

Bay Area Macroinvertebrate Bioassessment Information Network (BAMBI)*Issue Paper #5***I. Topic: Standards and guidelines for managing, reporting and sharing bioassessment data****II. Background Information**

Bioassessment sampling involves several types of data (listed below). The California Stream Bioassessment Procedure (CSBP) for wadeable streams provides standardized datasheets and protocol recommendations concerning the following:

- a. Field data associated with the sampling event and site. CSBP requirements include sampling date, site mapping coordinates, individual sample locations and chain-of-custody information.
- b. Associated data about conditions at the sampling site. The CSBP and most other protocols include basic water quality chemistry, measurements of physical parameters at sampling locations and other physical habitat assessments at reach level.
- c. Laboratory processing records and data qualifiers. The CSBP provides a model bench datasheet with space for recording subsampling information and some issues such as numbers of unidentifiable organisms. The protocol also recommends that specialized benchesheets be customized to the Standard Operating Procedures of the laboratory. Additional QA/QC information may include the results of taxonomic validation, resamples, validation reanalyses or other checks.
- d. Raw counts of identified taxa per subsample, summarized from bench data sheets.
- e. Metrics or other summary statistics derived from raw taxonomic counts. The CSBP suggests

In the past, and for individual small projects, spreadsheets have generally been used to store and present these data in a flatfile or tabular format. The CDFG's Aquatic Bioassessment Laboratory (ABL) has moved to a relational database application for storage of data generated through its own projects and contract work.

Why is it important?

Data storage and reporting formats have direct bearing on the types of analyses and interpretations that can be accomplished. Method development of bioassessment as a regional tool requires integration of data from different sources, which generates concerns relating to both data content and the manner of its storage and presentation. Because taxonomic categories and lists of preferred metrics are likely to vary over time, it's extremely important to preserve raw source data if a dataset is to be useful for reviewing trends in the future, or for comparisons with other bioassessment projects in nearby watersheds.

Regional or national considerations also play a role. Federally sponsored programs including grants require data reporting in formats compatible with STORET, an Oracle database maintained by USEPA for surface and ground water data collected by federal, state and local agencies, Indian Tribes, volunteer groups or individuals, etc. The "modernized" (1999)

STORET supports data maintenance and browsing via the Internet and easier conversion from a variety of sources. This data model is a “warehousing” concept in which access to data is controlled by the “primary owner” agency or entity (source).

Current status in the Bay Area

- One of the largest single CSBP projects is the Surface Water Ambient Monitoring Program (SWAMP) program conducted by the SFBRWQCB (72 sites in 2001), with data stored in Cal-EDAS format at the ABL.
- Stormwater programs subcontractors are typically in Excel formats generated by contracting laboratories. Some of these were based on Excel templates previously distributed by ABL. There is probably variation in formats and amount of site or QA/QC data included. Stormwater CSBP datasets are dominated by the Marin Co. dataset (with Alameda, Vallejo and some San Mateo samples in the same format.
- Many other agencies and organizations are using “in-house” applications and formats; range of variation has not been surveyed

III. Objectives of issue paper:

1. Review general issues for selection of data storage formats and implications for management, sharing and reporting.
2. Present some options for improving consistency of data formats to facilitate data sharing
3. Identify potential next steps

IV. Current Issues

Reporting:

The CSBP recommends that standard reporting of results include a table of taxonomic counts and a suite of representative metrics as well as coefficient of variation among replicate samples from a single sampling event. CDFG collection permit provisions also encourage electronic submittal of CSBP metrics, taxonomic counts and site data to meet the reporting requirements. Sampling site locations and event data should also be documented, as well as any protocol variations.

Sharing:

Past data sharing has been mainly through hard copy reports, but there is a trend towards electronic transmittal of metrics and/or taxonomic counts. These are typically in Excel files which may or may not include all site data. Associated site data and QA/QC flags should also be included, along with metadata and QA/QC information (owner/creator of data, whether the data has been checked, date of original file creation and last modification, etc.) Ideally this information is integral to the datafile but can also be generated from reports or accompany the data as a readme file. If datasets are to be combined for reanalysis, formats and field definitions

should also be compatible as much as is practicable, and areas of inconsistency need to be identified.

Formats and management:

Excel files are typically used by many contractors and project managers. Advantages include:

- The software is widely used and supported, on different systems and understood by many people.
- Once a template is set up, data entry is simple. Some spreadsheet templates incorporate calculation of metrics.
- A generic metadata worksheet can be inserted in Excel files (appended)

Disadvantages of this Excel-based situation include:

- Multiple templates in use limit compatibility
- Limited ability to recalculate metrics-can be cumbersome
- Large datasets can become unwieldy
- Format may need conversion for inputting to some statistical procedures

An alternative format is the relational database, in which data resides in multiple tables linked by specific fields common to two or more tables. This structure reduces duplication of information by its handling of many-to-one relationships. A well-designed database application can facilitate data entry and support more flexibility in querying and reporting, as well as exporting to other formats. However, initial investment in design and programming is high, and some ongoing technical staff time will probably be needed for support and management of the database in most organizations.

Several initiatives attempt to address the need for bioassessment data storage:

Data Management Tool (Datamon) for the Central Coast Ambient Monitoring Program (CCAMP) was developed by Dave Paradies initially for the Morro Bay watershed program. While not a true database, Datamon is a functioning set of linked Excel spreadsheet that gives many of the capabilities of a relational database. Data types incorporated in the file include water quality, sediment and tissue chemistry, benthic macroinvertebrates, stream transect data, habitat and geomorphological parameters. (CCAMP, 2001)

Ecological Data Application System (EDAS); Tetra Tech, 1999) is a relational database in MS Access, designed and distributed by Tetra Tech under contract for US EPA. Its original purpose was to manage input and reporting for biological data, including calculation of metrics. Data tables for fish, macroinvertebrates, and riparian vegetation as well as basic chemical, habitat and weather parameters were based on existing assessment practices in various states.

Cal EDAS is an ABL modification of the EDAS database. As modified, it is a fully functioning database for most aspects of a bioassessment program based on the CSBP. At this point, CalEDAS is strictly an in-house ABL database and aspects of it are still being refined. ABL's long range goal is to make the database easily transferred to others, but there are no current plans to provide tech support for it. ABL intends to maintain compatibility with STORET and any

future California monitoring database (see below). At present CalEDAS is in Access 2002 format; although it can be saved in Access 2000 format, further testing is required to confirm its functionality and it is not compatible with pre-2000 versions of Access.

System for Water Information Management Phase II (SWIM II) As part of the statewide SWAMP program, this is a project for a refined MS Access database to replace the existing pilot system. Implementation was planned for 2003, to include an Internet interface and STORET compatibility. Development is part of a wider contract to Moss Landing Marine Laboratories.

Data Categories and Fields

Consistency of data fields is extremely important both internally and among datasets. This applies at several levels:

- Names of fields should be internally consistent. If conversion is expected between formats, length of fieldname and presence or absence of spaces are constraints
- Definitions of fields including type of data and the exact parameter description
- Text entries must be consistent in spelling and use of abbreviations. Pull-down menus improve reliability but require advance programming
- Fields for comments, qualifier codes or other notes should be used to identify differences in protocol or data quality. This becomes much more important when combining datasets from different projects or organizations.

STORET is constructed with data fields grouped into several major categories:

- Organizations (creating, owning or sponsoring)
- Projects and Surveys (contacts, goals, data quality objectives)
- Sites (station)
- Samples (all data associated with a particular sampling event, including physical, chemical or biological data and any qualitative observations)
- Results (findings or interpretations, including metrics and metadata)

Related BAMBI topics:

An important general consideration is the amount of interest in sharing data at different levels of intensity. Data requested by the RWQCB to support regional analyses for reference conditions or biocriteria development may need conversion to CalEDAS. Datasets collected with different protocols should not be pooled for analysis, but results may be evaluated in parallel if there is adequate documentation and metadata.

Standardization of protocols and QA/QC will influence the content and number of data fields; Supplementary or Level II suites of physical habitat parameters would also need to be accommodated.

V. Options for improvement

Several levels of improvement are possible, depending on the interest levels of BAMBI participants:

- Maintain individual data management applications, while improving documentation and reporting to facilitate interpretation and comparison of results
- Promote common standards for data field definitions, QA/QC information and metadata
- Promote use of common formats in Excel, with guidance and/or tools to assist conversion to CalEDAS format
- Promote conversion to a common relational database structure, and seek funding for regional support for users.

VI. Suggested next steps

Possible activities, in increasing order of effort and coordination:

1. Disseminate models or templates for basic spreadsheets and/or reporting standards.
2. Assess existing formats in use, and priorities of users for compatibility with others in region.
3. Develop guidelines for fieldnames and definitions.
4. Explore potential conversion tools between common Excel formats and Cal-EDAS or other relational database.
5. Develop committees or partnership for converting to and maintaining a common data format, and seek funding.

VII. References

Central Coast Ambient Monitoring Program (CCAMP), 2001. User Manual for CCAMP Data Management Tool for Microsoft EXCEL 97, 98, and 2000. Central Coast RWQCB. Contact: Karen Worcester

Tetra Tech, 1999. Ecological Data Application System Version 2.0 for US EPA Region IX. 410-356-8993

The Bay Area Macroinvertebrate Bioassessment Information Network (BAMBI) includes scientists, watershed managers and regulators interested in local applications of *bioassessment* --the use of biological community data for assessing the condition of waterbodies and watersheds. BAMBI's focus is on *benthic macroinvertebrates* (bottom-dwelling animals without backbones, visible to the naked eye) which are present in most aquatic environments and are useful indicators of ecosystem function because their community composition responds to a wide range of ecosystem variables. BAMBI Issue Papers provide background and discussion of technical areas important to the development and improvement of bioassessment in the Bay Area. These are provisional workproducts for BAMBI discussion in January 2003, contributed by members of the Bay Area Stormwater Management Agencies Association (BASMAA), and the SFBay RWQCB. Forward comments or questions to BAMBI c/o watersheds@acpwa.org