Section B10. Data Management

One major challenge in conducting a statewide monitoring effort is development of a unified data system. For instance, in many cases the participating SWAMP organizations have previously developed data management systems of their own, or for their own specific objectives. These systems vary in the types of data captured, the software systems in which they are stored, and the degree of data documentation. In order to meet the SWAMP Program goal of centralized data management, a cooperative information management system is necessary to ensure that the collected data can be shared effectively among participants.

Information management needs to occur on several levels. First, a process must be developed to ensure the quality, compatibility, and timeliness of the data each organization collects. Once collected and organized, it must be available in as timely a manner as possible to the Regional Board SWAMP staff and others for review, analysis and ultimately for interpretation. Ultimately, one of the major goals of the use of this information, once interpreted, is to make it available to other interested organizations and the general public. The SWRCB SWAMP Program is also in the process of creating and maintaining an official website for the SWAMP Program, upon which such reports and data and other information would become available.

Appendix J to this SWAMP QAMP, the Interim SWAMP Information Management System (SIMS) Plan, documents and describes in detail the information management system (IMS) in that will support data capture and reporting during the initiation of SWAMP, although several elements are still being documented and finalized. The Interim SIMS Plan focuses on four major functions of the SIMS:

- The standard protocols each participating agency or laboratory will use to transfer data from their internal date generators to the SWAMP IMS.
- The process by which data will be submitted to the SWAMP data managers, including the path and quality control procedures the data will follow until it has been accepted.
- The technical specification (guidelines) of how the data will be organized in the SWAMP database.
- The milestones and mechanisms by which the data will be made accessible to project participants, other organizations, and the general public.

APPROACH TO INFORMATION MANAGEMENT

The Information Management System has several purposes, most importantly to provide a mechanism for sharing data among project participants. Data sharing is required if the SWAMP goal of producing an integrated hydrologic unit assessment of the State’s surface waters is to be achieved. While this is the primary focus, the IMS has been developed in recognition that SWAMP represents an initial effort toward data standardization among regions, agencies and laboratories and that protocols adopted here may be later used for other data sharing purposes beyond this project. Thus, the system was designed to be flexible to future adaptation. In addition, the system was constructed primarily to serve the RBS and technical committees, but it has also been designed to supply data to non-project scientists and the interested public.

The IMS will be based on a centralized data storage model (see Figure 6). A centralized system was selected because SWAMP is an integrated project and the typical data user will be interested in obtaining synoptic data sets from discrete hydrologic units or large geographical regions of the state. The centralized system was also selected over the alternative of a distributed system linked through a server or series of FTP sites because sophisticated tools would need to be developed and implemented for users to access those sites. There is also valid concern over the difficulty of maintaining a linked-distributed system for an extended number of years. Current budget allocations make the centralized system a more achievable model for handling data in the SWAMP program

The centralized database will be developed using standardized data transfer protocols (SDTP) for data exchange and Data Entering/Editing Forms for field data and observations. The SDTP details the information to be submitted with each sample collection or processing element, the units and allowable values for each parameter, and the order in which that information will be submitted. They are necessary to ensure that data submitted by the participants are comparable and easily merged without significant effort or assumptions by the organization responsible for maintaining the centralized data system. Use of SDTP allows each participating organization to retain data they generate in their local data management system while providing a mechanism for data exchange among project participants and a means for populating a centralized database.

The SWAMP database will be organized through a relational structure. The central database will be called the Replicate Master and will contain a temporary and permanent side, which are further described in the Data Flow Section below (and in Figure 6). The relational structure involves use of multiple data tables linked through one or more common fields or primary keys. A relational structure allows data created at different times (e.g. lab data vs. field data) to be entered at the time of data production, minimizing the possibility of data loss. This relational structure also minimizes redundant data entry, by allowing data that are recorded only once (e.g. station location) to be entered into separate tables rather than to be repeated in every data record.

The data table structure of this database was designed around a sample driven model. One distinct feature of this database captures a “nominal” position of the station (lat/long) which is stored in the stations table while still capturing a “actual” position of each sample. This is important because many different organizations will be occupying a station at different times to collect different samples. An example would be one group collects water samples, another group would deploy and retrieve bivalves, while yet another would collect stream bioassessment information at a station. This database structure was also designed with surface water sampling in mind, however, it is also built to capture information collected at multiple depths in the water column more commonly observed in marine and freshwater lake sampling systems.

It is imperative that station failures are documented in the sample table of the database to insure that a value is not missing from the database but was indeed documented as not being sampled. An example would be a station not being sample because it was "dry". This will be further described in the Master Table Structure in Appendix J of this QAMP.

ROLES AND RESPONSIBILITIES

SWAMP is a cooperative effort among eleven organizations (SWRCB, nine RWQCBs, CDFG) plus numerous additional subcontractor labs which have limited experience working together. Effective implementation of the SWAMP Information Management System Plan requires clearly defined roles for participants (Figure 6). For the purpose of defining roles, there will be four types of participants in SWAMP:

- Data generators - Field crew leaders (Key Data Entry) and laboratory supervisors who will be responsible for compiling data their organization generates and entering the data into the Data Entering/Editing Forms or the SDTP tables.
- SWAMP IM Coordinator (SIMC)- Responsible for working with Data Generators and leading the Data Management Team (DMT) at Moss Landing Marine Laboratories to develop SDTP, and for creation and management of the centralized SWAMP database.
- SWAMP Data QA Coordinator (SIMQA) - Responsible for overseeing quality assurance during migration of completed datasets to permanent data in the SWAMP database.
- SWIM IM Coordinator- Responsible for accepting data from SWAMP, placing it in the SWRCB SWIM database, and transferring it to other EPA databases, such as STORET.

Data Flow (still under development)

Official submission to the database can occur in two ways, 1) by form entry and 2) by batch loading (Figure 3) using SDTP. These data will reside in the temporary side of the Replicate Master. Data will be considered draft as it is loaded into the database (meets statewide comparability criteria) and compared with the task orders and found to be complete. Completeness checks will be accomplished with coordination between the SIMC and the Regional Board Staff requesting the work to be done as they have the best understanding of the study design. Data is considered complete when all results are entered in the database for a specific sample. Any SWAMP participants can have access to this draft data by requesting a replicate from the SIMC.

Once the draft data is certified complete, it will be transferred to the permanent side of the Replicate Master and sent to the SIMQA (QA Officer replica) for quality assurance checks. Once the data is validated by the SIMQA it will be considered final data.

Following certification of all portions of the data by the SIMQA, the SIMC will submit the integrated across-state data set to be stored in a Manger's replica of the SWAMP database. The SIMC will be the point of contact for data requests about the integrated data set. The SIMC will also be responsible for making the SWAMP data available to other data centralization functions such as the SWRCB SWIM database. The SWIM IM Coordinator (SWIMC) will be responsible for maintaining the current version of the SWAMP database within SWIM, and transferring it to other databases, such as STORET.

General Structure of Database

The SWAMP database currently contains 20 data tables (Figures 7 and 8). There are 10 entry level data tables and 10 permanent level data tables, both containing similar content. The main table is the Sample table, which includes a single data record for sample taken. Samples created can be 1) laboratory samples (lab generated), 2) analytical samples (field generated), 3) field observations or 4) field results. The Sample table includes all fields necessary to uniquely describe a sample. This sample is linked in a one:one or one:many relationship with all subsequent data tables. It is imperative that the *StationCode, SampleDate*, and *SampleTime* remain the same for all the field-generated samples, observations and results in order to link the information.

The combination of the fields *StationCode, EventType, SampleDate, SampleTime, SampleTypeCode, Duplicate, DepthSampleCollected, DistanceFromBank*, and *AgencyCode* will ensure that each record in the Sample table is unique. Sample records need to be linked with all results data and thus become the foundation of the database. The chemistry and toxicity results tables, all laboratory and analytical data are captured at the level of individual replicate, rather than in a summarized form. It is essential that the laboratories receiving samples be supplied with the information in this table for each sample.

Form Entry/Editing Protocols

Key Data Entry people (limited number per RWQCB) will enter field data into a replicate of the central SWAMP database data entry/editing forms provided to them by the DMT. Limited analytical data can also be entered through the form entry system. The DMT will provide training and support for use of these forms. The individual replicates will be synchronized with the central SWAMP database (Replicate Master). Recommended QA for form-entered data include double checking of data, or at minimum 20%, and range checks of the Field Results table. Data will next be submitted to the SIMC for synchronization to the Replicate Master and QA of data types.

Standardized Data Transfer Protocols

The data formats for the SDTP table submissions are detailed in App J, Chapter V, Section C (SWAMP Data Formats). These data formats include Lookup lists that are required to use in order for the data to be loaded into the database. The DMT will work with analytical labs on an individual basis to make this process as seamless as possible. Fields for summary quality assurance information are also included. A detailed laboratory QA report will be required and addressed in detail in the SWAMP QAMP.

Upon receipt, the DMT will update a data submission log to document the data received from each submitting organization. The DMT will then initiate a series of error checks to ensure the data: 1) are within specified ranges appropriate to each parameter measured, 2) contain all required fields, 3) have encoded valid values from constrained look-up lists where specified, and 4) are in correct format (text in text fields, values in numeric fields, etc.). If there are only a few, easily correctable errors, the DMT will make the changes. Changes will only be made with the consent of the data generator, with a list sent back to the data generator documenting the changes. If, there are numerous errors, or corrections difficult to implement, the DMT will send the data file back to the submitting organization with a list of necessary corrections. The submitting organization will make the corrections and resubmit the file to the DMT, who will subject the file to error checking once again. Each of these paths will be documented by the DMT as part of the submittal tracking process.

Data revisions (still under development)

Data can be revised in several ways depending on the stage of the data. When data is in the temporary side of the database, key data entry people will have the ability to revise data using the Data Entry/Editing Forms. When data is synchronized with the Replicate Master these edits will be committed to the database. It is important to note that the key data entry people or the DMT who make these edits bare the responsibility of making sure they are valid. Data deletions at this stage could have severe consequences to the database and should be used with care. Data being submitted using the SDTP can either be revised before or after it is submitted to the DMT. Once the data is transferred to the permanent side of the Replicate Master, only the DMT, Designated Regional Board Staff and SIMQA will be able to edit it.

Schedule

The schedule for data submission varies by data type. Data collected in the field will be due first, while data produced through extensive laboratory analysis will be produced on a schedule consistent with nominal laboratory processing times. Key data entry people should provide their data to the DMT so that there is sufficient time for the DMT to resolve any data discrepancies and to ensure the data are in the proper format for the addition of the batch input data.

Data Sheets

To assist organizations in meeting the data entry forms and improve the efficiency of data input, the DMT has created a series of data sheets. These sheets follow closely with the data entry forms, however data gatherers are not required to use them.

Figure 6: Flowchart of Statewide SWAMP Data Generators/Roles and Responsibilities

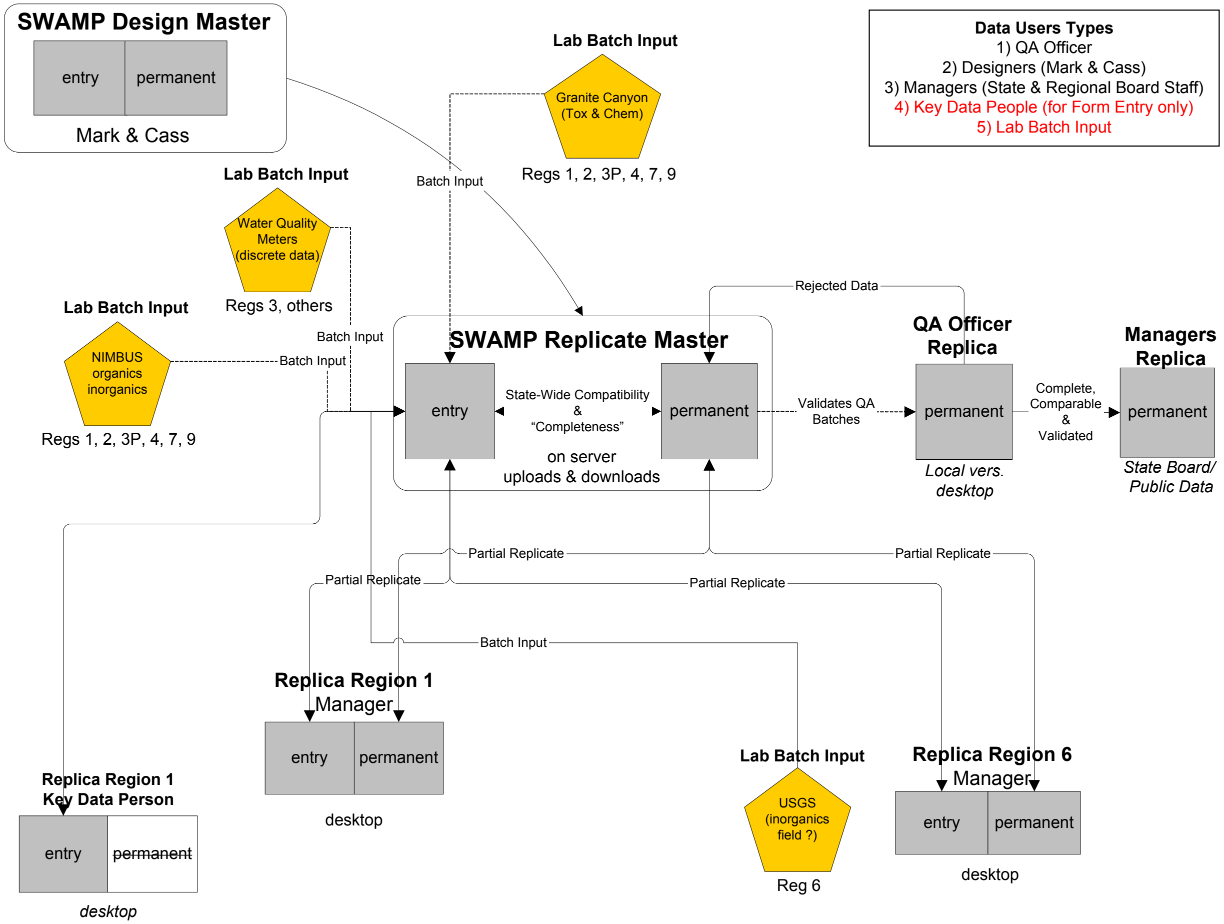
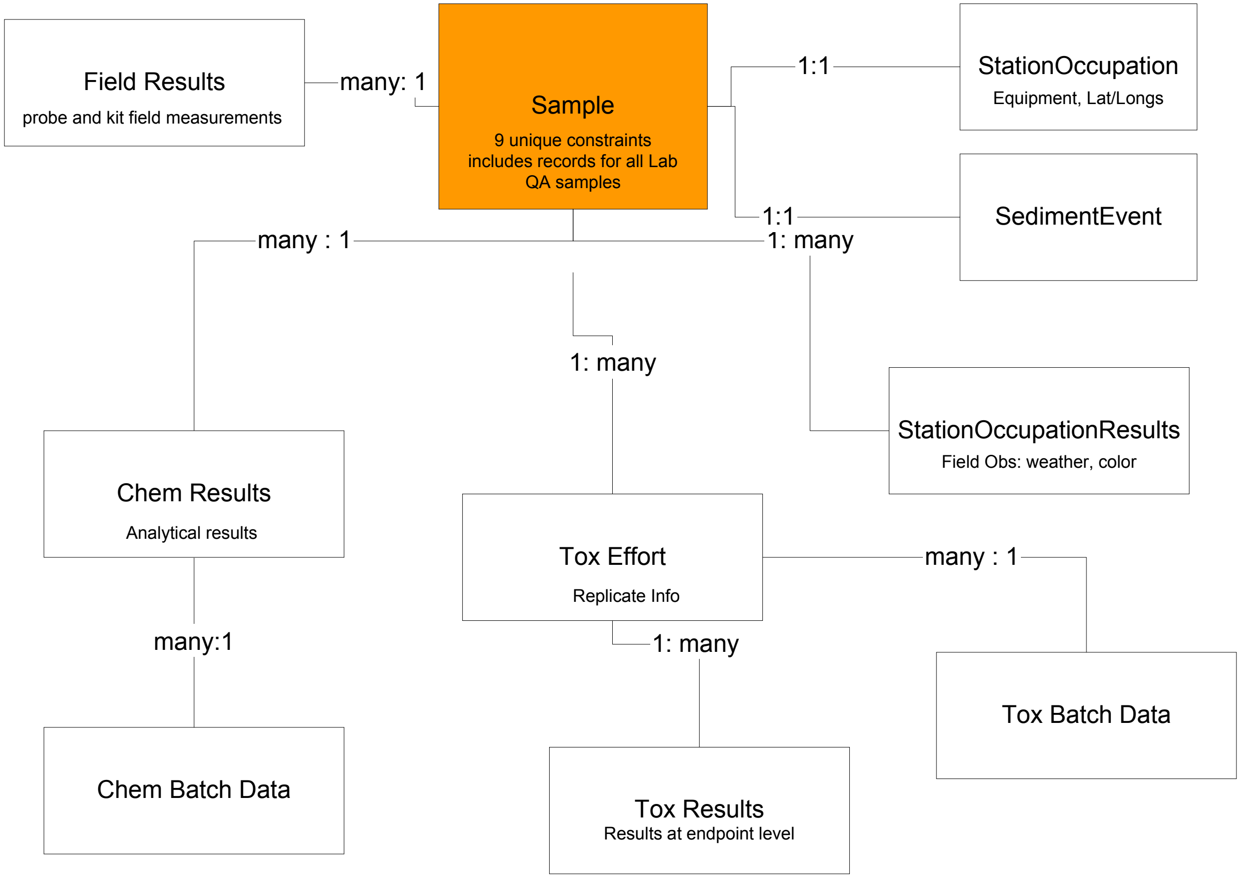
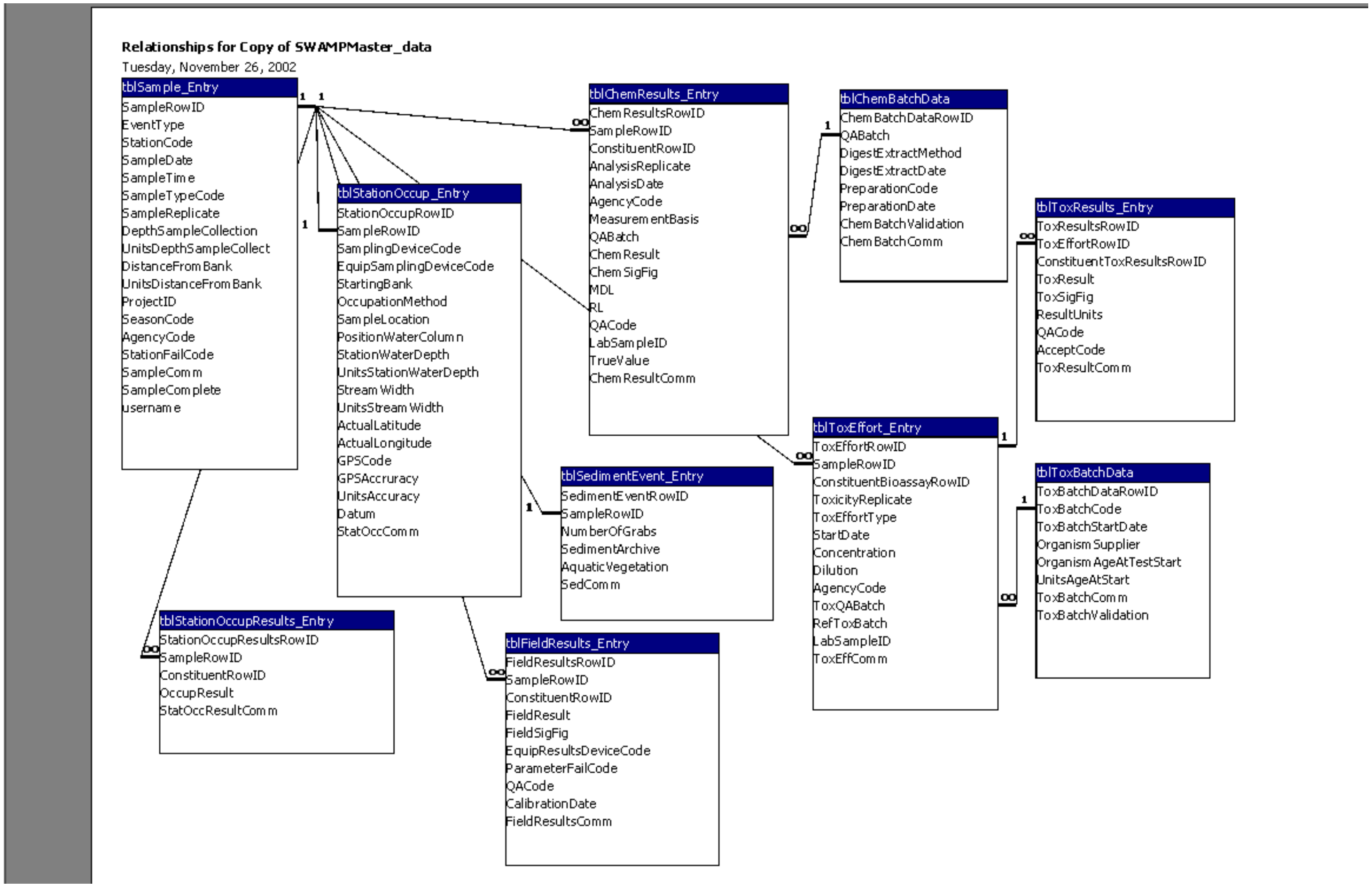


Figure 7: Outline of SWAMP Standardized Data Transfer Protocol Tables



September 18 2002

Figure 8: Contents of SWAMP Standardized Data Transfer Protocol Tables



DATA ACCESS (still under development)

All measurement and supporting data gathered during SWAMP will be made available to all participating organizations and to the general public, though the schedule of availability and point of contact will vary by user. The different schedules reflect the differing levels of quality assurance and data documentation that will have been completed at various stages in the project.

The first location of data availability will be the SIMC, who will be responsible for the SWAMP database generated within the state. The SIMC will be free to distribute SWAMP data collected within the state, at any point after the data has approved as complete by the SIMQA and submitted to the final SWAMP database. Data released prior to having been transmitted and accepted by the SIMC and SIMQA should be identified as DRAFT data, not SWAMP data, because SWAMP quality assurance procedures will not yet have been performed. If Draft data is released, all filenames will include the word "DRAFT". If hardcopies of Draft data are released, the pages must be stamped "Draft". It is highly recommended that data released prior to its submittal to the SIMC be limited to organizations directly participating in the SWAMP project, rather than to outside agencies or the general public. Releases to the general public are not recommended until quality assurance has been performed by the SIMQA and metadata documentation is completed.

Nodes (planned for future implementation, if funding allows)

The second location of data availability will be the SWIMC, who will be responsible for integrating SWAMP and other state program data sets into the SWIM database. These data sets may be made available through other centralized or distributed databases, as coordinated by the SWIMC. It is the responsibility of the SWIMC to obtain express permission of the individual Program Managers prior to distribution of their respective program's data outside of the SWIM database.

Metadata

Each release of data to the public will include comprehensive documentation about SWAMP and the accompanying data sets. Referred to as metadata, this documentation will include database table structures (including table relationships) and lookup tables used to populate the fields in each table. It will also include quality assurance classifications of the data and documentation of the methodologies by which the data were collected.

A second type of metadata will document changes made to the data over time. As the data are used, we anticipate that errors will be found. As changes to the data are made, they will be documented in a file organized by date and data table. Including this file with each data download will allow users to reconcile potential differences in analysis output that result from using different versions of the data.

Metadata will follow guidelines from the Federal Geographic Data Committee, Content standard for digital geospatial metadata, version 2.0. FGDC-STD-001-1998 (FGDC 1998), including the Biological Data Profile and the Biological Names and Taxonomy Data Standards developed by the National Biological Information Infrastructure (NBII 1999). For tabular data, metadata that meet the FGDC content standard are contained by a combination of the SWAMP Data Directory and the SWAMP Data Catalog. For Arc/Info coverages, the metadata are in the .DOC file embedded in the coverage. This file stays with the coverage. When the coverage is moved to a public web site, it will be duplicated to an ASCII text file.

Contact information for SWAMP information management:

SWAMP IM Coordinator (SIMC):	SWAMP Data QA Coordinator (SIMQA)
Cassandra Roberts *Moss Landing Marine Laboratories* *Marine Pollution Studies Lab* 7544 Sandholdt Road Moss Landing, CA 95039 Ph. 831 771-4163 Fax 831 633-0128 roberts@mlml.calstate.edu	Not filled at this time--anticipated to be hired within 2003 if SWRCB funding allows.
Data Management Team (DMT) Mark Pranger *Moss Landing Marine Laboratories* *Marine Pollution Studies Lab* 7544 Sandholdt Road Moss Landing, CA 95039 Ph. 831 771-4176 Fax 831 633-0128 pranger@mlml.calstate.edu	

For Future Implementation, if budget allows for development of nodes(these entities are currently providing extensive technical assistance into the database development):

Regional IM Coordinator (RIMC)- California Central Valley:
San Francisco Estuary Institute:Bruce Thompson
Department of Water Resources: Carl Jacobs

Regional IM Coordinator (RIMC)- - Southern California:
Larry Cooper
Southern California Coastal Water Research Project