



A Database for Water Resources Discipline Aquatic Community and Habitat Data

**Peter M. Ruhl, Mitchell A. Harris, and Alex K. (Sandy)
Williamson, U.S. Geological Survey National Water-
Quality Assessment Program**

Today

- **The Situation**
- **The Solution Pursued**
- **Managing Taxonomic Data for the Long Term**



The Situation

Aquatic ecology has become an integral part of the hydrologic science done by USGS Water Resources Discipline (WRD)

- **National Water-Quality Assessment (NAWQA) Program Ecology (since 1994)**
 - **Macroinvertebrates, algae, fish, stream habitat**
 - **Approx. 16,000 samples at > 2,000 sites**
- **Non-NAWQA WRD Ecology Projects (2000 – 2006)**
 - **80 % (41 of 51) reported collected biological data**
 - **147 projects with ecology/bioassessment component**
 - **Approx. 15,000 samples**

80 percent (120 of 147) of the WRD ecology projects included aquatic macroinvertebrate, fish, algae, or in-stream habitat components

Collecting macroinvertebrates



Most of the data are stored electronically but are likely to have a short lifespan and are difficult to discover and access

- 47% in Excel
- 13% are in EMAP databases
- 19% in home-grown relational databases



Consequences

- Inability to distribute WRD bioassessment community and habitat data to national audiences (internal or external)
- Missed opportunities for
 - Leveraging existing data within similar (or same) spatial and temporal contexts
 - Reusing existing data to enhance project analyses
 - Building new program based on previous work
- Jeopardizes ability to support USGS strategic goals related to stream ecology



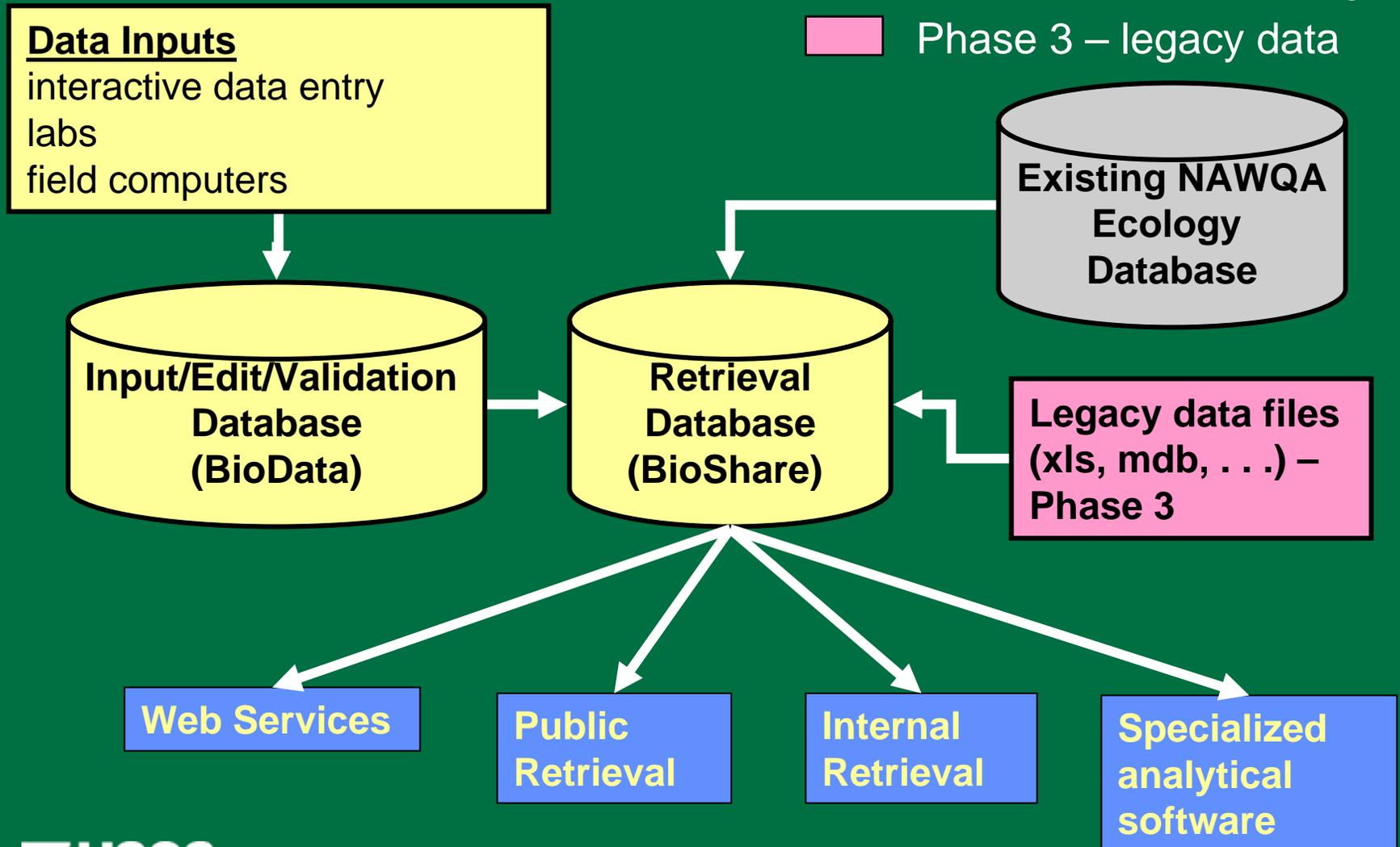
The Solution Pursued

The Fundamental Objective

- **Storage and retrieval capability for the most commonly collected aquatic community, population, and in-stream habitat data collected by WRD.**
 - **Benthic Macroinvertebrates**
 - **Algae**
 - **Fish**
 - **Reach-scale in-stream habitat data**

The Big Picture

- Phase 1 – new systems
- Phase 2 – NAWQA migrates
- Phase 3 – legacy data



Big Picture Goals

- **Internet-oriented**
 - Input and retrieve data using web browser
 - Capture and distribute data via web services
- **Extensible/Adaptable**
 - Quickly adapt to new requirements for sample-collection and lab-analysis protocols
 - Gradually increase support for capturing digital field data (e.g. electronic field forms, photo's, GPS, deployed instrumentation).
- **Play Well With Others**
 - Support for data exchange standards and mechanisms (E.G. USEPA Water Quality Exchange Network (WQX))

Capabilities – Version 1 (summer 2010)

- Support for the most commonly used field collection protocols (60% of samples reported by WSC projects, all of the NAWQA samples)
 - USGS NAWQA protocols
 - USEPA NRSA protocols
- Multiple taxonomic labs (internal and external)
 - Create/Transmit orders to labs
 - Electronic transfer and batch load of lab results
- Simple set of data review and retrieval capabilities for project staff and general public

Capabilities – Version 2 and Beyond

- Additional field and lab protocols
- Data exchange services
- Additional data review and QA capabilities
- Additional sample status tracking/notification capabilities
- Data Visualization (graphs and maps)
- Support for lab evaluation and quality assurance data

Challenges

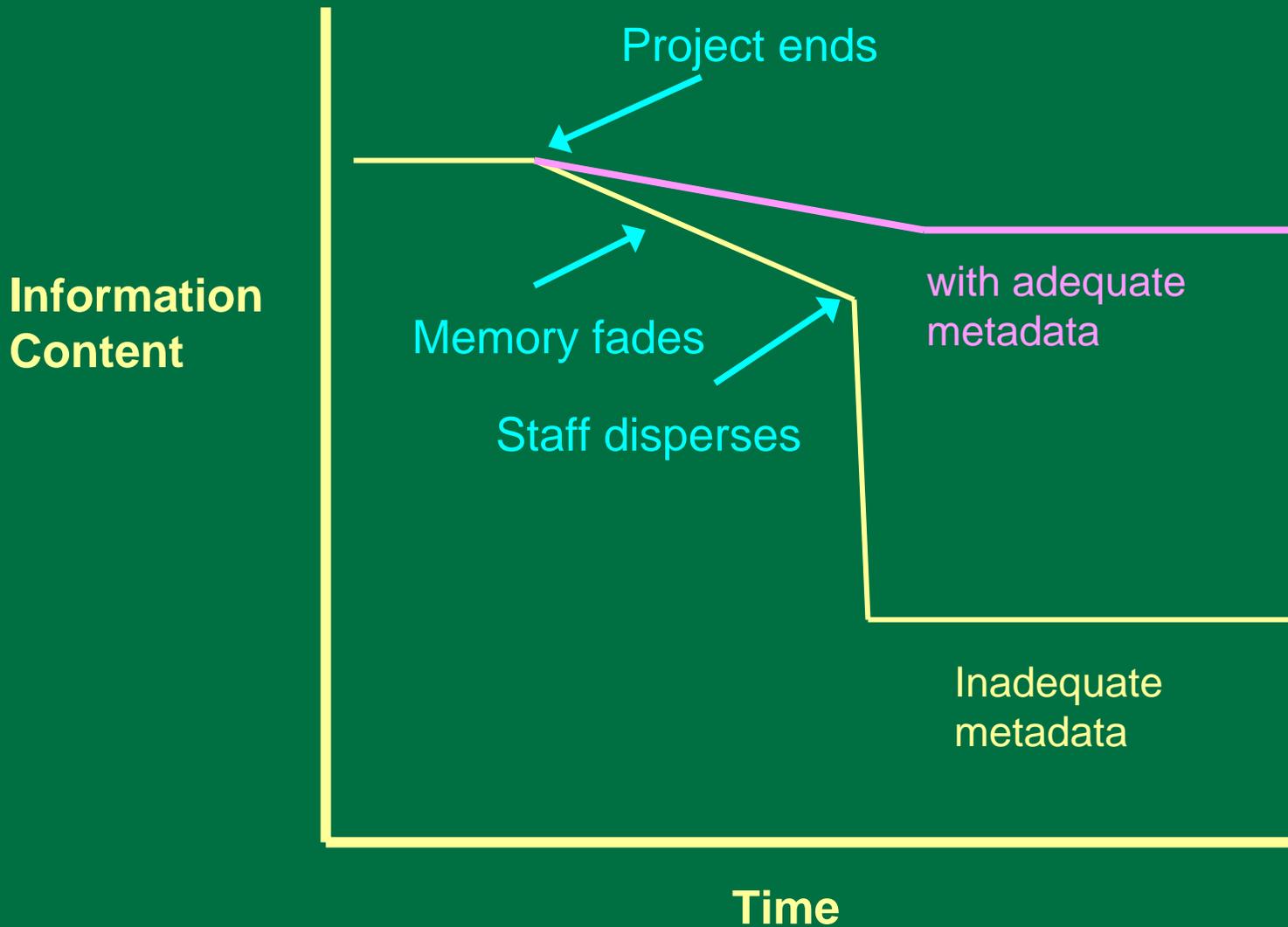
- **Multiple Field and Lab Protocols**
- **Multiple Labs Providing Data**
- **Managing Taxonomic Data for the Long Term**

Taxonomic Data

- **“At the Bench”**
 - Someone assigns a taxonomic identity or name to a specimen or group of specimens (makes an assertion that, “this is that taxon.”)
- **In the Database**
 - We record the assertion (the identity assigned) and information about the assertion (metadata).

Taxonomic Data for the Long Term

- **What We Try To Keep In View**
 - How can we preserve the information content of a taxonomic “result” (assertion) over a very long time-frame (e.g. 100 years)? What information (metadata) do we need to capture now?
 - How can we ensure the immediate usefulness of a taxonomic assertion for people who were not involved in the project?



Taxonomic Data for the Long Term

- **Finding the happy medium**
 - **Metadata requirements too high = too much effort and too much cost. Won't happen anyway.**
 - **Metadata requirements too low = usefulness of data is limited, information content deteriorates too much over time.**

What should we record other than the taxon identity (name?)

- Who (lab, taxonomist, taxonomist certification level, etc)
- When
- Verification/QA information or metrics
- Bench qualifiers (e.g. specimen was damaged)
- Specimen archival / images (if done)

The Trouble With Names

“Genus 1, Species A” in 2030 might not mean the same thing that “Genus 1, Species A” did in 2010.

What !!! ?

The Trouble With Names – Split E.G.

- **2010 Morphologic Key**
 - Genus 1, species A
 - This is the only species in the genus according to the key used.
- **2030 Morphologic Key**
 - The newer key recognizes that species A has been split into 3 species.
 - *Genus1 species A* (note that this name continues)
 - *Genus1 species B*
 - *Genus1 species C*

The Trouble With Names - Analysis

- In 2031
 - An analyst retrieves a data set that has samples with id's based on both keys (a 2010 and a 2030 sample)
 - Sample 1 (2010 key) contains record for
 - Genus1 speciesA
 - Sample 2 (2030 key) contains record for
 - Genus1 speciesA
 - Genus1 speciesB
 - Genus1 speciesC
 - Analyst does not know whether sample 2 is really more diverse than sample 1. Species B, C might have occurred in sample 1, but the 2010 key would not have known of split.
 - Adding the authority to the name won't help

The Missing Metadata Element

To retain all of the information needed to fully qualify the identification, the database needs to record

“Genus 1, species 1 *according to the description found in reference x.*”

Where reference x might be:

- A key
- An original description in the literature
- A museum reference specimen
- Etc.

Timeframe

- High-level requirements gathering – essentially done
- Design/Development
 - Begin October, 2009
 - Version 1 release in summer, 2010

For More Information

Pete Ruhl – pmruhl@usgs.gov

Project Wiki –

<http://privusgs2.er.usgs.gov/display/biodatadb/Home>