

Embedding metadata in the data

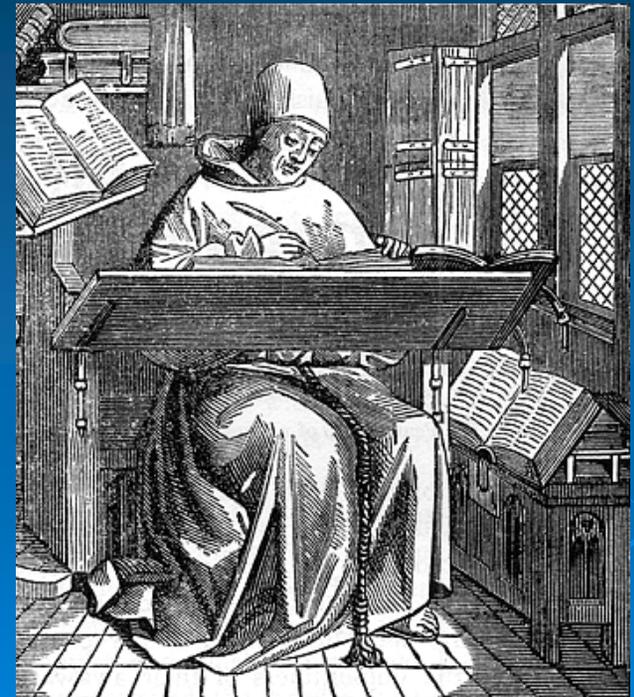
Toward self-documenting databases

*By Marc Vayssières,
California Department of Water Resources*

The background of the slide is a solid blue color. In the lower half, there are several decorative elements consisting of concentric circles, resembling ripples in water. These circles are centered at various points and vary in size and opacity, creating a subtle pattern.

What is Metadata?

- ❖ The answer to basic questions:
 - who, why, what, when, where, and how
- ❖ The information needed for integrating data from different sources while **ensuring comparability**
- ❖ The data elements for reporting water quality results (www.nwqmc.org)
- ❖ What still needs to be done when you think your work is finished...



Metadata about my program

❖ Who?

- The “Environmental Monitoring Program” (EMP)
- Part of the Interagency Ecological Program
- Joint effort of CA Department of Water Resources and US Bureau of Reclamation.
- With assistance from CA Department of Fish and Game and US Geological Survey.



Metadata about the EMP

❖ Why?

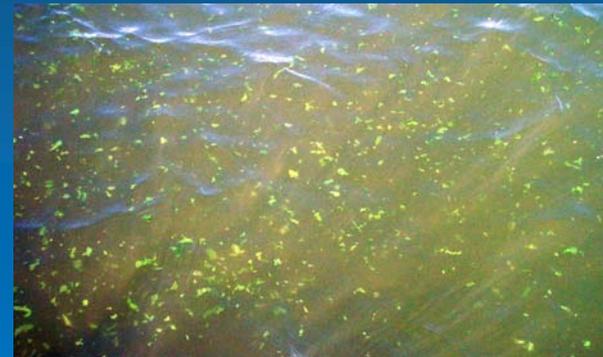
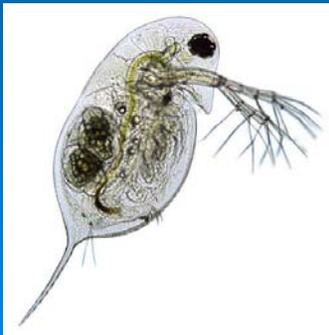
- Monitoring mandated by the State Water Resources Control Board
- Tied to water right permits issued to DWR and USBR for exports of water to southern California
- To determine compliance with water quality standards
- To document effects of diversions and flow manipulation



Metadata about the EMP

❖ What?

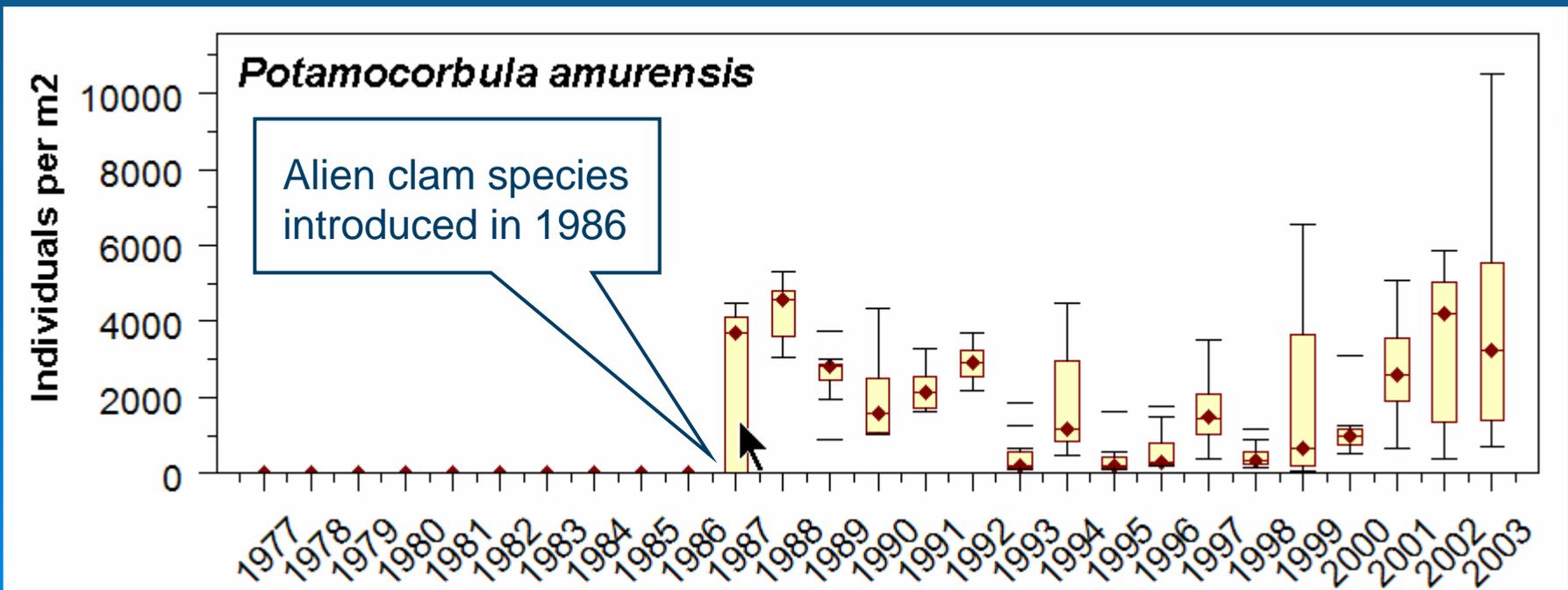
- Environmental Water Quality
- Phytoplankton
- Zooplankton
- Benthic macro-invertebrates



Metadata about the EMP

❖ When?

- Began in 1972, revised in 1978 and 1996
- Monthly discrete sampling and automated continuous monitoring stations
- Databases cover 1975 to present



Metadata about the EMP

❖ Where?



Metadata about the EMP

❖ How?



Ponar dredge grabs



Continuous monitoring stations



Vertical Profiles



Field measurements



Research Vessel



DWR chemical lab

Data sharing 1994

WQ1992.csv

```
"RSAN1 12" "C10" "19920114,1325,03," "9," "11," "16798," "0," "....."
11," "971," "10.5," "7.1," "24," "2," "....."
0.6," "0.29," "3.33," "65," "574," "1.40," "1.3," "0.29," "0.13," "135,"
17.0," "6.16," "3.33," "65," "574," "1.40," "1.3," "0.29," "0.13," "135,"
"RSAN1 12" "C10" "19920225,1215,03," "15," "21," "18824," "0," "....."
31," "896," "7.9," "6.8," "98," "....."
0.9," "0.21," "1.60," "1.5," "0.51," "0.22,"
14.0," "11.39," "9.49," "55," "550," "....."
"RSAN1 12" "C10" "19920312,1125,03," "17," "21," "18182," "3," "135"
22," "1050," "8.5," "7.5," "86," "....."
0.6," "0.04," "2.00," "1.4," "0.50," "0.19,"
17.0," "21.12," "10.93," "66," "829," "....."
"RSAN1 12" "C10" "19920323,1010,03," "17," "18," "18188," "5," "90"
28," "....."
0.8," "....."
27.88," "15," "....."
"RSAN1 12" "....."
14," "....."
0.4," "....."
14.0," "21," "....."
"RSAN1 12" "....."
21," "....."
0.5," "....."
33.61," "9," "....."
"RSAN1 12" "....."
23," "....."
0.2," "....."
13.0," "30," "....."
"RSAN1 12" "....."
13," "....."
0.3," "....."
"RSAN1 12" "....."
15," "....."
0.3," "....."
10.0," "140," "....."
"RSAN1 12" "....."
```

Metadata.txt

Parameters such as water temperature, pH, turbidity, dissolved oxygen Winkler titration method), and electro-conductivity are analyzed in the field.

On board instruments are standardized before sampling and checked against references after each sampling run. Bryte Laboratory processes field blank samples and runs a duplicate of every fifth water sample for QA/QC purposes.

TABLE III EPA TEST IDENTIFICATION:

| Parameter | EPA Test Number: |
|--------------------------------|------------------|
| BOD, 5 DAY MG/L | 405.1 |
| DISSOLVED SOLIDS, 105 C MG/L | 160.1 |
| SUSPENDED SOLIDS MG/L | 160.2 |
| VOLATILE SUSPENDED SOLIDS MG/L | 160.4 |
| ORG N TOTAL MG/L | 351.2 |
| ORG N DISS MG/L | 351.2 |
| NH3+NH4- N DISS MG/L | 350.1 |
| NH3+NH4- N TOT MG/L | 350.1 |
| NO2 - N, DISS MG/L | 353.2 |
| NO3- N, TOTAL MG/L | 353.2 |
| NO2&NO3 N, DISS MG/L | 353.2 |
| NH3&ORG N, TOTAL MG/L | 351.2 |
| PHOSPHORUS, TOTAL MG/L P | 365.4 |
| PHOSPHORUS, DISS. ORTHO MG/L | 365.1 |
| CHLORIDE, DISS MG/L | 325.2 |

Number of variables in each water qual

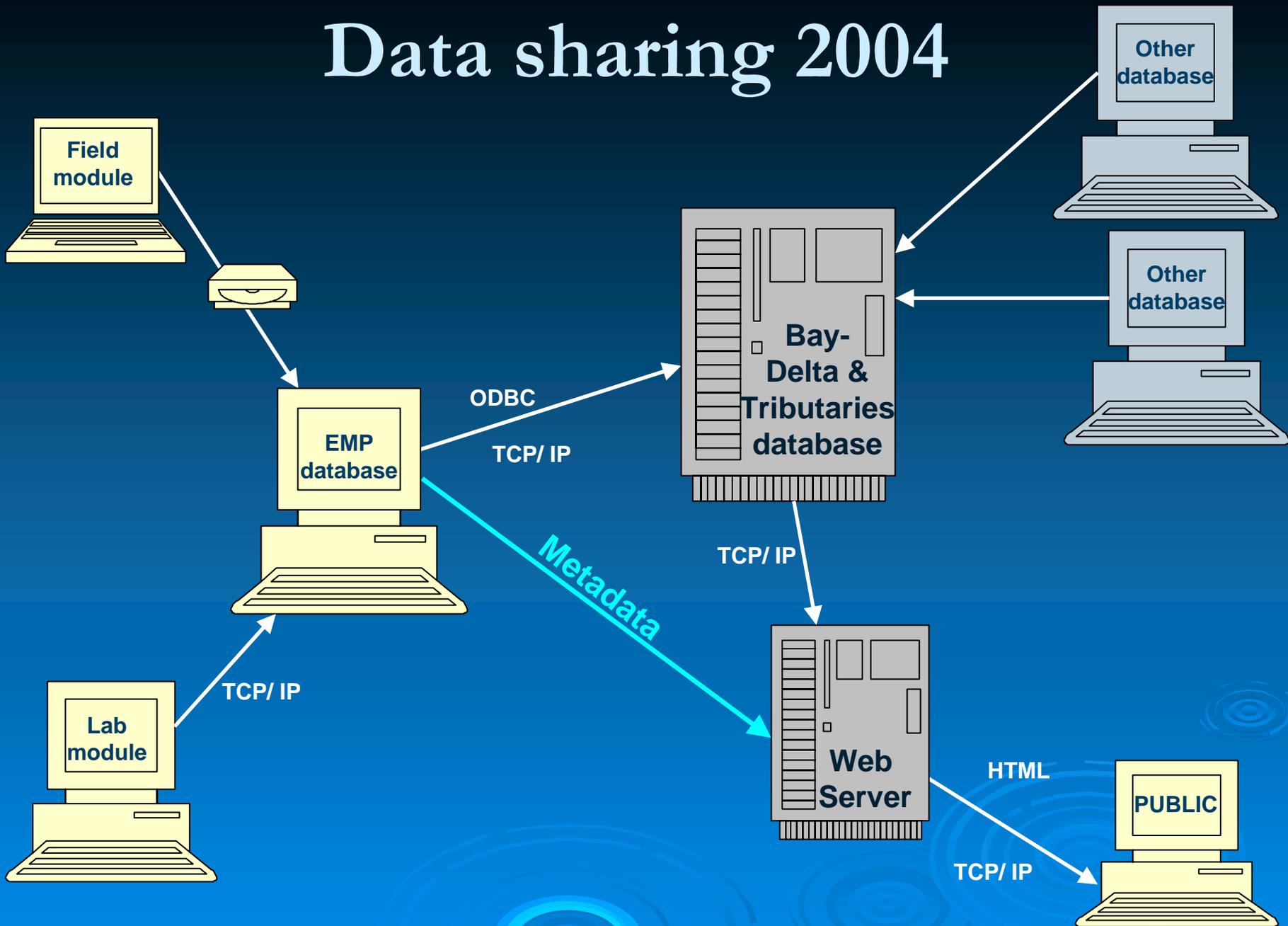
| COLUMN, | WIDTH | VARIABLE |
|---------|-------|------------------------|
| 1 | 17 | RKI LOCATION |
| 24 | 6 | STATION |
| 26 | 9 | SAMPLE DATE (YYYYMMDD) |
| 35 | 4 | SAMPLE TIME (PST) |
| 40 | 2 | SAMPLE DEPTH (FT) |

DATA

| COLUMN, | WIDTH | REMARK, | WIDTH | VARIABLE |
|---------|-------|---------|-------|--------------------|
| 43 | 8 | 52 | 3 | WATER TEMP CENT |
| 56 | 8 | 65 | 3 | AIR TEMP CENT |
| 69 | 8 | 78 | 3 | FIELD IDENTIFICATI |
| 82 | 8 | 91 | 3 | WIND VELOCITY MPH |
| 95 | 8 | 104 | 3 | WIND DIRECTION FRO |
| 108 | 8 | 117 | 3 | STREAM STAGE, FEET |
| 121 | 8 | 130 | 3 | TIDE STAGE CODE |
| 134 | 8 | 143 | 3 | TURBIDITY, HACH FT |
| 147 | 8 | 156 | 3 | SECHI, METERS |
| 160 | 8 | 169 | 3 | CONDUCTIVITY AT 25 |
| 173 | 8 | 182 | 3 | DEPTH-M 1% LIGHT R |
| 186 | 8 | 195 | 3 | DO, MG/L |
| 199 | 8 | 208 | 3 | BOD, 5 DAY MG/L |

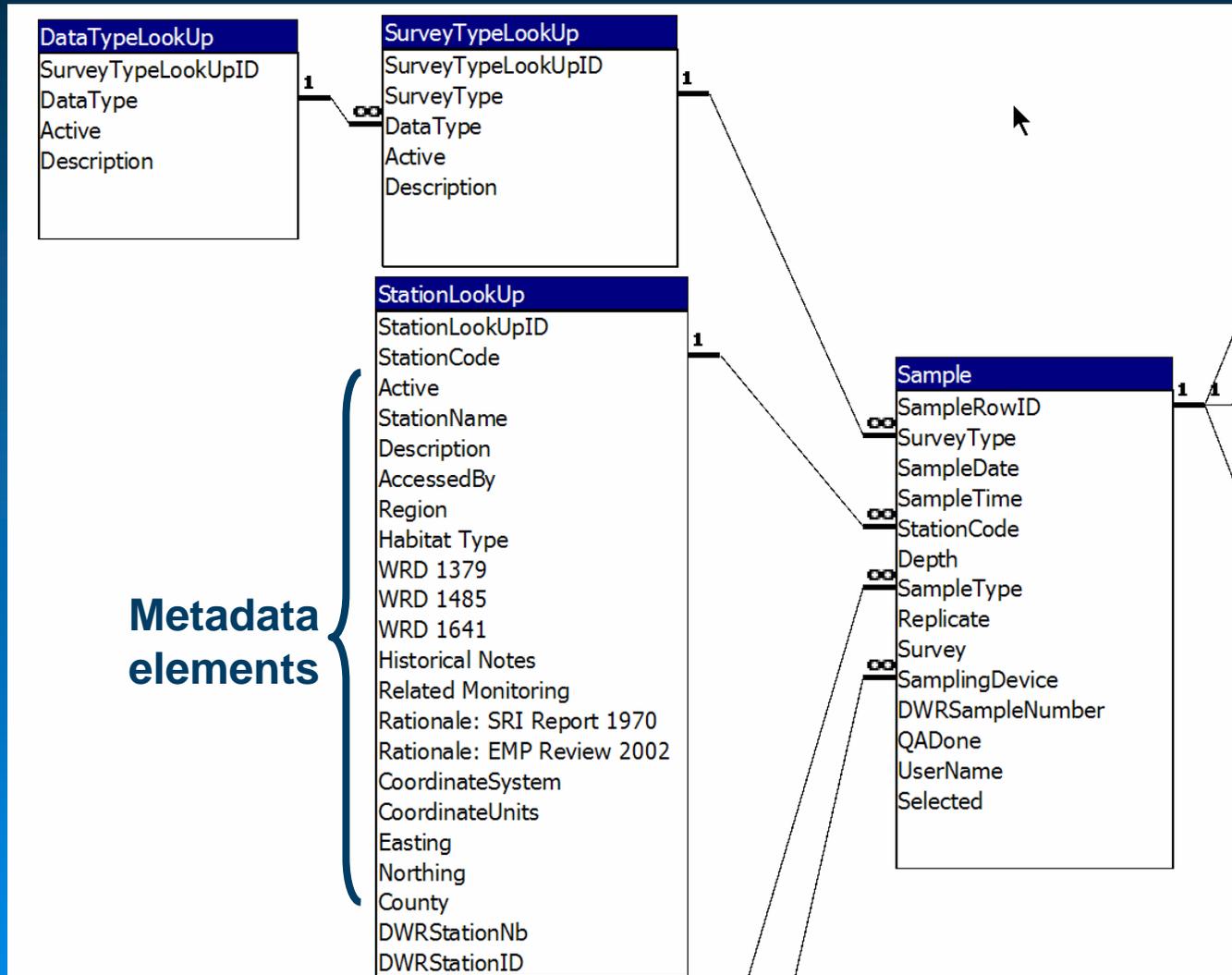
WQFormat.txt

Data sharing 2004



Beyond flat files

❖ Relational databases and lookup tables



Beyond text files

❖ Html documents

B. Name and Location Information for Discrete Water Quality Sampling Sites

- [Currently Sampled Stations](#)
- [Historically Sampled Stations](#)

Notes:

- Coordinates are in decimal degrees, Geographic coordinate system, North American Datum 1983 and have been verified to be accurate for 1:24,000 scale mapping.
- Habitat types are based on ecologically important physical and chemical habitat characteristics.
- Regions are based on cluster analyses of monthly water quality variables. For specific analyses see Lehman 1996 and Lehman and Smith 1991; Jassby and Cloern 2000;

C. Sample Sites' History and Rationale for Monitoring

- [Sampling Rationale and Historical Notes](#)

Notes:

- "Rationale-1970" is from the Stanford Research Institute report to the State Water Resources Control Board of 1970.
- "Rationale-2002" is from the [EMP Review of 2001-2002](#).

IV. Period of Record

Most stations and variables monitored by the EMP were established based on recommendations by a 1970 Stanford Research Institute (SRI) study commissioned by SWRCB, or derived from the preceding "Delta-Suisun Bay Surveillance Program" and the "Delta Fish and Wildlife protection study." The original set

Beyond static tables

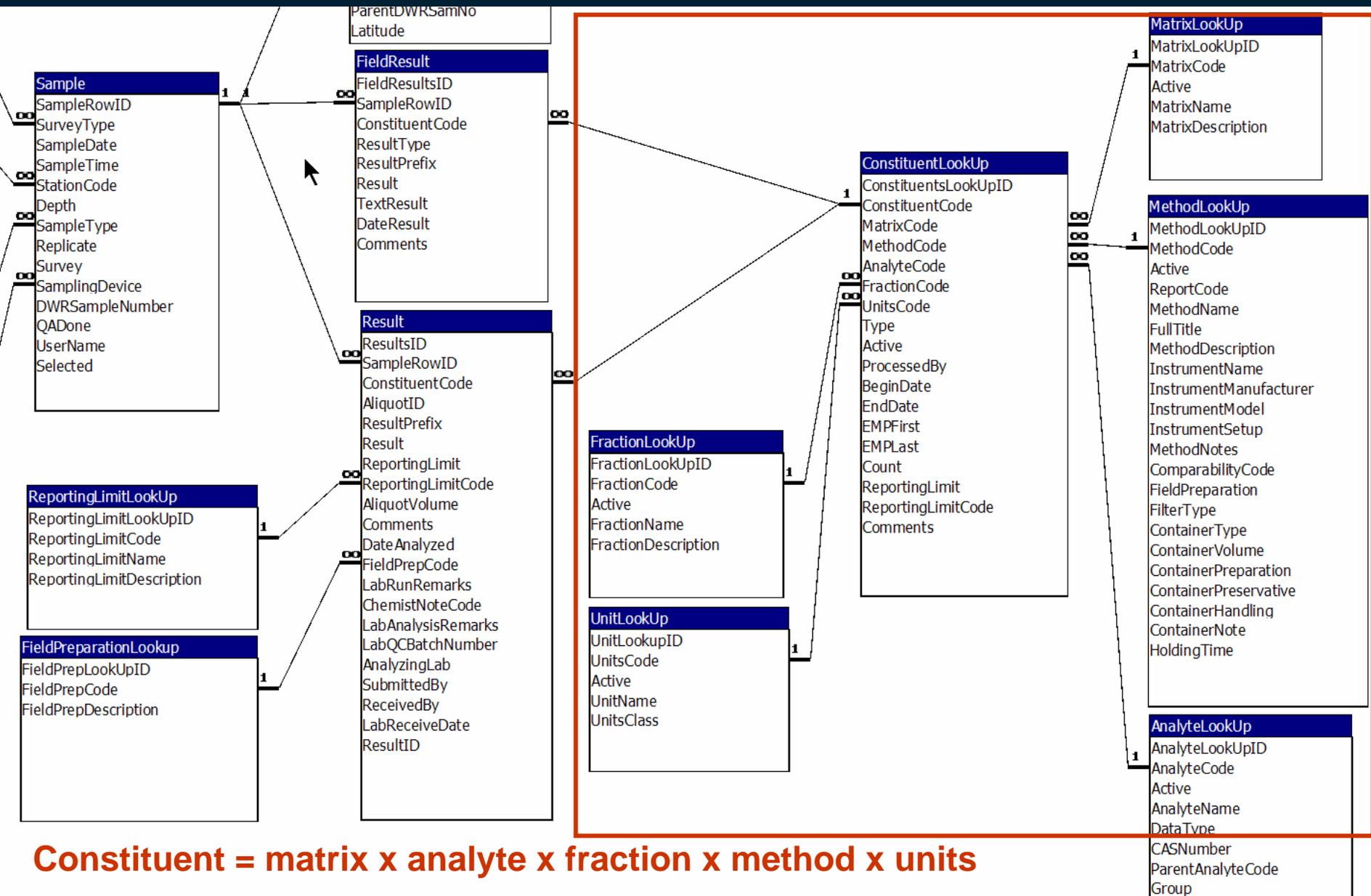
❖ Queries from the database linked in html document

Last revised on Monday, May 03, 2004

Metadata - stations rationale and history

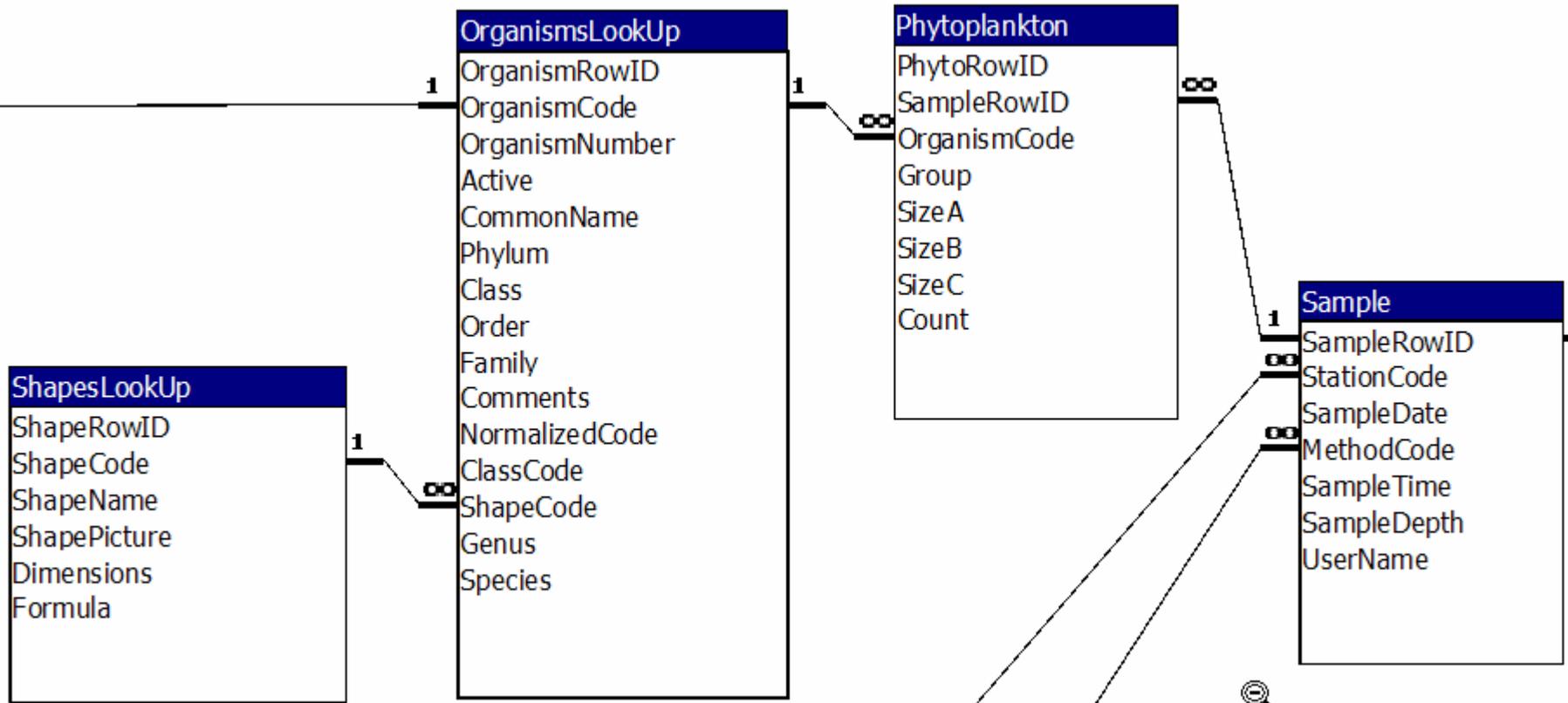
| Code | Name | Data Availability | Rationale-1970 | Rationale-2002 | Historical Notes | Related Monitoring |
|-------|------------------------------------|-----------------------|----------------------------------|---|---|---|
| Blank | No station | Jan. 1998 - Apr. 2003 | n/a | | n/a | n/a |
| C10 | San Joaquin River near Vernalis | Jan. 1975 - Apr. 2003 | n/p | "Rim" station with a long, comprehensive, highly utilized data record and an important flux station (imports into the Delta from the San Joaquin watershed) with high productivity. D-1641: flow rate and Specific conductivity water quality objectives. | 1938 - Site initiated for Central Valley Operations Sampling Program. 1968 - station for the San Luis Drain Program within the combined USBR-DWR Delta-Suisun Bay surveillance Program. Ongoing discrete monitoring. | At nearby location C10A: Proposed automated continuous water quality monitoring. |
| C3 | Sacramento River @ Greenes Landing | Jan. 1975 - Apr. 2003 | Principal northern Delta inflow. | "Rim" station with a long, comprehensive, highly utilized data record and an important flux station (imports into the Delta from the Sacramento watershed) with low productivity. | 1952 - Site initiated for USBR Central Valley Operations Sampling Program. 1969 - station for the San Luis Drain Program within the combined USBR-DWR Delta-Suisun Bay surveillance Program. Ongoing discrete monitoring. | At nearby location C3A: Automated continuous water quality monitoring in compliance with D-1641 |
| C7 | San Joaquin River @ | Jan. 1975 - Dec. 1995 | Lower Delta location; shallow | | 1968 - station established by USBR | At nearby location C7A: Automated |

More complex structures



Non-text elements

❖ Phytoplankton shapes and bio-volume formulas



Phytoplankton Metadata

Algal biovolume can be calculated from the dimensions using for example the formulae given for different algal shapes by Kellar et al. (1980).

[Shapes and biovolume formulas](#) used in EMP's phytoplankton database

The list of all species of phytoplankton identified, their shape codes and the total number of algae counted can be found in the [Phytoplankton Dictionary](#)

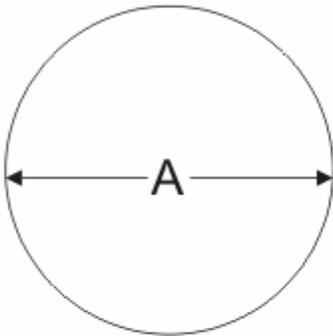
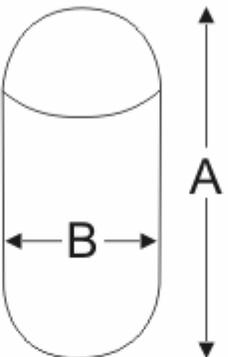
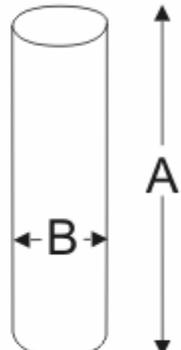
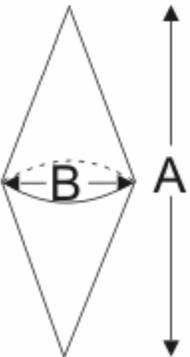
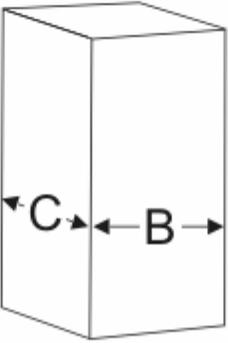
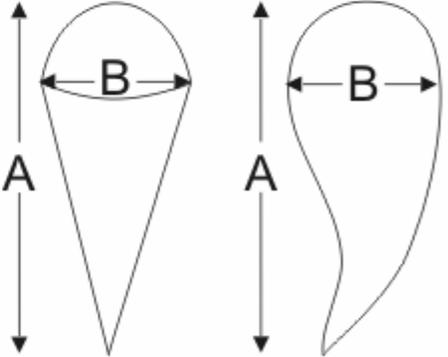
IX. Data Management

The results of identification, enumeration and size measurement are entered in EMP's Phytoplankton database directly at the lab. After reviewing of the results for accuracy and completeness, data are exported

Metadata - Phytoplankton Dictionary

| Organism Code | Family Common Name | Family | Genus | Species | Shape Code | Total Count |
|---------------|--------------------|-------------------|------------|-------------|------------|-------------|
| ACHN BREV | Diatom | BACILLARIOPHYCEAE | Achnanthes | brevipes | 5 | 1 |
| ACHN DELI | Diatom | BACILLARIOPHYCEAE | Achnanthes | delicatula | 9 | 153 |
| ACHN EXIG | Diatom | BACILLARIOPHYCEAE | Achnanthes | exigua | 9 | 1 |
| ACHN GIBB | Diatom | BACILLARIOPHYCEAE | Achnanthes | gibberula | 9 | 582 |
| ACHN LANC | Diatom | BACILLARIOPHYCEAE | Achnanthes | lanceolata | 9 | 346 |
| ACHN LINE | Diatom | BACILLARIOPHYCEAE | Achnanthes | linearis | 9 | 2 |
| ACHN MINU | Diatom | BACILLARIOPHYCEAE | Achnanthes | minutissima | 9 | 5 |
| ACHN | Diatom | BACILLARIOPHYCEAE | Achnanthes | sp. | 9 | 3552 |
| ACHN TEMP | Diatom | BACILLARIOPHYCEAE | Achnanthes | temperei | 9 | 2 |
| ACTE | Diatom | BACILLARIOPHYCEAE | Actinella | sp. | 9 | 80 |
| AMPC MIRA | Diatom | BACILLARIOPHYCEAE | Amphicampa | mirabilis | 9 | 1 |

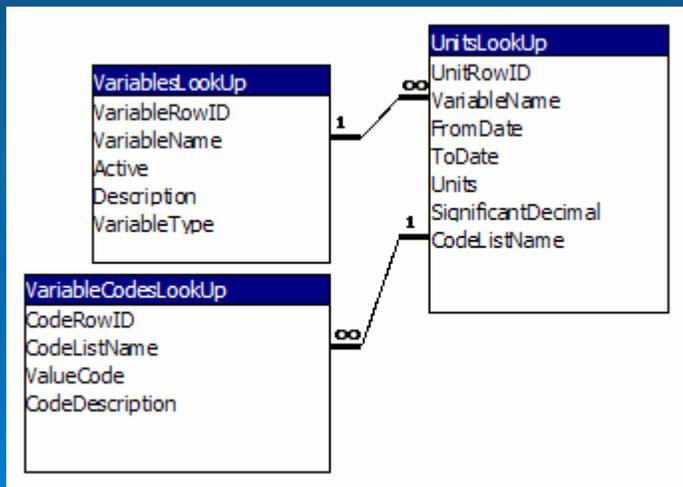
Metadata - EMP's Phytoplankton Shapes and Bio-volume Formulas

| <i>Code - Name Dimensions and Formula</i> | <i>Code - Name Dimensions and Formula</i> | <i>Code - Name Dimensions and Formula</i> |
|--|---|---|
| <p>1 - Sphere</p>  <p>Volume = $\pi A^3/6$</p> | <p>2 - Ellipsoid</p>  <p>Volume = $\pi AB^2/6$</p> | <p>3 - Cylinder</p>  <p>Volume = $\pi AB^2/4$</p> |
| <p>4 - Two cones</p>  <p>Volume = $\pi AB^2/12$</p> | <p>5 - Cuboid</p>  <p>Volume = ABC</p> | <p>6 - Ellipsoid + cone</p>  <p>$V = (\pi B^2 (A+B/2))/12$</p> |

Can't justify a lookup table?

❖ Include a data dictionary structure for:

- variable names
- description
- minor codes and units



Editing tables VariablesLookUp, UnitsLookUp and VariableCodesLookUp

Data Dictionary

Add New Main Menu

Variable Name: RunName Active? Find (in current field) Find Next

Description: Common name of sampling run.

VariableType: Quantitative Qualitative **LookUp code** RowID, UserName, etc.

Quantitative variable properties

Qualitative variable properties

Code list name: UserRunNameList

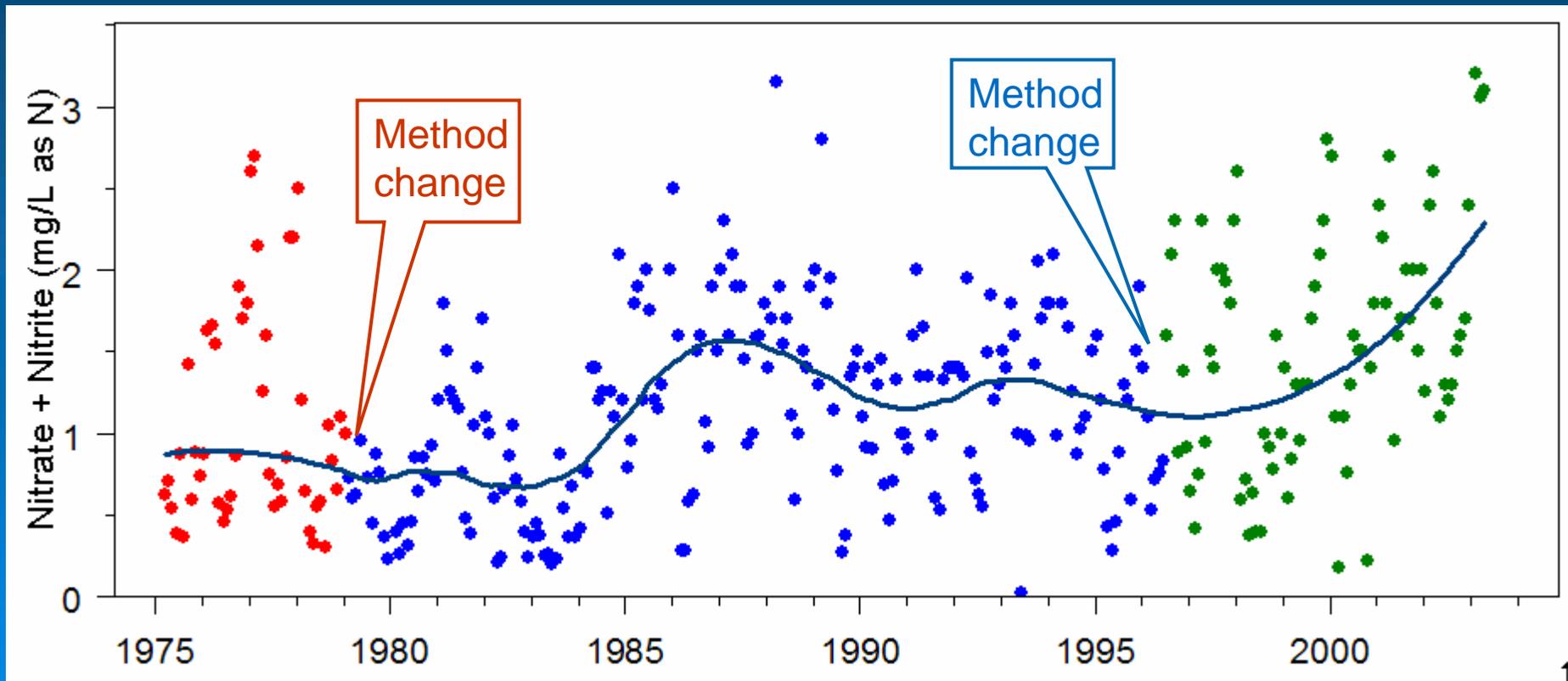
| Value Code | Code Description |
|-------------|--|
| ▶ Van Run | Lab van sampling run |
| Mid Delta | Mid Delta sampling run |
| Zoo Tow | Zooplankton sampling run |
| Suisun Bay1 | First of two Suisun Bay sampling runs |
| Suisun Bay2 | Second of two Suisun Bay sampling runs |
| * | |

LookUp code properties

Record: 34 of 49

Advantages of Embedding

- ❖ The metadata is distributed with the database
- ❖ Changes in time-series are made obvious to the data user while maintaining flexibility in querying



Advantages of Embedding

- ❖ Use metadata elements to summarize data:
 - by phytoplankton (or benthic) family, genus, guild
 - by type of station (region, habitat, etc.)
- ❖ Use metadata elements for quality assurance:
 - Compare results by lab, by sampling crew, etc.
- ❖ Update elements once use many times:
 - for metadata page, for yearly data report, etc.
 - consistency is ensured across documents
- ❖ Keep metadata pages updated more easily:
 - queries + macros + time stamp => updated html tables

Advantages of Embedding

- ❖ produce cross-tabulations: what x where x when
 - Groups of lab constituents per station per year

Last revised on Monday, May 03, 2004

Metadata - lab results per station per year

| StationCode | 1975 | 1976 | 1977 | 1978 | 1979 | 1980 | 1981 | 1982 | 1983 | 1984 | 1985 | 1986 | 1987 | 1988 | 1989 | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 |
|-------------|-----------|-----------|-----------|-----------|-----------|-----------|----------|-----------|-----------|----------|-----------|-----------|----------|-----------|-----------|-----------|-----------|-----------|-----------|------|------|------|------|------|------|------|
| C10 | MP BNO | MP BNO | MP BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO |
| C3 | M BNO | MP BNO | M BNO | MP BNO | M BNO | MP BNO | M BNO | M BNO | MP BNO | M BNO | M BNO | M BNO | M BNO | MP BNO | MP BNO | MP BNO | MP BNO | MP BNO | MP BNO | BNO |
| C7 | BNO | BNO | BNO | MP BNO | MP BNO | M BNO | M BNO | M BNO | M BNO | M BNO | M BNO | M BNO | M BNO | MP BNO | MP BNO | MP BNO | MP BNO | MP BNO | MP BNO | BNO | BNO | | | | | |
| C9 | M BNO | M BNO | MP BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | | | | |
| D10 | M BNO | M BNO | M BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | B | B | B | |
| D11 | BNO | BNO | BNO | MP BNO | M BNO | M BNO | M BNO | MP BNO | M BNO | M BNO | MP BNO | MP BNO | M BNO | MP BNO | MP BNO | MP BNO | MP BNO | MP BNO | MP BNO | BNO | BNO | | | | | |
| D12 | MP BNO | M BNO | M BNO | MP BNO | MP BNO | M BNO | M BNO | MP BNO | MP BNO | M BNO | MP BNO | MP BNO | M BNO | MP BNO | MP BNO | MP BNO | MP BNO | MP BNO | MP BNO | BNO | BNO | B | B | B | | |
| D14A | M BNO | M BNO | MP BNO | MP BNO | M BNO | M BNO | M BNO | MP BNO | MP BNO | M BNO | MP BNO | MP BNO | M BNO | MP BNO | MP BNO | MP BNO | MP BNO | MP BNO | MP BNO | BNO | BNO | | | | | |
| D15 | M BNO | M BNO | M BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | BNO | | | | |

Groups are M = Metals, P= Pesticides, B= Biological, N = Nutrients, O = Other

Advantages of Embedding

- ❖ produce cross-tabulations: how x when
 - Laboratory methods over time

Last revised on Thursday, May 13, 2004

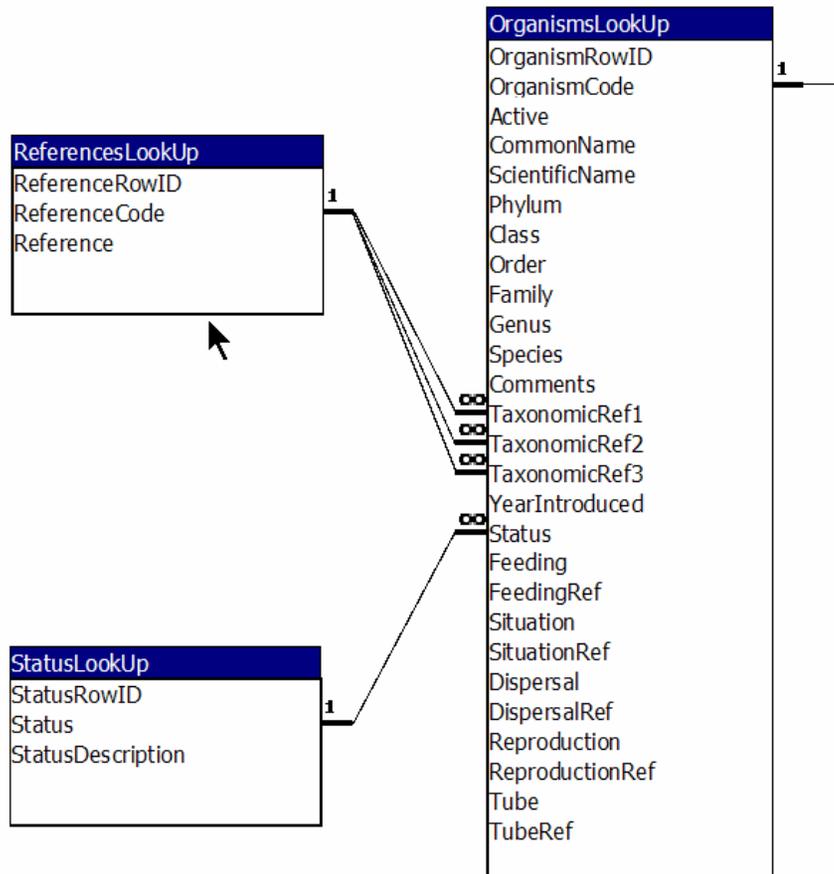
Metadata - Laboratory methods

| Current? | Code | Name | Description | Comparability notes | Data from | Data to |
|----------|-------------------------------|--|---|---|-----------|-----------|
| Yes | Std Method 10200 H | Spectrophotometric Determination of Chlorophyll | Measured samples (500 to 1000 ml depending on the concentration of suspended solids) are filtered in the field and frozen. Pigments are extracted in aqueous acetone with tissue grinding, centrifugation and incubation at 4°C. Chlorophyll a and Pheophytin a | Data comparable to other standard methods in this category, within accuracy and precision limits. | 2/2/1998 | 4/14/2003 |
| No | Std Method 10200 H (modified) | Spectrophotometric Determination of Chlorophyll (Sonication) | Measured samples (400 ml) are filtered in the field and frozen. Pigments are extracted in aqueous acetone with warm water bath (58°C), sonication followed by incubation at room temperature. Chlorophyll a and Pheophytin a concentrations are measured with | A study was conducted in 2001-2002 and found good agreement between the between the modified and unmodified chlorophyll extraction methods. (See Triboli, K., Mueller-Solger, A. and Vayssières, M. 2003) | 1/7/1975 | 1/9/1998 |
| Yes | Std Method 2540-C | Total Dissolved Solids (TDS) | Total Dissolved Solids, total filterable residue dried at 180 degrees Celsius. | Data comparable to other standard methods in this category, within accuracy and precision limits. | 7/19/1996 | 4/11/2003 |
| No | Std Method 3111 B | Cations by Flame AA | Cations: Ca, Mg, Na, and K analyzed by Direct Air-Acetylene Flame AA. | Data comparable to other standard methods in this category, within accuracy and precision limits. | 1/7/1975 | 9/11/1986 |

Advantages of Embedding

❖ Make metadata better

- e.g. linking taxonomic references to benthic organisms made us discover some inconsistencies



Last revised on Tuesday, April 20, 2004

Metadata - Taxonomic References

| ReferenceCode | Reference |
|---------------|--|
| T01 | Allen, R. K., and C. M. Murvosh. 1987. Mayflies (Ephemeroptera: Tricorythidae) of the southwestern United States and northern Mexico. <i>Ann. Ent. Soc. Am.</i> 80: 35-40. |
| T03 | Bousfield, E. L., Research Associate, Royal British Columbia Museum, 675 Belleville Street, Victoria V8V 1X4, B.C., Canada, personal communication. |
| T04 | Bowman, T. E., Curator Emeritus, Department of Invertebrate Zoology (Crustacea), NHB 163, National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20560, personal communication. |
| T05 | Brinkhurst, R. O. 1986. Taxonomy of the genus <i>Tubificoides</i> Lastockin (Oligochaeta: Tubificidae): species with bifid setae. <i>Can. J. Zool.</i> 64: 1270-1279. |
| T06 | Brinkhurst, R. O. 1986. Guide to the freshwater aquatic microdrile oligochaetes of North America. <i>Can. Spec. Pub. Fish. Aquat. Sci.</i> 84: 259 pp. |
| T07 | Brinkhurst, Ralph O., Director, Aquatic Resources Center, P.O. Box 680818, Franklin, TN 37068-0818, personal communication. |
| T08 | Burch, J. B. 1975. Freshwater Sphaericean Clams (Mollusca: |

Advantages of Embedding

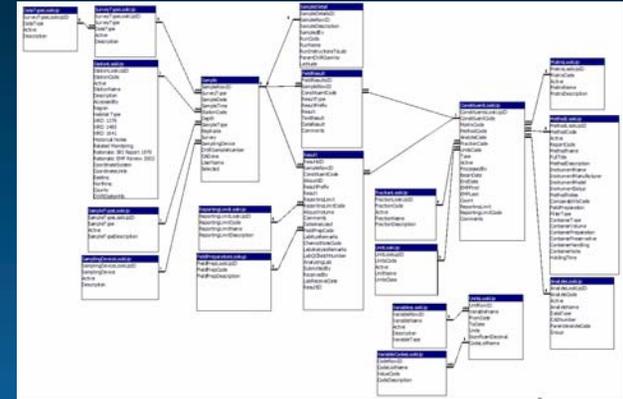
❖ Make metadata routine

Editing Constituents

| Constituent Builder | | Select Constituent |
|--|---|--|
| | Constituent Component Name | |
| Matrix: | Water <input type="checkbox"/> | [Wat] Chlorophyll a (µg/L) - Extraction |
| Method: | On Site Analysis <input type="checkbox"/> | [Wat] Conductivity (µS/cm) - Foxboro EC |
| Analyte: | Conductivity <input type="checkbox"/> | [Wat] Conductivity (µS/cm) - Schneider EC |
| Fraction: | n/a <input type="checkbox"/> | [Wat] Conductivity (µS/cm) - Schneider EC 2 |
| Units: | µS/cm <input type="checkbox"/> | [Wat] Conductivity (µS/cm) - YSI Multi-Probe |
| Equipment: | YSI Multi-Probe <input type="checkbox"/> | [Wat] Fluorescence (FU) - SCUFA Fluoro. |
| Active?: <input checked="" type="checkbox"/> | Use check boxes to show active items only. ^ Uncheck them to show all items. | [Wat] Fluorescence (FU) - Turner 10 |
| New Constituent <input type="button" value="New Constituent"/> | | [Wat] Pheophytin a (µg/L) - Extraction |
| Delete Constituent <input type="button" value="Delete Constituent"/> | | [Wat] Stage (ft (MSL)) - Shaft Encoder |
| Cancel <input type="button" value="Cancel"/> | | [Wat] Temperature (°C) - ASTM Thermometer |
| Close <input type="button" value="Close"/> | | [Wat] Temperature (°C) - Schneider WT |
| | | [Wat] Temperature (°C) - YSI Multi-Probe |
| | | [Wat] Turbidity (NTU) - SCUFA Turbidity |
| | | [Wat] pH (pH Units) - Schneider PH |
| | | [Wat] pH (pH Units) - YSI Multi-Probe |
| | | [Wat] Dissolved Chloride (mg/L) - Schneider Cl |
| | | [Wat] Dissolved Oxygen (mg/L) - Schneider DO |
| | | [Wat] Dissolved Oxygen (mg/L) - Winkler |

Downsides of Embedding

- ❖ higher upfront effort:
 - set up the database structure
 - build forms for easy data entry
 - input historic information.
- ❖ Corrections may be more onerous
 - e.g. if some results were mapped to the wrong method
- ❖ May not be right for all metadata elements



Don't Forget the Nots

- ❖ Document what is **NOT** in the database:
 - Water quality results below Reporting Limit.
 - Samples accidentally lost, not correctly preserved
 - Benthic grabs without any macro-invertebrates
 - Sampling events planned but missed for any reason.
- ❖ Will make the data manager's life easier
- ❖ Will make your data more credible and useful to users

Be explicit about the Nots

- ❖ To compute benthic invertebrates abundance per unit area one must be able to distinguish between:
 - missed sampling events – marked “Not Sampled”
 - lost grab samples – marked “Sample lost”
 - and “empty” grabs – identified by a single pseudo-species “none” and a count of zero

VI. Data availability in EMP's benthic database

A. Sampling events

- [Number of samples per station per year](#): all conducted sampling events
- [Number of missed samples per station per year](#): scheduled sampling events that were missed entirely (see [note](#))

B. Grabs

- [Number of grabs per station per year](#): grabs with valid data, including those without organisms (see [note](#))
- [Number of lost grabs per station per year](#): lost grabs, no data available

More @ <http://iep.water.ca.gov/emp/>



Interagency Ecological Program

Environmental Monitoring Program



- [About the EMP](#)
- [Data Base Access](#)
- [Metadata](#)
- [Water Quality Project Work Team](#)
- [Sampling Schedules](#)
- [EMP Review Final Reports](#)
- [EMP Review 2001-2002](#) *

(* Password Required)

Questions? Please e-mail [Anke Mueller-Solger](mailto:Anke.Mueller-Solger)
or call her at (916)227-2194.