Tables

Appendices

Appendix A: *Statistical Methods for Analyzing Censored Water Quality Data Sets*

Appendix B: STORET Project Establishment Form

Appendix C: RLWD Laboratory Information for STORET Data Entry

Appendix D: STORET Monitoring Station Establishment Form

Appendix E. Example Spreadsheet for Submitting Data for Entry into STORET.

1.0 Introduction

The original objective of this handbook is to report/explain the procedures used and products developed from the Red River Watershed Assessment Protocol Project. This document was also created to provide guidance to water quality staff from the Red Lake Watershed District and other groups and agencies conducting water quality monitoring programs. The information and methods contained in this document were pulled together from a large number of sources in order to provide a very robust methods document. It serves as a methods handbook for water monitoring project development, water quality data collection, and data management. This document helps ensure continuity in data analysis, even throughout changes in personnel. Although it is, at times, focused on the Red Lake Watershed District and the Minnesota side of the Red River Basin, it is intended to also be useful to other agencies collecting water quality data. This will be a living document. Changes in methodology, newly developed data analysis methods, or any methods overlooked by this document will be included in future editions. Hopefully, the time spent creating this handbook will help save time in the future and prove to be an efficient resource for its users.
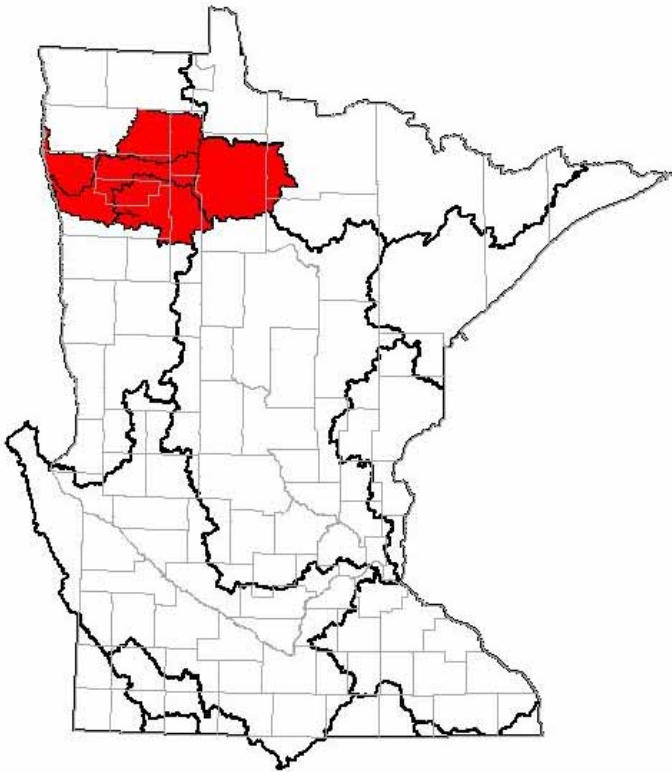
Figure 1. Location of the Red Lake Watershed District
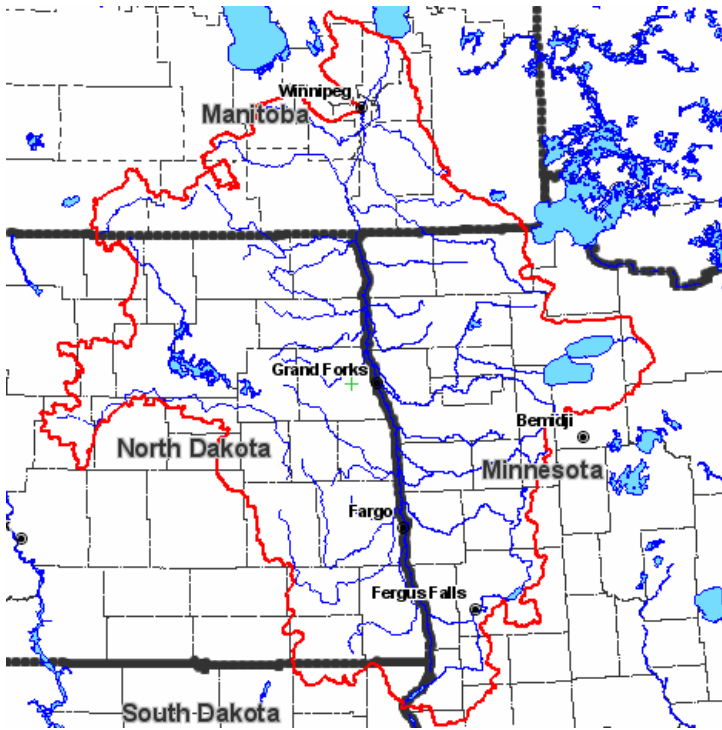
Figure 2. Red River Basin.

The purpose of the Red River Watershed Assessment Protocol Project was to establish procedures for developing water quality reports, field and lab standard operating procedures, quality assurance project plans, and statistical analysis techniques for the Red River Basin, providing needed coordination as identified in county water plans. The project was funded by a Minnesota Board of Water and Soil Resources Challenge Grant. There are many organizations that are monitoring water quality within the Red River Basin. However, until recently, the sharing of data among agencies was limited. The Red River Watershed Assessment Protocol project is meant to help agencies take a step in the right direction towards better coordination of monitoring efforts and comparability of data. This project recommends the use of standard methods by all these agencies so that data is comparable due to similar collection and analysis methods. The coordination of data collection efforts among agencies will lead to less duplication of sampling efforts, and greater number of sites that will be monitored across the RLWD by one agency or another. Other products of the Red River Watershed Assessment Protocol Project include the RLWD website and water quality database, *Standard Operating Procedures for Water Quality Monitoring in the Red River Valley*, *Statistical Methods for Analyzing Censored Water Quality Data Sets, 2004 Red Lake Watershed District Water Quality Report*, *River Watch Quality Assurance Project Plan* (QAPP), and the RLWD QAPP.

2.0 Data Storage

2.1 Database Design and Agency Coordination

Some of the RLWD's needs that were fulfilled by the Red River Watershed Assessment Protocol Project were the needs for a website for public outreach, a central database for the storage of water quality data, a tool for viewing GIS data and creating maps, and data analysis tools. Houston Engineering was contracted to create the RLWD website, which meets all of the aforementioned needs. Along with the other features of the website that were created (see Section 6.3), a central Microsoft Access database was created. It is stored, along with all other files related to the website, on a Houston Engineering-owned server. Data is stored in a set of interrelated tables. There are tables within this database for water quality data, site information, organization information, and site pictures. The tables are linked by site ID number and organization name. A set of web pages are used to display the data within these tables.
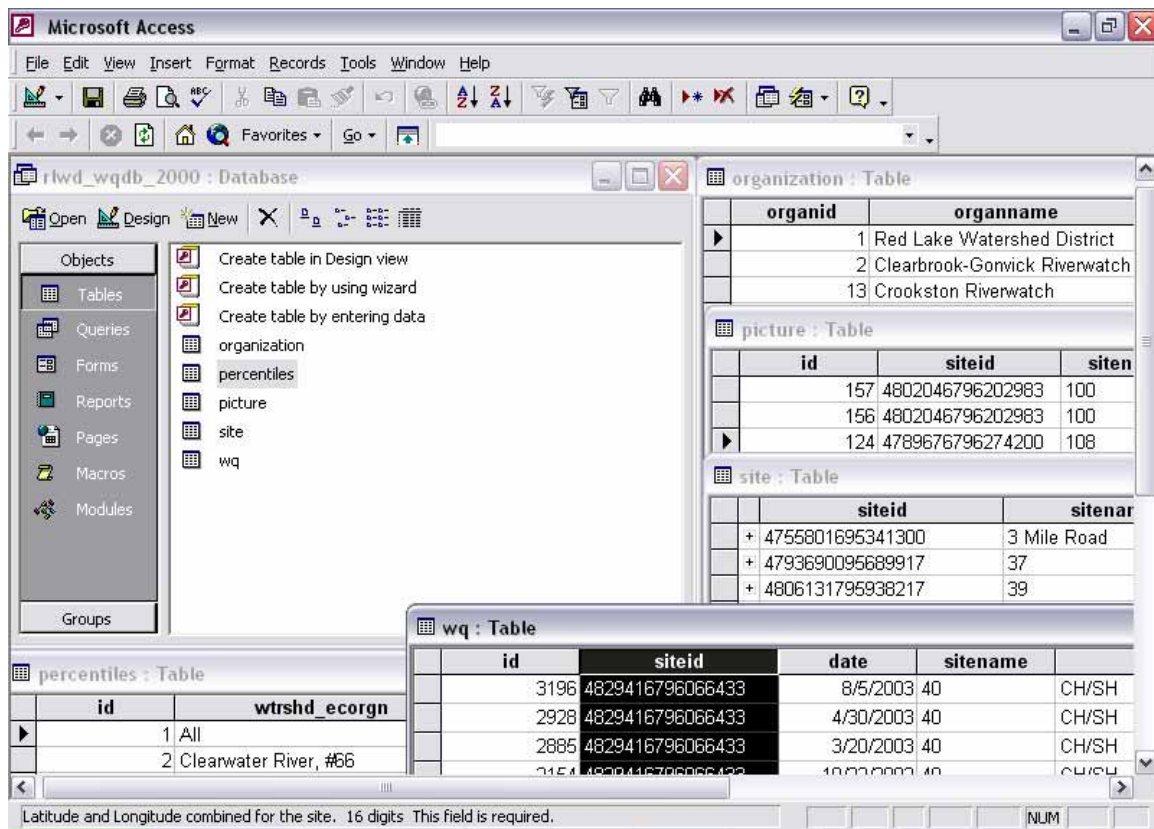
Figure 3. RLWD Water Quality Database.

The database was originally updated RLWD staff using an online data entry form. The data entry pages were password protected so only RLWD staff can enter data. Data is entered into the database by the RLWD. This page features a blank cell and a flag cell for each water quality parameter that may be entered into the database. For the online data entry form to work properly, a numerical value must be entered in every cell. There are cases, however, when data results are not represented by a number. A method was needed for distinguishing among results below the reporting limit, zero values, and missing values. For results that fall into one of these categories, a zero is entered into the cell and a value is selected from a flag field that specifies whether the value is below the detection limit, equal to zero, or if there is no value for the field. The online data entry format was not as convenient as it was intended to be. The RLWD has gotten rid of the online data entry form and has switched to a more direct and simpler method of data entry.

Figure 4. Online data entry form that was used by the RLWD.

Data entry forms have recently been added to the Access database itself that make data entry even easier than online entry. After data has been added to the Microsoft Access database, the database file located on the ftp server is simply replaced with the new, updated version.

Figure 5. Microsoft Access Database Entry Form.

Data in the database can be downloaded by anyone visiting the RLWD website (www.redlakewatershed.org). After a successful search for water quality data from a particular monitoring site on the RLWD website, this set of web pages will appear for the site. These pages include a report card page, site information page, data viewing page, analyze and download data page, and a site location map. Information displayed on the site information page is stored in the site and picture tables in the database. The report card page performs calculations using data in the water quality data table (entitled "wq") and compares the results to the standards in the percentiles table in order to produce a letter grade for each monitoring site. The View Data and Analyze and Download Data pages link to the water quality data table to display the data, calculate summary statistics, create time series plots of the data, and load data into the StatCrunch program for additional statistical analysis options.
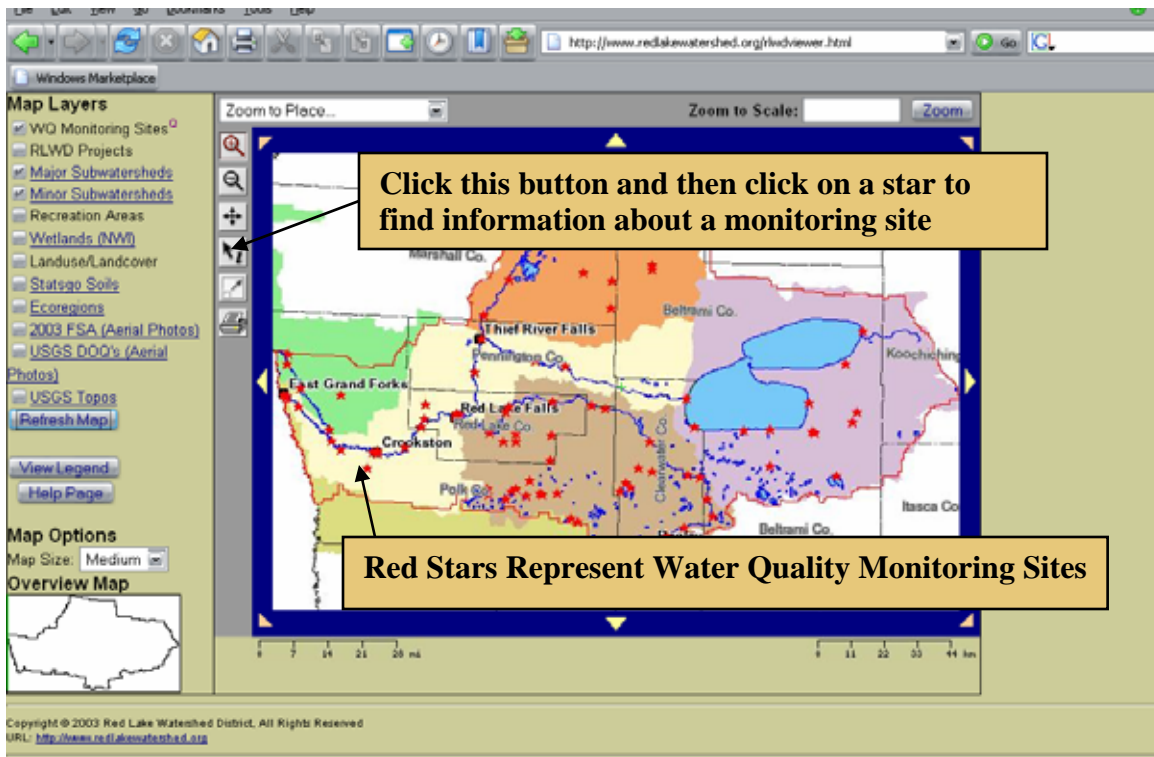
Figure 6. Interactive Map on the RLWD Website.

Figure 7. A report card webpage is linked to the water quality data table in the database to create "grading curves" and give a site a grade based upon the curves.
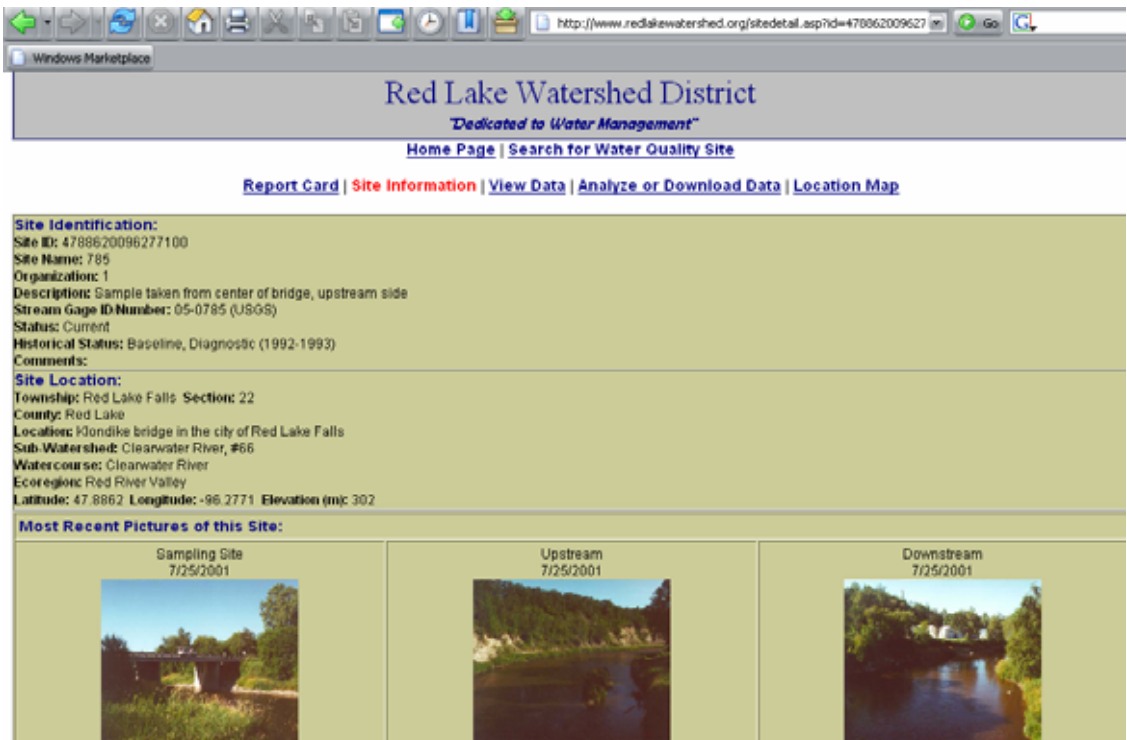
Figure 8. A site information page is generated for each site and is linked to a site information table in the database.
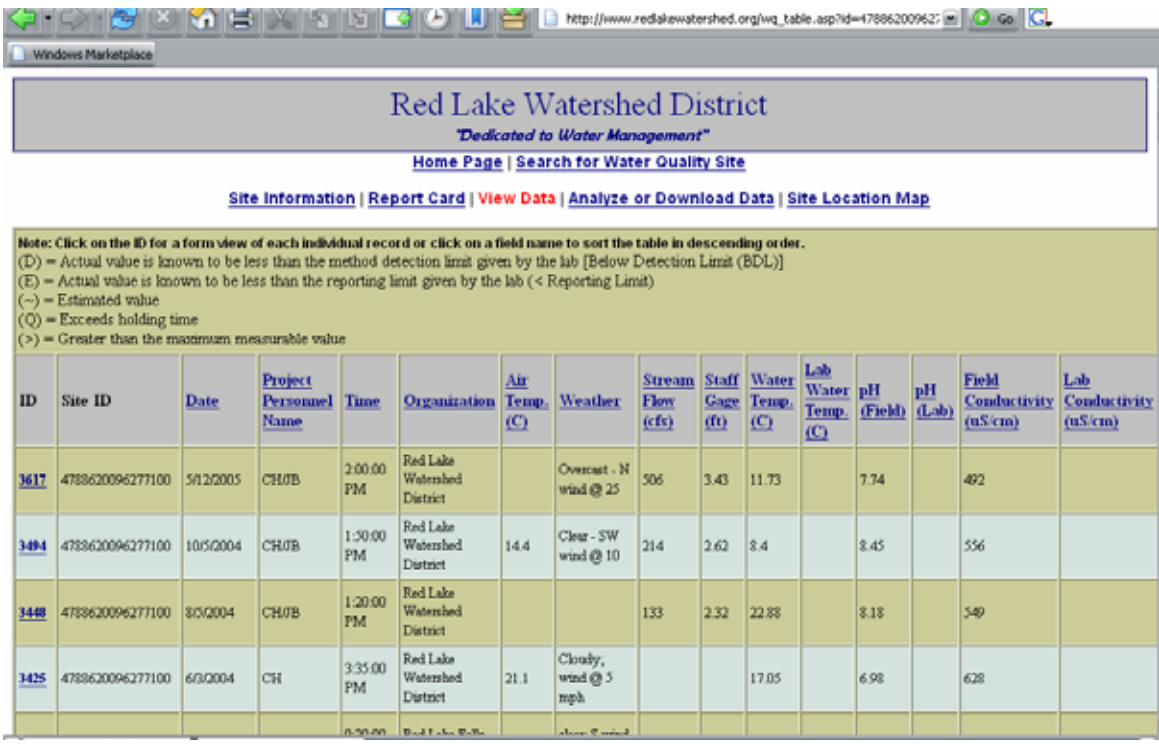
Figure 9. The View Data webpage simply displays data from the selected site - it is linked to the water quality data table within the database.
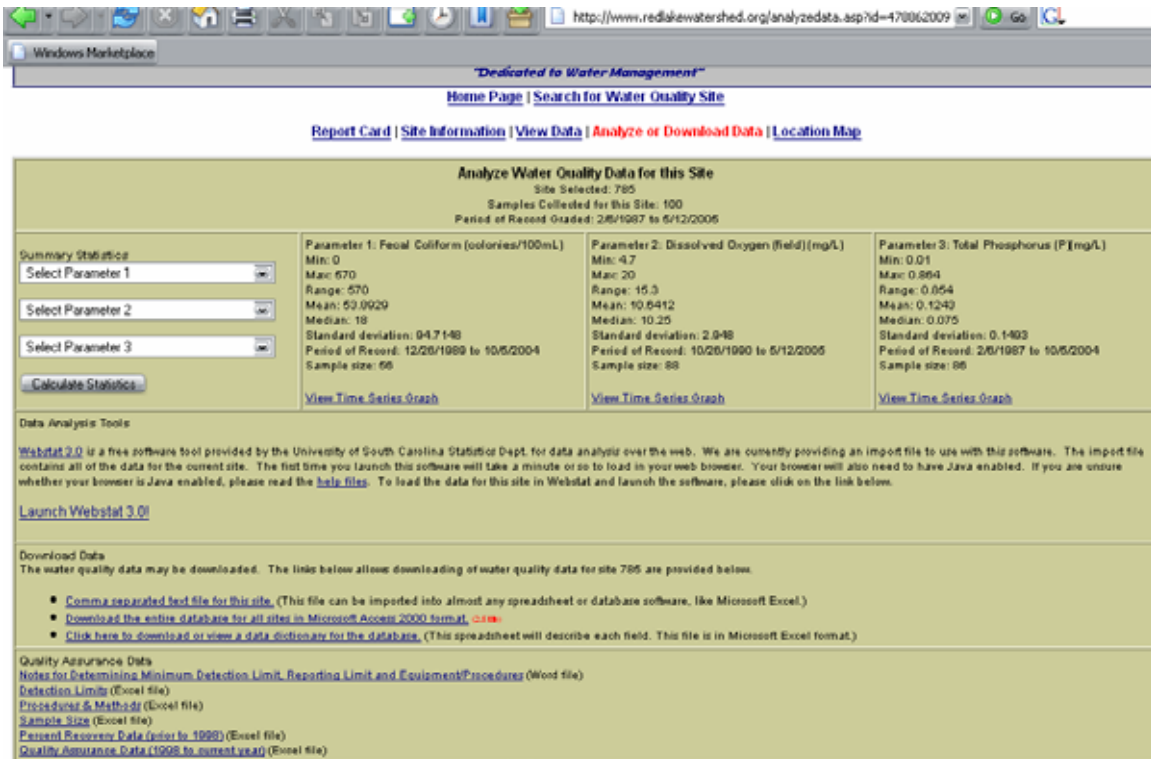
Figure 10. Analyze or Download Data Page.

Quality control measures can be incorporated into an Access database. Examples include a range of allowable values for a data entry cell, or even special procedures for entering data into a database. The Red Lake Department of Natural Resources uses a data entry system that involves duplicate data entry and data verification. If any data entered during the second round of entry does not match data from the first round of entry, an error message is displayed and the user must double-check the original data sheet to verify the correct value. This helps to eliminate data entry errors. Some common types of data entry errors include entering data into the wrong cell or field, misplacing a decimal point, adding an additional digit (accidentally hitting too many keys), and omitting data altogether.

In a data entry form like the figure below, validation rules can be added to each field. Number fields should accept only number values. Fields can also be made to reject or question values that do not fall within a specified range. For pH, for example, the range of possible values is 0-14, so if a value of 72.6 is entered, the value will be rejected and the user will have to check the results and enter the correct value. So, for example, a user may have misplaced the decimal point on the first try and, after receiving an error message, has a chance to replace 72.6 with the correct value of 7.26.

A normal range can be defined for a parameter as well. This function would question the data entry personnel about whether a value is correct or not if it falls outside a normal range for the parameter. Abnormally high values can still be recorded by verifying the number, but false values caused by extra or misplaced keystrokes will be checked and corrected. Validation rules can be added within the design view of the database. Right click on the cell and select Properties to access the window, shown below, in which validation rules and other controls can be added to the cell.
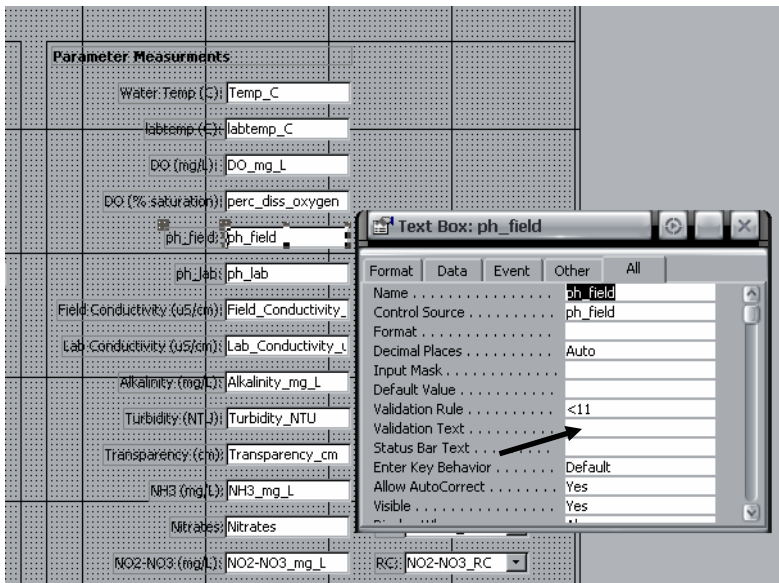
Figure 11. Adding a validation rule to a data entry form cell.

Finding these errors during data entry is important for getting data into STORET since the database will reject data that is out-of-range. Data will then have to be corrected and re-submitted, thereby delaying the entry of the data into STORET. See Section 4.3 for more tips on getting data into the STORET database.

2.2 Data Storage in Microsoft Excel

Although Microsoft Access is one of the best options for storing a large amount of data, especially for linking tables and querying data, Microsoft Excel is the program most likely to be used for data analysis. There are two ways to have data available in Microsoft Excel for the purposes of data analysis. It can either be entered directly or it can be imported from another program. An alternative method of data storage and analysis would be to export data from Microsoft Access and import it into Microsoft Excel or another statistical analysis program to analyze the data. The RLWD has entered data directly into both Microsoft Access and Microsoft Excel. Having updated Microsoft Excel spreadsheets on-hand is valuable because data analysis can be performed quickly.

The RLWD will be switching to Excel spreadsheets that are linked to the Access database. This way, there will be updated Excel spreadsheets available for analyzing data, but data can be imported from the RLWD's Access database instead of entered cell by cell.

1. Know the location of the Microsoft Access database from which you will be importing data.

2. Begin a new query: Data => Get external Data => Database Query (you may need to have your Microsoft Office CD ready in order to install this feature)
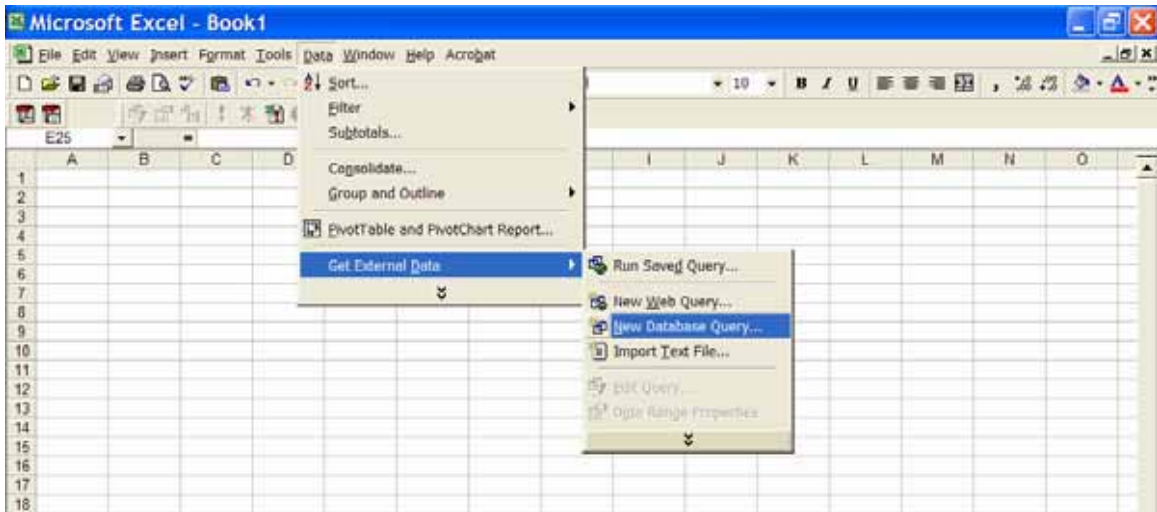
3. Select MS Access Database in the Choose Data Source window and click OK.
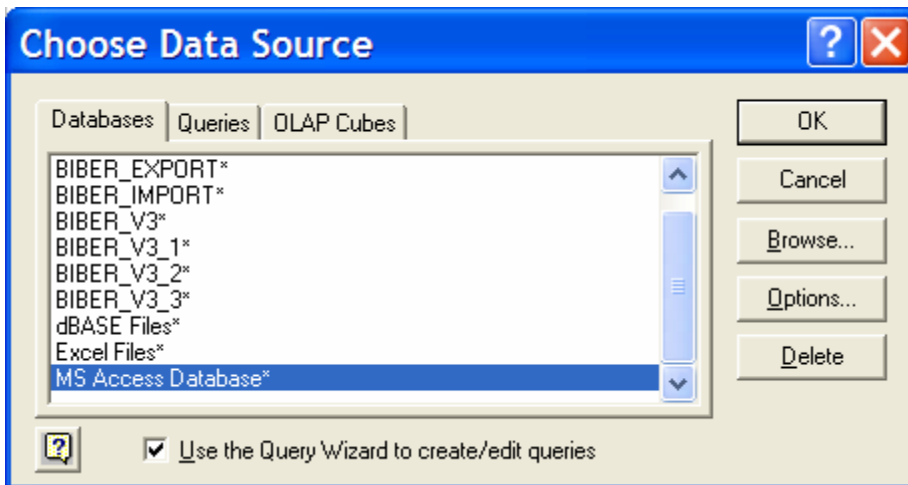
4. Browse to the location of the database from which you will be importing data and click on the OK button.
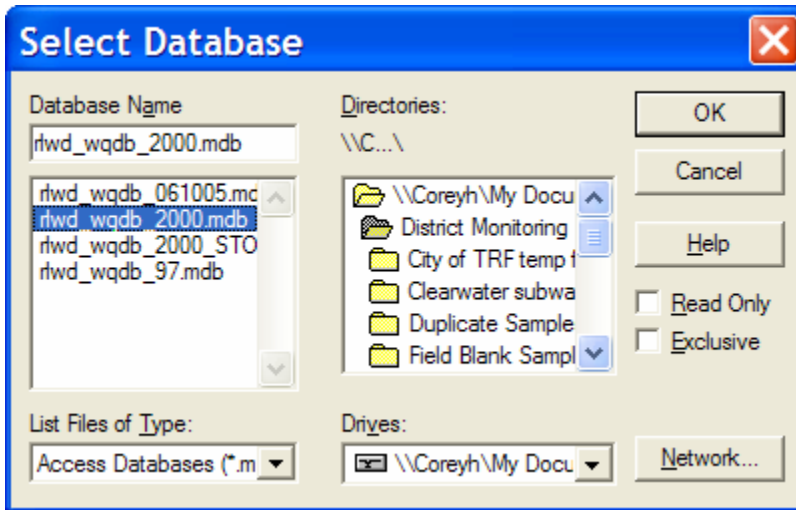
5. In the Query Wizard – Choose Columns window, choose the table and columns that you want to import into your spreadsheet. Click on the Next button.
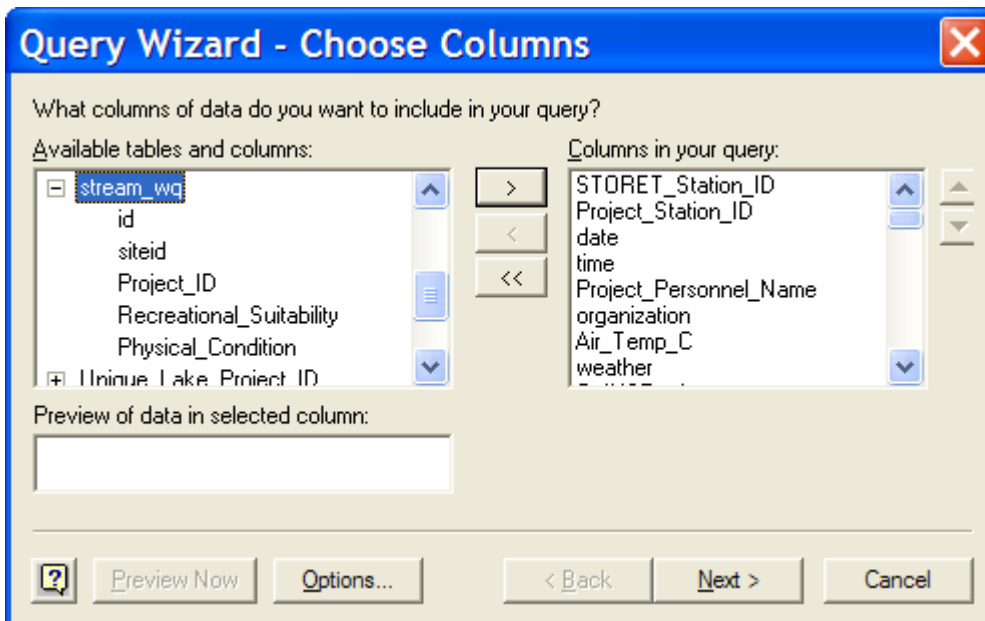
6. In the next window (Filter Data), you may choose to filter the data by date, site, etc. If your water quality data table within Access contains data for more than one site, for example, you may filter the data by site name and only import data from one particular site.
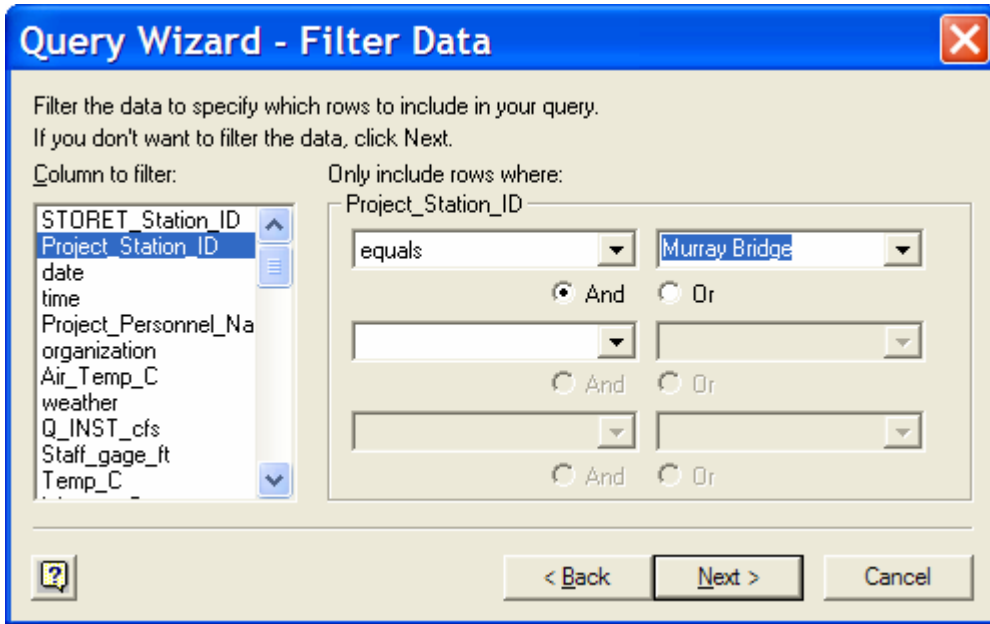
7. The Sort Order window of the Query Wizard is where you can signify how the data should be organized within the table. The example below will place the data in a chronological order.
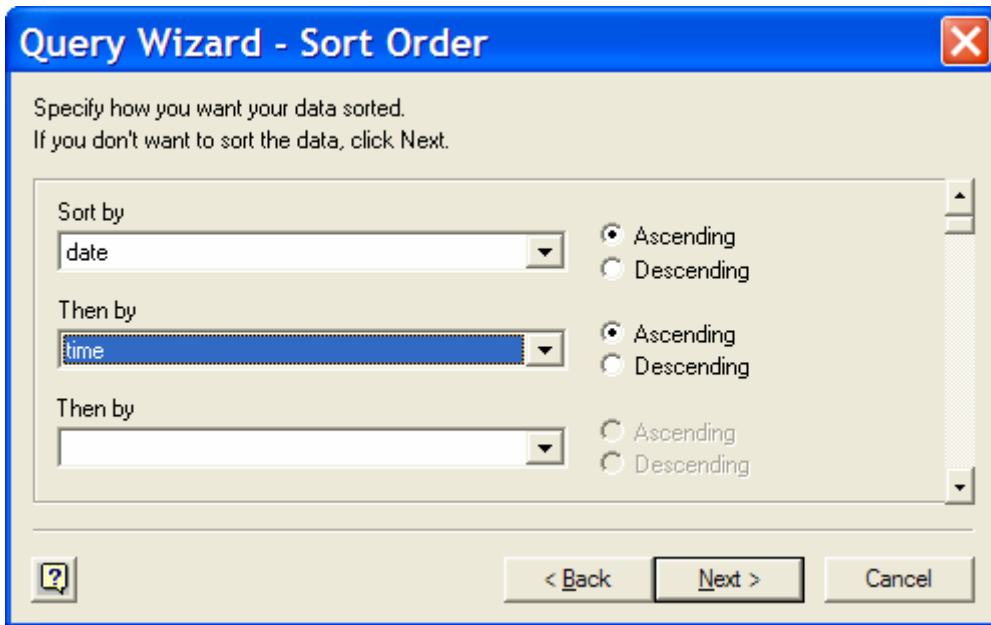
8. In the next window, leave "Return Data to Microsoft Excel" selected and click Finish.
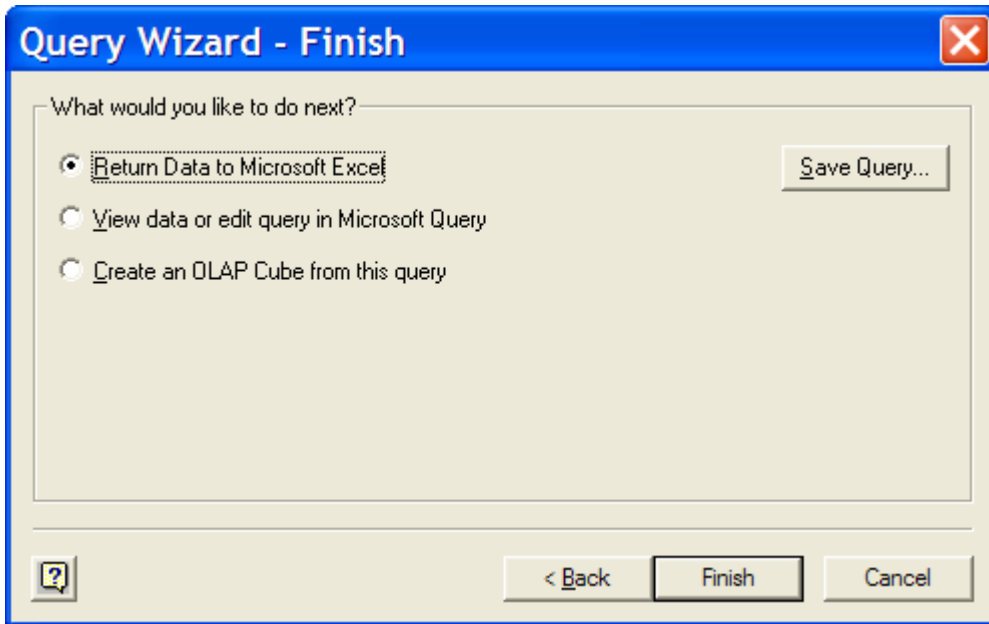
9. Choose a worksheet as a destination for the data.
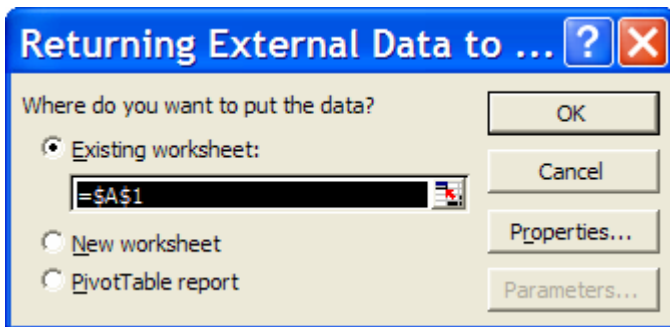
10. You now have an excel spreadsheet that can be updated from the Access database with a push of a button.

11. Display the External Data toolbar to make refreshing data easy:
    View=>Toolbars=>External Data
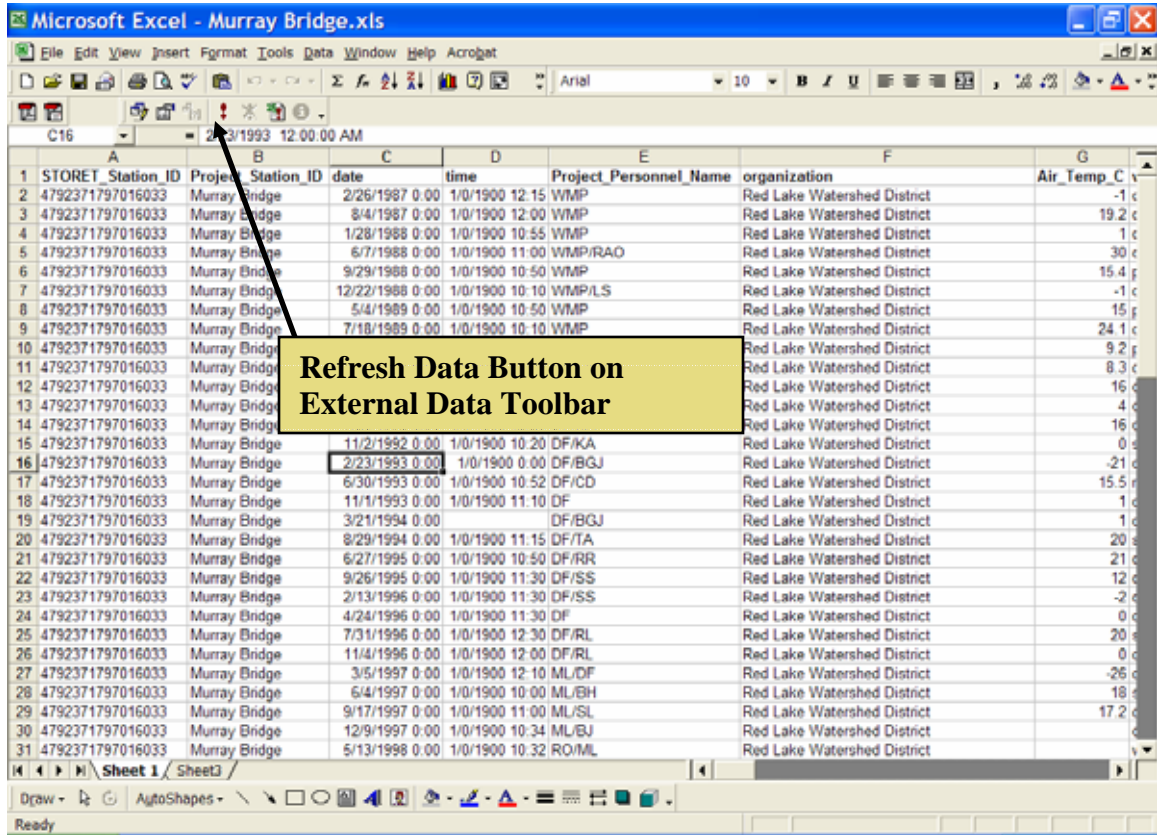
No strict protocols will be established by this document for the organization of data in Microsoft Excel due to the different needs of different monitoring projects and the flexibility of the program. There are, however, some relatively universal tips that help make a clean, useful Microsoft Excel spreadsheet for water quality data. The name of each parameter in its respective column (or row) heading should be clearly stated. Units (mg/L, NTU, ft, etc.) should be indicated if applicable. Dates should be in Excel format (mm/dd/yyyy or mm/dd/yy). A Microsoft Excel workbook (entire file) can contain many worksheets (separate spreadsheets). Each worksheet is represented by a tab at the bottom of the window (defaults = Sheet 1, Sheet 2, and Sheet 3). Some user may choose to have only one workbook for all their monitoring sites, or a separate workbook for each site with multiple worksheets dedicated to data analysis results. The RLWD uses a separate workbook for each long-term monitoring site, but will also combine sites into one workbook for smaller, short-term monitoring projects. Within a workbook, raw data should be stored in one worksheet. Other worksheets can be used for pivot tables, summary tables, assessments, graphical analysis, and statistical analysis. Methods for conducting these different types of analysis can be found in the following chapter.

One thing that can cause problems with data entry and analysis is water quality parameter data that isn't represented in numerical format. This may include lab results that are below the minimum detection limit (MDL). These results are reported with a < symbol in

front of the value of the MDL. Lab results that are too numerous to count are another example. If you wish to use your data for analysis, it will be necessary to create modified columns next to the original data columns into which data can be transformed into useable numeric data. If your data is going to be submitted to a database such as the EPA's STORET database, fields containing flags or remark codes will need to be added next to the original data. When data that is below the minimum detection/reporting limit or greater than the maximum detection limit is submitted to STORET, the detection limit is entered into a column under the parameter and units heading and the remark code is placed in a column directly to the right of this one (with a heading of RC of FLAG). See Section 3.1 for more information on using censored data.

A limitation of Microsoft Excel is its storage efficiency for large amounts of data. Microsoft Access can efficiently handle a larger amount of data than Excel. Even Access has its limitations and large scale databases will require programs such as Microsoft SQL and Oracle. The EPA's modern STORET water quality data, for example, is stored using an Oracle database. These databases are generally only used by agencies that need to store a very large amount of data (USEPA, USGS) and large companies that need to store a large amount of transaction data.

3.0 Data Analysis

Before beginning data analysis, think about what questions you want to answer. Here are some examples:

- Are designated uses generally supported in the watershed?

- Did the levels of pollutants violate state water quality standards? How many times or what percent of the samples at each site? Where? When?

    o See section 3.53 for directions on assessing water quality data for the determination of impairment.

- How does the water quality compare with ecoregion water quality standards? Ecoregion values are often expressed as percentiles, so you will need to calculate the corresponding percentiles for your results in order to compare them to the ecoregion values.

    o See Section 3.54 for ecoregion values and Section 3.21 to learn about calculating percentiles.

- How do results compare over time? How might any changes be explained?

    o See Section 3.3 to learn about trend analysis.

- How does one parameter compare to another?

  o See Section 3.25 to learn about measures of association.

- How do sites compare spatially (upstream vs. downstream)? How might any changes be explained?

  o This can be done by comparing summary statistics (Section 3.21).

- Are specific stressors affecting the health or human use of the water body?

- How do the results upstream of a suspected source of pollution compare with the results from downstream?

- Would any of the monitoring streams qualify as reference (unimpacted, pristine) streams?

- What is the natural background water quality like in the watershed?

- Did you collect the required number of samples from the minimum number of sites (completeness)?

  o See Section 3.22 to learn about quality assurance calculations.

- How will the sensitivity of the methods and equipment you used affect the results? (Section 3.22)

- How did quality assurance results (from split, duplicate, spiked, replicate, known, unknown, and blank samples) compare with expected results? Did they meet your data quality objectives? (Section 3.22)

- Did you sample frequently enough and at the right times?

- What is the degree of change that is significant for each parameter, considering natural baseline and variability?

- Do the field notes coincide with the data? Are there any data entry errors?
  o See Section 2.0 on data storage.

- How much of a particular water quality parameter (i.e. sediment) is being transported past a monitoring site?

  o See Section 3.4 to develop load estimates.

- How healthy is a particular lake? How suitable is it for recreation or aquatic life?

  - See Sections 2.51 and 2.52 to learn about the Carlson's Trophic State Index and water column temperature and dissolved oxygen profiles.

3.1 Using Censored Data

One thing that can cause problems with data entry and analysis is water quality parameter data that isn't represented in numerical format. This may include lab results that are below the minimum detection limit (MDL). Laboratory analysis techniques have a limited accuracy. The smallest amount of a parameter such as nitrates, total suspended solids, or fecal coliform that laboratory methods can detect is referred to as the minimum detection limit (MDL). Results that fall below this limit are reported as either BDL or < a number. These values are not useable when calculating summary statistics such as the mean or median. Removing this data from the data set is not a good option because the statistical analysis results would be biased and misleading. Since the value of these measurements is unknown, questions arise as to what should be done with this data so that it can still be used for statistics.

Lab results that are too numerous to count are another example. Transparency tubes are also recorded in such a way that analysis cannot be performed on raw, untransformed data. There are two readings taken for each measurement and sometimes transparency values are greater than the highest reading possible on the tube as well.

In order to be able to use this data for analysis without losing the original results, a modified column can be created to the right of the original data column for each parameter. The modified column is a numerical representation of the original data. While the modified field is needed for analysis, a different field, the flag or remark code field is required for the submission of data to the MPCA's STORET database.

If you plan on using your Excel spreadsheet for storing data in a STORET acceptable format, you will need to insert a flag field (or remark code) column to the right of any data columns that include any results that are MDL, BDL, > than detection limit, etc. Place the value of the minimum/maximum detection/reporting limit in the data column and, in the flag field column, input the appropriate flag character. See Section 4.3 for more details on these flag fields and entering data into STORET. If you will be using the spreadsheet for analysis only, then follow the directions in the following paragraph.

Lab results that are less than the minimum detection limit (BDL , <.02, <1, etc.) can be transformed to a numerical format in the modified column. This allows the censored data to be used in data analysis. The value in the modified column should be equal to one half of the minimum detection limit. The same value should be used in place of every BDL result for a parameter even if the reporting limits change over time.

A study entitled *Statistical Methods for Analyzing Censored Water Quality Data Sets* was completed by Houston Engineering, Inc. for the Red Lake Watershed District in 2002.

This study was conducted as a part of the Red River Watershed Assessment Protocol Project. The purpose of the study was to find the best method for dealing with censored data. The study examined the simple substitution method; distributional methods such as the probability plot, maximum-likelihood estimation (MLE), and fill-in with expected MLE values techniques; and the Helsel's Robust Method.

The study recommended using the simple substitution method when dealing with BDL values. For the simple substitution method, the BDL result is replaced by an actual value. This value may be 0, the MDL, or a value equal to one-half the MDL. Since substituting 0 or the MDL may still bias the results of statistical analysis. Therefore, the study recommends using either the ½ MDL value, or calculating summary statistics from the substitution of both 0 and 1 and averaging these results. The study is included in this document in Appendix A. This topic is also covered in Sections 2.1 and 4.3.

Detection limits may change over time and may differ among laboratories, equipment, and methods. If detection limits, for example, get smaller over time and different ½ BDL values are entered into the modified column for use in data analysis, the decreasing BDL values may impart a false decreasing trend. The reason this trend would be false is because, whether the reporting limit is .4 mg/L or .1 mg/L, the actual value is unknown, so one cannot automatically assume that the actual concentration of a sample is higher with a MDL of .4 than it is with a MDL of .1 mg/L. So, if there are multiple reporting limits, what value should be used for all the results? The censored data study completed by Houston Engineering (see Appendix A and/or Section 3.1) recommends applying the highest MDL to all data, while *We Have Stream Data, Now What?!* recommends applying the smallest MDL to all the data. The justification for using the smallest MDL is that ½ of the larger MDL may be equal to an actual reading that was recorded while the smaller MDL was in use. The justification of using the larger (less sensitive) MDL is that it is necessary to censor quantified values that are less than the largest MDL in order to prevent artificial trends. The RLWD will follow the recommendations of the Houston Engineering censored data study.

Now, what is to be done with results that exceed the highest value that can possibly be measured? A value can be entered into the modified column that is equal to the highest possible reading plus one. So, if a transparency reading is recorded as >100 cm, it may be recorded as 101.

However, we run into a problem with changing maximum detection limits that is similar to the problem we have with minimum detection limits. The solutions discussed in the following paragraphs will use transparency tube readings as an example since they are a widely used water quality measurement device and there are several different tube lengths available. The concepts discussed in the following paragraphs can also be applied to other parameters such as turbidity.

Since there are different lengths of transparency tubes, there may be data sets that contain values of 60+ cm, 100+ cm, or even 120+ cm. For these, *We Have Stream Data, Now What?!* recommends using the lower of these two numbers and even excluding data from

the longer tube. This method has some merit because some of the actual transparency conditions recorded as 60+ cm may not have been greater than 100 cm. So, this method avoids any false statements by not changing 60+ cm to 100+ cm. Also, when the lower maximum value is used for all measurements, any results from the 100 cm tube that are greater than 60 cm must be transformed from their original value to 60+ cm. If it is necessary to transform 100+ cm readings to 61 cm, than all readings greater than 60 cm must be transformed to 61 cm, not just the "100+" readings. This does avoid false statements or assumptions about the data. For example, results from the 100 cm tube of 65, 80, or 100+ cm <u>are</u> greater than 60 cm.

Censoring all the data that is greater than the maximum value of the shortest tube used in a dataset may prevent the appearance of false trends, but may prevent the determination of any trend at all. For example, a stream was monitored for 5 years with a 60 cm tube and then for five years with a 100 cm tube. If the water quality in this stream has been deteriorating over the last 10 years from an average transparency of <100 to an average transparency of 70 cm, this trend wouldn't be detectable if all values were changed to 61 cm. With this method, you are losing data for both periods of time. An argument for this method would be that all the values would be true (100 <u>is</u> greater than 60). This method would work better for streams with transparency values that are normally below the maximum of the shorter tube than it would for cleaner streams with transparencies that are normally greater than the shorter tube's maximum.

Increasing all the "greater than the detection limit" values to the maximum height on the taller tube would allow for more of the data from the taller tube to be used. Data from at least one of the tubes will be completely represented in the analysis data set. No data censoring occurs in this method beyond the limitations of the equipment at the time that the data was recorded. This method may be helpful in cleaner waters that exhibit transparencies that are close to the maximum value on the taller tube – where trends would be masked if all results are reduced to the maximum of the smaller tube (plus one). This method may create false assumptions about the data from the shorter tube, unfortunately. If values are rarely near or above the maximum of the taller tube and/or are frequently below the maximum on the shorter tube, this method definitely should not be used. Using a value of 101 cm in the modified column for a reading of 60+ cm for a stream with an average transparency of 45 would be unacceptable. Using a value of 101 cm in the modified column for a reading of 60+ cm for a stream that has an average transparency of 99 cm may be more acceptable.

The method you use for your analysis may depend upon your data. You may even have to try multiple methods for you may find a trend with one method that you couldn't find with another. The best solution to the problem, however, is to use consistent methods and equipment so the problem of multiple maximum readings is not encountered.

Another option is to conduct separate trend analysis for different monitoring methods or equipment. This, perhaps, may be the best method to use if more than one type of tube has been used and values are frequently greater than the lesser of the maximum detection limits that were used.