Chapter 4. The Fox River Watershed Water Quality Database

Water chemistry data from various sampling activities in the Fox River watershed were compiled into a single database for analysis. The FoxDB is a relational database that contains information on sample sites, parameters measured, and the results of laboratory analysis of the samples, as well as the sampling agency or group. A data quality rating scheme was developed to assign a numerical grade to the data rating reliability and the comparability of the sampling and analysis methods to contemporary standards. This chapter describes data sources, database design, and the grading system. A data dictionary for the database is provided (Appendix 2) in addition to a description of the conversion of major datasets to the FoxDB (Appendix 3) and an interface program for data loading and viewing (Appendix 4).

4.1. Purpose and Goals

The objective of developing a water quality database for the Fox River Watershed Investigation was to compile information on specific parameters that define the nature of the stream and river environment: water chemistry data, sediment chemistry data, and the physical parameters such as temperature, DO, and pH. Streamflow data are included as an integral part to interpretation of constituents reported in units of concentration. Data related to biotic measures were not compiled; however, the database could be expanded to include those parameters in the future.

A variety of monitoring activities have been pursued in the Fox River watershed over the years. Some monitoring efforts were designed to collect long-term datasets to monitor ambient water quality conditions, others for short-term projects, compliance or permit monitoring, or collection by volunteer citizen groups. The database serves as a central repository for the data, stored in a consistent format for retrieval and comparison. The structure and attributes of the original datasets were reviewed and translated to a common format in the Fox River database, FoxDB. The “quality” of the data, collection protocol and laboratory analyses were reviewed to assign a “grade” to the datasets for comparability and reliability in consistent manner.

4.2. Data Description

The FoxDB is populated with data from several sources. Regular monitoring programs of the U.S. Environmental Protection Agency (USEPA), Illinois Environmental Protection Agency (IEPA), and U.S. Geological Survey (USGS) represent a major portion of data available for the watershed. These data were acquired from the USEPA Legacy Data Center (formerly STORET), the USGS National Water Information System (NWIS), and the USGS National Ambient Water Quality Assessment (NAWQA) databases (all available online). The IEPA data collected after 1998 were not available from the new STORET system and were acquired directly from the agency. Some local governments and facilities carry out regular monitoring in their area of interest. There also have been a few special studies investigating water quality-related issues in the Fox River watershed; however, the scope and scale of these studies vary significantly. This section describes individual data sources, their original structure, attributes, and any special considerations.

4.2.1. Data Sources

USEPA. The STORET (short for STOrage and RETrieval) is a repository for water quality, biological, and physical data. The system is used by state environmental agencies, USEPA and other federal agencies, universities, private citizens, and many others. The USEPA maintains two data management systems containing water quality information for the nation's waters: the Legacy Data Center (LDC), and the new STORET. These data may be accessed through the Internet from the STORET home page (USEPA, 2003f, 2003g).

The LDC contains historical water quality data dating back from the early part of the 20th Century to the end of 1998. The STORET system contains data collected beginning in 1999, along with older data that have been properly documented and migrated from the LDC. Both systems contain raw biological, chemical, and physical data on surface water and groundwater collected by federal, state, and local agencies, Indian Tribes, volunteer groups, academics, and others. All 50 states, territories, and U.S. jurisdictions, along with portions of Canada and Mexico, are represented.

USGS. Water quality data from the USGS are available through the NWIS. The NWIS is a distributed database in which data can be processed over a network of workstations and fileservers at USGS offices throughout the United States. The system has four components: the Ground-Water Site-Inventory System, the Water-Quality System, the Automated Data-Processing System, and the Water-Use Data System.

The Water-Quality System contains results of more than 3.5 million analyses of water samples that describe the chemical, physical, biological, and radiochemical characteristics of both surface water and groundwater. The Web site provides current and historical data (USGS, 2003c).

Data from NAWQA are stored in a separate database (USGS, 2003e). The USGS began its NAWQA program in 1991, systematically collecting chemical, biological, and physical water quality data from study units (basins) across the nation. The data warehouse currently contains and links the following data through September 30, 2001: chemical concentrations in water, bed sediments, and aquatic organism tissues; site, basin, well, and network characteristics; daily streamflow information for fixed sampling sites; and groundwater levels for sampled wells. The database overlaps to a certain extent with the NWIS database. However, each of the two databases contains unique data that the other database does not have.

The Urban Land Use Gradient Study was conducted by the USGS as part of the upper Illinois River basin study of the NAWQA program. Physical, chemical, and biological data were collected at 46 sites in the Fox and Des Plaines River basins in July 2000 for habitat, geomorphic characteristics, water discharge, water chemistry (nutrients, major ions, wastewater indicators, pH, and specific conductance), and aquatic communities (algae, invertebrates, and fish). Water temperatures were collected at most sites continuously from approximately May 2000 to June 2001. Stream cross sections were surveyed from November 2000 to May 2001. Fish were collected in August 2000 or July 2001 at sites not previously sampled by other agencies (Adolphson et al., 2002).

IEPA. The IEPA conducts a wide variety of water quality monitoring programs. Stations are sampled for biological, chemical and/or in-stream habitat data, as well as streamflow. Water quality monitoring programs consist of a combination of fixed station networks and intensive or facility-related stream surveys in specific watersheds. The IEPA operates an Ambient Water Quality Monitoring Network (AWQMN) of fixed stations to support surface water chemistry data needs. Integrated water column samples are collected on a 6-week sampling frequency and analyzed for a minimum of 55 universal parameters, including field pH, temperature, specific conductance, dissolved oxygen (DO), suspended solids, nutrients, fecal coliform bacteria, and total and dissolved heavy metals (IEPA, 2002b).

Intensive river basin surveys are conducted on a five-year rotational basis in cooperation with the Illinois Department of Natural Resources (IDNR). These intensive surveys are a major source of information for annual 305(b) assessments. Water chemistry and biological data (fish and macroinvertebrates) and qualitative and quantitative in-stream habitat information, including stream discharge, are collected to characterize stream segments within the basin, identify water quality conditions, and evaluate aquatic life use impairment. Fish tissue contaminant and sediment chemistry sampling also are conducted to screen for the accumulation of toxic substances (IEPA, 2002b).

Fox River Study Group. The Fox River Study Group (FRSG) initiated its monitoring in April 2002. Seven stations on the Fox River mainstem are sampled bi-weekly. Samples are analyzed for nutrient-related parameters such as DO, temperature, chlorophyll *a*, nitrogen, and phosphorus. The FRSG sample collection and analysis program operates under the guidance established in the Quality Assurance Project Plan (QAPP) approved by the IEPA in March 2002. Samples are collected at seven sampling locations along the Fox River from the Johnsburg Bridge north of McHenry to the Route 47 Bridge at Yorkville. The sample sites are approximately ten miles apart and located on bridges crossing the river at locations both above and below the dams. Volunteers from the wastewater treatment facilities and representatives from local environmental groups are responsible for collection of the samples and performing the required field testing. The sample teams were trained to handle samples in an identical fashion, following the guidelines in the QAPP to ensure reproducibility of techniques. The samples are collected every other Tuesday at approximately 10 a.m. and transported to the Fox River Water Reclamation District (FRWRD) in Elgin for distribution to the analytical laboratories at the City of Elgin, Fox Metro Reclamation District (FMRD) and FRWRD.

The sampling program is closely aligned with the techniques used by the IEPA. The samples sites and the transect composite samples are similar to the sites and procedures used by IEPA. This program was designed to augment the IEPA sampling program. Sample collection, and analytical and quality assurance procedures in this program ensure that all data generated are fully comparable with data collected and analyzed by IEPA.

Huff & Huff, Inc. A Huff & Huff, Inc. study evaluated ammonia levels in a 40-mile stretch of the Fox River from Yorkville to Carpentersville (Huff and LaDieu, 1995). Grab samples were taken weekly at 27 monitoring sites for 12 months beginning in September 1994. The Fox River was sampled at 19 locations and its tributaries at 8 stations. Samples were analyzed for DO, temperature, pH, total ammonia, and carbonaceous biological oxygen demand ($CBOD_5$).

Max McGraw Wildlife Foundation. A two-year study conducted by Max McGraw Wildlife Foundation (MMGWF) investigated the environmental effects of dams on fisheries, macroinvertebrates, physical habitat, and water quality in a 100-mile stretch of the Fox River between the Fox Chain of Lakes and Dayton, Illinois (Santucci and Gephard, 2003). Summer low-flow conditions were sampled at 40 sites located in a free-flowing river directly below dams, impounded river directly above dams, and free-flowing or impounded mid-segment areas between dams. Samples included sediment, ambient water, fish, biological communities, and information on land use. Continuous measurement of selected parameters was carried out over 16-, 40-, and 96-hour periods (15-minute intervals).

Illinois State Water Survey. The Illinois State Water Survey (ISWS) conducted a study of oxygen regime in St. Charles Pool in 1993-1994. Short-term intensive water quality data were collected during two separate time periods, three days in August 1993, and six days in June 1994 (Singh et al., 1995). Data for DO, temperature, conductivity, and pH were collected at five stations at 15-minute intervals. Grab samples collected at the beginning of each event were analyzed for basic physical, chemical, and biological parameters. Sediment oxygen demand was measured at five sites where sediment samples also were collected. Biological sampling consisted of macroinvertebrates and algae.

Local Monitoring. Limited monitoring was conducted by local government or water treatment facilities. The McHenry County Health Department surveyed 14 stations from 1981 to 1997 with varying frequency and for different parameters. The FRWRD provided their data from January 1991 to February 2002. A total of eight stations were sampled: six stations on the Fox River and two stations on tributaries (Poplar Creek and Tyler Creek). Stations typically were sampled weekly to bi-weekly and analyzed for up to 20 parameters. The FMWRD samples three stations in its vicinity weekly for DO, temperature, pH, and total ammonia. In addition, samples from two of these stations are analyzed quarterly for a variety of parameters, including trace metals. Other facilities carry out limited water quality monitoring. The USEPA's Permit Compliance System Database was searched via the Envirofact Data Warehouse Web portal (USEPA, 2003c), to identify entities with National Pollution Discharge Elimination System (NPDES) permits within the study area. The list was reviewed and the 25 largest permitted discharges were identified. A letter was sent to each of these permit holders requesting ambient (in-stream) monitoring data. Ten responses, both written and by phone, were received. For the most part, ambient monitoring is not required and most responses did not reveal existence of additional water quality data.

4.2.2. Streamflow Data Sources

Streamflow (discharge) is sometimes measured and recorded as a parameter result when a water quality sample is taken. However, streamflow information was not included with a majority of water quality data sources. Streamflow thus was estimated from data collected by the USGS at their regular gaging stations network. Daily discharge data from continuous stations and stage data were used to estimate daily flow data for sample sites along the Fox River's mainstem downstream of Stratton Dam. Calculated and measured streamflow data were maintained in separate tables rather than added to the sample parameter results in the FoxDB.

4.3. Database Design

4.3.1. Conceptual Design

Monitoring and testing results do not stand alone: the location, time, methodology, and other information also must be documented. The purpose of a database is to store information in a useful way. The USEPA and the USGS maintain the most comprehensive national water quality databases. Formerly, the USEPA database STORET and the USGS database WATSTORE contained essentially the same data that was collected under a joint agreement. These databases were the standard until recent years. The USEPA has developed a new STORET database that stores data collected since 1999 in a new format. Data collected prior to 1999 are warehoused in the old STORET format or LDC. The structure of the new STORET system differs dramatically from the former system, making it difficult, if not impossible, to import LDC data (sometimes referred to as Legacy STORET) into the current STORET system. The IEPA has been migrating data collected since 1999 to the new STORET system. The USGS has developed the NWIS portal to a variety of surface and groundwater data, and water quality data. Other data sources described above were typically in the form of spreadsheets, with text documentation or hard copy only.

The STORET system is designed so that a registered user can install the software on a personal computer (or a network system), input data, and then upload data to the national warehouse. A user also can download data. The STORET structure provides many avenues for complete and detailed data documentation, a strength that is also problematic for historical data that tend to have an insufficient level of detail to populate the database. The single greatest difference between STORET, LDC, and NWIS is the use of parameter codes. The LDC and NWIS systems use a five-digit code to identify a parameter that also embeds information on the units, medium, and procedures. Table 4.1 provides an example of parameter codes from the NWIS Web site. Parameter codes used in the LDC and NWIS are essentially identical, although the USGS has added a few specialty codes for their purposes, which STORET does not use. Rather during data entry, various fields are coded as to medium, units, and collection method to incorporate the information. To date, the USEPA has not provided a translator between parameter codes and STORET attributes.

The FoxDB mimics the conceptual structure of STORET and, where possible, the same codes and field names were used. Because the majority of data in the database was retrieved from the LDC and NWIS, the FoxDB retains the use of the LDC/NWIS five-digit parameter

Table 4.1. Example of USEPA/USGS Five-Digit Parameter Codes (USGS, 2003c)

Parameter code	*Parameter definition*
00910	Total calcium, in milligrams per liter as calcium carbonate (unfiltered-water sample)
00915	Dissolved calcium, in milligrams per liter as calcium (filtered-water sample)
00916	Total calcium, in milligrams per liter as calcium (unfiltered-water sample)
91051	Total calcium, in micrograms per liter as calcium (unfiltered-water sample)

codes. Data from other sources were reviewed, and appropriate five-digit parameter codes were assigned. The IEPA data collected since 1998 retained enough linkage with the LDC to determine assignment of parameter codes. Documentation accompanying other data sources was reviewed to assign parameter codes.

4.3.2. Relational Database

The FoxDB is a relational database. The following sections describe basic principles of relational databases, introduce the FoxDB data model, and present an implementation of the model in Microsoft Access/SQL Server.

A relational database is a collection of formally described tables that can be edited or expanded in many different ways without having to reorganize the database tables. A new table can be added to the database without modifying all existing tables. Data are entered into tables based on subject and related by a key that makes the records within any given table unique. The columns of a table are called fields; the rows are called records.

Information about each station (sample site) is recorded in the table *TBLStation*. Each record (row) contains information about one station. Fields include station name, a unique identification number, location description, latitude, longitude, etc. The table *TBLSample* contains information about samples collected and has a record for each sample; the fields include a unique sample number, date, time, method, and unique station number. These unique numbers or keys provide the link from one table to the next. Information about the station is linked to each sample taken at that station without repeating the station information for each sample.

In the same way, each sample is related to the results table *TBLResults* by a sample number that is uniquely assigned when the sample and results records are added. Five-digit parameter codes are used to identify individual constituents analyzed in the sample. The parameter table *TBLParameter_Codes* then may be combined with the results table to view the full name for the parameter using the parameter code.

The process of removing redundant data from a relational database by separating information into smaller tables is called normalization. A normalized database generally improves performance, lowers storage requirements, and makes it easier to change the application to add new features.

A data model is a conceptual representation of data structures required by a database. Data structures include data objects, associations between data objects, and rules that govern operations on the objects. The data model focuses on required data and how it should be organized rather than on what operations will be performed. A data model is independent of hardware or software constraints. Rather than try to represent the data as a database would see it, the data model focuses on representing the data as the user sees it in the real world. It serves as a bridge between the concepts that make up real-world events and processes, and the physical representation of those concepts in a database.

4.3.3. Data Model Description

The FoxDB data model describes water quality monitoring and data as a complex but related process. Figure 4.1 shows the conceptual representation of the data model implemented in the FoxDB. Monitoring stations are located along rivers, streams, and lakes. Selected stations are sampled as part of a specific monitoring project. Individual samples are collected and shipped to a laboratory for analysis of specified parameters. The results of the analysis are the numerical values of each parameter analyzed. Results also include the values of field-measured parameters, such as temperature and streamflow.

Each arrow in the diagram designates a separate table in the FoxDB. Individual tables are related through unique identifiers. As described above, a sample is identified by a sample number, and attributes include information about the monitoring station and monitoring project in addition to sample descriptors such as sampling date, sampling depth, medium, etc. The sample number is included in a table of laboratory and field data results linking the values to a particular sample.

A discussion of the main features of the FoxDB follows to give the reader an overview of the FoxDB and its structure. Appendix 2 provides the fields, definitions, and formatting details for each table. All tables, fields, and links to illustrate the database configuration are shown in Appendix 8.

The diagram in Appendix 8 includes tables organized in five major groups corresponding to the entities shown in Figure 4.1: station, sample, project, results, and parameters. Rivers as spatial features are part of a geographical coverage, and the link to stations is established by spatial location. Laboratories are not included at this stage because the information is often unknown and not readily available from original data sources. The table *TBLIDLocations* is part of the database, but it is not included in any of these categories. It describes the source from which data were acquired for this project. The information also is used for the database maintenance and batch data import. For discussion purposes, actual table names in the FoxDB are italicized and actual field names are within quotation marks.

A station is described in the table *TblStation_Information.* Station locations may be displayed in a Geographical Information System (GIS) environment using latitude and longitude, which were determined for each station from the original data source or from the station description and 1:100,000 scale topographic maps. In addition, the station location in the stream network is established by river name and both National Hydrography Dataset (NHD) and Reach File Version 3 (RF3) codes from USGS and USEPA river geographical coverages, respectively. Other attributes include various station codes: “Station_ID” represents a unique identifier within the FoxDB, USGS and USEPA codes are included for stations where available, as are special station codes used by any other agency or sampling program. Other fields describe the station’s attributes. For example, “Station_Type” identifies by a code whether the station is located on a river, lake, wetland, canal, etc. The description of the code used in “Station_Type” is given in a lookup table, *TBLSTation_Type*, which provides the station’s “Primary_type” and “Secondary_type.” The lookup table also indicates whether the station is located on a natural or an artificial water body. Primary and secondary station codes are identical to those used in the USEPA’s new STORET database (http://www.epa.gov/storet/).

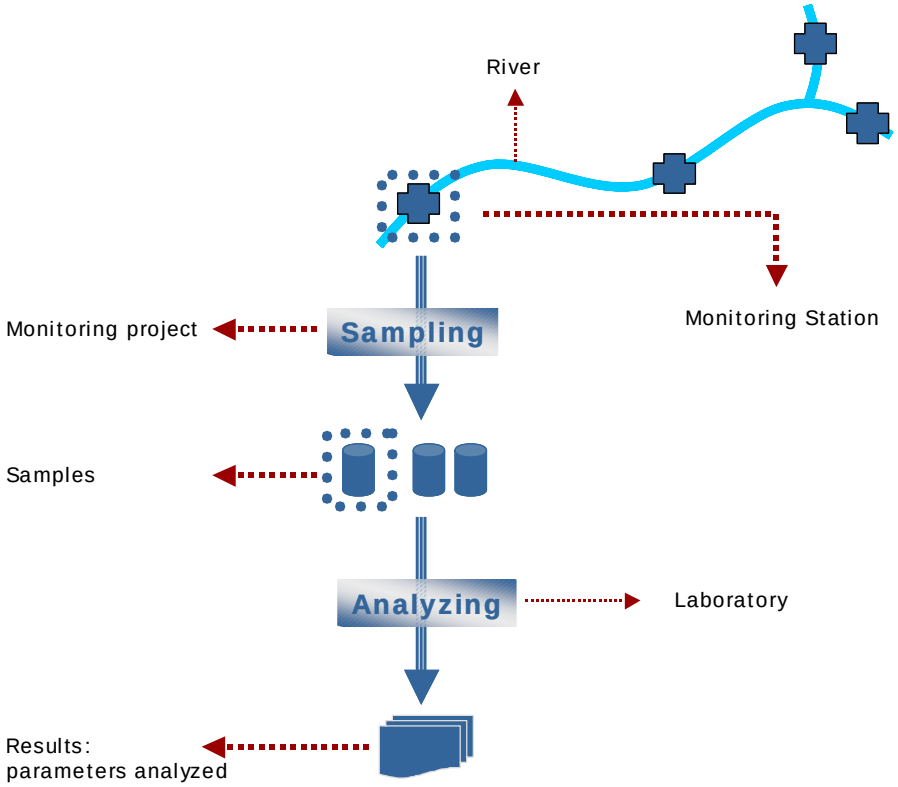

Figure 4.1. Schematic representation of water quality data model used in development of the FoxDB

A sample is described in the table *TBLSample* by the station where it was taken, sampling date and time, sampling depth and a monitoring project under which it was collected. Sample "Medium" indicates what was sampled: water, sediment, biota, physical characteristics (including habitat), etc. "Sample_Type" further describes sampling methods (composite, grab, fish tissue, etc.). "Composite_statistic_code," a field preserved from the USEPA's Legacy STORET database (http://www.epa.gov/storpubl/legacy/ref_tables.htm), indicates whether a summary value was stored rather than an individual value, for example, an average of several samples. The code has not been completed for all data coming from other sources. Lookup tables explain the codes used in each field to describe medium, sample type, and composite statistic codes. A comment field is used for any comments relevant to the sample, for example, existence of replicate samples, quality assurance concerns, etc.

Parameter codes are defined in the table *TBLParameter_Codes*, which includes a verbal description, both full and abbreviated, and reporting units. Additional related tables associate parameters with a parameter group. The database includes two schemes for grouping parameters: the original USEPA parameter groups (used in the Legacy STORET database) and the Quality Assurance Project Plan (QAPP) groups developed specifically for this project. The QAPP groups were created to facilitate evaluating the quality of imported data. The QAPP scheme groups parameters on two levels: first, by medium sampled, and second, by constituent analyzed. The first number of the QAPP code indicates the medium; the second number indicates

the main parameter group (basic inorganic, nutrients, metals, organics, etc.); and the third number indicates the constituent subgroup (for example, nitrogen in the nutrients group, or pesticides in the organics group).

A result is defined by a sample code (linking it to the sample analyzed) and a parameter code (the constituent measured). Original five-digit parameter codes from Legacy were associated with most data, and the FoxDB uses these Legacy codes. A result value is accompanied by a remark code explaining mostly quality assurance issues. For example, a value reported may be below a detection limit, calculated from other parameters, estimated, etc. An optional grade can be used to flag any questionable data identified during analyses. Numerical and non-numerical results are stored in tables *TBLResults* and *TBLResults_Vol_NonNumeric*, respectively, to ensure integrity of the value field. All replicate results are kept in a third table, *TBLReplicates*. Some of the datasets imported did not differentiate clearly when results were from replicate samples.

A project is described in the table *TBLProjects_Programs*, which includes a project name or title for which monitoring was performed, a code for the monitoring organization, project study area, project purpose, beginning and ending dates, and contact information. The organization is described by its full and abbreviated names, and category (federal, state, facility, or other). The address, contact person and phone number, and the organization Web site are given, if available. A project can be assigned a quality assurance (QAPP) grade and a comparability-usability (CU) grade for any QAPP parameter group.

4.4. Implementation and Navigation

The FoxDB was developed and tested using a Microsoft SQL Server. The database was converted to Microsoft Access format for distribution. Both Microsoft Access and the SQL Server support relational databases. The complete Microsoft Access database is available for download from the ILRDSS Fox River Watershed Investigation Web site (http://ilrdss.sws.uiuc.edu/fox/).

The Microsoft Access database includes both core and lookup tables with all established links. Data can be imported to many applications using an Open Database Connectivity (ODBC) interface. This interface enables accessing data among various software applications regardless of vendors. For example, data can be imported to Excel or Statgraphics (statistical software) for analyses.

Two queries have been designed and included with the FoxDB. These queries are recommended for casual users with some experience with relational databases and Microsoft Access and may be used as examples for construction of additional queries. Advanced users are encouraged to build custom queries.

4.4.1. Example Queries

Query name: CountPhosphorusResult

Figure 4.2 illustrates a query that generates a list of stations for which there are result values for any form of phosphorus, the medium (water or sediment) and the number of samples at the station. The FoxDB tables involved are:

TBLStation_Information
TBLSample
TBLResults
TBLParameter_Codes
TBLQAPPGroups
TBLQAPP_Group_Codes

The selection is performed by specifying:

TBLQappGroups_Codes, *Parameter Group* = 10 (lists all results for any form of Phosphorus)

The expression in the first column: CountofResults: Count(*), counts the number of results for each parameter.

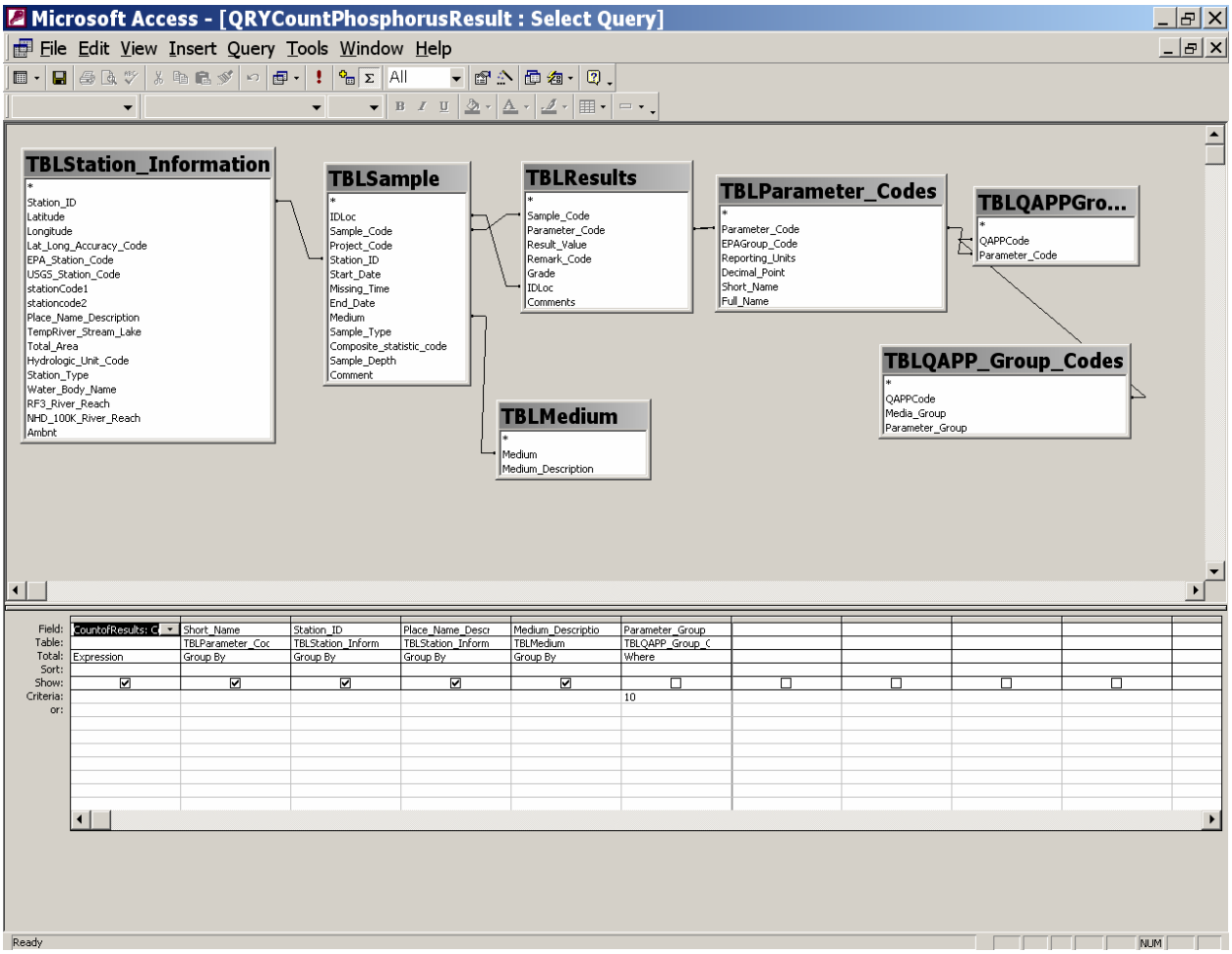

Figure 4.2. FoxDB query name: QRYCountPhosphorusResult

Query name: Phosphorus Results by Station

Figure 4.3 illustrates a query that generates a list of all the phosphorus result values for samples collected at all stations. The tables from FoxDB involved are:

TBLStation_Information
TBLSample
TBLResults
TBLResults_Remarks
TBLParameter_Codes
TBLQAPPGroups
TBLQAPP_Group_Codes
TBLParameter_Group

The selection is performed by specifying:

TBLQappGroups_Codes, *Parameter Group* = 10

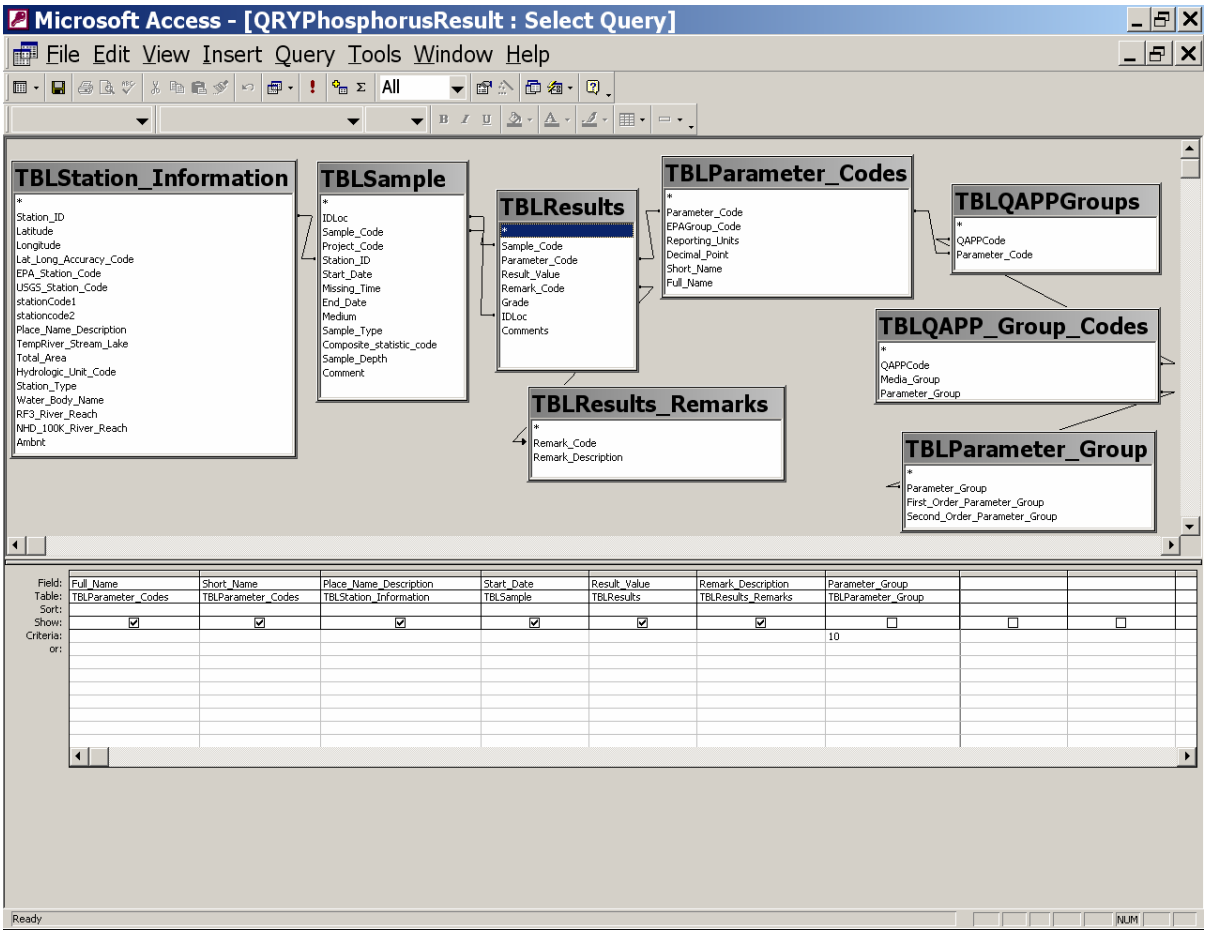

Figure 4.3. FoxDB query name: QRYPhosphorusResult

4.4.2. Importing Data and Future Updates

A program has been developed for loading and viewing data in the FoxDB. This program provides a user-friendly interface to explore the database content and to enter new data. The interface is designed for entering data one parameter result value at a time, as would be necessary when creating an electronic copy of data from laboratory sheets. An experienced database manager can import large electronic datasets to the database. The interface setup program may be downloaded from the Fox River Watershed Investigation Web page (http://ilrdss.sws.uiuc.edu/fox/). Appendix 4 describes program navigation. The Data Loader & Viewer program can be installed, with the full Microsoft Access database on multiple, independent personal computers. However, a "primary" or "master" database copy needs to be maintained.

Database maintenance is essential to extending the useful life of a dataset. Protocols must be established for data entry and maintenance of a "master copy" of the database. One option is to identify a single location where the master database is maintained, and all data entry is at that point. Alternatively, where there are multiple data entry sites, such as the various water reclamation districts, the location (personal computer or server) where the master copy of the database is maintained should be designated as well as the primary data manager. Data may be entered at remote locations, files sent to the primary data manager, and loaded upon review and acceptance.

4.4.3. Special Considerations

Station Location. In the FoxDB, locations of monitoring sites are identified by latitude and longitude coordinates as well as a detailed description. Latitude and longitude included with the original data were used to display the stations in ArcMap GIS software. The location based on latitude/longitude was checked individually against the description for every station. Additional geographical layers, such as river network, roads, or topographical maps, were used to verify the location. Geographical information and description for 427 stations (85%) matched. Of the remaining stations, locations of 30 stations (6%) identified by the IEPA station code were determined from a geographical coverage of IEPA stations provided by that agency. Locations of the remaining stations were determined from the description and 1:100,000 scale topographic maps. Latitude and longitude then were identified in the ArcMap environment.

First, the Legacy and USGS stations were displayed. Each location was verified and corrected when necessary. All stations from other sources were first checked against the existing stations to avoid duplicate locations. In such cases, descriptions were combined to include all keywords from both sources. Source of latitude/longitude is reflected in the "Latitude_Longitude_Accuracy" field of the table *TblStation_Information*. A two-letter code indicates the use of original or corrected data, and the accuracy level of locational data, respectively, when known. The location was verified only for Illinois stations. All Wisconsin stations have retained their original latitude and longitude.

Parameters. Parameters are uniquely identified by a five-digit USEPA Legacy code. Data from USEPA, IEPA, and USGS already included the proper code. Other data sources described the parameters analyzed within their respective projects verbally. These descriptions were compared against STORET descriptions, and codes for the appropriate parameter were selected. The organization providing the data was contacted to clarify descriptions, methods, or units when several STORET codes matched the provided description. For the few cases where the organization did not respond, professional judgment was used to assign a code with the best matching description. If a code could not be determined, data were not imported into the FoxDB. Only a limited set of data was not imported (e.g., metals and organic data from McHenry County Health Department).

Translation of Attributes/Codes. The FoxDB structure is, in essence, a fusion of STORET and Legacy plus some elements from USGS datasets. Each original dataset came with a specific set of attributes. Thus, a translation key was developed for converting original attributes and codes into FoxDB attributes (see Appendix 3). Attributes of a sample were retained when present in the original datasets.

Removing Duplicates. All imported data were checked for duplicate entries: samples taken at the same place and time analyzed for the same parameter. Such duplicates were identified between different data sources, as well as within the same data source. Duplicate entries were moved to a separate table *TblReplicates* structured after the results table *TblResults*. Replicate values include both the original sample code and that of the corresponding result. In addition, a comment identifies the existence of a replicate for respective sample.

Most duplicate entries showed the same numerical result for a parameter in question. Entries with different results were examined individually. Most of these cases were caused by a different rounding process; only a limited number showed distinctively different numbers. When a different number was reported, the record indicating worse water quality condition was retained in the results table; the other record was considered a replicate. This procedure was necessary because original data do not contain detailed information on duplicate sampling or analyses. The comment by the sample flags these data so that the user can check the replicate value and rerun analyses, if desired.

Most apparent duplicate results originated from an overlap between various databases. For example, the USGS maintains two water quality databases: NWIS and NAWQA. The data overlap to a certain extent, but distinct data exist in each database. Data in the NWIS database represent rounded values of data in the NAWQA database. Similarly, duplicates exist between the Legacy data and the USGS data (both NWIS and NAWQA). In this case, Legacy data were recorded with higher precision than NWIS data. Although it was reported that all USGS data were removed from Legacy database, the data from joint projects between the USEPA and the USGS remained in the Legacy database. Data recorded with higher precision were retained in the FoxDB.

Other duplicates were found when several samples were collected at the same time but for different projects (USGS data). This was usually the case for basic physical parameters

(temperature, pH, flow, etc.) measured separately when ambient water, sediment, and biological samples were collected simultaneously.

4.4.4. Data Quality

Water quality data imported into the FoxDB were collected by a variety of agencies and research organizations over the years. Procedures for data collection and analysis have changed. Different laboratory techniques have different levels of precision and detection limits. While it is desirable to take advantage of the wealth of information available, it is essential that reliability, accuracy, precision, and comparability of data be documented. Data not meeting contemporary standards for collection and analyses may yield valuable information about trends but should not be included in an actual result value comparison. Previous studies that are not fully documented but performed by reputable organizations may not be appropriate for some uses, but may close data gaps, and should not be excluded from consideration.

The following rating criteria were devised to provide documentation of the grade or the confidence in the quality of the various datasets. Assignment of grades to the data provides a simple mechanism to perform queries on the composite data with screening levels appropriate to the analysis. Full documentation of the data collection procedures, as available, are provided in the original reports.

The data rating criteria developed uses a two-tiered approach to determine the quality of the data received for the FoxDB. The first tier is to determine QAPP availability and acceptability. Sampling design, analytical protocols, and comparability of a dataset with others also are evaluated in this tier. This level of evaluation determines if datasets are documented adequately to provide some assurance as to the accuracy and precision of the information. The second tier is performed by using statistical analyses on the datasets to determine data consistency and reliability.

This procedure was applied to all samples taken on or after January 1, 1998 (last five years). Historical data may be used to evaluate trends or to supplement analysis when present data are not sufficient for evaluation. Changes in analytical methods and their detection limit, as well as changes in sampling protocol, are of major concern when evaluating long-term data, regardless of the reputation of an agency.

First Tier. The QAPP integrates all technical and quality aspects of a project, including planning, implementation, and assessment. The USEPA requires a QAPP to include certain elements (USEPA, 2001d). These elements are arranged into the following categories (see Tables 4.2 – 4.5 for a listing of individual elements):

A. Project Management: The elements in this group address the basic area of project management, including the project history and objectives, roles and responsibilities of the participants, etc. These elements ensure that the project has a defined goal, that the participants understand the goal and the approach to be used, and that the planning outputs have been documented.

Table 4.2. Project Management Elements

ID	*Element name*	*Evaluating*
A1	Title and Approval Sheet	Presence
A2	Table of Contents	Presence
A3	Distribution List	Presence
A4	Project/Task Organization	Presence
A5	Problem Definition/Background	Presence
A6	Project/Task Description	Presence
A7	Quality Objectives and Criteria	Presence
A8	Special Training/Certification	Presence
A9	Documents and Records	Presence

Table 4.3. Data Generation and Acquisition Elements

ID	*Element name*	*Evaluating*
B1	Sampling Process Design (Experimental Design)	Presence
B2	Sampling Methods	Acceptability
B3	Sample Handling and Custody	Acceptability
B4	Analytical Methods	Acceptability
B5	Quality Control	Acceptability
B6	Instrument/Equipment Testing, Inspection, and Maintenance	Acceptability
B7	Instrument/Equipment Calibration and Frequency	Acceptability
B8	Inspection/Acceptance of Supplies and Consumables	Presence
B9	Non-direct Measurements	Presence
B10	Data Management	Presence

Table 4.4. Assessment and Oversight Elements

ID	*Element name*	*Evaluating*
C1	Assessments and Response Actions	Presence
C2	Reports to Management	Presence

Table 4.5. Data Validation and Usability Elements

ID	*Element name*	*Evaluating*
D1	Data Review, Verification, and Validation	Acceptability
D2	Verification and Validation Methods	Presence
D3	Reconciliation with User Requirements	Presence

B. Data Generation and Acquisition: The elements in this group address all aspects of project design and implementation. Implementation of these elements ensures that appropriate methods for sampling, measurement, and analysis, data collection or generation, data handling, and quality control (QC) activities are employed and are properly documented.

C. Assessment and Oversight: The elements in this group address the activities for assessing the effectiveness of the implementation of the project and associated quality assurance (QA) and QC activities. The purpose of this assessment is to ensure that the QA Project Plan is implemented as prescribed.

D. Data Validation and Usability: The elements in this group address the QA activities that occur after the data collection or generation phase of the project is completed. Implementation of these elements ensures that the data conform to the specific criteria, thus achieving the project objectives.

First, a score is assigned based on a level of compliance with the USEPA document (Tables 4.2 – 4.5). When available, the QAPP is searched for all required elements. Elements are evaluated either based on their acceptability or based on their mere presence, depending on the importance of the particular element. For example, the description of *Problem Definition/Background* is sufficient to satisfy the requirement, and it would receive a score for presence. On the other hand, *Sampling Method* needs to be up-to-date to receive the high score. The basis for evaluating a QAPP element is included in Tables 4.2 – 4.5. The QAPP element receives the highest score if it corresponds in quality to the IEPA requirements (IEPA, 1994; see also Schumacher and Conkling, 1991).

If a QAPP document is not available, sampling procedures and analytical methods used in the monitoring program are investigated. A score is assigned to each of the selected QAPP elements listed in Table 4.6. A maximum score of 40 can be assigned based on the QAPP elements (QAPP score).

As acceptability of some rating factors varies for different parameters, a project may be evaluated several times if necessary. For example, if sample handling methods are up-to-date for basic inorganic analysis but unacceptable for dissolved trace metals, the relevant QAPP elements will be evaluated twice as they pertain to specified parameter groups. This prevents mislabeling acceptable data and warns about quality of specific parameters measured within a project.

An additional score assigned for selected elements evaluates the comparability with present methods. Several factors describing the data usability and its comparability between different sources are included in addition to the QAPP elements. Table 4.7 shows the various elements inclusive in the rating. A maximum score of 16 can be assigned based on data comparability and usability (C/U score).

The QAPP and C/U scores are rated individually (Table 4.8). The final grade assigned to a project and a parameter group reflects the acceptability of data as compared to present expectations set by the IEPA. A grade of zero represents data of acceptable quality.

Table 4.6. Tier 1 Rating Factors: Evaluating a QAPP

QAPP	*Rating factors*	*Possible values*	*Score*
Available	Presence/acceptability of individual components	Present	1
		Not present	0
	(Tables 1-4)	Up-to-date (IEPA, 1994)	3
		More lenient but acceptable	2
		Unspecified & unacceptable	0
	Approval	IEPA Approved	2
		Internal documents	1
		Nonexistent & unknown	0
Not available	Training and certification	Trained sampling crew	6 or 0
		Certified laboratory	6 or 0
	Documents and records	Required and available	4
		Required but not available	2
		Not required & unknown	0
	Sampling methods	Up-to-date (IEPA, 1994)	6
		More lenient but acceptable	4
		Unspecified & unacceptable	0
	Sample handling and custody	Up-to-date (IEPA, 1994)	6
		More lenient but acceptable	4
		Unspecified & Unacceptable	0
	Analytical method	Standard methods (approved by the USEPA)	6
		Non-standard	2
		Unknown	0
	Quality Control	Up-to-date (IEPA, 1994)	6
		More lenient but acceptable	4
		Unspecified & unacceptable	0

Table 4.7. Tier 1 Rating Factors: Evaluating Data Comparability and Usability, C/U Score

QAPP	*Rating factors*	*Possible values*	*Score*
Available/not available	Sampling frequency	Continuous	6
		At least biweekly	4
		At least monthly	2
		Less than monthly	0
	Sampling period	Long term	6
		Year	4
		Season	2
		Less than a month	0
	Sampling method	2-D composite	4
		1-D composite flow weighted	3
		1-D composite regular grid	2
		Grab or unknown	0

Table 4.8. Tier 1 Evaluation Scale

Class	*Min QAPP score*	*Min C/U score*	*Data rating*
Excellent	32	14	2
Good	27	10	1
Acceptable	22	6	0
Poor	17	4	-1
Very Poor	0	0	-2
No Information	NI	NI	

Second Tier. Possible outliers in data were identified using statistical methods. Data from individual projects are first evaluated separately for consistency within an individual sampling site. Statistical evaluation of individual datasets used the following techniques:

1. Basic statistics (mean, median, and standard deviation)
2. Probabilistic distribution plot, quantile plot, test for normal or log-normal distributions
3. Time-series plots
4. Scatter plots (change of parameter with flow etc.)
5. Statistical tests for suspected outliers

Data reported as “below detection limit” or “nondetects” were treated according to the USEPA recommendation (USEPA, 2000c). For statistical purposes, data were separated into three categories depending on percent nondetects: (1) less than 15 percent data, (2) between 15 and 50 percent data, and (3) more than 50 percent data. The proportion of nondetects above 15 percent affects the distribution fitting and special procedures need to be applied. Some statistical characteristics cannot be properly estimated when more than 50 percent of the data are reported nondetects.

When a sample result is suspected to be an outlier, additional data are analyzed to seek possible explanations for the unusual value. This may include, but is not limited to, preceding flow and rainfall data, relevant chemical constituents (pH, temperature, suspended solids, etc.), and available biological data (fish and macroinvertebrates). If this fails to provide a reasonable explanation, additional effort is used to gather information from the original data source, such as the field and laboratory reports.

Outliers were treated on a case-by-case basis. Outliers associated with typographical or measurement errors were marked as *identified outliers*, and every effort has been made to correct the result values in the FoxDB. Measurements identified by statistical procedures as outliers were marked as *suspected outliers* when additional data do not provide an explanation of the problem. The data analysis and all outliers found were properly documented and flagged in the FoxDB. Water quality analysis should be carried out both with and without outliers for comparison purposes.

Usability of datasets may be greatly enhanced by combining data collected from identical locations and matching time periods but for different projects, provided the data quality (QAPP score and Sampling Method under C/U score) determined in tier 1 is comparable. Such data were evaluated for consistency between datasets to verify whether these datasets may be merged. The datasets were compared using the following techniques:

1. Quantile-quantile (q-q) plots
2. Two-sample tests for population means
3. Two-sample tests for population distribution

After merging the data, the C/U score can be recalculated to reflect a change in sampling frequency of combined dataset.