DATA AND INFORMATION MANAGEMENT

Introduction

The Long Island Sound Monitoring Program (LISMP) is designed to optimally utilize existing local, regional and federal monitoring programs. A necessary consequence of this design is that data from a range of sources will have to be combined, analyzed, interpreted, and disseminated. While utilizing a combination of data already being collected by other monitoring agencies presents obvious economical advantages, the coordination, aggregation, QA/QC and archiving of these aggregated data will require specific economic investments. It is essential to pay careful attention to the data management/quality issues. Inappropriate treatment of the data may result, in the long range, in the waste of the large amounts of money invested in the monitoring program. This can happen if the data from the different agencies are not compatible, are of poor quality, are poorly documented, or are difficult to retrieve.

Data management functions are complex procedures just as integral to a project success as are field and laboratory operations. The design and implementation of the data management part of the LISMP must be conducted by qualified people who should be constituted as a steering committee. They will decide on the scope of the data management plan, and on the modalities of its execution. Consensus within the committee on the long-range strategy of the plan, as well as on the details of its implementation, is crucial for the LIS Monitoring Program to succeed. It also is necessary that each participating agency feel that it is benefiting from this alliance. The proposed plan stresses the need of the

construction of a Sound-wide, coordinated data management program. There are several reasons why this is necessary.

1. Similar QA/QC criteria need to be applied to data from all contributing agencies.

2. Similar data documentation criteria need to be established for all collecting agencies.

3. Back-up and security must be provided for all the data.

4. An archive for historical and orphan data must be ensured.

5. Assemblage/synthesis of analogous data collected by different agencies is necessary.

6. Easy retrieval of data collected by separate entities is necessary for generation of LISMP reports and other information products.

Background

Because Long Island Sound and the New York-New Jersey Harbor are interconnected systems, management of monitoring data from both systems must be coordinated. As a result, the Long Island Sound Study and the New York-New Jersey Harbor Estuary Program have entered into a voluntary agreement to coordinate data management efforts. Both

programs have adopted the EPA's Ocean Data Evaluation System (ODES) as a common repository of historical data and have combined resources to hire a data manager to ensure that data are organized and stored into ODES. Both programs have also agreed to coordinate long-term data and information management approaches.

Important Issues To Be Considered

Data management must be a high priority in the overall design of the composite monitoring program. Data quality, ease of use, and tools to aid in data integration are not features that can be added readily to the data flow at a later date. They must be built into the system as essential components of all data generation and handling activities. This can only be achieved if management, technical and analytical personnel understand the fundamental importance of data management and are committed to the development of an effective plan. The most effective data management plans thus incorporate information, priorities, and insights from all participants and give them an active role in the design and implementation of the plan.

Standardization of Entries

Comparing results or combining data from different studies always introduces problems of standardization. Conflicts and/or differences in such areas as taxonomy, nomenclature, spelling, data formats, measurement units, detection limits, and procedures for handling missing values all create problems when merging data from different programs.

These issues must be resolved in the early stages of the project by the scientists involved in the projects. Some facilities for coordinating the views of all participants should be established and a formal method for making decisions should be defined. In the proposed plan, this task will be defined at the inception of the program by the Steering Committee.

Data Documentation

Documentation must be available to make effective use of data. Just as with the data themselves, documentation has to be continuously maintained. Current, accurate documentation, including field and laboratory procedures and the description of the data management procedures, should be accessible to all users. Brief descriptions of the data sets should also be included with maps of the station locations, sampling frequencies, etc., the history of the sampling program itself, and the history of the data management, along with the identification of contact points for each set of data. Documentation should be viewed as an integral part of the data management plan and should be made accessible to all data users. In the proposed plan the documentation format will be defined by the Steering Committee at the inception of the plan, and its verification during the monitoring itself will be handled by a data manager.

Data Quality

The primary quality control issues are the prevention, detection and correction of errors. A basic principle is to detect errors as early in the data path as possible. The earlier errors are detected, the lower the

overall costs of the quality control effort. Error prevention requires establishing data handling procedures and data entry programs that can detect many errors at the time of input. It also requires training as a means of building quality into the process. It requires an in-depth understanding of how the data are collected and how they are manipulated. Generalized techniques do not work. Effective prevention requires working closely with the users and making them part of the solution, since errors are often introduced into the data as a result of how the users work with them. Errors often indicate flaws in procedures or lack of understanding about the characteristics of the data or the intended use of the data.

Error detection depends on systematic procedures, but it is also somewhat of an art which is developed with experience. Users of data, especially if they are not the original source of the data, often develop sophisticated techniques for detecting errors. These include comparing the data to the accompanying documentation, checking for internal consistency, range errors, duplicate records, missing data, and other more subtle errors. If any errors or questionable data are detected, the recipient then contacts the data source and works out the corrections. Large, complex data sets can present special challenges because error checking requires examining the data from a variety of perspectives. Errors or discrepancies are often interconnected so that one set of errors only becomes visible after another set is found and corrected. Formalized procedures and experienced personnel are a requirement for this process. In the proposed plan this task is one of the primary responsibilities of the data manager.

Flexibility and Continuity

A data management plan must be flexible enough to handle the changing needs of users. As survey methods evolve and new fields of study are added, the system must be able to respond quickly and efficiently to incorporate these changes. However, at the same time the program must be responsible for providing continuous, long term data sets. The Steering Committee must devise a management plan that has both these qualities.

Accessibility and Ownership

Data storage and retrieval are, of course, the main functions of a database management system. Regardless of the specific approach to data management that is taken, it is essential that the data be available to other users in this project with minimal labor and cost. However, the issue of the data ownership must be resolved at the beginning of the plan, to the satisfaction of all participating agencies. In the proposed plan, these issues must be defined by the Steering Committee at the inception of the program.

The Proposed Plan

There are several principal technical approaches to managing data in a regional monitoring context. The one chosen here as a minimalist plan relies on the assumption that all the data collectors (e.g. the States,

NYCDEP, EMAP, NOAA's S&T, etc.) will maintain their own data bases and will accept a commitment to distribute those data as needed by the monitoring program. For the regional monitoring program to be successful, ongoing consensus among all participants is essential. This is why the first priority is the formation of a Steering Committee. This committee will be composed of representatives from all participating data collectors and will have the power to alter and control the actual sampling and data storage procedures. The Steering Committee will be assisted by a part-time data manager, who will implement its decisions on a daily basis.

The Steering Committee will have four primary tasks; all requirements for the success of the monitoring program. It is recommended that an expert on regional data management be consulted during this initial phase. The four tasks are:

1. Prepare a management plan that synthesizes both the long-term strategy for coordinated data management and the detailed implementation needed. It is suggested that, at this stage, the Steering Committee should review how a number of successful large scale programs (e.g. global change program) are dealing with issues of centralized data management of data from multiple sources. The management plan will spell out the issues discussed above, such as standardization, data documentation, flexibility, ownership, etc.

2. Identify priority data sets to which the program must ensure continued access. These sets may be historical or ongoing data collections. The Steering Committee will be responsible for assuring that these data sets are archived in a way that both ensures longevity of the data and provides easy access. The minimalist plan proposes using existing services. Several options are ODES, state libraries, telephone companies, and information systems such as Compuserve. While using existing data bases presents an economic advantage, it has to be noted that these data bases were not designed for the specific needs of the LISMP. As a consequence, their management functions may not be flexible and user-friendly, and the input/retrieval of the data may be complex and/or costly. Therefore, the Steering Committee should carefully evaluate the economic advantages of using the existing data management services against the impracticality of their usage. In the event that the Steering Committee deems that a data management function specifically tailored to the needs of the LISMP is needed, then the first of the desirable options described below, the formation of a data center, should be included in the minimalist program.

3. Execute the data management plan. This task includes the establishment of task forces to advise the program on specific issues such as: modeling; QA/QC; standards for taxonomy; methods; mapping; documentation; data exchange; archiving

data; data ownership and data distribution policies. As each issue is resolved, that task force should be dissolved.

4. Assure that the data base format is designed in the best way to meet the needs of those responsible for data syntheses and informational products.

Once the long-range plan for data management of the Long Island Sound Monitoring Program has been completed, the Steering Committee should proceed to the execution of the data management plan. The Committee will be assisted by a part-time data manager who will take care of the daily tasks. She/he will be responsible for:

1. Maintaining liaisons with the involved parties.
2. Implementing the Steering Committee's decisions.
3. Ensuring that appropriate QA/QC is done on all the data by the collaborating agencies.
4. Archiving and backing-up the data in a "Sound-wide" data base.
5. Maintaining and distributing a directory and index of available data, including contact points (names, addresses, telephone and fax numbers, etc.) of the data owners.
6. Making the data easily accessible to the participating agencies, researchers, and the people who will be in charge of priority informational products.

The estimated budget needed for this information management plan is $100,000 per year.

Desirable Additions

In case additional funds become available for the data management portion of the monitoring program, the following additional elements are recommended.

1. Formation of a data center designed specifically to address the needs of the participating agencies. The data center would be a repository for all relevant data and provide easy, prompt, and inexpensive access to the data. The Steering Committee will decide on the level that data (e.g. raw data, processed data, etc.), should be stored in this system to provide a favorable cost/benefit ratio. This option provides a system that is specifically tailored to the needs of the LISMP and of the participating agencies.

2. Data entry system with automated QA/QC. This is done using a data entry program that is capable of performing various tasks both in data acquisition and in data analyses. Some examples in the data acquisition process are double data entry acquisition, range checks, valid values checks, and internal consistency checks. In the analyses stage, the system evaluates data integrity, which means the detection of errors

created during the manipulation of the data. An example of such manipulation is the process of combining data sets.

3. On-line index of the available data including documentation. This system provides a computerized data directory which includes information on what data are available, where they are located, how to extract them, and what the data may be good for.

4. Establishment of transfer formats. These are software processes that, with a simple keystroke, transfer data from the original agency's format to the required format for the archival data base, whether it be that an existing one (such as ODES), or a specific Sound-wide data base. In essence this process establishes a "common language." Furthermore, this improves data integrity and quality because the data conversions are made electronically. The use of transfer formats has an economic value, since it eliminates the burden of performing these conversions from the data providers or requesters, and in the long range it may mean significant savings. The integration of transfer formats into minimalist base program should be carefully evaluated by the Steering Committee.

5. Brokerage of data, which includes acquiring data collected by agencies not represented in the Steering Committee (e.g. pathogens, or contaminants data from NOAA's Status and

Trends). The program may assume responsibility for some processing of the data or for reformatting specific program elements. This element may also include analyses of the data which will result in the production of statistical analyses, GIS files, etc.

6. Development of liaisons with other regional monitoring programs so to extend the range of the data or to provide and receive guidance on how best to address common problems.

The estimated cost of an information management program that includes the minimalist plan plus the listed additions is $100,000-150,000 per year.

Step by Step Description of the Data Management Process

The following section provides a summary of the steps to be taken in the construction of an efficient and effective data management program for the Long Island Sound Monitoring Program.

1. Form a Steering Committee.
2. The Steering Committee, possibly with the assistance of a regional data management expert, determines needs, preferences, and expectations of the data providers.
3. The Steering Committee proceeds with the four tasks identified in the minimalist program.
4. The Steering Committee selects a data manager and provides a detailed description of his/her responsibilities.

The data management portion of a monitoring program is essential for the program as a whole to achieve its goals. A monitoring program should be dynamic and should be modified as time passes and requirements change. When this occurs, the data management plan, under the direction of its "leading team" -- the Steering Committee, should also be modified to accommodate the changes.