Data Management Implementation Plan

Prepared by Vikram Vyas
CRESP-Amchitka Data Management Component
Draft Version 2
10 June 2004

1. Introduction

1.1. Objectives and scope

The Data Management (**DM**) component of the Amchitka project will develop a geo-referenced database, in order to serve the following objectives:

- Compilation and synthesis of information from the sampling campaign and laboratory analyses of the samples;
- Tracking of specimens from point of collection to laboratory analysis;
- Quantitative analysis of information collected by the project and communication of the findings of the project; and
- Continued monitoring and improvement of data quality

The DM team will work with scientists conducting the sampling and analysis campaigns to build a system for assembling, synthesizing, and analyzing the information generated through this project. The primary requirement in initiating data management activities is a protocol for standardizing activities related to data reporting, processing and archival; the issues related to these steps are discussed in subsequent sections. Please note that as these sections evolve through interactions within the project team, this implementation plan will involve into a data management protocol for the project.

This implementation plan outlines some issues that have to be addressed by the DM component in collaboration with individual project components, in order to develop a consistent and robust data repository for the CRESP-Amchitka project. Appendices include templates of metadata files and table structures borrowed from federal data management standards documents that should be considered for use within this project.

The current version (of 10 June 2004) does not include a data exchange standard. This standard is dependent upon the format and contents of data files produced by individual components, and details of the standard will be developed through interaction between DM and other components.

2. Data Reporting Conventions

Data reporting conventions relate to the following: data file names, specimen and sample identifiers, data format, data units, space and time coordinate information, and qualifier flags. These conventions will enhance data sharing, improve the efficiency of data transfers, and facilitate analyses. The DM team will interact with project scientists to develop or select applicable reporting conventions and implement appropriately to the project's needs. It is expected that many of these data files may be generated automatically, especially for the measurements of physical parameters. It is therefore essential to have templates of data files available before the actual sampling commences, to finalize details for the data management protocol. Individual components are requested to forward their expected data reporting conventions and template files to Vikram Vyas and Lisa Bliss for development of an across-the-board system for data reporting.

2.1. Data File Names and Specimen Identifiers

The data file name or specimen id should indicate the name of biological specimen or physical parameter, date of record generation, and any other information necessary to uniquely identify each specimen or measurement.

2.2. Data File Format and Contents

Data files may be generated automatically, or may be manually entered. Some sample excel spreadsheets are attached separately, and a template for an ASCII (simple text) file is attached as an appendix, to serve as an example of the type of information that may be required to build a metadata document for the raw as well as processed data.

Generating metadata documents to describe the contents of the data files should not be viewed as an optional refinement. Their use is required by a Presidential directive. Metadata files become essential for data archival purposes, and also for communicating the details of data collection and dataset characteristics to other agencies or stakeholders. Metadata generation should therefore be taken up from the very beginning of the information generation process.

When information is manually entered into excel spreadsheets, including extra fields for describing the details (as shown in the attached template spreadsheet) will help include metadata descriptors in the initial files themselves. Likewise, a system must be established for ensuring the generation of metadata for automatically generated observation files. The metadata accompanying the raw data will be used to build up the metadata files for the final validated and quality assured data that will be submitted as the product of this project. It is recognized that this project will collect data of multiple types and varying formats, and that the process of creating corresponding metadata will be unique for each data type. Development of metadata templates will be a key issue in interactions between DM and investigators, because information that is not included at the very beginning may be impossible to obtain at a later stage.

The final metadata documents for GIS coverages produced by the project will conform to Federal Digital Geospatial Metadata Standards of 2000.

2.3. Data Units

As a general practice, SI reporting units will be used. However, it is possible that some parameters may have to be reported in non-SI units, and the data management team should be made aware of these exceptions. Further details of reporting units and consistency of units across data will be worked out as template files become available for the different components. In all cases, it is imperative that measured parameters be fully described and units explicitly stated for all data files.

2.4. Space and Time Coordinates

The reporting of spatial and time coordinates deserves a separate section because of the room for inconsistency in their reporting. Spatial coordinates will be measured by GPS

systems of varying accuracy, and reporting the likely error in spatial measurement will be essential to prevent inconsistency between different components in the project.

Likewise, local time is often reported by field personnel, and this may cause confusion or inconsistency in analysis at a later date. Therefore, sample dates and times must be reported for all measurements or specimens, with the following pieces of information: (1) local time (either clock time or standard time for the local time zone), and (2) time zone.

In addition, the local time coordinates used and the time zone are specified for every record, for time-dependent information. Both the begin time and end time must be reported for every record, and in both time formats. Begin time must be reported as time at the beginning of the averaging period. End time must be reported as the time at the end of the averaging period.

The daily time cycle runs from 00:00:00 to 23:59:59 (24:00:00 is not a legitimate value). Sampling times should be reported as hh:mm, or as hh:mm:ss, when practical and possible. The colon(s) must be used. Analysis dates should be reported for all measurements where laboratory analyses are done on dates other than the sampling date (this is standard practice, but the statement is included for the record). Reported dates must include the year, month, and day, and be formatted as yyyy/mm/dd (e.g., 2004/07/15 or 2004-07-15). Character values may not be used to denote sampling or analysis months and leading zeros should be used for day and month entry values less than ten (i.e., 08 to represent August, not 8 or AUG). This requirement helps with software that may need special programming to handle characters in date records.

Time-averaged data must include times of the beginning and the end of the time-averaging period. Further details about submission of time averaged data and their relationship to non-detects will be obtained from the laboratory analysis QA/QC protocols.

2.5. Qualifier Flags

Measurement results below the minimum detectable limits (MDL) of the instrument must be reported as measured and to the level of precision of the instrument, but flagged accordingly. Data values (e.g., averages) derived from any MDL data should be flagged. Information on Instrument Detection Limit and Estimated Quantitation Limit must also be included where necessary and applicable.

3. Data Validation

Data validation is the process of determining and denoting the quality of a dataset (data having either a common method of collection or data collected by various methods in one location). The validation process consists of evaluating the internal, spatial, temporal and physical consistency of each data set for invalid data and for outliers (data that are physically, spatially, or temporally inconsistent). During validation, physically unrealistic data are invalidated, biases and instrumental drift are noted, and gross errors are identified. The objective of this process is to produce an archive with values that are of known quality. The function of data validation will have a considerable overlap

between different CRESP-Amchitka components, and each component will have significant responsibility for documenting and implementing its primary QA/QC procedures. DM will serve as a secondary party for implementing QA, in collaboration with individual components, and DM will help coordinate and track validation status of individual datasets.

In order to track the data validation status of different data sets, it is recommended that all data should include a flag designating their level of validation with values ranging from zero (0) to three (3). Initially, data will be at the lowest level of validation. In response to further quality checks, the validation flag will be upgraded. Revisions to flag codes will be made at the component level (with notification to DM and CRESP management board) and records of these changes will be maintained. When additional validation work is performed on existing data, the component leader will send the DM a revised data set reflecting the new validation status or provide the DM with specific instructions to revise validation codes in archived data sets. Details on validation activities and actions can be included with the metadata. Because not all activities are performed concurrently, at any given time the archive may contain data at different validation levels.

Recognizing the potential for CRESP-Amchitka to generate research products other than "data" is important. Examples of other possible products are statistical models, GIS coverages, and reports. Each of these products must be validated. A validation level and status discussion must be included in the metadata record or information associated with the dataset or research product. The following general levels of validation can be used; however, the details of specific validation checks that will be performed for any given research product will have to be based on the type of information generated and the method of producing the information by each component, and therefore need to be worked out. For some components, specialized software or analytical procedures may be required for assuring the quality and reliability of data or data products, and the general levels described below may not be directly applicable.

Level 0 validation indicates a reasonably complete data set of unspecified quality that consists of research products subjected to minimum processing in the field and/or in the laboratory by individual component staff. Level 0 designations will be given to raw data and other research products that have not been audited or peer reviewed. Level 0 status will remain in force until all audits or peer reviews associated with Level 1 validation of the product have been completed and the investigator's response recorded.

Level 0 data contain all available measurement data and may also contain data in the form of quality control checks and flags indicating missing or invalid data. Level 0 data consist of instrument outputs expressed in engineering units using nominal calibrations. Missing data from on-site backup data loggers or strip charts have been filled in. Level 0 data may include flags indicating QC check data, power failures, excessive rate-of-change, insufficient data for the averaging period, or other automatically generated/programmed occurrences.

Level 1 Validation indicates a complete data set of specified quality that consists of research products subjected to quality assurance and quality control checks and data management procedures. As part of the Level 1 process, site documentation is reviewed for completeness and performance compared with other locations. Compliance with documented data quality objectives, standard operating procedures (SOPs), and research protocols is evaluated in the Level 1 process. Audit and peer review reports have been evaluated (and necessary corrective actions taken) for all research products designated Level 1. The comparison and crosschecking activities done under Level 1 may be conducted by the component staff in coordination with the DM components and scientists from other components.

Level 1 data are generated by groups of investigators from different components. In response to audits, data may have been adjusted. The DM will adjust the data in the database under the guidance of the project component responsible for submitting the data; the project group assigned the data validation task will be responsible for determining precision and accuracy, and will perform consistency checks with other data within the same data set. These internal consistency checks might include diurnal analyses to look for expected patterns or time series analyses to detect outliers, extreme values, or time periods with too little or too much variation. Level 1 designation will be assigned after the project group has performed all quality control activities identified in their QIWP and addressed all quality issues stemming from audits and reviews.

Level 2 Validation indicates a complete, externally consistent data set of specified quality that consists of research products that have undergone interpretative and diagnostic analysis by CRESP investigators. A validation level and status discussion should be included in the metadata record associated with the research product.

External checks might include correlation by scattergram, comparison of data with other similar data (if available), and comparison of a measurement made by two different methods or laboratories. If comparisons are not within the precision of the measurements, then measurement records and other information will be reviewed. If a check of measurement records uncovers a process error, the value will be corrected or invalidated. If such errors are not found, then an annotation will be entered. If the value is invalidated, it will be deleted from the database and replaced by a missing value and flagged appropriately. A record of changes will be permanently retained in the database. Level 2 designation will be assigned after CRESP investigators and/or data users have performed comparative tests and addressed the quality issues and have evaluated the test results and supporting QA documents. Authority for Level 2 designation lies at the management board level.

Level 3 Validation consists of data that have received intense scrutiny through analysis or use in modeling by CRESP review groups. As analysis of the data proceeds, analysts may raise questions about portions of the Level 2 data set. Additional checks and tests will be performed on such data and the Level 3 code will be affixed to data passing these tests. If this scrutiny reveals an inconsistency that appears to be caused by a measurement error, the entire chain of evidence for the measurement will be reviewed.

This includes reviewing site logs and quality control test data as well as reviewing performance audit results and any other relevant documents. The importance of having a robust database for tracking data records becomes apparent. The data users will recommend a Level 3 designation to project staff on the basis of the reevaluations. Alterations to the data or its validation codes, if warranted, will be made by DM under the guidance of CRESP PI or management board.

Level 4 data are those that have been subjected to all applicable quality checks and have been deemed to be valid. These data will be cleared for release to the general public or for external review, upon authorization by CRESP PI.

The table below summarizes the characteristics and status of data and research products at each level of the validation process.

	Level 0	Level 1	Level 2	Level 3	Level 4
Review Status Source	Raw Data and Un-reviewed Products	QA'd Data and Peer Reviewed Products	Data Analyses Completed and Products Assessed	Project Use	Public Release
Processing and Reviews Performed by	Component leaders	CRESP	CRESP	Peer Scientists	Anonymous reviewers, stakeholders, general public
Metadata Records	Incomplete	Corrections Documented and Review Comments Resolved	Assessment Issues Documented and Metadata Files Available	Applicable Records Reviewed (e.g., site logs and performance audits)	Completed metadata documents
Access	Component leader	CRESP	CRESP	CRESP and peer reviewers	Unrestricted
Change Control Point of Contact	Component leader	Component Leader, DM and CRESP-PI	Component leader, DM, and CRESP-PI	CRESP-PI	CRESP-PI

4. The Database Management System

4.1. Project Database Documentation

Project specific databases include spreadsheets, data sets, and databases (e.g., MySQL, SQL Server) to manage the project data. The project specific databases will be documented. The database documentation should identify the commercial database product used, the database name, structure, and computer locations. The minimum database documentation will consist of:

- name and version of commercial software used;
- names of project databases created;
- database structure definitions, including field names and descriptions; and storage
- location and media.

4.2. Project Software Documentation

Project specific software includes scripts written by the DM group for data management activities and programs written by the project team for the production of data products. Data products are defined as any extraction, summary, or analysis of tabular data that result in an electronic data extraction summary or a hard copy product such as tables, graphs, statistics, or maps.

Software documentation should include the software program name, description, special requirements, author, revision, completion date, and documentation of the QA review. Data products documentation should also include all information to uniquely describe how the data product was produced, including the sources used, the manipulations made, and the tools used to produce the data product. Software documentation can be maintained in hard copy or it may be included as comment blocks embedded within the project software program. The minimum software documentation should consist of

- name and version of the commercial software used;
- name and version of the software program written by the project;
- author;
- date;
- revision;
- system requirements; and
- storage location.

4.3. Project Software Quality Assurance

The DM component will define the QA requirements for project specific database management software. At a minimum, scripts to load data, calculate statistics reported in project deliverables, and produce data products should be reviewed to ensure they meet the desired objectives. Ideally, the reviewer should be someone other than the person who wrote the software program.

4.4. Project Specific Software Configuration Control

Project specific software should be protected from unauthorized modification or deletion. This can be accomplished by administrative controls or file security options provided by computer operating systems. Changes to project software should be documented and a history of revisions which impact the results or data products should be included in the project file. Commercial products are available to maintain a record of software revisions [e.g., Revision Control Software (RCS)]. Another way to do this is to keep the initial or baseline software in a storage area separate from the working software. Then, when the software changes, the new software can be moved to this separate area; with copies of all revisions saved in a backup location. Project specific configuration and revision control will be documented. The project software configuration control documentation should include

- commercial software used;
- program names;
- revisions (including dates of revision); and
- storage locations.

Developed software application programs will be tested and validated to ensure compliance with all user requirements and to provide confidence that the software will perform satisfactorily in service. The technical adequacy of results generated by these applications will also be reviewed by another person, tested and validated (e.g., dose calculations, sample management, risk assessment). Configuration management of the developed software application programs will be conducted.

4.4.1. Software Configuration Control

Changes and enhancements to software, both during the development process and after formal acceptance, are inevitable. Problems detected during testing must be corrected, existing system requirements often change or are reallocated to improve processing, and new requirements are often added. For any software development effort, this change process is controlled to ensure that only authorized changes or enhancements are incorporated and that system integrity is maintained. Configuration control is applied to project baselines. A Software Configuration Management process must be identified and implemented. The flow of configuration control is managed by a reporting documentation system geared to track software changes. All changes or enhancements made to software applications will be documented.

4.4.2. Database Structure Configuration Control

Database configuration control includes many of the same requirements as software configuration control. Database configuration control ensures that the physical database design and implementation are properly protected from unauthorized changes or destruction and that authorized changes are identified and tracked. Database structures are controlled in the same way as software (see previous section). Before any changes or enhancements are made to the project's database structure, appropriate approval must be granted and documented. Access control requirements will be adhered to, and all changes or enhancements made to database structures will be documented.

4.5. Day-to-Day Operation of Data Management Systems

This section addresses the day-to-day operations of a data management system, including backups, access, security, data entry, data control, and data archival.

4.5.1. System Backups

Project data will be protected from loss through preventative database backup and recovery mechanisms. Database backups will be performed on a periodic basis, with a greater frequency of backup during peak data acquisition periods. This frequency will be selected to minimize the extent of consequences of data loss and time required for data recovery. The working database will be stored on a hard disc that gets backed-up daily, and periodic backups on CD-ROMS will also be made. Recovery procedures will be developed and documented in preparation for the event of hardware or software failure.

4.5.2. Data System and Database Access

DM will protect systems and data from unauthorized access by implementation of administrative and procedural controls. Access controls will be managed based upon specific data user roles that will be defined by the types of data and functionality required (e.g., DM may have the capability of creating and updating data from field logs; a component leader or investigator may require read-only access to perform on-line queries).

4.5.3. Data Entry

Data entry, transfer, and transformation activities will be verified to ensure that data integrity is maintained. This includes all movement/copying of data from one storage medium to another and transformation from one format to another. All data, including analytical data produced and reported by a laboratory, should conform to this requirement. This verification encompasses all data recording media, such as handwritten or hard copy produced via electronic means, as well as automatically generated data from data loggers. It also includes all data collection methods (e.g., electronic collection through real-time monitoring instrumentation, bar coding equipment, and handwritten log entries). If a record modification or transfer activity has occurred before receipt of the data by DM (i.e., between creation and final reporting), the verification may be performed by the reporting party but sufficient evidence to document the process must be provided by the reporting party.

4.5.4. Database Content Configuration Control

DM will establish configuration control requirements for the contents of the project database. The requirements should ensure traceability of field and laboratory data from the original reported values through authorized data changes to current values stored in the database. The changes will be made under the guidance and approval of concerned component leader, and the configuration control will document the approval process required for making changes to the database. The minimum information maintained for each database change will include

- a description of the change;
- the reason for the change;

- the name of the individual making the change;
- the name of the component leader approving the change;
- the date of the change; and
- a copy of the data before the change took place.

4.5.5. Data Archival

The final version of the database for public release will be saved separately from the original database, as an additional backup measure.