DATA AUTOMATION TO MEET PERMIT SCHEDULES AND QUALITY OBJECTIVES FOR THE WATER QUALITY MONITORING OF SECONDARY SEWAGE EFFLUENT IN MASSACHUSETTS BAY

Ellen Baptiste-Carpenter[1]
Andrew M. Parrella[1]
Scott A. Sauchuk[1]
John Hennessy[2]

Andrew M. Parrella, *Researcher*
Mr. Parrella is a researcher in Battelle's Environmental Data Systems group. He has a BS in Biochemistry from SUNY Binghamton and an MS in Marine Environmental Science from SUNY Stony Brook. He was the laboratory manager and senior technician for a tropical marine phytoplankton laboratory at SUNY Stony Brook's Marine Sciences Research Center. Since joining Battelle, he has worked on several large database projects and now maintains Battelle's copy of EM&MS for the MWRA Harbor Outfall Monitoring project.

Abstract

The Massachusetts Water Resources Authority (MWRA) is undergoing a transition from managing a baseline water quality monitoring program to managing compliance with their recently issued discharge permit. One impact of this change is the need to greatly accelerate the throughput of monitoring data to enable the identification of environmental perturbations in near-real time. For the past year, Battelle's automated data flow process has satisfied this need. Sample collection data are transferred electronically from the field to the central database. Key elements of these data are supplied to the laboratories in data entry/loading applications. Quality control checks built into the applications allow error detection and resolution to occur upstream where the problems are introduced and the answers lie. Laboratory data, returned in the same entry/loading applications, are entered into the central database in one simple step. Values are automatically checked for violation of environmental thresholds during loading. Once in the database, data are available to the client and to principal investigators involved with the monitoring program through direct database access and Web access. This approach has cut data delivery times by more than half while reducing costs.

[1] Battelle
397 Washington Street
Duxbury, MA 02332

[2] Resource Data, Inc.
1205 E. International Airport Road
Anchorage, AK 99518

BACKGROUND

Since the late 1980s, the Massachusetts Water Resources Authority (MWRA) has collected tens of thousands of environmental samples and taken many millions of *in situ* measurements in Massachusetts Bay and Boston Harbor. The purpose of all this sampling is to establish baseline environmental conditions prior to the startup of a new offshore sewage treatment outfall. Once the outfall is operational, secondary effluent that was previously discharged into Boston Harbor and nearshore environments of Massachusetts Bay will instead be discharged 9.5 miles offshore. The decision to transfer the discharge offshore has created intense public concern among those who believed the effluent would be transported to Cape Cod Bay and towards other coastal communities. There was also great concern about the impact on the marine mammal sanctuary at Stellwagen Bank. Because of these concerns, the discharge from the new outfall pipe will be monitored under an NPDES permit that, at time of issuance, was the most comprehensive permit in the United States. The permit has strict requirements to report suspected problems. The goal for reporting possible violations is 90 days after sample collection, with a final deadline of 150 days. Thus MWRA needs to access and analyze results well in advance of these deadlines.

Throughout the baseline monitoring program, the data generated from sample analysis and *in situ* measurements were loaded into a central database between one month and three years after collection. Clearly, this had to change if permit conditions were to be met. Beginning in 1996, Battelle began to map data flow from sample planning through collection, analysis, and data entry. Battelle uncovered several bottlenecks in the data processing stream including:

- Mismatched database codes between MWRA and its contractors
- Incomplete or missing sample collection data
- Incomplete or improperly formatted data deliverables from laboratories
- Inefficient and inconsistent methods for loading and checking the database
- Poor database performance hampering data analysis and report-writing efforts
- Database access limited to data management staff

This data flow map was used to identify opportunities for automation that would yield shorter data turn-around times. The implementation of our data management methods led to a dramatic decrease in data delivery time within one year. This paper presents the technical details behind Battelle’s system.

SYSTEM INFRASTRUCTURE

During the multi-year baseline monitoring period, MWRA worked with various contractors to develop the Environmental Monitoring and Management Information System (EM&MS). The EM&MS is based on an Oracle RDBMS and ARC/INFO Geographic Information System (GIS). The EM&MS is a mature system with about 36 data and reference tables and over 350 referential and integrity constraints. As the prime monitoring contractor to MWRA, Battelle is required to report data as Oracle exports that can be readily imported into EM&MS.

Battelle maintains a copy of EM&MS on our Oracle database server with all constraints enabled. We are responsible for populating the database with new monitoring results while maintaining the baseline data. Concurrently, MWRA loads data from other in-house monitoring programs into their local copy of EM&MS. Experience has shown us that under such circumstances, maintenance of a common code list can be very difficult. Battelle solved this problem by having only one master code list table, maintained by MWRA on their copy of EM&MS. A simple and automated procedure was developed that allows MWRA to export a new code list to Battelle nightly via the Internet. This allows the two databases to stay synchronized and provides Battelle with the latest codes daily. Since the inception of this procedure, several other reference tables have been included in this nightly update. These tables are also maintained by MWRA only.

In addition to providing Battelle with the EM&MS system definition, MWRA provided their data check scripts. These scripts, along with others generated by Battelle, are run against every data set prior to export. This ensures that MWRA receives only data that meet all their quality control objectives.

Battelle drew from a broad range of experience in managing environmental monitoring data to further expand EM&MS functionality (Figure 1). Sample collection data are recorded electronically in the field and transferred to EM&MS. Data loading applications are pre-populated with sample IDs and planned analysis information directly from EM&MS. These applications provide the laboratories with data entry capabilities that eliminate most common errors and shift error detection and resolution back to the laboratory where the problems were introduced and the answers lie. Along with improved data loading capabilities, Battelle created expanded data reporting and access. Principal investigators and MWRA can access current data and project tracking information over the Internet. The production of standard reports and graphs for data and synthesis reports is automated. These tools are discussed in more detail below.

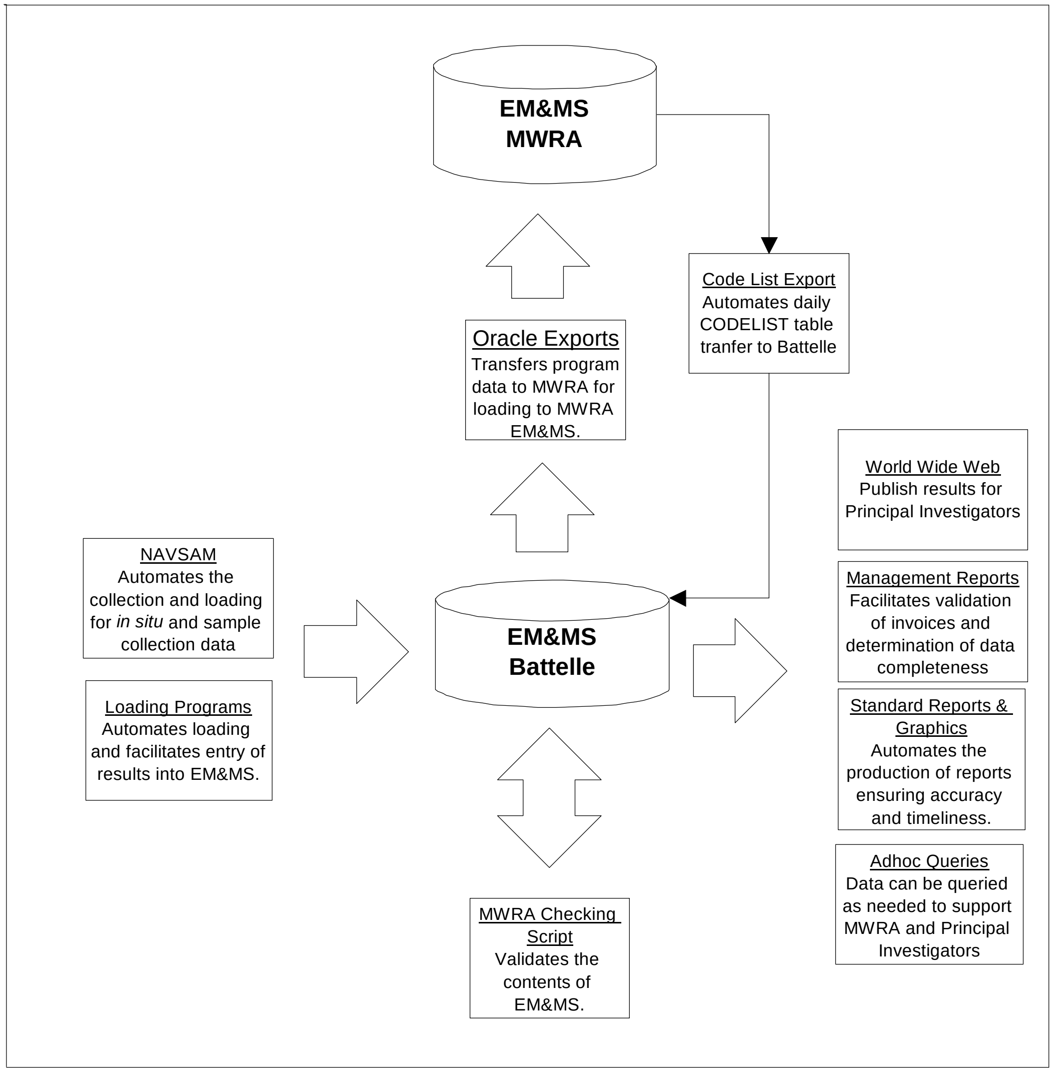
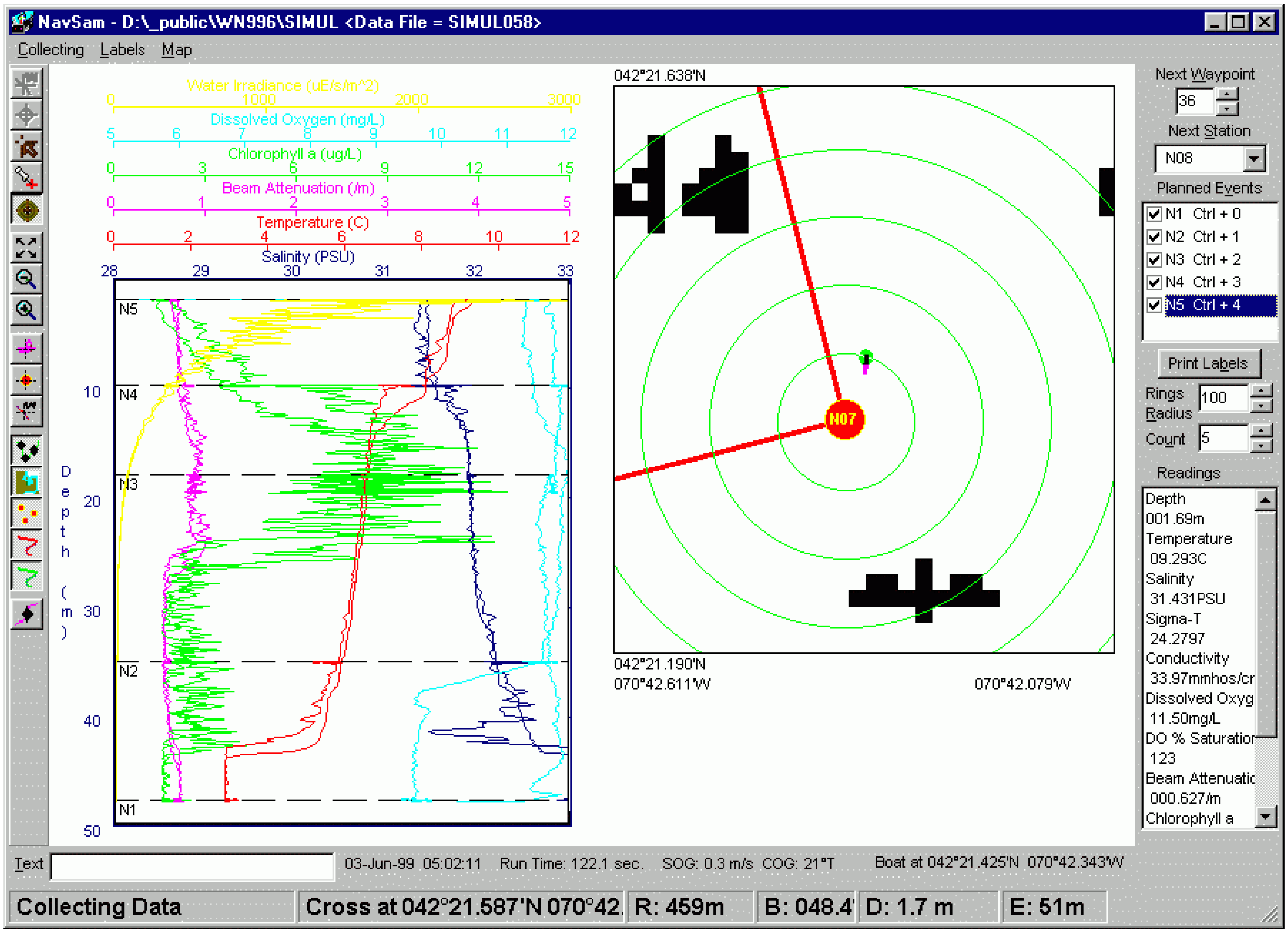

EM&MS
MWRA
Code List Export
Automates daily CODELIST table tranfer to Battelle
Oracle Exports
Transfers program data to MWRA for loading to MWRA EM&MS.
World Wide Web
Publish results for Principal Investigators
NAVSAM
Automates the collection and loading for *in situ* and sample collection data
EM&MS
Battelle
Management Reports
Facilitates validation of invoices and determination of data completeness
Loading Programs
Automates loading and facilitates entry of results into EM&MS.
Standard Reports & Graphics
Automates the production of reports ensuring accuracy and timeliness.
Adhoc Queries
Data can be queried as needed to support MWRA and Principal Investigators
MWRA Checking Script
Validates the contents of EM&MS.

Figure 1. Data flow diagram.

FIELD DATA COLLECTION

Collection and tracking of field data can be a complex process, with different stations requiring different types, sizes, and numbers of samples collected for various parameters. To simplify this process, Battelle created a Windows™-based software package called NavSam (See Figure 2). Working with digitized NOAA charts, NavSam allows the chief scientist to plan each part of the survey, from station selection and course plotting to the number and types of samples to be taken at each station. When the vessel arrives at a station, NavSam displays the list of required samples, preventing the field crew from missing a sample or taking the wrong type of sample. As each sample is taken, whether a Niskin bottle firing, a plankton net tow, or anything else, NavSam records navigational and environmental information. When a sampling rosette is being used, the *in situ* data from the CTD are also integrated into the record for that sample. NavSam generates unique sample identifier codes as samples are taken, with the codes based on the survey ID and a unique sample marker.



NavSam - D:_public\WN996\SIMUL <Data File = SIMUL058>

Next Waypoint: 36

Next Station: N08

Planned Events:

- ☑ N1 Ctrl + 0
- ☑ N2 Ctrl + 1
- ☑ N3 Ctrl + 2
- ☑ N4 Ctrl + 3
- ☑ N5 Ctrl + 4

Rings Radius: 100

Count: 5

Readings:

Depth
001.69m
Temperature
09.293C
Salinity
31.431PSU
Sigma-T
24.2797
Conductivity
33.97mmhos/cr
Dissolved Oxyg
11.50mg/L
DO % Saturatior
123
Beam Attenuatic
000.627/m
Chlorophyll a

03-Jun-99 05:02:11 Run Time: 122.1 sec. SOG: 0.3 m/s COG: 21°T Boat at 042°21.425'N 070°42.343'W

Collecting Data | Cross at 042°21.587'N 070°42. | R: 459m | B: 048.4' | D: 1.7 m | E: 51m

Figure 2. NavSam in transit to the next station. CTD instrument traces are shown on the left; ship's position relative to station shown at right.

Once the samples are on board, the software prints adhesive labels containing the unique sample IDs and a bar code for each sample container. The automatic generation and printing of sample IDs eliminates transcription errors because sample IDs need never be written down. In the event that a sample label is damaged, another can be printed, and the duplicate indicates that a copy was made. NavSam also allows the crew to take unplanned samples and to record environmental observations such as sightings of whales or surface slicks. At the end of the survey, chain-of-custody forms are printed facilitating sample tracking from the field through multiple subcontractor laboratories. The sample collection data are output as a Microsoft Access database and delivered to data management on a single floppy disk.

FIELD DATA LOADING

Field data loading is the final stage of managing sample integrity. It is critical to incorporate the field results quickly and simply. The improved process enables sample collection data to be loaded into the central database automatically. At the click of a button, a Microsoft Access application populates the five upper-level tables in the database (Event, Station, Sample, Bottle, Sample Depth Class), setting the stage for the subsequent insertion of analytical data. Integrity constraints on these tables ensure data validity, including verification of the time zone. Within minutes a table of samples collected can be produced for inclusion in a survey report. This automatic system of loading reduces costs by virtually eliminating human error and allowing the task to be leveraged down to junior staff.

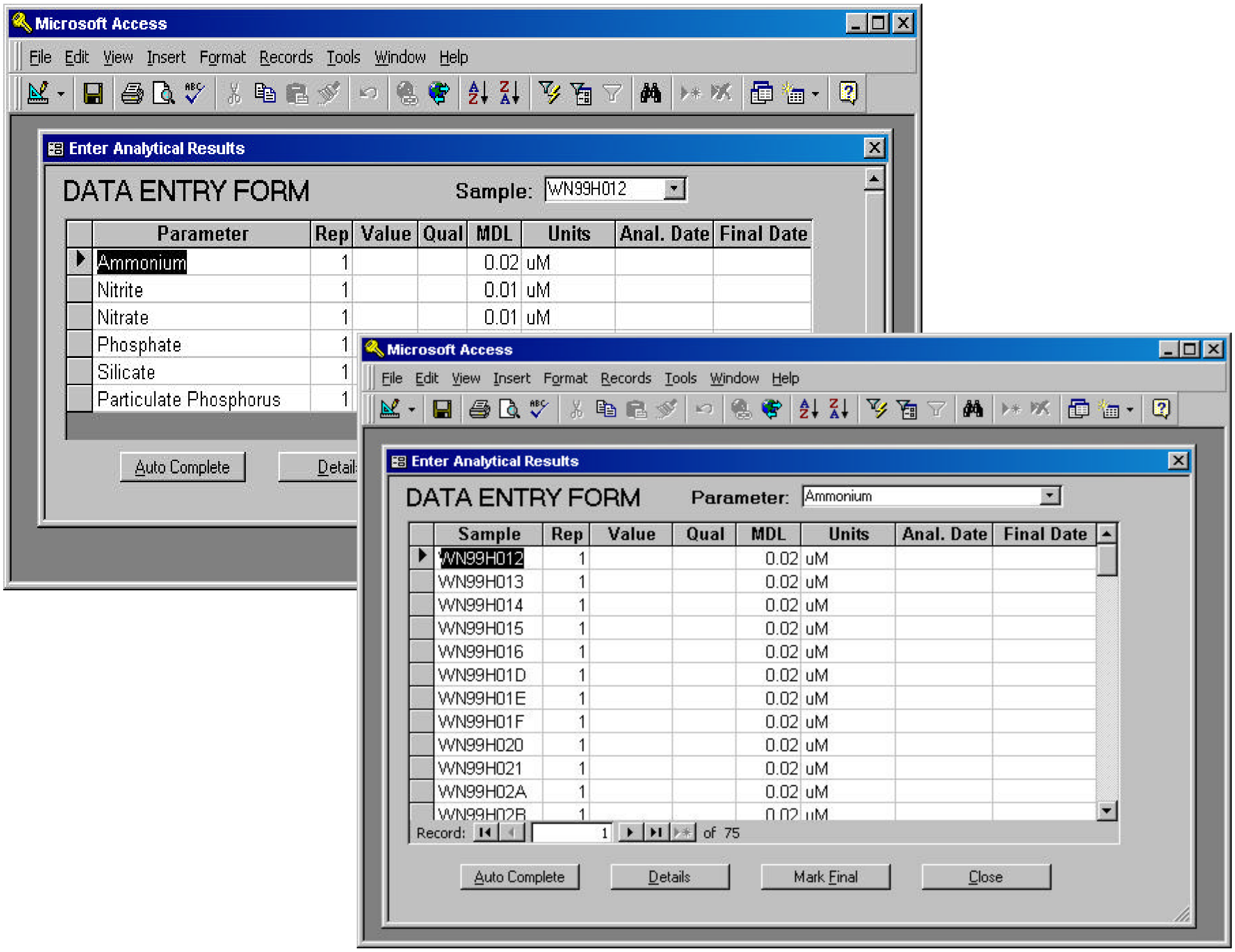
LABORATORY DATA COLLECTION

The complexity of the MWRA monitoring program necessitates the use of a variety of subcontractor laboratories with different degrees of data automation. Battelle has developed two separate methods for acquiring laboratory data. The first method was designed for laboratories lacking a Laboratory Information Management System (LIMS). For these laboratories, Battelle created custom data loading applications, configured to record data generated using the analytical method specified by MWRA. Various data checks were built into the application (*e.g.*, missing values, valid qualifiers). Following a survey these templates are linked to the new sample collection data, populating a data loading application with only the sample IDs needed for that analysis and laboratory.

By pre-populating the loading application, we eliminated several sources of error. In previous projects, we often received subcontractor data with simple, yet costly, errors such as data with sample IDs that did not match any of the samples taken. Both the data management and the laboratory staff would have to stop and trace these IDs back to their true sample ID. Sometimes the laboratory would omit the results for a certain analysis for some samples, and this oversight would not be caught until report production. The new loading applications make it very clear which results are expected for each sample. The populated applications are sent to the outside laboratories via email, FTP, or overnight carrier. Laboratory personnel can enter their data by sample ID or by parameter (See Figure 3), while the application ensures the proper use of parameter, species, and qualifier codes.

To help identify errors in real-time, a variety of quality control reports are available to the laboratory. These reports include a data report, exception report, deliverable list, and a sample summary. The data report provides the laboratory with a hardcopy data report to check data entry and to submit with the electronic data submission. Battelle requires the laboratory to run the data exception report showing any new species added, missing or suspect results, and other useful information. This allows error detection and resolution to be performed at the laboratory where the solutions are more readily available. An annotated hardcopy of this report is required with each data submission. The deliverable report provides the laboratory with a checklist to ensure a complete submission to Battelle. The sample summary report provides a list of analyses used to verify laboratory invoices.

The second method of data acquisition was developed for laboratories employing a LIMS. This method is very flexible and requires little change for most laboratories. The laboratory submits a spreadsheet in standard cross-tabular format that includes all the fields required by EM&MS (*e.g.*, sample id, analysis method, and units). A Battelle generated Microsoft Access application, named Flipper, is used to prepare these spreadsheets for loading. Flipper parses the spreadsheet and produces normalized data tables containing all the necessary fields for automated loading into EM&MS. The application is extremely flexible and can normalize a wide variety of spreadsheet formats amenable to each laboratory's automated system. Since this method is generic to allow it to be applied to a variety of data types and laboratories, it lacks most of the data checks that were built into the custom data loading applications. These checks must be run separately once the data have been imported into the working Oracle database.


Microsoft Access
File Edit View Insert Format Records Tools Window Help
Enter Analytical Results
Auto Complete
Detail

DATA ENTRY FORM Sample: WN99H012

Parameter	Rep	Value	Qual	MDL	Units	Anal. Date	Final Date
Ammonium	1			0.02	uM		
Nitrite	1			0.01	uM		
Nitrate	1			0.01	uM		
Phosphate	1						
Silicate	1						
Particulate Phosphorus	1						

Microsoft Access
File Edit View Insert Format Records Tools Window Help
Enter Analytical Results

DATA ENTRY FORM Parameter: Ammonium

Sample	Rep	Value	Qual	MDL	Units	Anal. Date	Final Date
WN99H012	1			0.02	uM		
WN99H013	1			0.02	uM		
WN99H014	1			0.02	uM		
WN99H015	1			0.02	uM		
WN99H016	1			0.02	uM		
WN99H01D	1			0.02	uM		
WN99H01E	1			0.02	uM		
WN99H01F	1			0.02	uM		
WN99H020	1			0.02	uM		
WN99H021	1			0.02	uM		
WN99H02A	1			0.02	uM		
WN99H02B	1			0.02	uM		

Record: 1 of 75

Figure 3. Nutrient data loading application showing data entry by parameter and by sample.

LABORATORY DATA LOADING

The system for loading analytical data into the central database is also automated. What once took hours or days to load now takes minutes. Upon receipt of the analytical data, data management personnel log the data into a separate login database. Each data set is assigned a unique login ID that allows data management staff to maintain chain-of-custody for each result which began in the field and continues through electronic handling. The data loading applications and the Flipper application have a button to transfer the data to a working Oracle schema after prompting for the login ID. This ID is loaded with the data. A single mouse click loads the data into a working schema where all the business rules (integrity constraints) of the final database are enforced. The MWRA quality control check script is run against the data prior to transferring the data into the production schema via a standard transfer script. Once in the production tables, data are immediately available to program principal investigators and to MWRA via the Internet. All these automated data management systems reduce costs because junior staff can perform them in a few minutes.

The MWRA is required to report exceedance of certain predetermined environmental threshold conditions, such as dissolved oxygen falling below a certain level. Battelle’s data loading system automatically performs these threshold checks as data are loaded, allowing critical environmental parameters to be reported within hours of the receipt of final data. Critical data, such as dissolved oxygen, can be fully validated and integrated into the database in less than two weeks. Each time a threshold check is run, key information is stored on line in a threshold-tracking table. These data make it possible for MWRA to create summaries of their compliance history.

Our deliverables to MWRA are data exports from Oracle. The diversity of data being reported to MWRA necessitates an orderly manner for submitting incremental monitoring results. Battelle created a script that prompts the user to identify the desired data, then recreates the pertinent database tables with only the identified data in an export schema. The MWRA check script and an internal Battelle check script are run against the data in the export schema. Use of the export schema limits the check scripts to examine only the data to be delivered, which allows the script to run much more quickly and with fewer confusing issues from other data sets.

DATABASE PERFORMANCE TUNING

With several million records occupying a gigabyte of storage space, EM&MS is a small data warehouse. After appropriate indexing, SQL optimization, and regular recalculation of optimizer statistics, the Oracle server still requires considerable time to process complex queries. Many of these queries involve summations and calculation of derived parameters, as well as compilation of information from many large tables.

In order to reduce the time needed for data queries, Battelle uses summary tables to pre-assemble the information needed by the scientists interpreting the data and writing reports. Summary tables are refreshed nightly when server CPU time is typically more available. Thus, rather than joining, aggregating, calculating, and sorting the data with each data request, these operations are performed once nightly so that this information is more quickly available to the user. The drawback to this approach is that summary tables produce a snapshot of the data, which may be up to one day old. However, this slight delay is not a problem within the framework of this program. If necessary, the summary tables can be refreshed at any time to provide more recent data.

REPORTING TOOLS

Reporting tools interfacing with the EM&MS include software tools that produce tables and maps suitable for incorporation into reports. Several commercial products are used to generate data tables for reports and for use in interpreting the data. Typically, a software product needs to bridge the gap between the normalized mode of data storage of a relational database and an easily understood cross-tabular, or pivot, format common in many standard spreadsheet presentations. Reporting tools need to be able to pivot multiple repeating columns, such as value/qualifier pairs common to environmental chemistry data reports. Battelle uses a product called BrioQuery (Brio Technology) that has excellent pivot table capabilities and can report data from the standard database tables or from views containing summary data.

Because reports with standard sets of figures are an ongoing requirement of the current three-year monitoring program, Battelle has invested in automating the production of these figures to reduce long-term costs. Report-quality maps are produced with Battelle's Autographer software application, developed with Microsoft Visual Basic 6. Autographer derives data directly from views and summary tables in the EM&MS database to produce maps of station locations, vertical profiles, and posted concentrations of chemical and biological conditions. Contour maps are handled with an interface to Surfer (Golden Software). The final products are Windows Metafile (WMF) images that are suitable for incorporation into report documents.

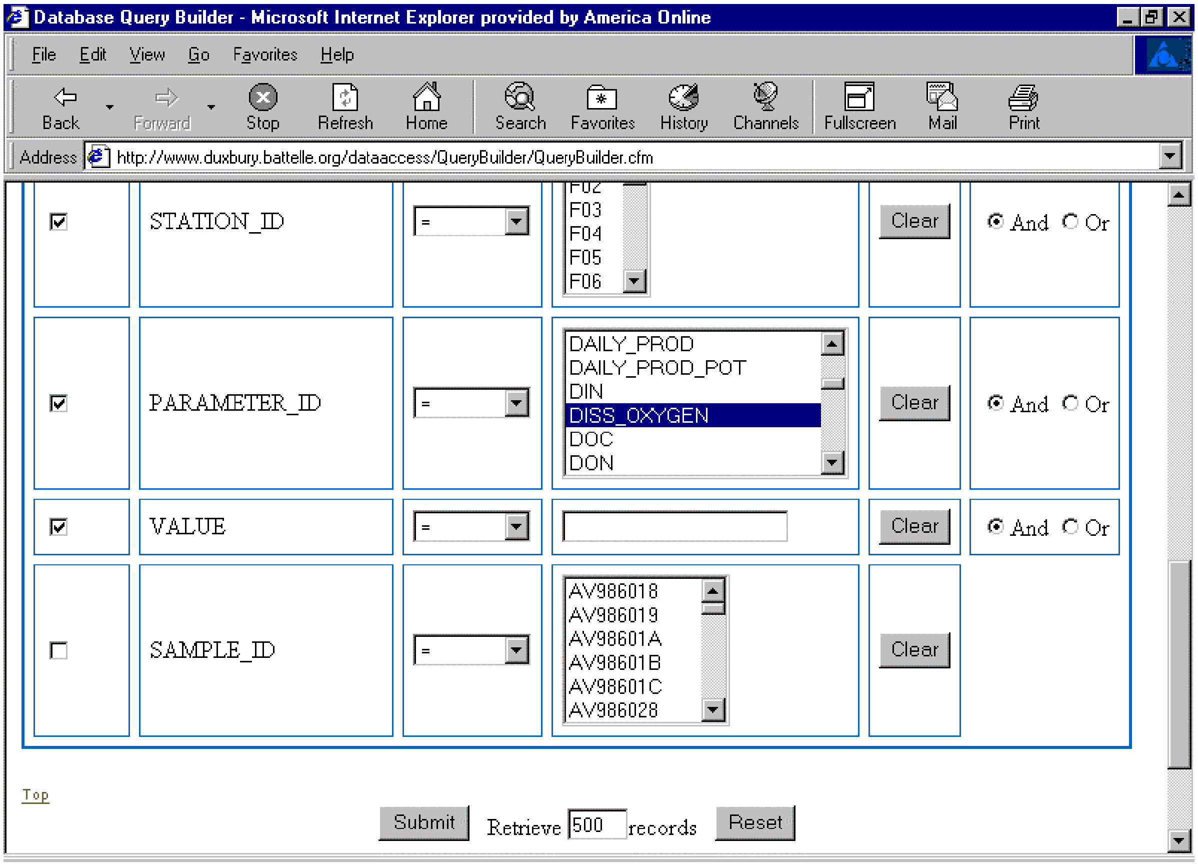
WEB-BASED ACCESS

A range of professionals, with different tools and skill levels, need access to the outfall monitoring results. The scientists at MWRA and Battelle have two methods of access to Battelle's version of EM&MS. The first method involves establishing a direct connection to the database through a dial-up connection to Battelle's remote access server (RAS) or through the Internet. This approach is useful for providing MWRA's database experts with real-time access to Battelle's version of EM&MS. However, the user must be familiar with the database structure and have Oracle client software installed locally. The second method is browser-based access to the database via Battelle's web server.

Battelle has set up a password-protected web site for database access by key clients, such as MWRA, and various scientists working for Battelle. Browser-based access presents the best available system architecture for general data access for the following reasons.

- No specialized client software is required (standard web browser and Internet connection).
- A wide audience is supported.
- Upgrades of web-applications are available immediately.
- User authentication and tiered access are supported.

Tabular data are presented through Battelle's generic web-based query builder. Battelle has set up a number of views in the database, and registered them to particular clients (*e.g.*, MWRA) and access levels (*e.g.*, manager and user). These views may either represent true Oracle views or summary tables presented in an easy-to- understand format requested by the user. Developed primarily with Cold Fusion (Allaire Corporation), the generic query builder reads the fields in the selected view and presents the user with a simple query interface (Figure 4).



http://www.duxbury.battelle.org/dataaccess/QueryBuilder/QueryBuilder.cfm

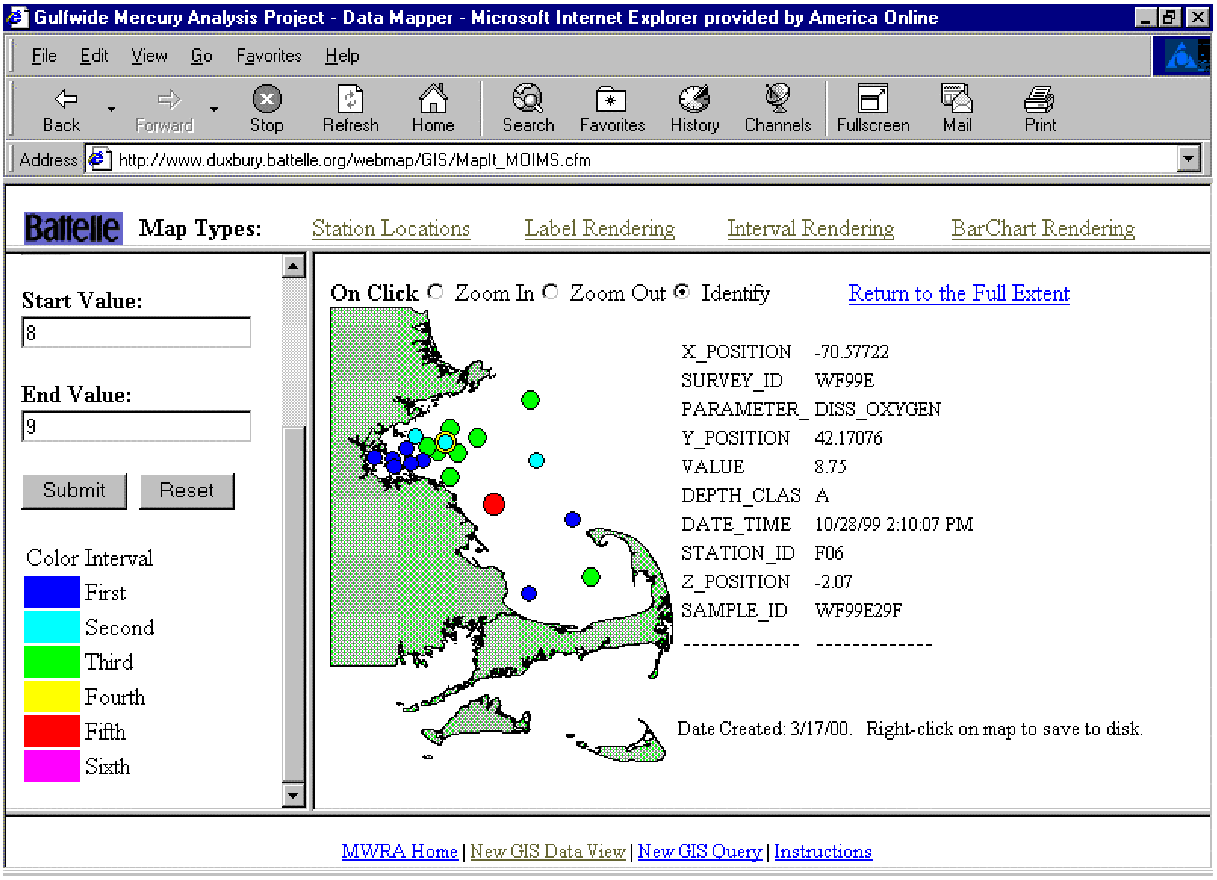
☑	STATION_ID	=	F02 F03 F04 F05 F06	Clear	● And ○ Or
☑	PARAMETER_ID	=	DAILY_PROD DAILY_PROD_POT DIN DISS_OXYGEN DOC DON	Clear	● And ○ Or
☑	VALUE	=		Clear	● And ○ Or
□	SAMPLE_ID	=	AV986018 AV986019 AV98601A AV98601B AV98601C AV986028	Clear	

Submit Retrieve 500 records Reset

Figure 4. Web-based query builder.

The user can select which fields to display, and qualify fields by selecting a comparison operator and one or more items from the distinct list of possible values. The tabular data may be downloaded to the user's computer for further processing. The generic nature of the query builder allows views to be added or modified quickly and easily by data management personnel with no web programming required.

Battelle has recently provided web-based mapping and GIS capabilities to the MWRA program. An interface similar to the generic query builder allows the user to select the data set of interest. A map of station locations is displayed by default. Stations may be labeled using any field in the view. Data may be represented graphically as symbol plot maps, where the size and/or color of the symbol are proportional to the numeric value, or as bar chart maps, where the multiple numeric parameters may be plotted together. The user may also zoom in or out and obtain information about a station by clicking on it. Figure 5 shows a symbol plot map of the concentration of dissolved oxygen in Massachusetts Bay, with the highest dissolved oxygen value identified and described by the text listing on the right side of the figure.



Address http://www.duxbury.battelle.org/webmap/GIS/MapIt_MOIMS.cfm

Start Value: 8

End Value: 9

On Click ○ Zoom In ○ Zoom Out ● Identify

```
X_POSITION    -70.57722
SURVEY_ID     WF99E
PARAMETER_    DISS_OXYGEN
Y_POSITION    42.17076
VALUE         8.75
DEPTH_CLAS    A
DATE_TIME     10/28/99 2:10:07 PM
STATION_ID    F06
Z_POSITION    -2.07
SAMPLE_ID     WF99E29F
------------- -------------
```

Date Created: 3/17/00. Right-click on map to save to disk.

Figure 5. Web-based mapping application.

CONCLUSION

Our goal of creating a system capable of supporting the MWRA's permit requirements has been achieved through the automation of the data management process. By examining the flow of data, we were able to eliminate the bottlenecks that had previously plagued this program. The automation also reduced the cost of data management and ensured a defined level of quality control. As a result, we can now use junior staff to perform this work and be assured that the results will be predictable.

Many of the bottlenecks found in the data flow review were addressed by our automation efforts. The nightly sharing of a central code list eliminates mismatched codes. The use of NavSam to plan and document sample collection ensures that the database has complete and accurate sampling information. The loading applications enable the analytical laboratories to check their data for common errors and transmit the data in the correct submission format. Performance and access have improved through the use of summary tables and Web-based queries.