Potentials & Pitfalls of Combining/Sharing Water Quality Data

"The best technique for scientists to use in mismanaging data is simply to hoard the data. One fringe benefit of hoarding is that it eliminates the need to write those annoying metadata files that are so useful to people trying to steal your results…Untold damage can result when people who cannot conceivably understand your data analyze them on their own & draw their own conclusions."

Stephen Hale, "How to Manage Data Badly" (2000)

Potential Benefits

- Increased spatial, temporal, or taxonomic scale of an assessment
- Increased pool of reference sites available to any one program
- Enhanced ability to make scientifically defensible judgements on water quality
- Provides decision makers with better information with which to devise and implement monitoring strategies
- Increased ability to use data produced by other programs & encourages collaborative efforts
- Reduced costs

Impediments/Challenges

- DATA COMPARABILITY
- DATA EXCHANGE
- DATA MGT
- DATA ACCESS
- DATA INTERPRETATION
- DATA DOCUMENTATION –metadata!!

What is Metadata?

- "Data about data."
- Includes characteristics such as content, accuracy, reliability, and the source
- Can be used to:
  - Concisely describe datasets & other resources using *elements* (e.g. name of the dataset, the quality, how to access the data, what is its intended purpose, whom to contact for more information about the data)
  - Enable effective management of data resources
  - Enable accurate searching and data resource discovery
  - Provide an online interface to a dataset and link to other information about it
  - Accompany a dataset when it is transferred to another computer so that the dataset can be fully understood, and put to proper use, and to duly acknowledge the custodian

Key Metadata Components

Metadata

- Who
- What
- Where
- When
- Why ("The objective of the study was…")
- How (Methods)
  - Data Collection
  - Data Processing
  - Data Analysis

Primary Data

- Samples/Events
- Locational Data
- Organisms
- Counts/Measure-ments

Derived Data

- Metric values
- LC50's
- Index scores
- Biocriteria

"Anything that requires math"

"If you think the cost of metadata production is too high - you haven't compiled the costs of NOT creating metadata: loss of information with staff changes, data redundancy, data conflicts, liability, misapplications, and decisions based upon poorly documented data."

FGDC (2000)

Real-World Examples

- DC
  - assessments (derived data) cannot be directly compared
  - raw data (primary data) cannot be located
  - metadata does not exist
- VA
  - "administratively decentralized" system
  - each region uses own variations in sampling method
- ICPRB
  - Potomac River Basinwide Assessments Project

"Omitting key metadata can be a foolproof way of rendering a data set useless."

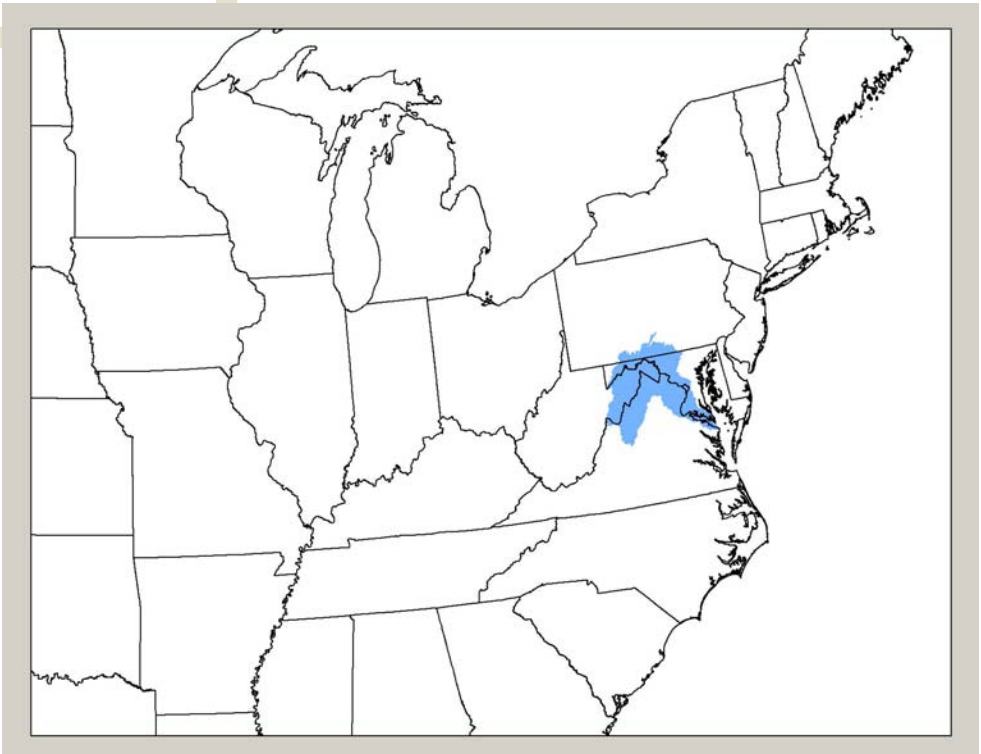
"How to Manage Data Badly," Stephen Hale (2000)

Potomac River Basinwide Assessments

Enhance state water quality monitoring & assessment programs with consistent, watershed-wide assessments of stream health

"The ability to combine datasets is desirable to make judgements on the condition of the water resource." (Barbour et al. 1999)

"Combining information from separate monitoring surveys improves understanding of the biological integrity of riverine systems." (Handcock et al.)



Real-World Example Across 3 Potomac Basin States

VA	PA	MD
•EPT	•EPT	•EPT
•Family	•Family	•Genus
•No formal training	•No formal training	•Training/certification required
•100-org subsample	•N/A; sort/ID in field and estimate RA's.	•120-org subsample
•Unknown (assume gridded pan)	•500 um kick net	•Tray w/100 5cmX5cm grids
•2 sq m area	•1 sq m area	•2 sq m area
•Multiple habitats sampled/composited	•Multiple replicates; not composited	•Multiple habitats sampled/composited
•BPJ	•Narrative only; attainment status determined in field	•Quantitative criteria
•Ecoregions?	•Subwatersheds	•Order, subwatersheds
•Targeted	•Targeted	•Stratified Random
•Judgement sampling	•Census sampling	•Probabilistic sampling

Metadata should be treated with the seriousness of a peer-reviewed publication and should include, at a minimum, a description of the data themselves, the study design and data collection protocols, any quality control procedures, any preliminary processing, derivation, extrapolation, or estimation procedures, the use of professional judgment, quirks or peculiarities in the data, and an assessment of features of the data that would constrain their use for certain purposes.

National Academy of Science (2002)