Data, Data and More Data:

Managing and Making Sense of Data from PNW Salmon and Habitat Monitoring Programs

Carol Volk
Integrated Status and Effectiveness Monitoring Program

Volk Consulting: Contractor to BPA
Carol.volk@noaa.gov

5/5/2009

- What is data?
- Where is all the data?
- What do we do with it?
- Is it useful?

    ISEMP

    Restoration Tracking Database

- Should we do something with it?
- Steps forward: making old and new data useful

What is data?

Data refers to facts usually collected as the result of experience, observation or experiment.

Data may consist of numbers, words, or images, particularly as measurements or observations of a set of variables.

Data are often viewed as a lowest level of abstraction from which information and knowledge are derived.

~Wikipedia

What kinds of data are out there?

Project Types

-Monitoring data

-Restoration data

Project sub types:

-Status and trend

-effectiveness

-implementation

-project information

What kinds of data are out there?

Project Types

-Monitoring data

-Restoration data

Project sub types:

-Status and trend

-effectiveness

-implementation

-project information

Data generators:

-Multiple agencies

Project scale:

-watershed (graduate projects)

-state-wide assessments

-multi-basins

Project designs:

-BACI

-GRTS randomized

-non-random

What kinds of data are out there?

Project Types

-Monitoring data

-Restoration data

Project sub types:

-Status and trend

-effectiveness

-implementation

-project information

Data generators:

-Multiple agencies

Project scale:

-watershed (graduate projects)

-state-wide assessments

-multi-basins

Project designs:

-BACI

-GRTS randomized

-non-random

Data types:

-spatial

-tabular

Data formats:

-projections, datums

-text files, Excel, databases

What kinds of data are out there?

Project Types
-Monitoring data
-Restoration data

Project sub types:
-*Status and trend*
-effectiveness
-implementation
-project information

Survey types:
-Fish
-Habitat
-Water quality
-Structure

Data generators:
-Multiple agencies

Project scale:
-watershed (graduate projects)
-state-wide assessments
-multi-basins

Project designs:
-BACI
-GRTS randomized
-non-random

Timeframe:
-multi-year projects
-single-year projects
-periodic assessments

Data types:
-spatial
-tabular

Data formats:
-projections, datums
-text files, Excel, databases

Collection methods:
-spawning surveys
-smolt trapping
-habitat evaluations
-multiple water quality parameters

How much data is out there?

3 data generators

2 project types

3 survey types

2 project designs

10 years

2 data types

5 data formats

5 collection methods

10 basins in Washington state

How much data is out there?

3 data generators

2 project types

3 survey types

2 project designs

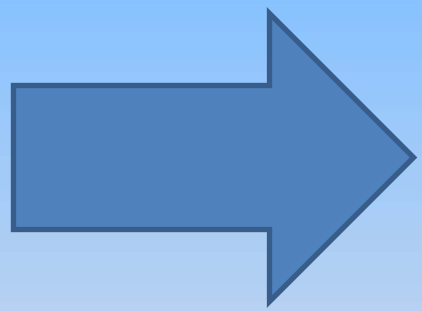
10 years

2 data types

5 data formats

5 collection methods

10 basins in Washington state



180,000
files of data
in Washington
state

HOW: The Data Acquisition Process

The Truth
- Projects in the Field

Project Reporting

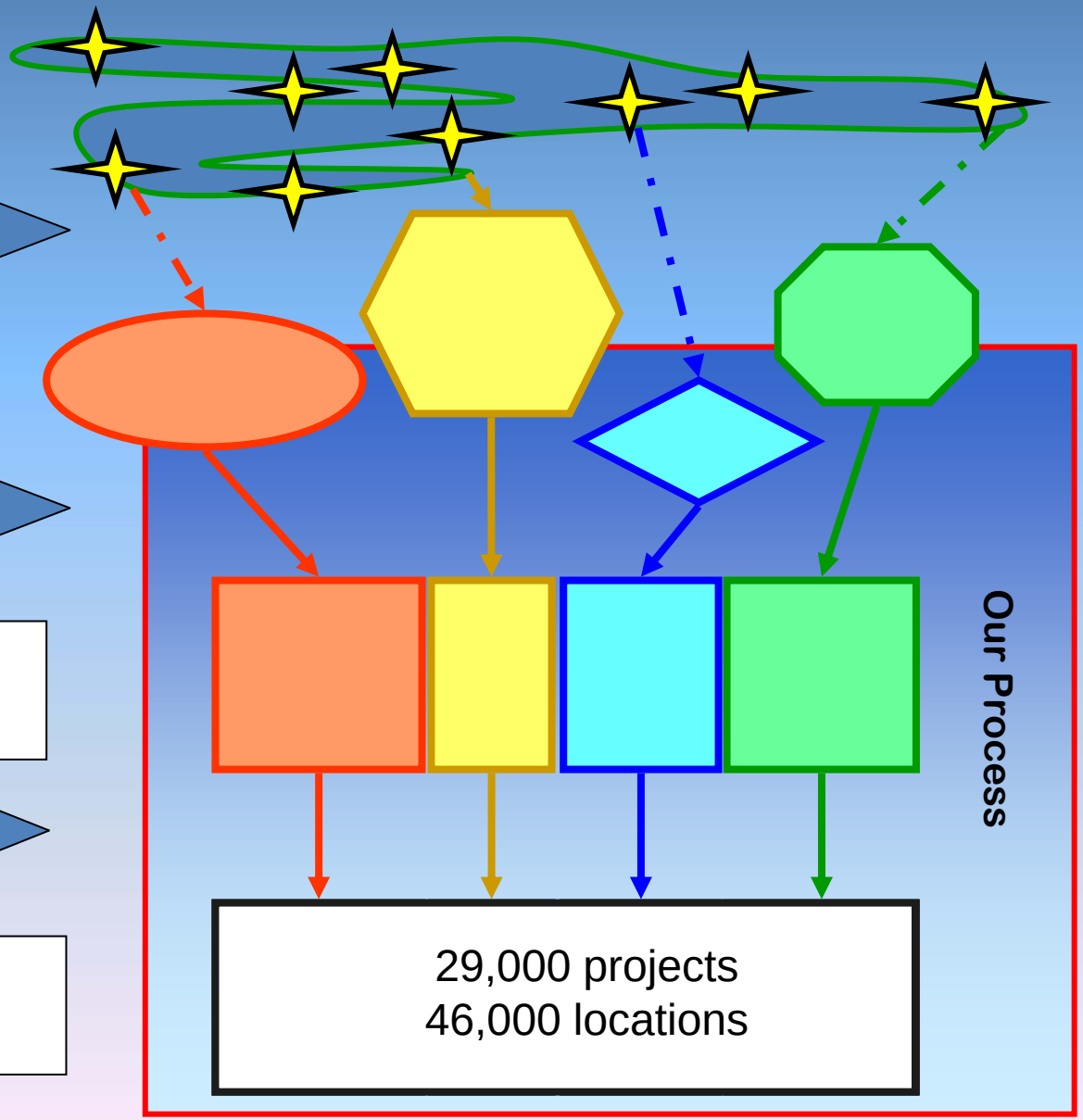
Source Datasets
- Different formats
- Different locations

Data Acquisition

Collected Datasets
- Different formats
- One location

Data Migration

Project Tracking DB
- One format
- One location



Where is all of the data?

- Regional databases
  - EPA STORET
  - DART (Data Access Real Time)
  - USGS (Real time and historic)
  - Metadata portals (REO, UW)

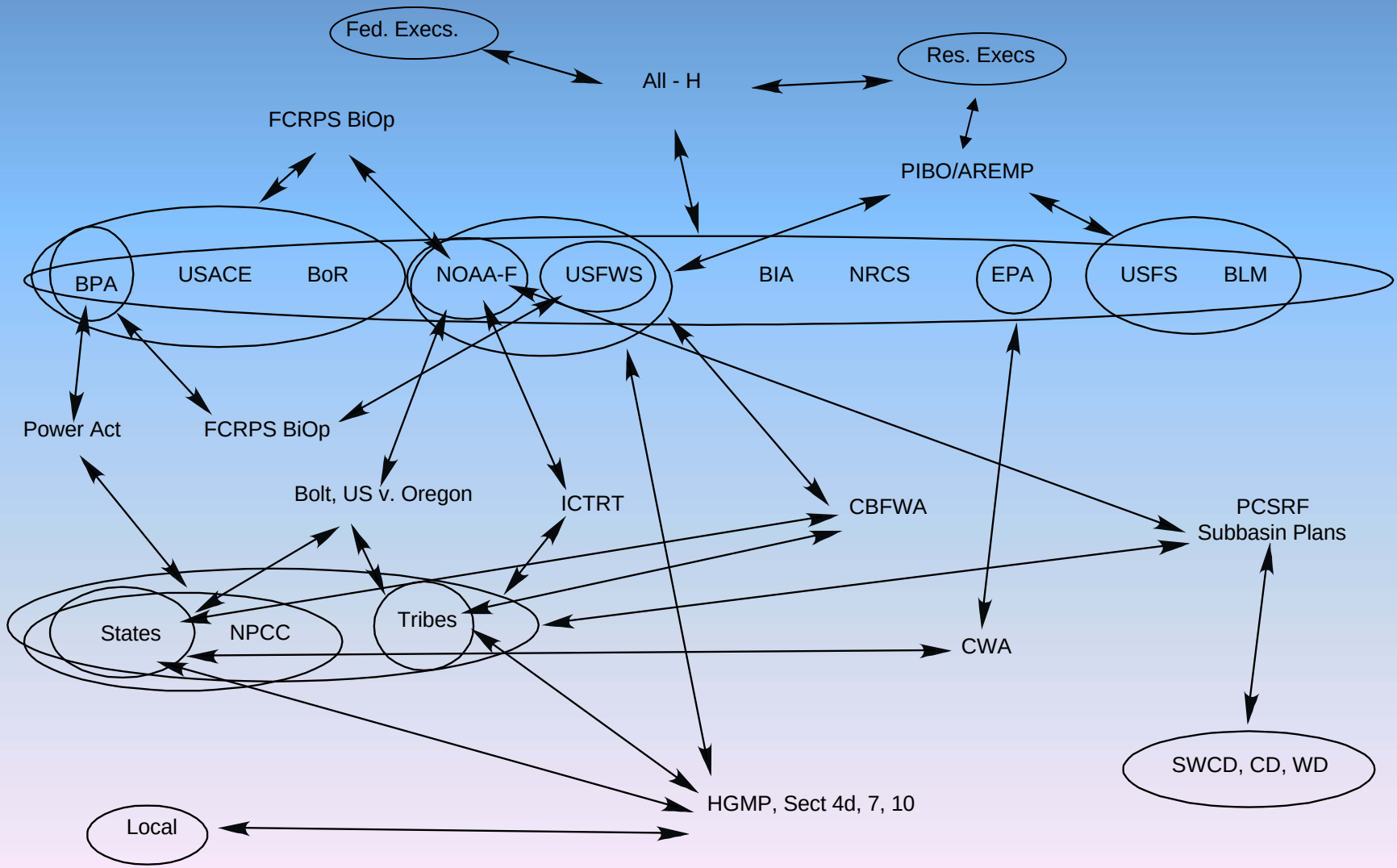
Where is all of the data?

- Regional databases
  - EPA STORET
  - DART (Data Access Real Time)
  - USGS (Real time and historic)
  - Metadata portals (REO, UW)
- Literature
  - Journal articles
  - Agency white papers

Where is all of the data?

- Regional databases
  - EPA STORET
  - DART (Data Access Real Time)
  - USGS (Real time and historic)
  - Metadata portals (REO, UW)
- Literature
  - Journal articles
  - Agency white papers
- “File cabinets”
  - **Possession of the collector**

Coordinating Monitoring, Evaluation and Recovery Planning in the Columbia River Basin



What do we do with data? Processing

- Collect and compile
- Deal with formatting
- Deal with data definitions and crosswalking
- Migrate data into desired format
- Summarize
- Analyze
- Provide instruction

- What is data?
- Where is all the data?
- What do we do with it?
- Is it useful?

  ISEMP

  Restoration Tracking Database

- Should we do something with it?
- Steps forward: making old and new data useful

Is this data useful?

Examples:

1. ISEMP John Day water quality project
2. Data dictionaries and the restoration project tracking database
3. Project coordination: Restoration and project placement in MFJD basin

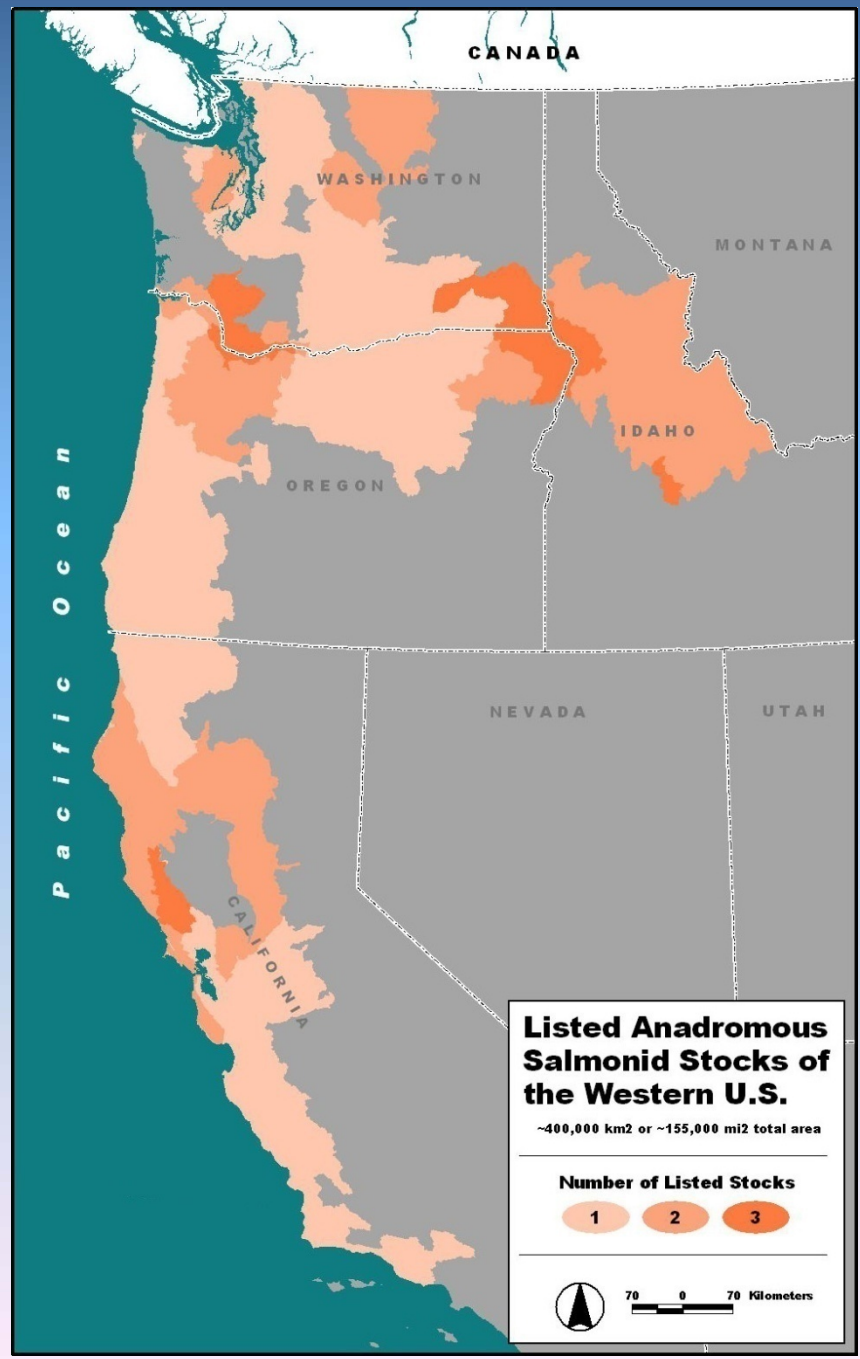
ESA listings of Anadromous Salmonid ESUs

-Recovery and Viability Criteria being developed

-Recovery Plans being developed

-Implementation of Recovery Actions underway in many areas

-5 Year Status Review cycle initiated



Integrated Status and Effectiveness Monitoring Program ISEMP

www.nwfsc.noaa.gov/isemp

Project goal is to develop a process for creating and evaluating monitoring programs for listed salmon

Developing:

-experimental designs for restoration projects

-experimental designs for status and trend monitoring

-metrics and indices

-plans for supplementing existing monitoring programs

-designs for a variety of life stages (smolt, juveniles, adults)

-designs impacted by variety of habitat variables (reach to landscape)

-DATA MANAGEMENT

ISEMP objective #1: Assess monitoring data in pilot basins

- Figure out what data is available in pilot basins
- Compile data
- Summarize data
- Evaluate data gaps
- Offer direction to monitoring programs

Compilation of Historic Data in the John Day Basin

Data collection in 2005-2006:
~80 hours in 3 month period

Processing data:
~1300 hours over 12 months 2006 (formatting)
~2000 hours in 2007-2008 (metadata updates)

Analyze data:
-ONGOING

What we found…

- Large quantity of historic data
- Multiple data types
  - water quality
  - habitat
  - fish abundance
- No consistent data format
- No central data storage within agencies
- Agencies struggling with multiple reporting responsibilities

30,000 Stream Temperature Records

Then:
-15 collecting agencies/groups

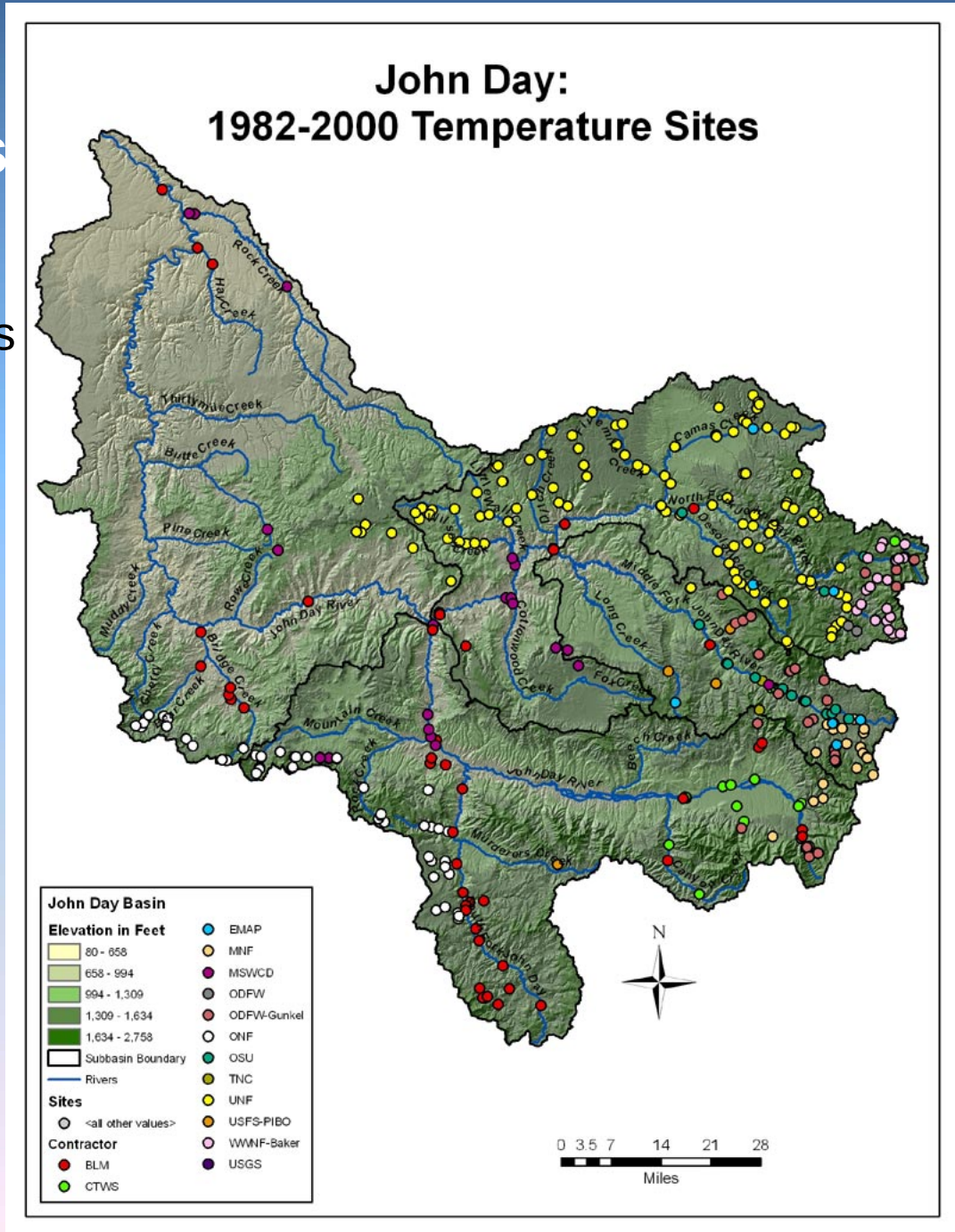
-20 years of data

-300 locations

-0 coordination and standardization

Now:
-1 Data Dictionary

-1 Water quality database



Limiting Factor for any M&E Program

‘Messy’ field data

Regional data analysis

-large volume

-multiple types

-multiple sources

-multiple formats

-centralized source

-standardized format

-metadata

-processing tool

-summary analysis

-distribution

ISEMP Data Management Strategy

- Data Storage and Analysis
  - RME Status and Trend Monitoring Database
  - Metadata tracking
    - Protocol manager tool
    - Data dictionary
- Data Processing and Reporting
  - Database templates
- Data Distribution
  - Salmon Data Management website
  - ISEMP website
  - Geo-databases

Why a data dictionary for salmon habitat restoration?

- Scale of salmon recovery- larger than the analysis scale the data were created for

- Need to combine current and historic data

↓

Necessitates the creation of tools to handle disparate data

Example #2: Creating a data dictionary for habitat restoration data:

Bottom up approach :

1. Look at the universe of available project data

Sources (federal, state, private etc.)

Project Attributes (time, contact, title, type, description)

Location Attributes (state, county, lat/long, stream)

2. Decide what information is important to the project purpose

Use the data to formulate future restoration and monitoring plans

Who, how, what, where, when, why

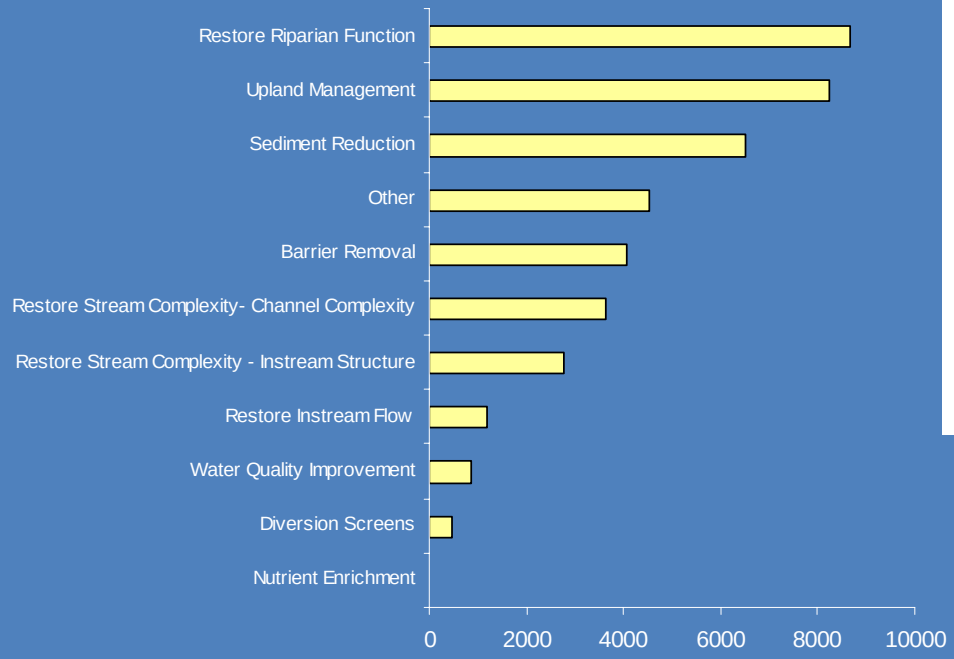
3. Define that information and resolve it into a common language using crosswalks

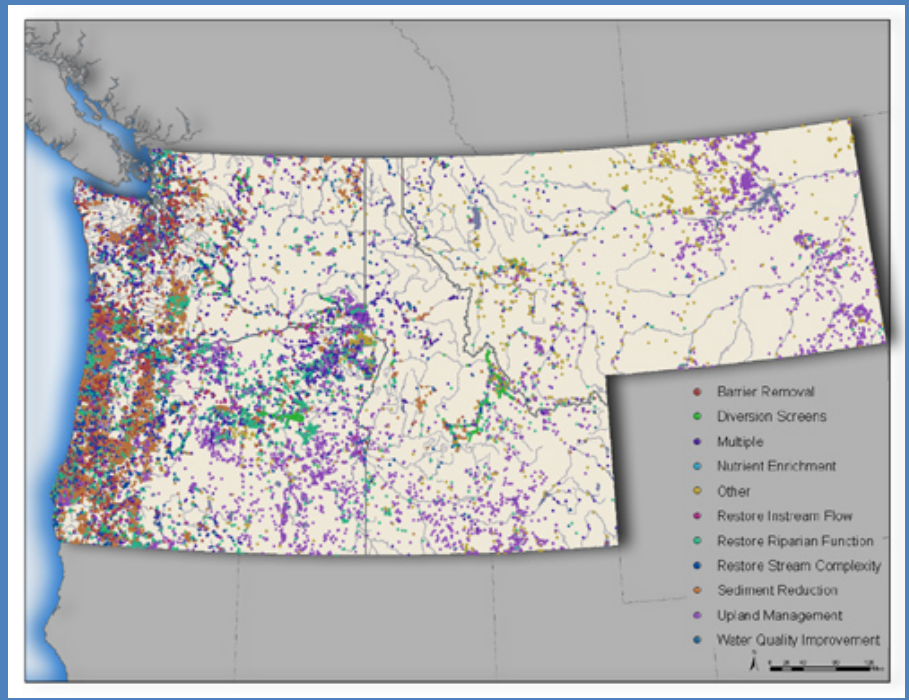
Need for effectiveness monitoring

Project type and location standardization

Pacific Northwest Salmon Habitat Project Database

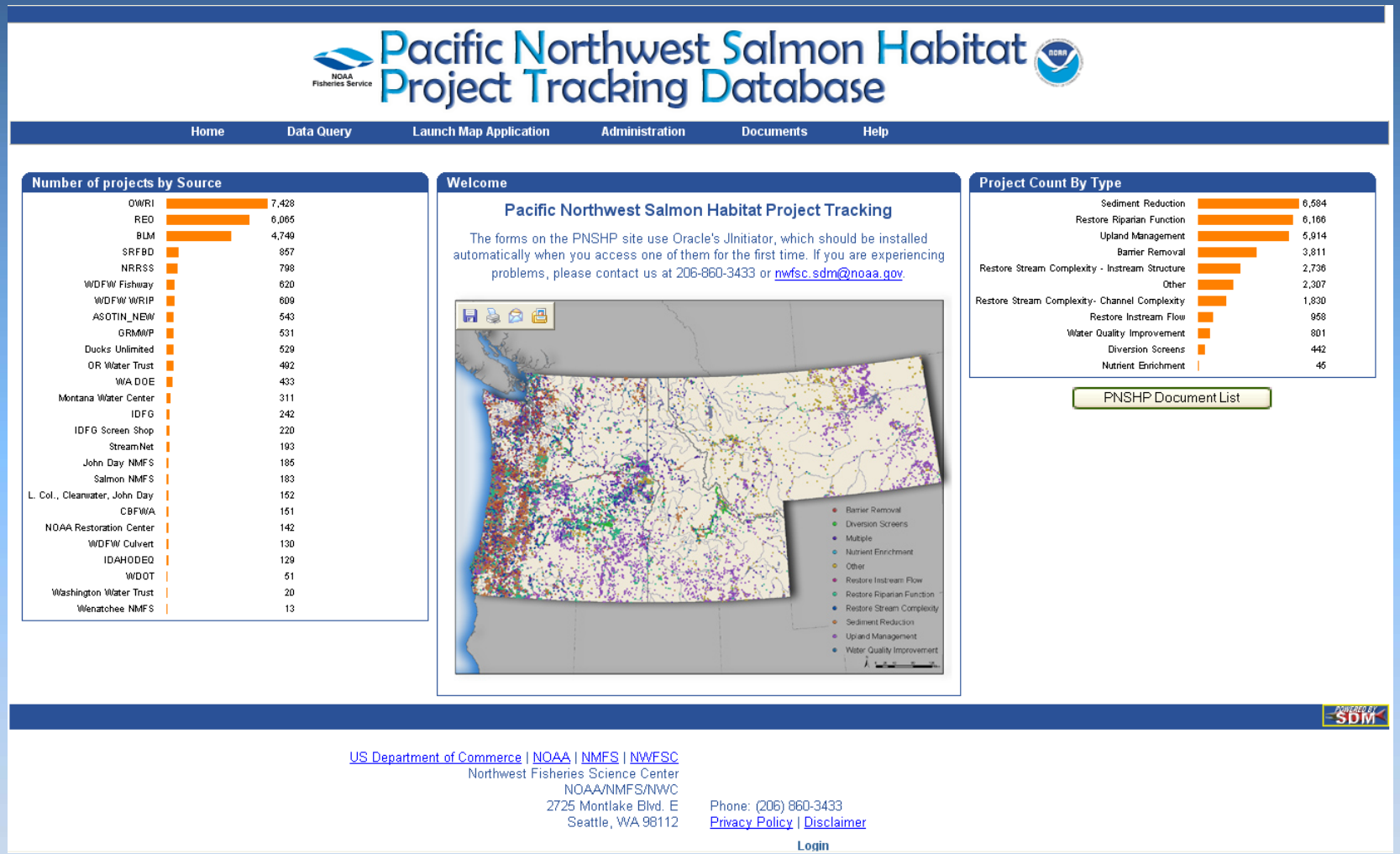
- 26 data sources
- 29,000 projects at 47,000 locations in WA, OR, ID, MT





http://webapps.nwfsc.noaa.gov/pnshp

(Katz et al. Restoration Ecology, 2007)



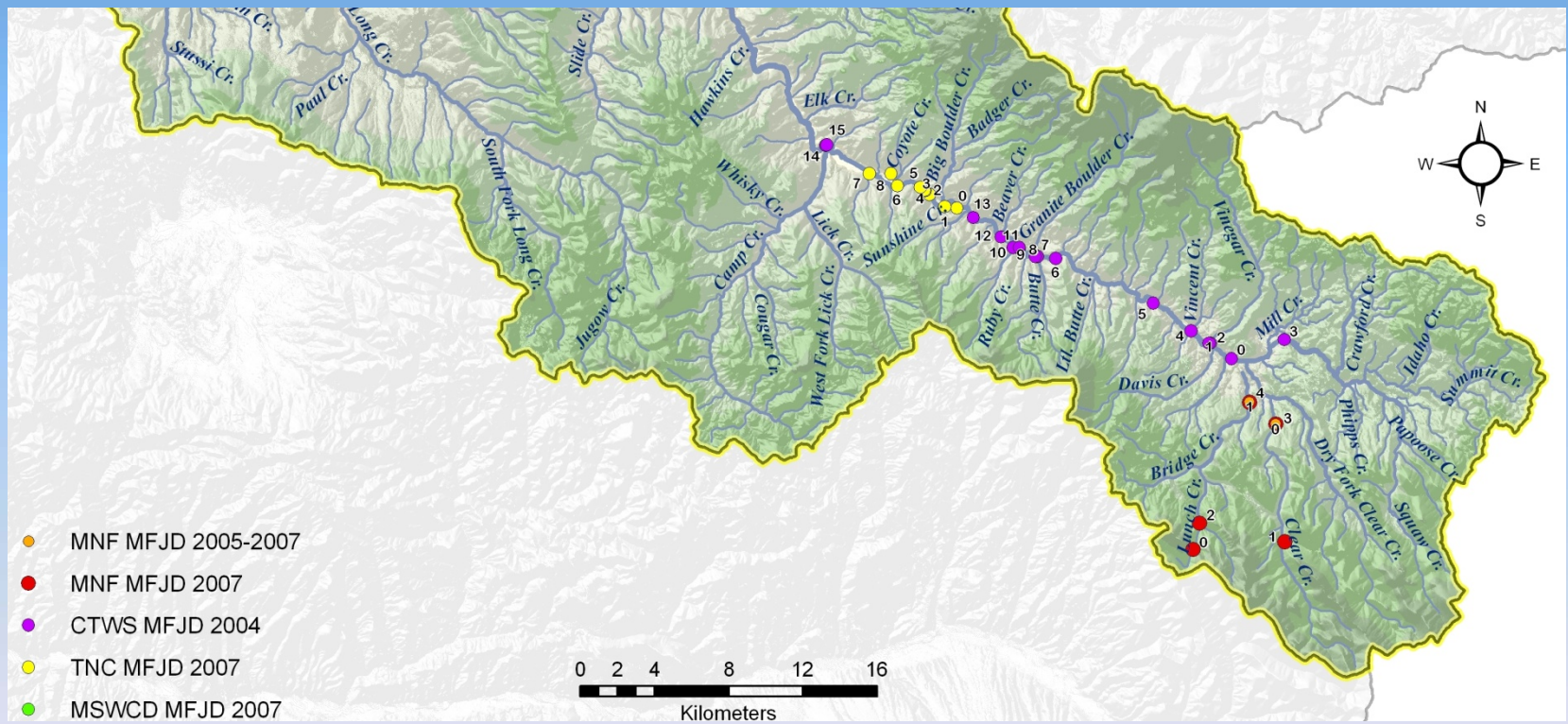
25,500+ Projects

40,000+ Locations

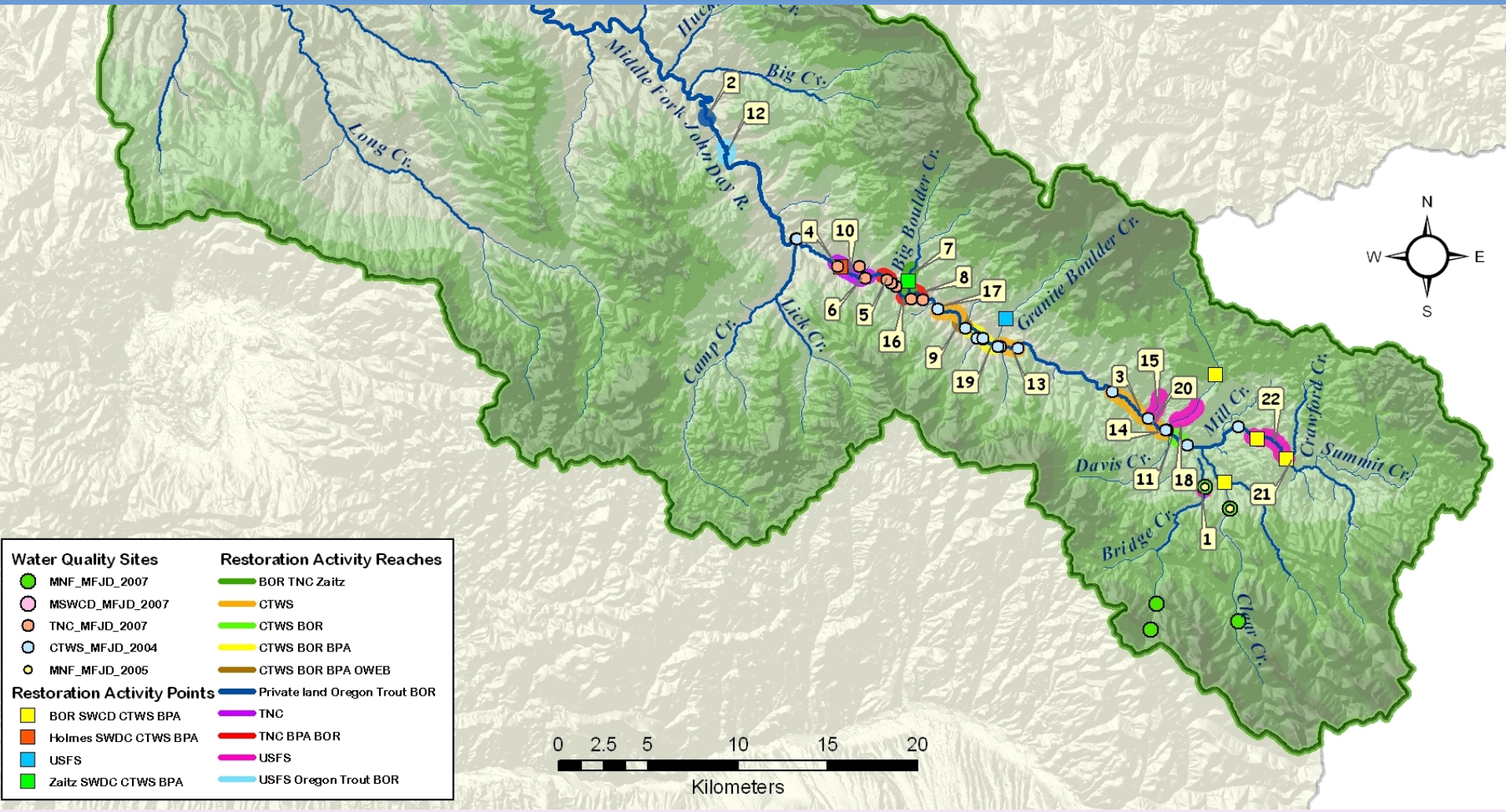
This Project was driven by the need for <u>monitoring</u> but very little monitoring data is available for restoration projects

Develop new tools to describe monitoring activities using a standardized set of terms

IMW restoration and temperature logger site coordination



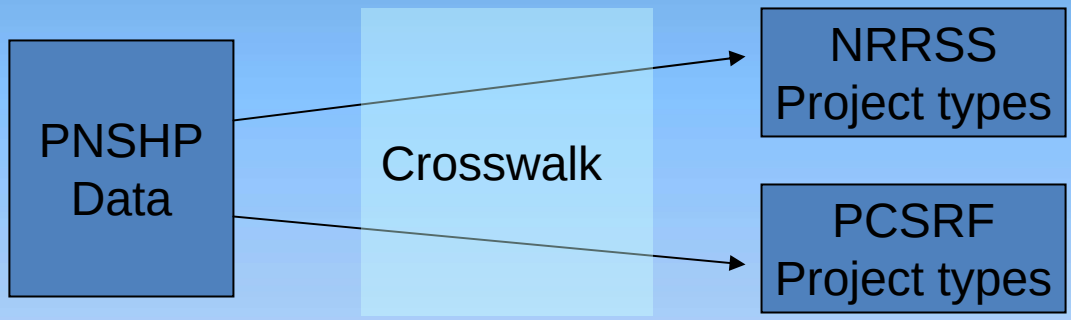
Restoration and logger sites



Should we use all of this data?

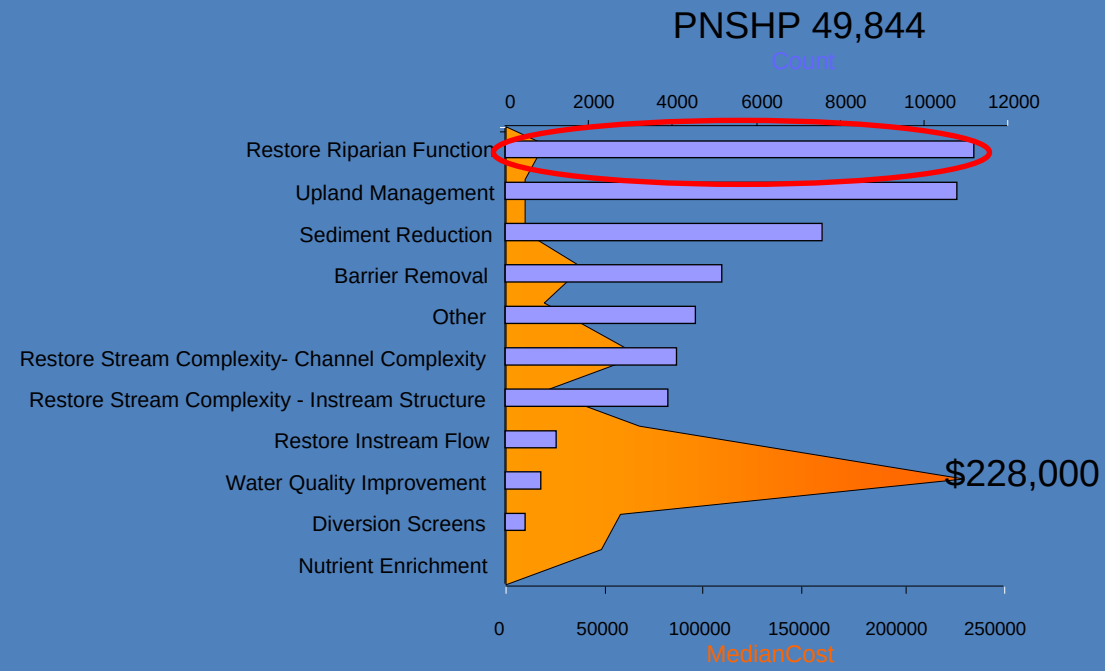
- What information is available about the data?
  - Can you translate it?
  - What was the study design?
  - What was the purpose of the study?
- Is data dense enough for the question of interest?
- What is the quality of the data?
- How long will it take to assemble the data?
  - Data requests?
  - Formats?
- Can you trust it?

How do different data dictionaries affect project metrics (type, cost)?

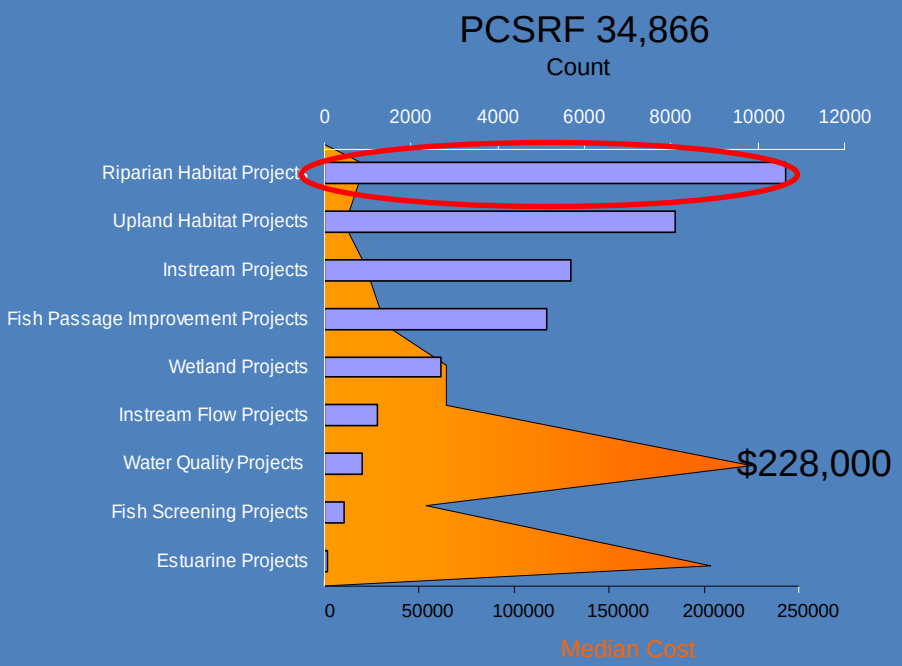


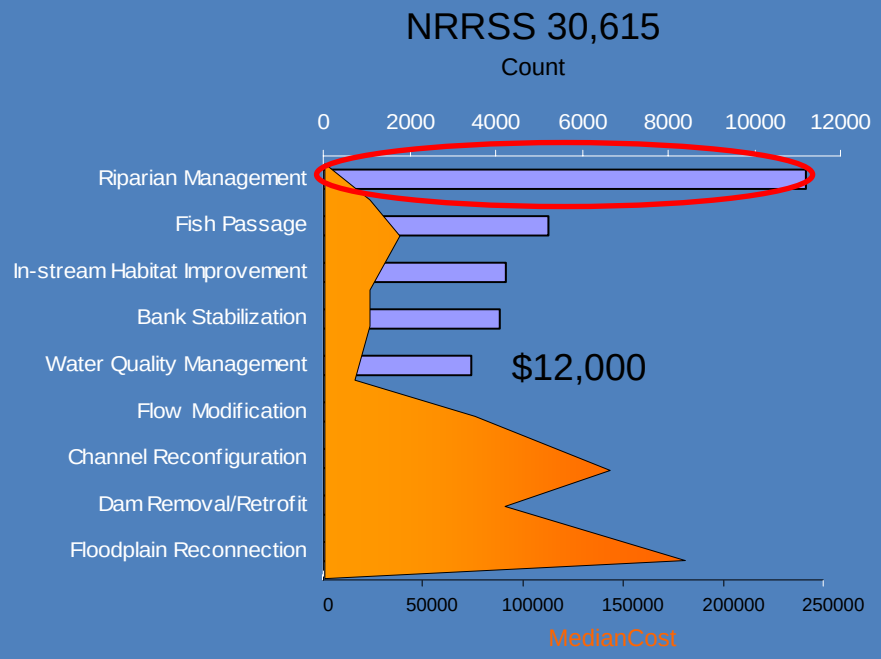
Crosswalks allow us to relate one data dictionary to another

PNSHP Type	PNSHP Subtype	PNSHP Type Definition	NRRSS Type	NRRSS Type Definition
Sediment Reduction	All Subtypes	*Projects that diminish excessive sediment transport into streams*	Water Quality Management	Practices that protect existing water quality or change the chemical composition and/or suspended particulate load. Remediation of acid mine drainage falls into this category as does CSO separation. Excludes urban runoff quantity management (see Stormwater Management).
Water Quality Improvement	All Subtypes	*Projects that improve stream water quality parameters.*		



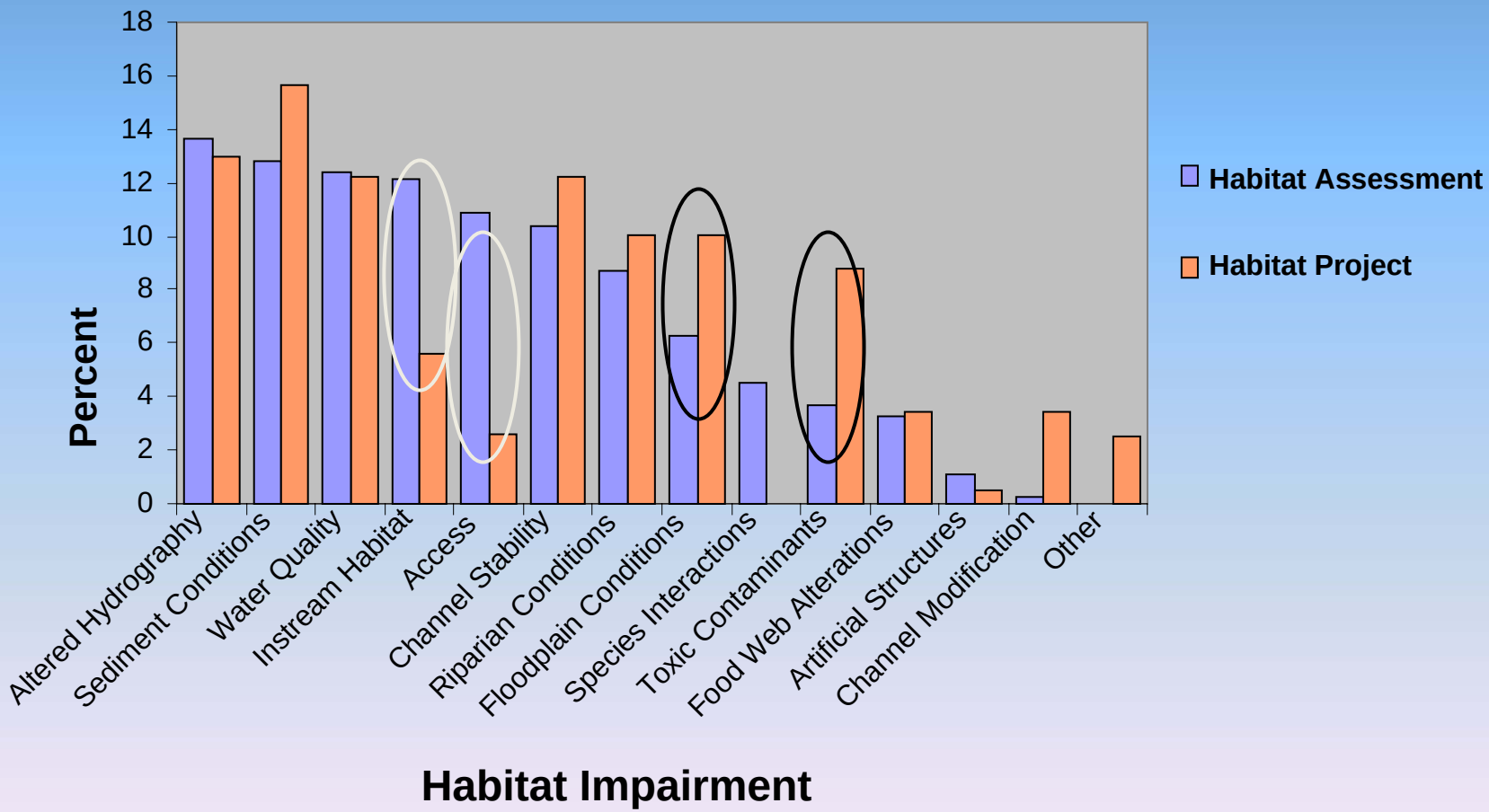
Does the choice of data dictionary change the results?



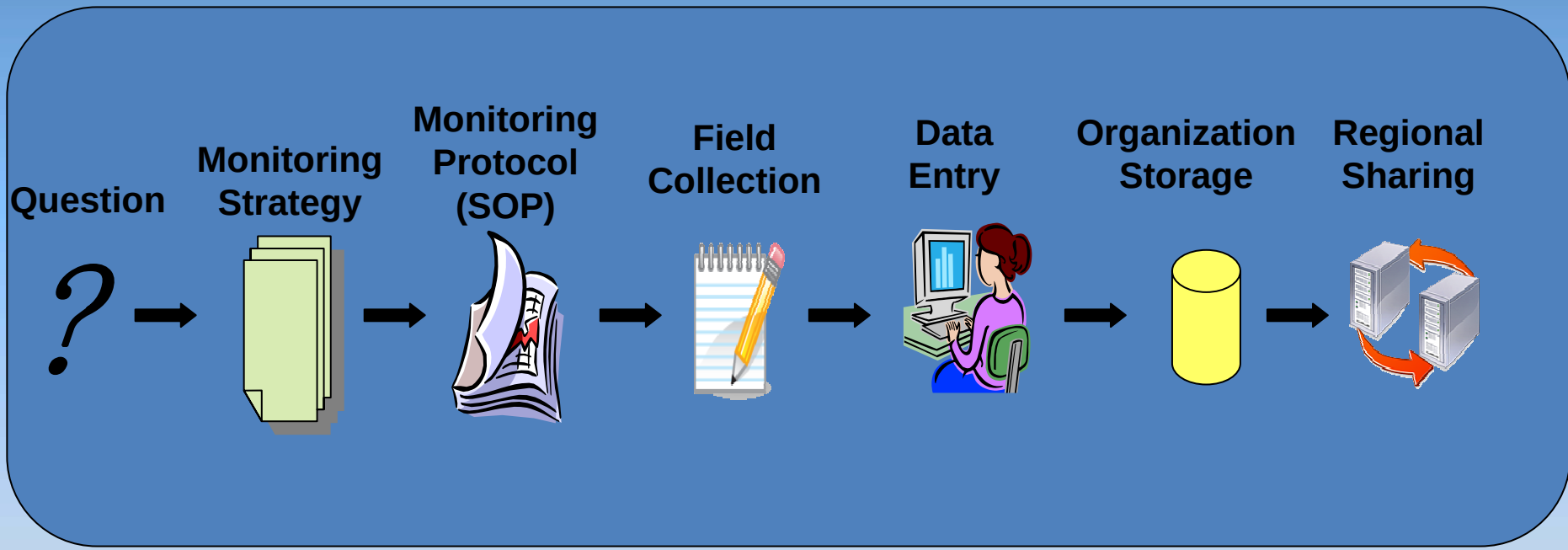


Level of detail of data dictionary should match the intended use

Do restoration projects address habitat need?



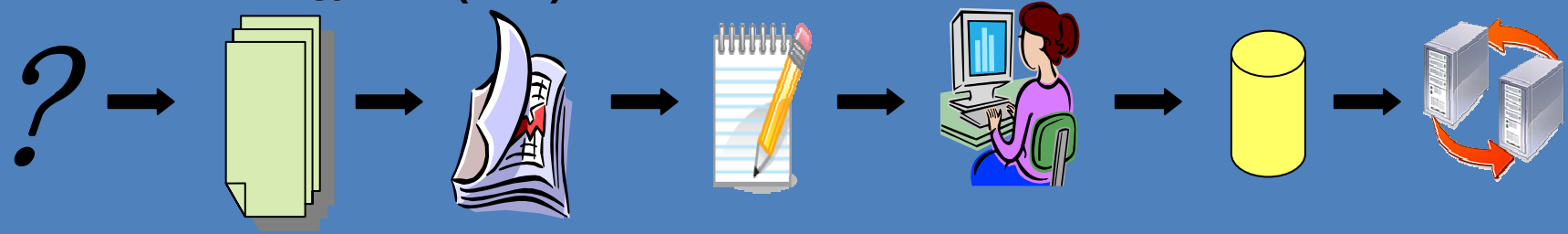
Work Flow



Organizations use workflows to produce products

Work Flow

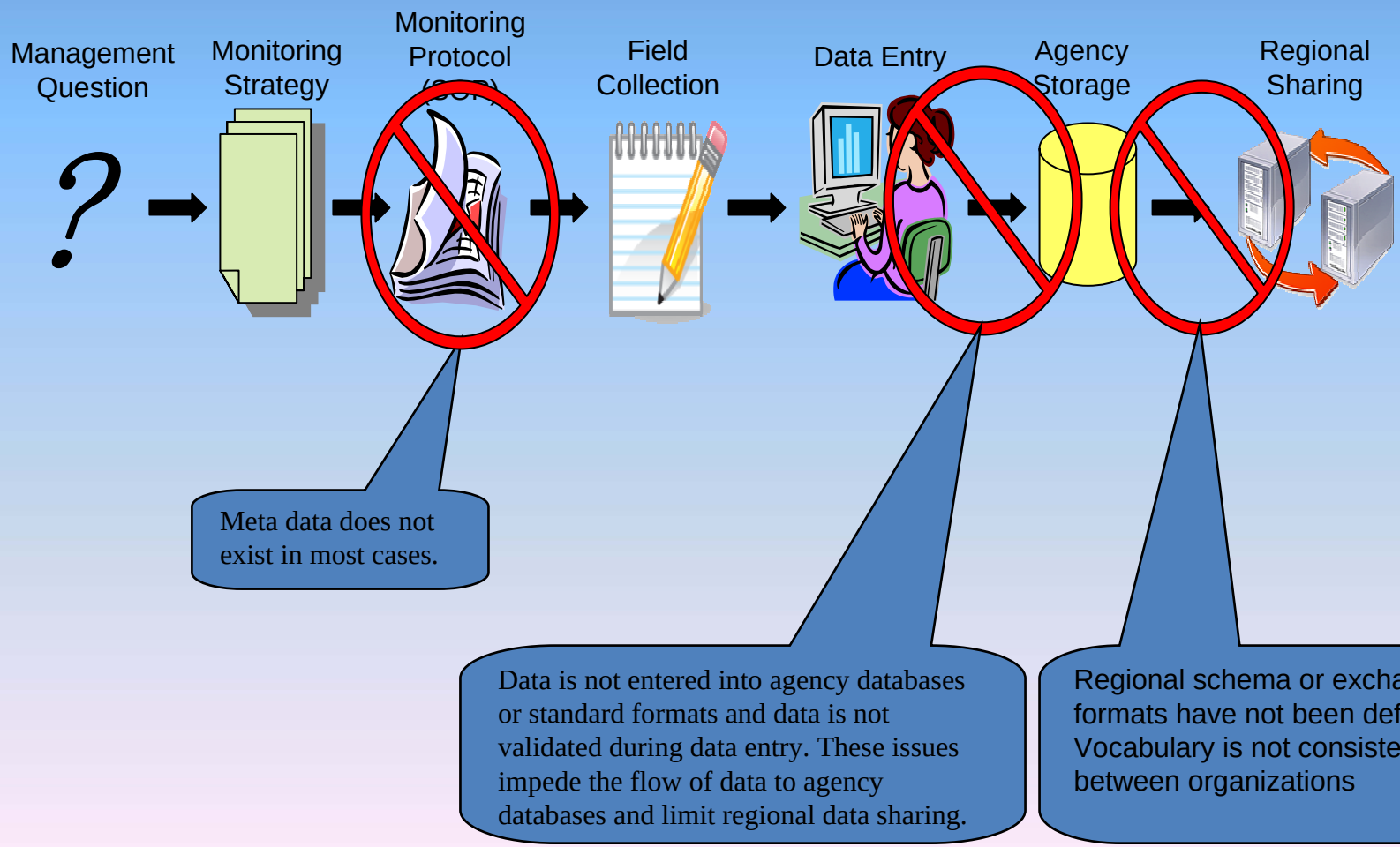
Question → Monitoring Strategy → Monitoring Protocol (SOP) → Field Collection → Data Entry → Organization Storage → Regional Sharing



Products

- Local Analysis
- Agency Analysis
- Regional Analysis

Work Flow Gaps



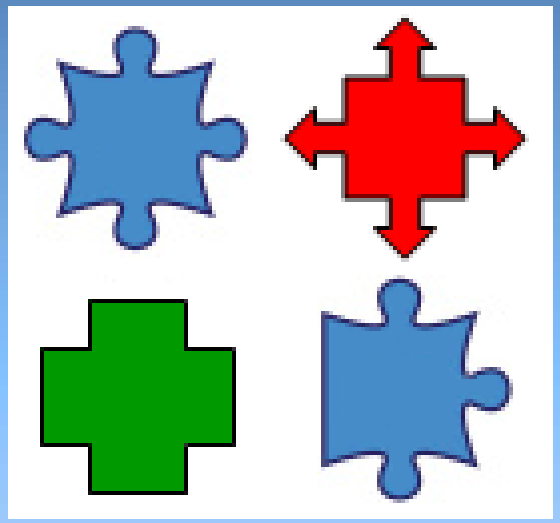
Road blocks

- Disparate data dictionaries
- Gaps data flow
- Objective issues:
  - Evolving questions and technology
  - Short term goals: get answers now
- Changing Personnel
  - Current structure relies heavily on “Legacy knowledge”
- Lack of training support/time

Making data more useful

- Why?
  - Easy to remember what you did
  - Increases chance that data will be used ‘properly’
  - Greater good/greater success of field

Data Sharing



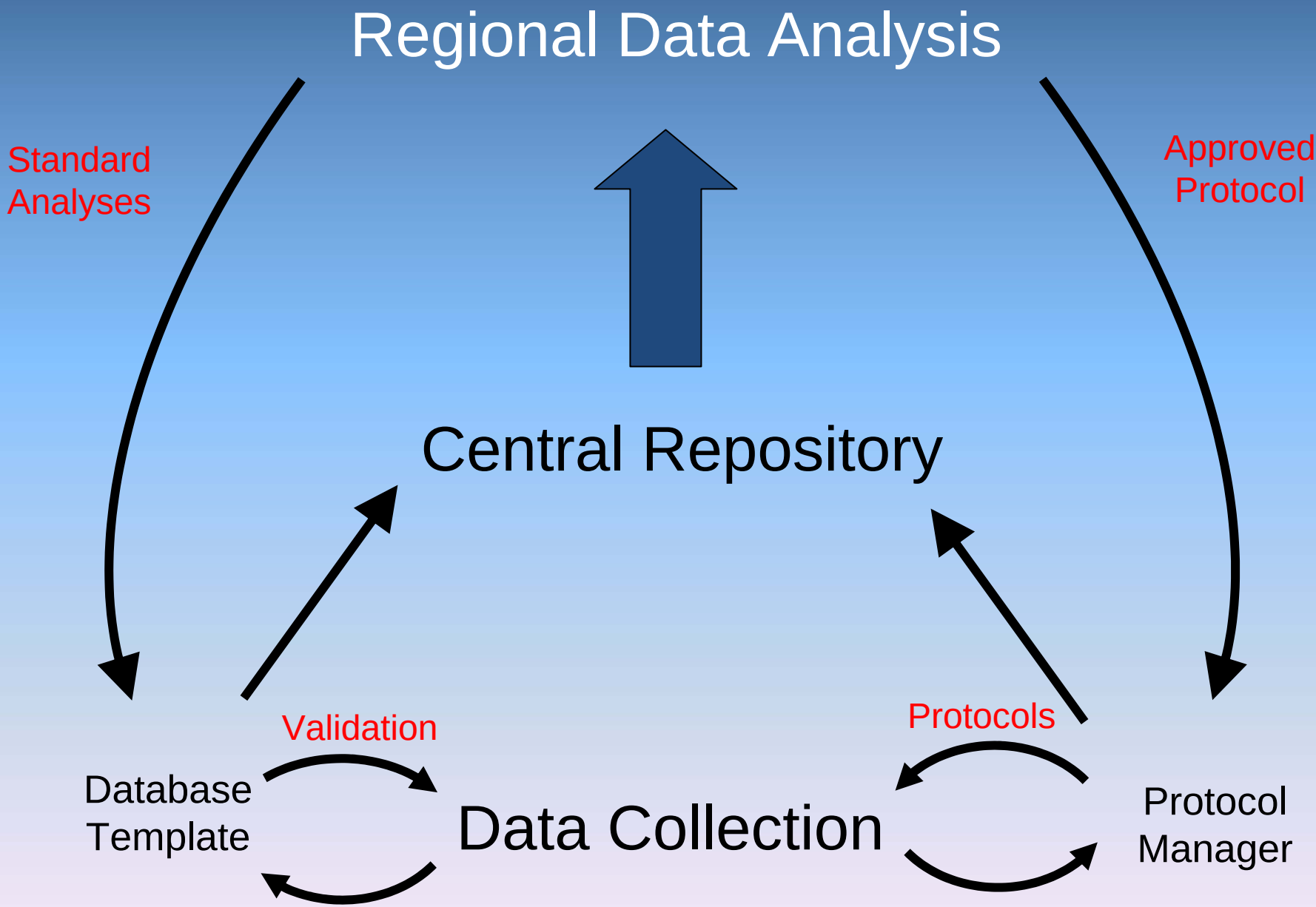
- Well documented
- Discoverable
- Accessible

- Inconsistent Formats
- Inconsistent Methodologies

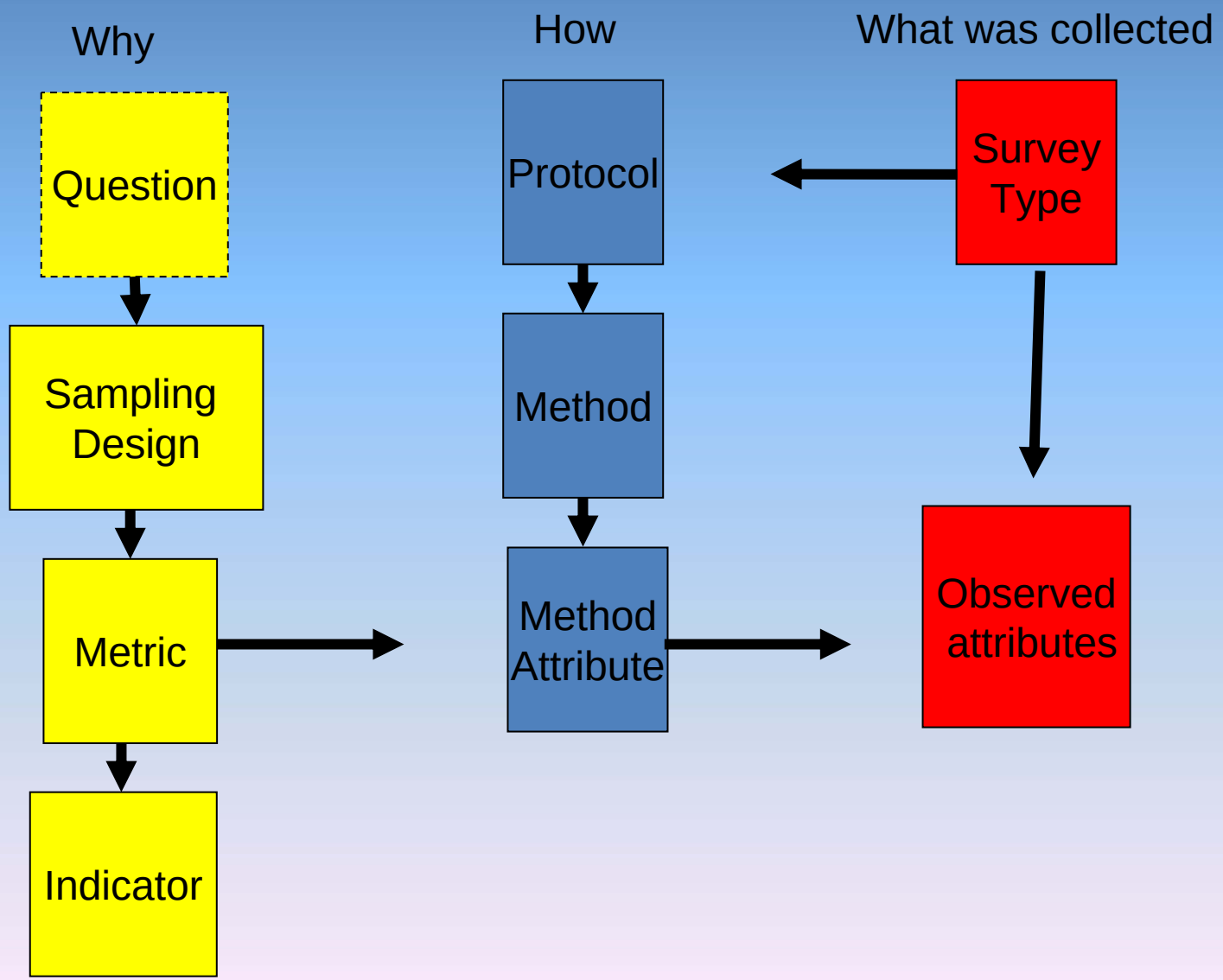
Data Interoperability

- Well documented
- Discoverable
- Accessible

- Consistent Formats
- Comparable Methodologies


Regional Data Analysis
Standard Analyses
Approved Protocol
Central Repository
Validation
Protocols
Database Template
Data Collection
Protocol Manager

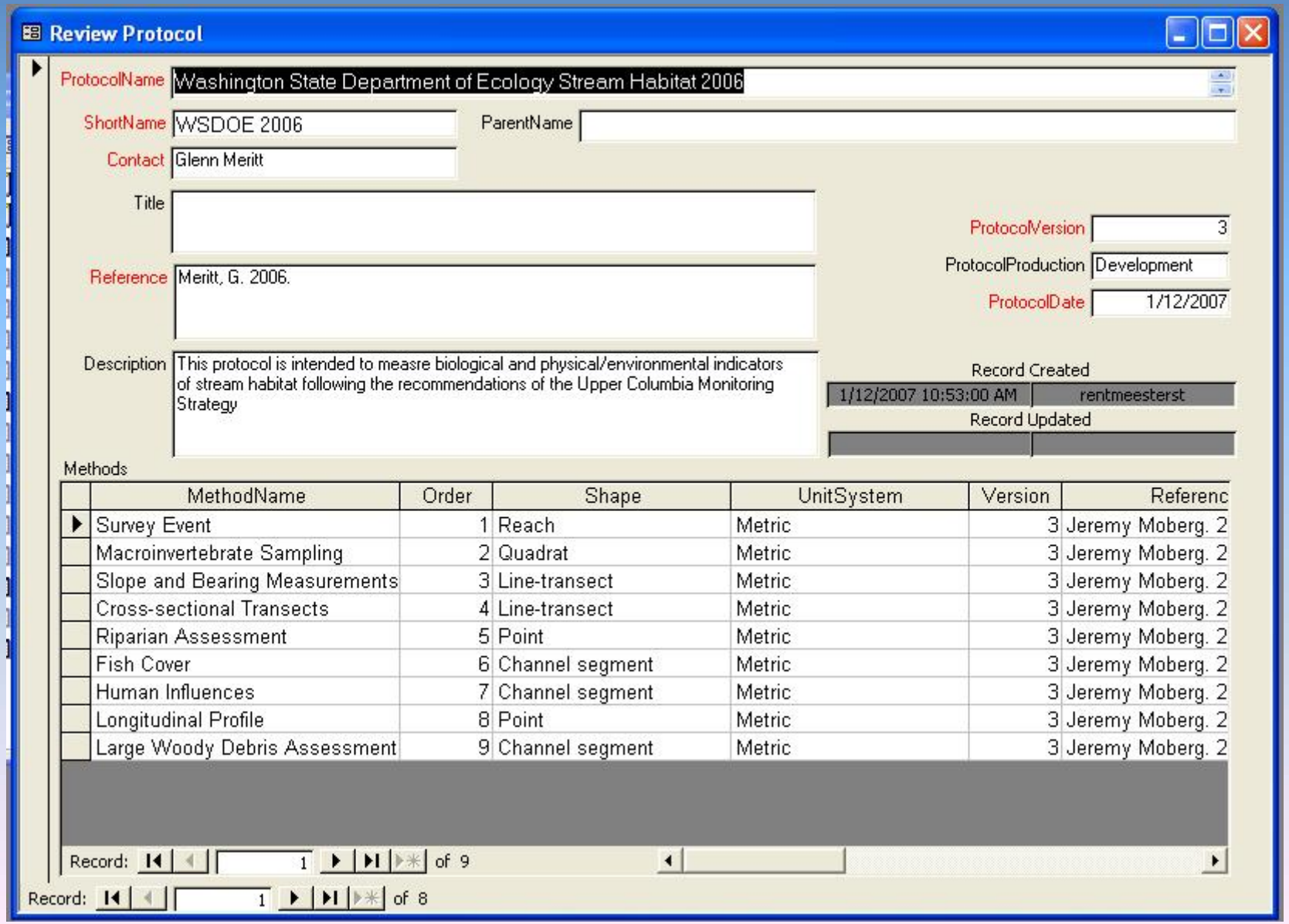
What needs to be recorded and managed?



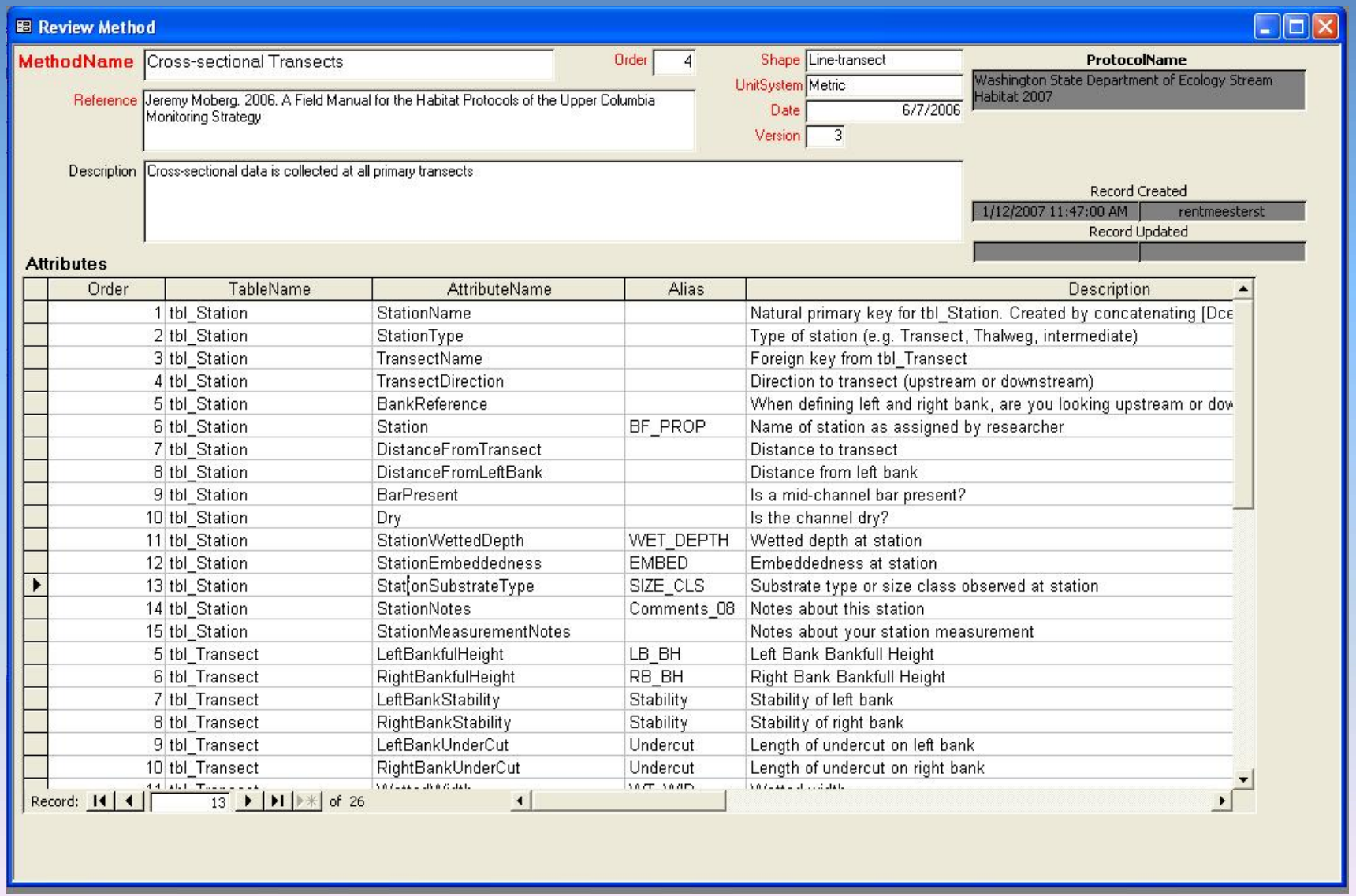
How do we do it?

- Metadata
  - Data dictionaries (PNAMP, NOAA-Fisheries)
  - Sampling designs
  - Protocols (References, Methodology)
    - Methods
    - Codes
- Managing observed tabular data
  - Normalized structure
  - Consistent
- Geodatabases (e.g. manages projections)

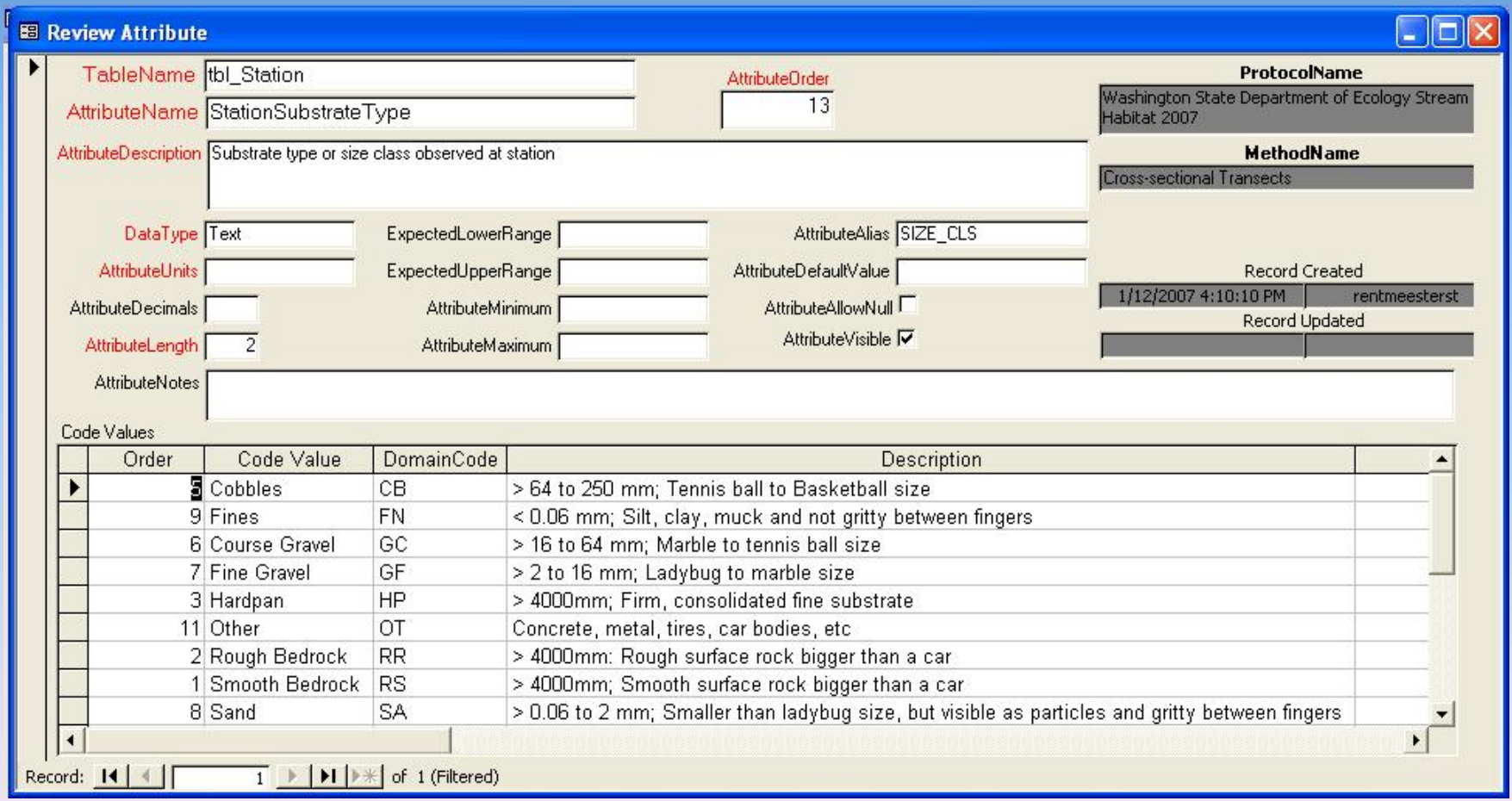
Metadata: Data Dictionaries can manage the community's data flow -- Protocols



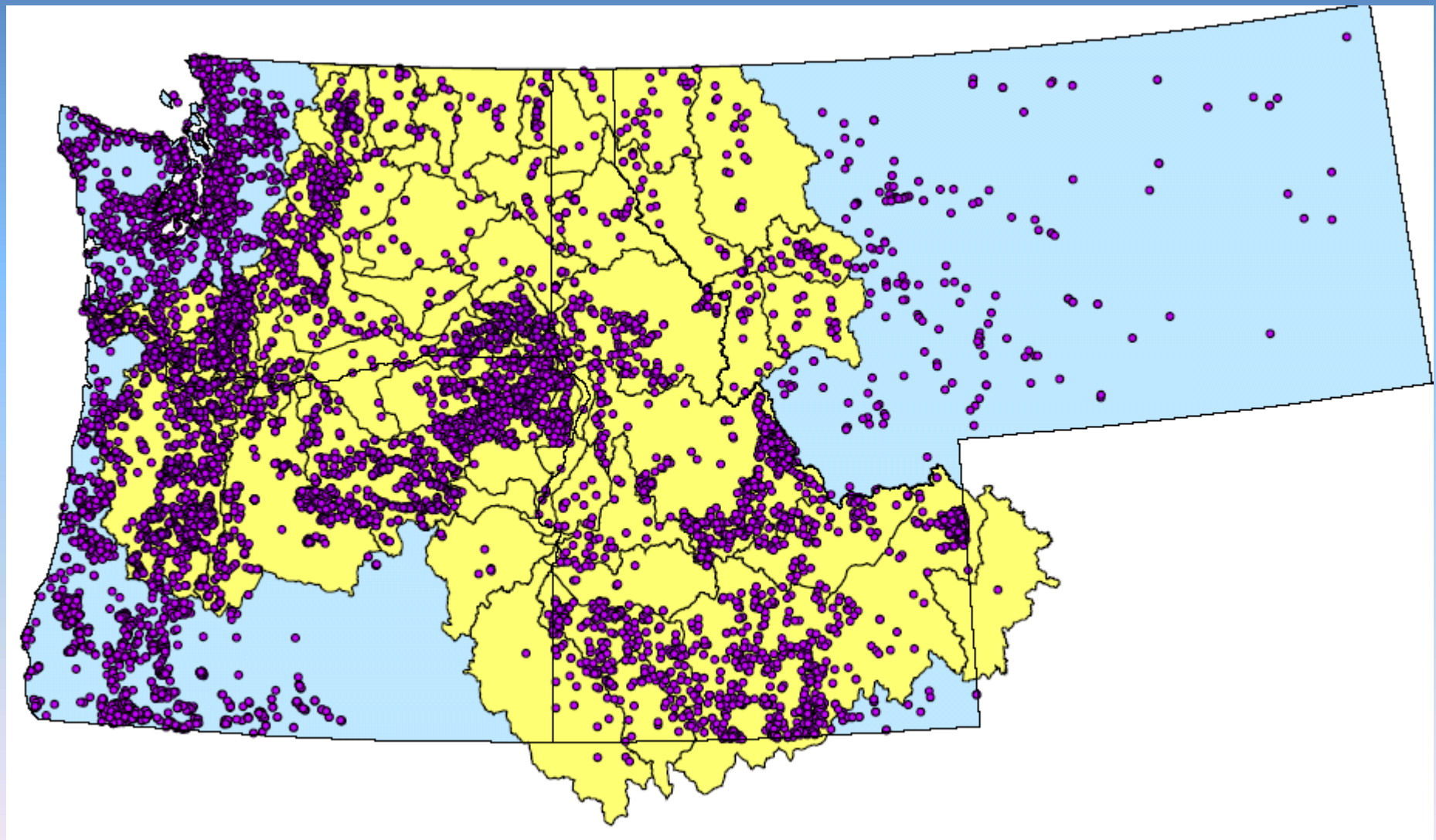
Metadata: Data Dictionaries can manage the community's data flow -- Methods



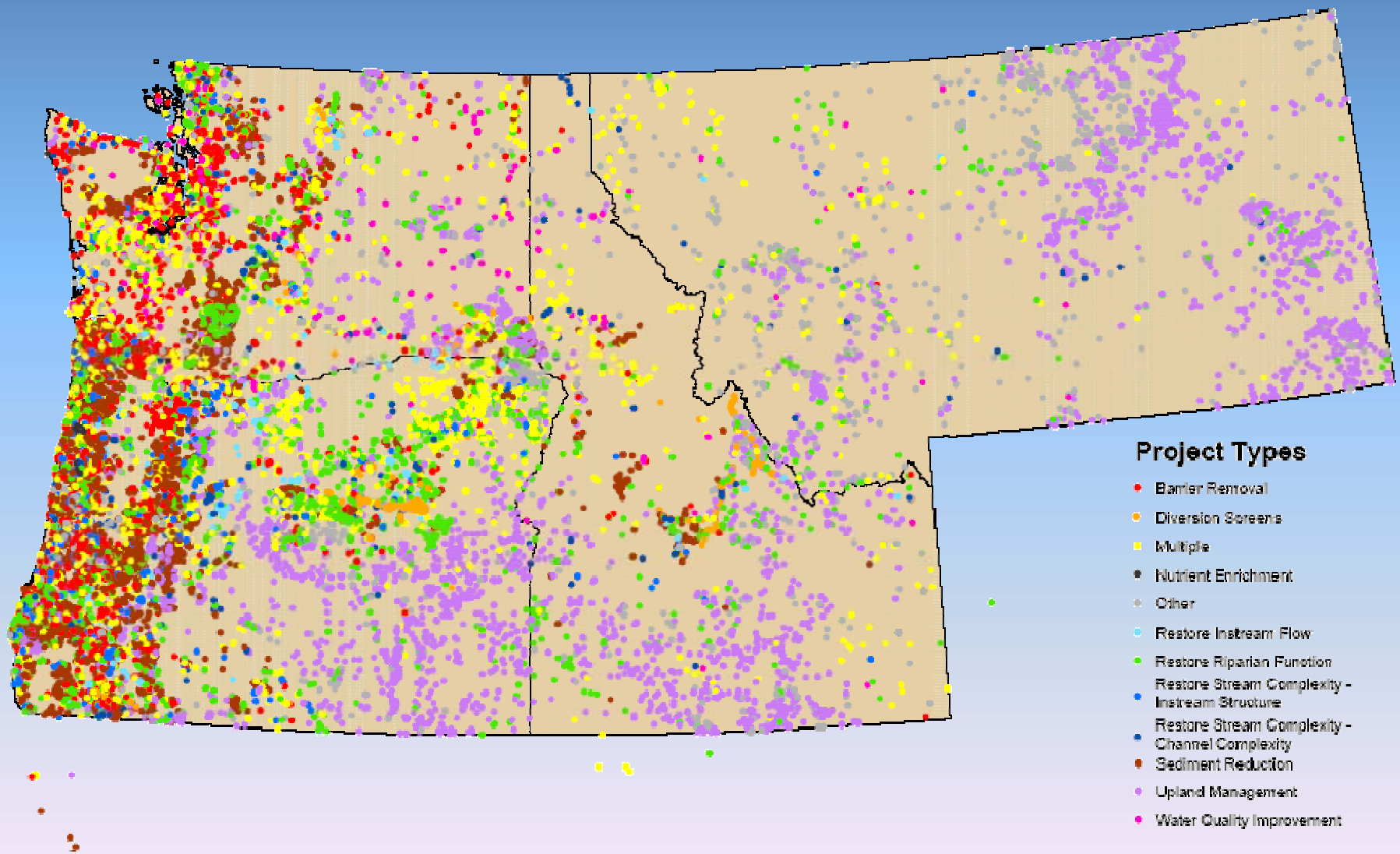
Metadata: Data Dictionaries can manage the community's data flow -- Attributes



25,000 Habitat Restoration Projects Across the Pacific Northwest



25,000 Habitat Restoration Projects Across the Pacific Northwest with a Data Dictionary



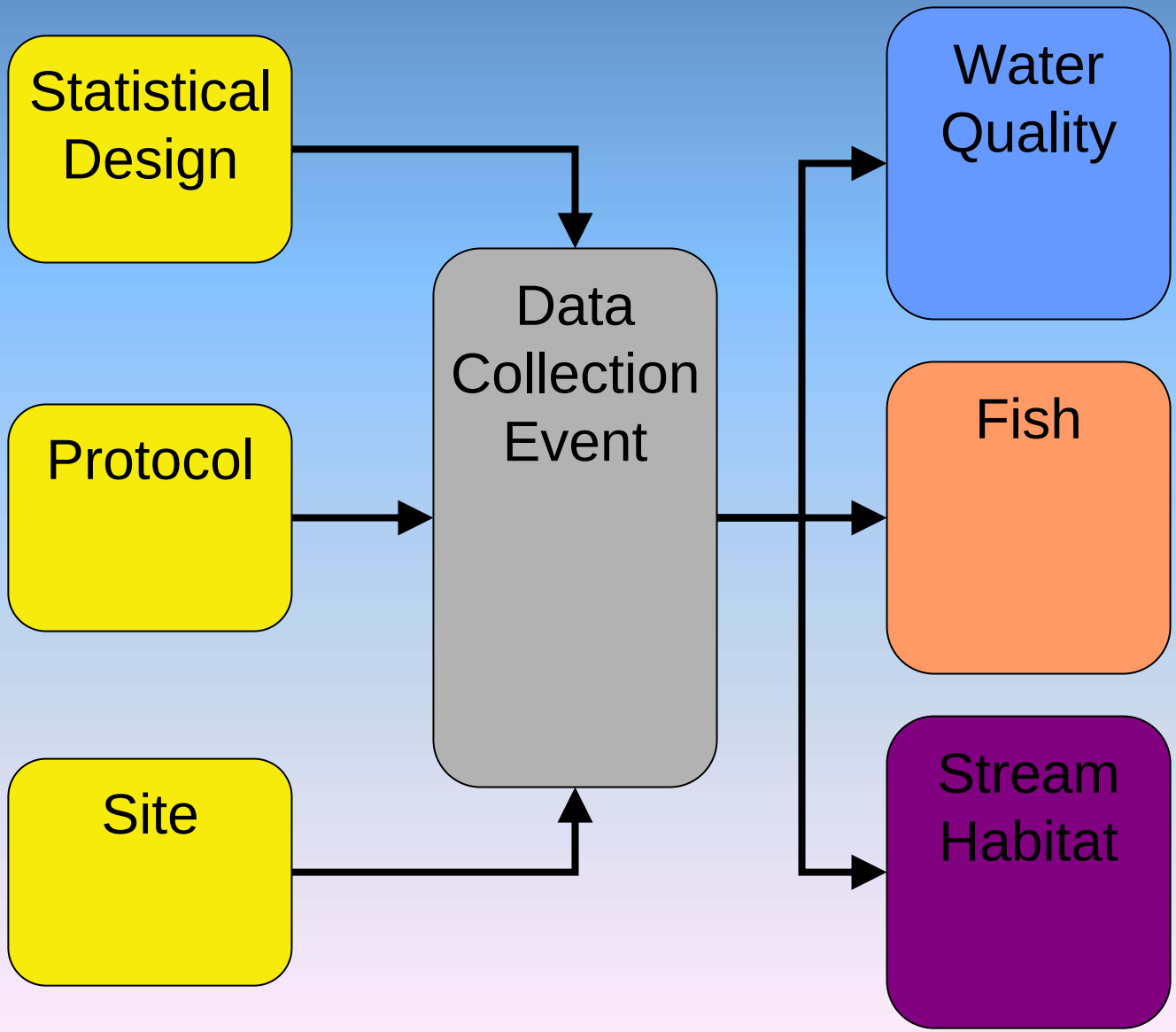
The Salmon World with and without Data Dictionaries

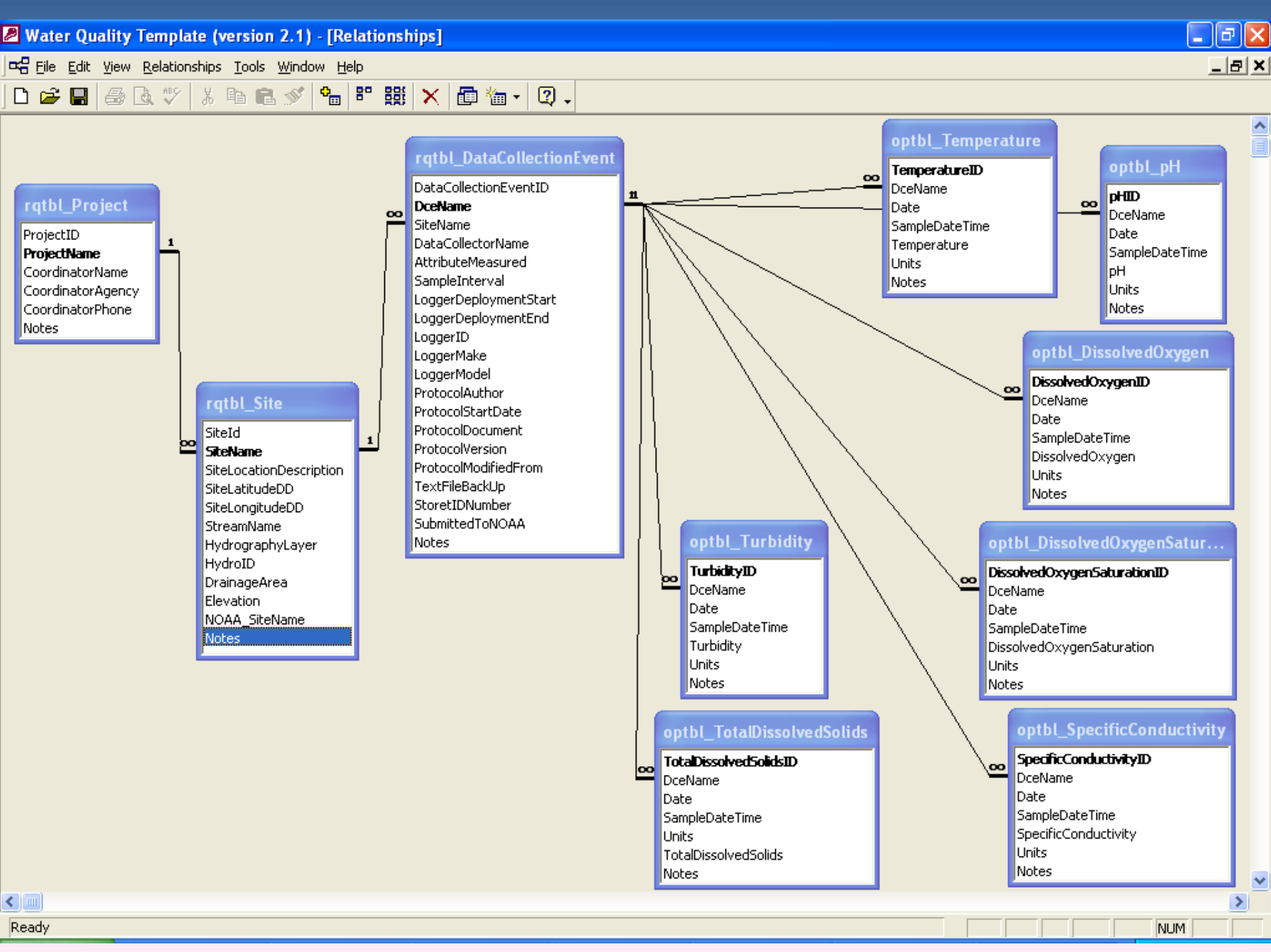
- Examples of the utility/power of going through the data dictionary process

- Doesn’t solve everything, but helps go a long way with the data -> information transition

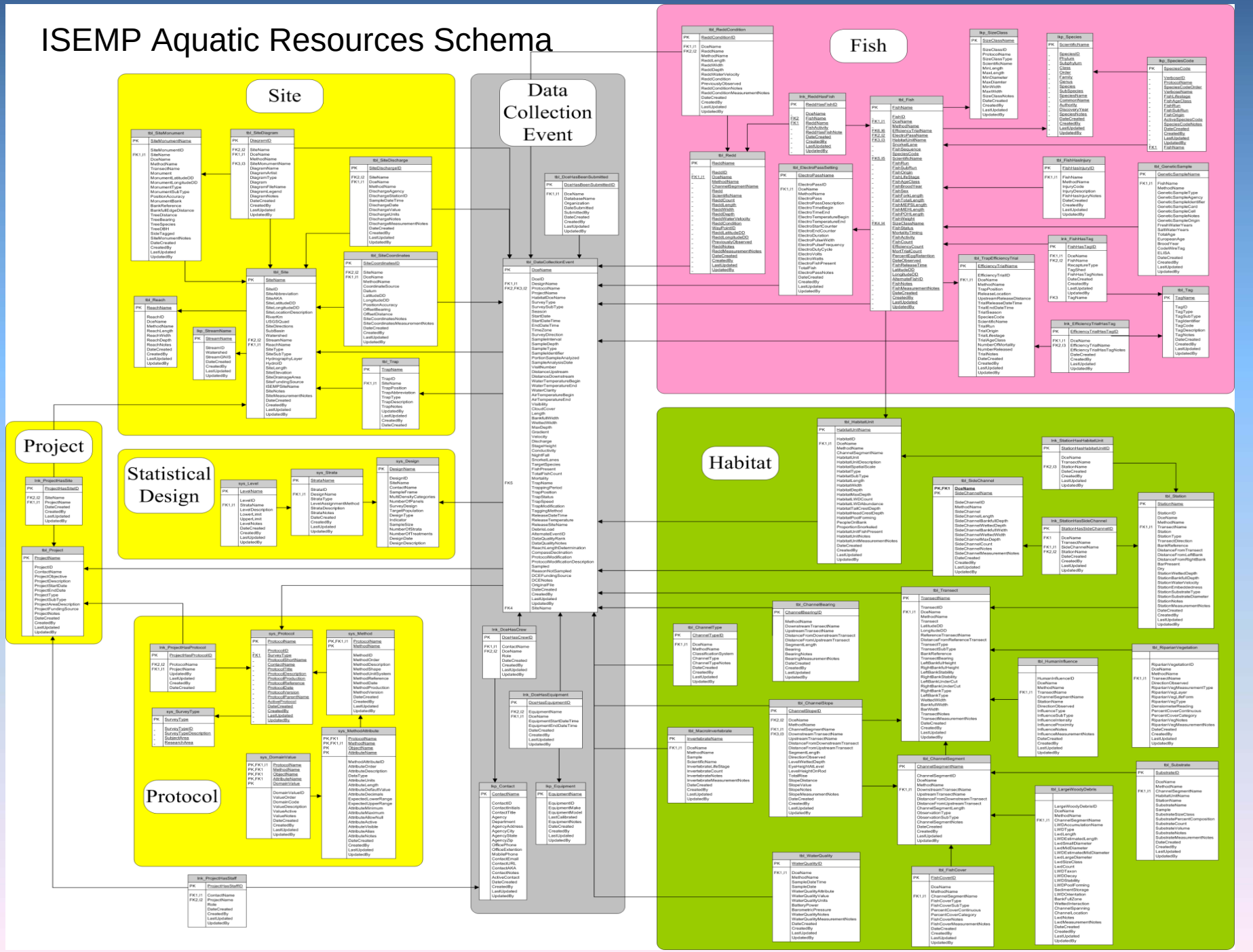
ISEMP's Aquatic Resources Schema

Normalizing obsevation data storage

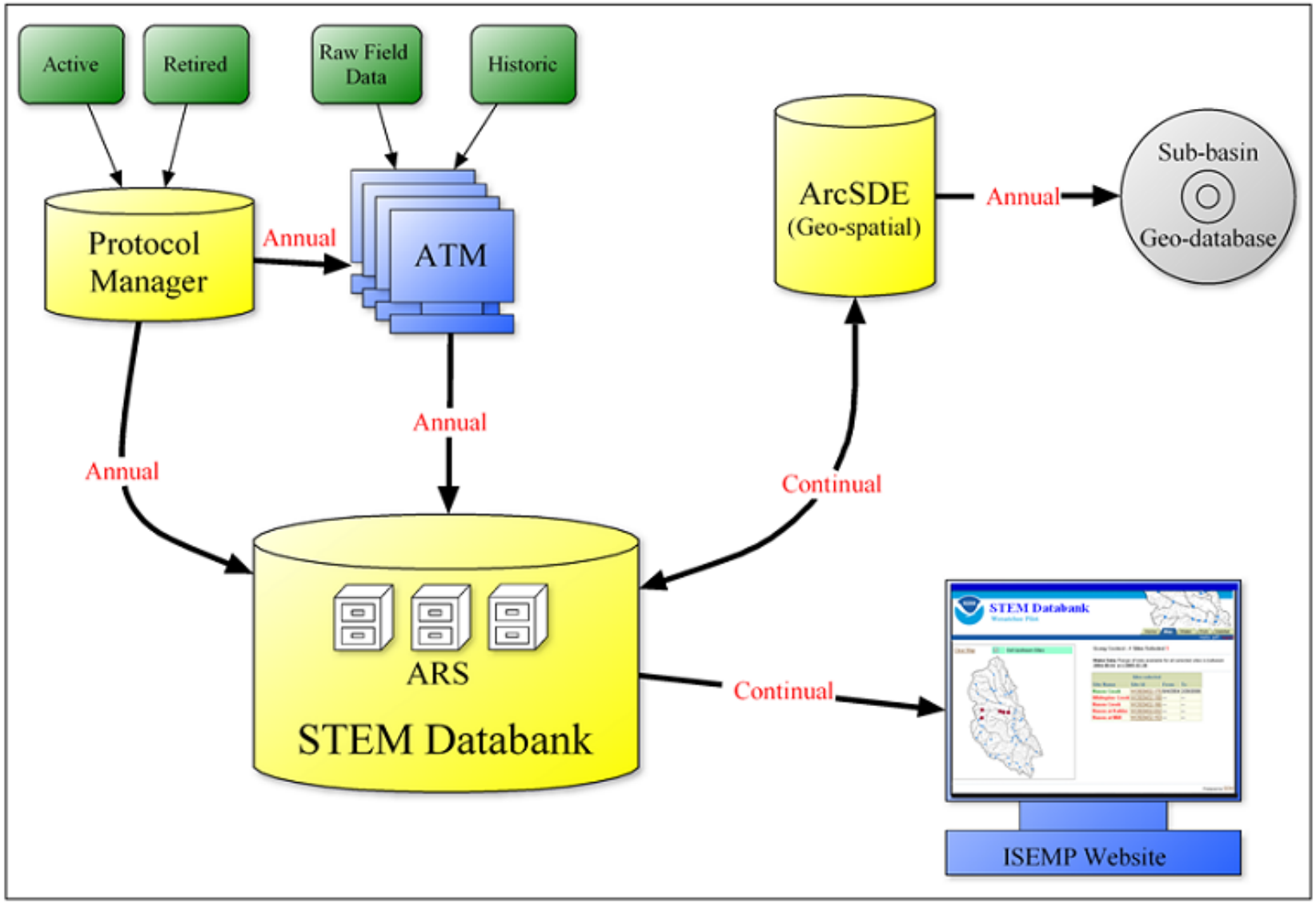



Water Quality Template (version 2.1) - [Relationships]
File Edit View Relationships Tools Window Help
rqtbl_Project
ProjectID
ProjectName
CoordinatorName
CoordinatorAgency
CoordinatorPhone
Notes
rqtbl_Site
SiteId
SiteName
SiteLocationDescription
SiteLatitudeDD
SiteLongitudeDD
StreamName
HydrographyLayer
HydroID
DrainageArea
Elevation
NOAA_SiteName
Notes
rqtbl_DataCollectionEvent
DataCollectionEventID
DceName
SiteName
DataCollectorName
AttributeMeasured
SampleInterval
LoggerDeploymentStart
LoggerDeploymentEnd
LoggerID
LoggerMake
LoggerModel
ProtocolAuthor
ProtocolStartDate
ProtocolDocument
ProtocolVersion
ProtocolModifiedFrom
TextFileBackUp
StoretIDNumber
SubmittedToNOAA
Notes
optbl_Temperature
TemperatureID
DceName
Date
SampleDateTime
Temperature
Units
Notes
optbl_pH
pHID
DceName
Date
SampleDateTime
pH
Units
Notes
optbl_DissolvedOxygen
DissolvedOxygenID
DceName
Date
SampleDateTime
DissolvedOxygen
Units
Notes
optbl_Turbidity
TurbidityID
DceName
Date
SampleDateTime
Turbidity
Units
Notes
optbl_DissolvedOxygenSatur...
DissolvedOxygenSaturationID
DceName
Date
SampleDateTime
DissolvedOxygenSaturation
Units
Notes
optbl_TotalDissolvedSolids
TotalDissolvedSolidsID
DceName
Date
SampleDateTime
Units
TotalDissolvedSolids
Notes
optbl_SpecificConductivity
SpecificConductivityID
DceName
Date
SampleDateTime
SpecificConductivity
Units
Notes
Ready
NUM

ISEMP Aquatic Resources Schema



ISEMP Data Management System



Observation Data Summary

John Day: thousands of Excel files ~ 10 MS Access databases

Wenatchee: 4 years of monitoring data into 7 databases (one per survey type)

Entiat: 100 potential restoration sites managed in one gdb

Spend our time improving metadata rather than reformatting datasets

Data management toolbox

1. Developed agency and regional tools for data processing:
   Aquatic Resources Schema and template databases

2. Centralized data repository (STEM)
   Data storage
   Metadata tracking
   Data analysis

3. Metadata tracking
   Library of protocols, methods and attributes

Data Management Toolbox

4. Queries and standards for data quality

5. Summary and analysis tools
   - Library of analysis queries and definitions

6. Tools for data distribution
   - Web interface database
   - General project website
   - Geo-databases

In Summary

- Current methods of storing data is inefficient for regional analyses and conclusions
  - Data flows
  - Data storage tools
- Current datasets can be useful but takes time
- Start small.

Organize once and analyze a thousand times.

Acknowledgements

ISEMP Team (Chris Jordan, Pamela Nelle, Nick Bouwes, Jody White, Chris Beasley)

- Especially Steve Rentmeester

Upper Columbia River Data Steward (James White)

Katie Barnas, NWFSC