Data Capture, Quality Management, and Storage Tools for Citizen Monitoring Groups

Revital Katznelson

CA State Water Resources Control Board

Today:

- The story of the four Functions
- Basic spreadsheet formats and database building block
- Examples of error calculation functions
- Advantages and disadvantages of Excel and Access
- Data flow

I wanted a data management system that has

- Tangible and user-friendly tools
- Stratified or tiered structure for different levels of detail
- Linkage between components
- Information retrieval and display tools
- Linkage to GIS, mapping options
- Compatibility with systems used by others at the Watershed, City, County, State, and Nation level
- Linkage to existing systems
- www Accessibility

I took a close look at available systems

STORET,
CCAMP, SFEI,
KRIS, CERES,
CALWATER, SINC, SWIM, SWAMP…

and discovered that we need to cater for four separate functions of a data management system

(1) documentation & QA/QC;
(2) storage & sharing
(3) retrieval, and
(4) interpretation & presentation.

Function (1) - Documentation & QA/QC

-most is done at the monitoring Project level by folks who know about the project,

- need a platform for data entry & documentation, error calculation, data verification and validation, etc.,

- it is easier to separate field measurements from lab analyses,

- need placeholders for all essential metadata and inventories, and

- can be done in MS Excel by most people, or in a combination of MS Access and Excel, if Access expertise is available.

Function (2) – Data Storage

- storage is very easy if all the information is already captured and can be stored as is, at the Project level,

- sharing data with others must be selective,

- only a sub-set of essential information will be uploaded onto the Project website or exported into a central database.

Function (3) – Retrieval

- requires that information is organized and interlinked in a way that allows any data user to sort, filter, group, and do any other query activity using anything from basic Excel tools to sophisticated Access or Oracle tools.

- good idea to implement basic database structure (I.e., parse information into “atomic” bits, have only one data type in a column, and avoid mixing of apples and oranges in drop-down menus). It is also good to provide for effective linkage between data tables

- if applied, any search engine and query tool can be used to retrieve your data from just about any relational database

Function (4) - Data Interpretation & Presentation

- this can be done ONLY after the retrieval tools have extracted the desired information from the database tables effectively,

- you will need additional tools for plotting, mapping, or running statistical comparisons

- if you have some programming-endowed folks who like to automate it in sync with the retrieval - the sky is your limit.

When you plan a monitoring effort you need to know…

-- what needs to be done (tasks),
-- who will do it (which role),
-- what will they use to do it (tools and platforms),
-- how much will it cost, and
-- can the Project afford it.

Building blocks of a database….

Start with Entities with Unique IDs

Station ID
Sample ID or ‘Activity ID’
Instrument ID
Project ID
Trip ID
Station-Visit ID

Unique IDs are used for tracking, sorting, grouping, filtering…

What do we need to capture about the Station?

- **Waterbody/sub-watershed/watershed**
- **Hydrologic unit (CalWater, HUCS, etc)**
- **Lat-Long Position AND datum**
- **Driving directions**
- **Nearest milepost**
- **Access to Station**
- **Verbal Description of Landmarks etc.**
- **USGS gauge # (if present)**
- **Pictures!**

(plus many other bits of information…)

Sample ID and 'Activity ID'

"Activity" can be an **Observation** (with verbal result), a **Field Measurement** (numeric result, done in Station), or a **Sample** (jar shipped elsewhere for analysis)

For a Sample, capture the following Sampling Log information:

- **Activity [or Sample] ID (helps tracking!)**
- **Station ID**
- **Date, Time**
- **Sampling Device**
- **Types and Number of containers**
- **Preservatives**

The project

What do we need to capture about the Project and the Project team?

1. **Organization Name**
2. Teams (Field Crews)
3. **People and roles**
4. **Contact Person**
5. **Contact information (address, email, phone, etc)**
6. **Project Duration (for STORET)**

Instrument ID and Standard ID

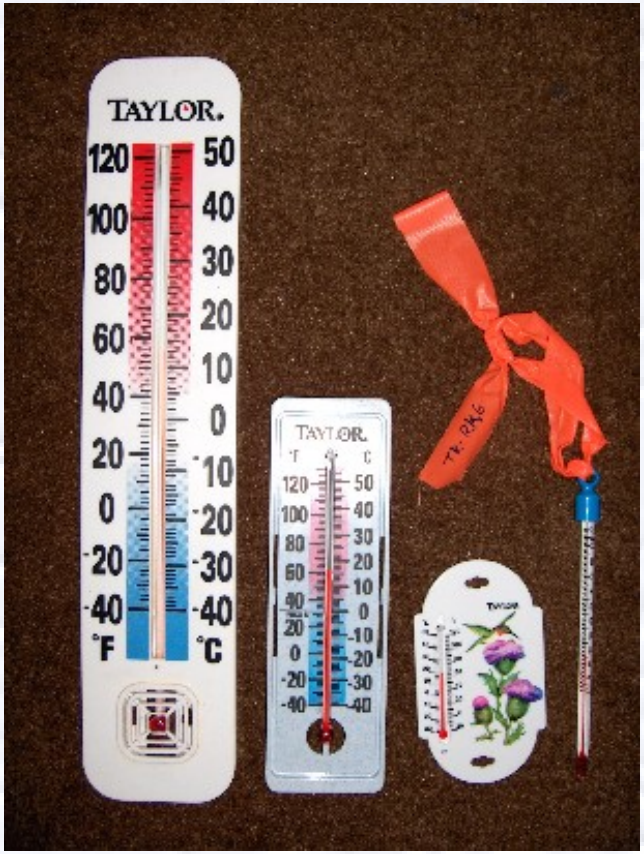
What do we need to capture about the Instrument?

Instrument ID, Serial number, or other unique identifier

Model; Type, features;

Range; Resolution;

Service records, etc.



Standards have unique LOT numbers that can be tracked, or you can create a Standard ID .

It DOES matter which one!

More building blocks of a database

What the users of your data want to know…

A. How good is your data: What is the accuracy and precision of your measurements and analyses?

B. What do your data represent in the environment?

b. When you plan a monitoring effort you also need to know what the Results will represent in the Environment

Spatial descriptors

Station Type : Creek, Outfall, Ditch

Station Selection Intent: Impact assessment, Source ID

Reach Selection Design: Systematic, Directed, Random, or Non-Deliberate (Anecdotal)

Station Selection Design: (same options)

Temporal descriptors

Flow Conditions: Storm runoff flows (wet) or base flow (dry) weather

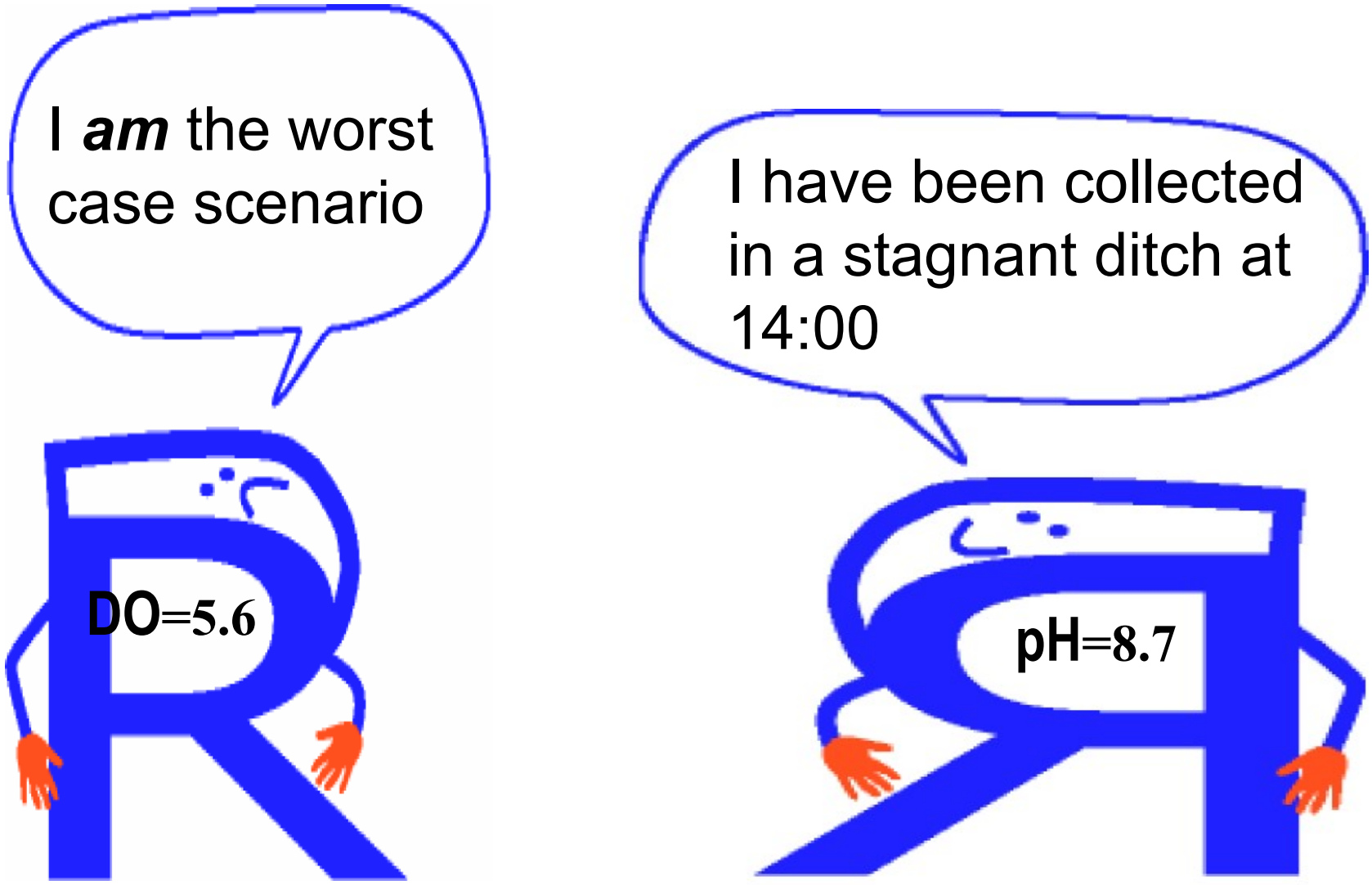
Sample Timing Intent: Worst case, Snapshot, Routine Monitoring

Seasonal Sampling Design: Systematic, Directed, Random, etc.

Diurnal Sampling Design: **(same options)**

Season of interest: Summer, Fall

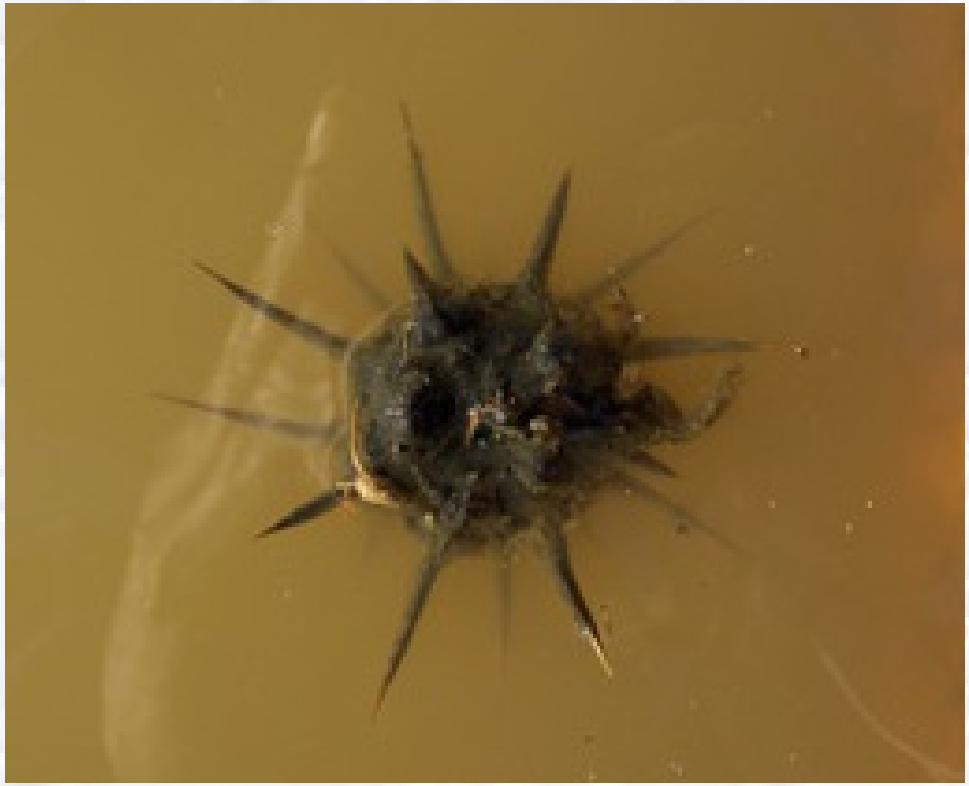
…And let your monitoring data speak for themselves!



Case Study: Field Measurements

Focus: Checking, recording, calculating, and communicating the accuracy and the precision of field measurements with probes and meters

(I am walking into murky waters with thorny issues here…)



Are you committed to deliver data of **known** accuracy and precision?

If you are… Here is what it takes

1. If you calibrated an instrument, collected data, and now you are ready to calibrate again, do an "accuracy check" first and record the reading before any calibration adjustments. [this is the same as "post-calibration" check].
2. Run periodic accuracy checks to all your non-adjustable instruments
3. Repeat discrete field measurements with each Instrument at least twice on every Trip
4. Write it all down, preferably with Instrument ID.

In other words…

-- Assign a unique Instrument ID to every measurement device

-- Link every Result with the Instrument that was used to measure it

-- Link every batch of Results with Instrument calibration and **accuracy checks** records, and Instrument **repeated measurement** records, for a given period of time

Formats for packaging information in tables

See handout: Spreadsheet formats

Redundancy happens!

It is inevitable, so you might as well put it where it looks into the future

Go Vertical!

But put in a manageable amount of records

Not all bits are needed in the database, but

For the number of information bits used at the project operations level (i.e., "on the ground"), the sky is the limit

Option 1: What was the actual accuracy and precision

Instrument ID	Characteristic (Parameter)	Results Units	Result	Accuracy (Percent)	Precision
TTP-STB01	Temperature, water	C	14.57	-1.4 %	0.06 %, RPD
ECP-STB01	Specific conductivity	uS/cm	758.7	-0.14 %	0.40 %, RPD
PHST-STB03j	pH	pH	8	0.5 Res.	0.5 Resolution
PHP-STB01	pH	pH	8.34	0.7%	0.12 %, RPD

Option 2: What MQOs for accuracy and precision were met

Instrument ID	Characteristic (Parameter)	Results Units	Result	Accuracy MQO	Precision MQO
TTP-STB01	Temperature, water	C	14.57	5 %	5 %, RPD
ECP-STB01	Specific conductivity	uS/cm	758.7	2 %	1 %, RPD
PHST-STB03j	pH	pH	8	0.5	20 %, RPD
PHP-STB01	pH	pH	8.34	5 %	5 %, RPD

How is the "% accuracy" generated?

From Post-event accuracy check (a.k.a. post-calibration) records: Reading of the instrument in Standard (before calibration adjustment), and the "true" value of the Standard.

This data quality indicator has to be calculated for both options, and compared to MQOs for Option 2

Essential post-event accuracy check records

Instrument ID	Characteristic (Parameter)	Units	Standard	"True" Value	Reading in Standard	Drift	Percent Accuracy
DOP-STB01	DO	% sat	humid air	100	97.3	-2.7	-2.7
DOP-STB01	DO	% sat	saturated water	100	95	-5	-5.0
ECP-STB01	Sp.Cond	uS	STB-EC10y	1412	1410	-2	-0.1
PHP-STB01	pH	pH	STB-PH20f	7	7.05	0.05	0.7
PHP-STB01	pH	pH	STB-PH29b	9	8.98	-0.02	-0.2
TTP-STB01	Temp	C	TR-STB43	21.5	21.19	-0.31	-1.4
TTP-STB01	Temp	C	TR-STB43	21	21.21	0.21	1.0

Differential = (Reading in Standard) – (True value)

$$\text{Percent accuracy} = \frac{((\text{Reading in Standard}) - (\text{True value})) \times 100}{(\text{True value})}$$

How is the "% RPD" generated?

From pairs of Repeated field measurements:
The difference between the two values expressed as a percentage of their average.

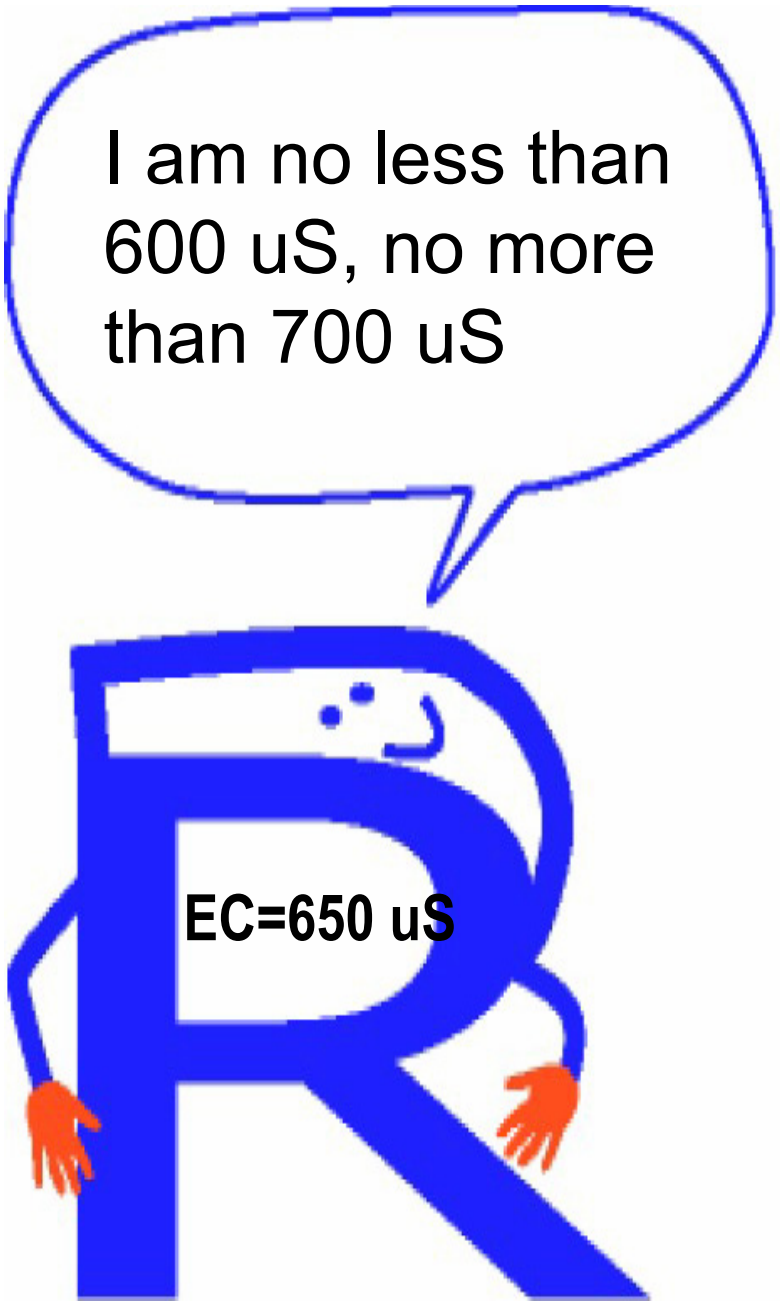
This data quality indicator has to be calculated for both options, and compared to MQOs for Option 2

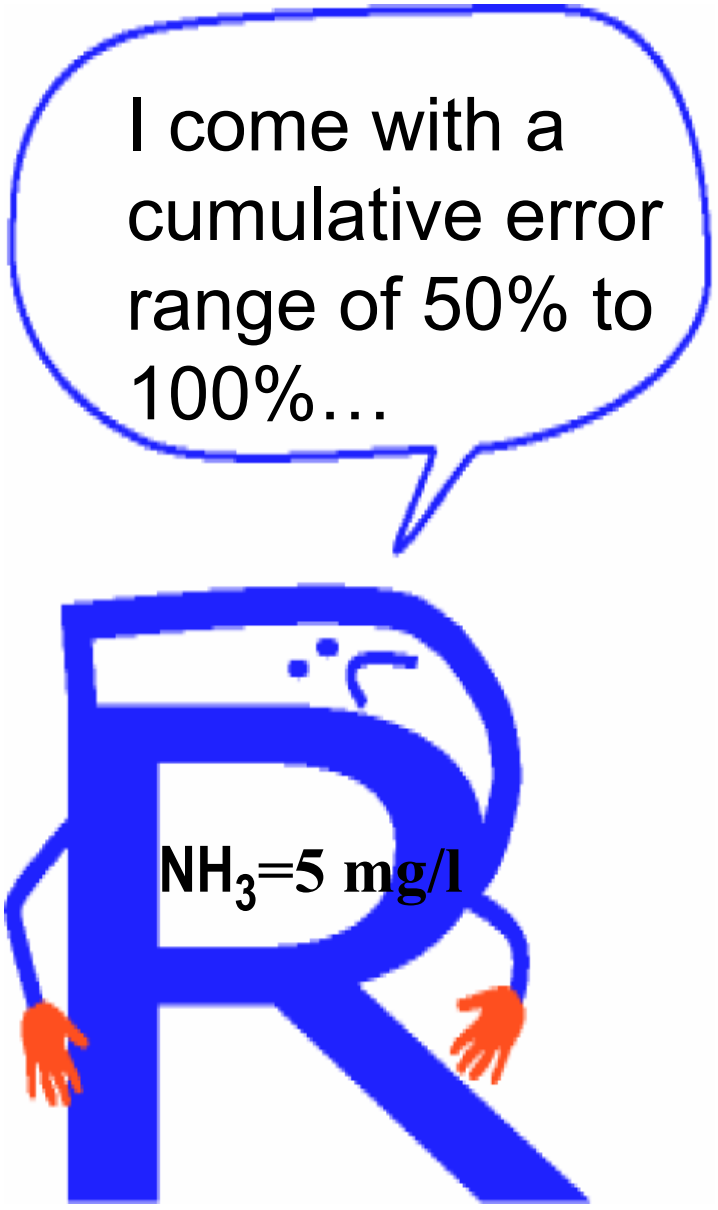
Essential Precision Worksheet columns

Instrument ID	Characteristic (Parameter)	Results Units	Result	Repeated Result	reproducibility (RPD*)	Max RPD*
DOP-STB01	DO	mg/l	2.84	2.65	6.92	
DOP-STB01	DO	mg/l	11.96	11.68	2.37	
DOP-STB01	DO	% sat	121.5	121.5	0.00	6.92
ECP-STB01	Sp.cond.	uS/cm	746.9	746.7	0.03	
ECP-STB01	Sp.cond.	uS/cm	648.4	651	0.40	0.40
PHP-STB01	pH	pH	8.61	8.62	0.12	
PHP-STB01	pH	pH	8.55	8.55	0.00	0.12
TTP-STB01	Temp.	C	15.97	15.97	0.00	
TTP-STB01	Temp.	C	16.19	16.2	0.06	0.06

*** RPD is the Relative Percent Difference**

$$RPD = \frac{((Result) - (Repeated\ Result\ Value)) \times 100}{((Result) + (Repeated\ Result\ Value))/2}$$

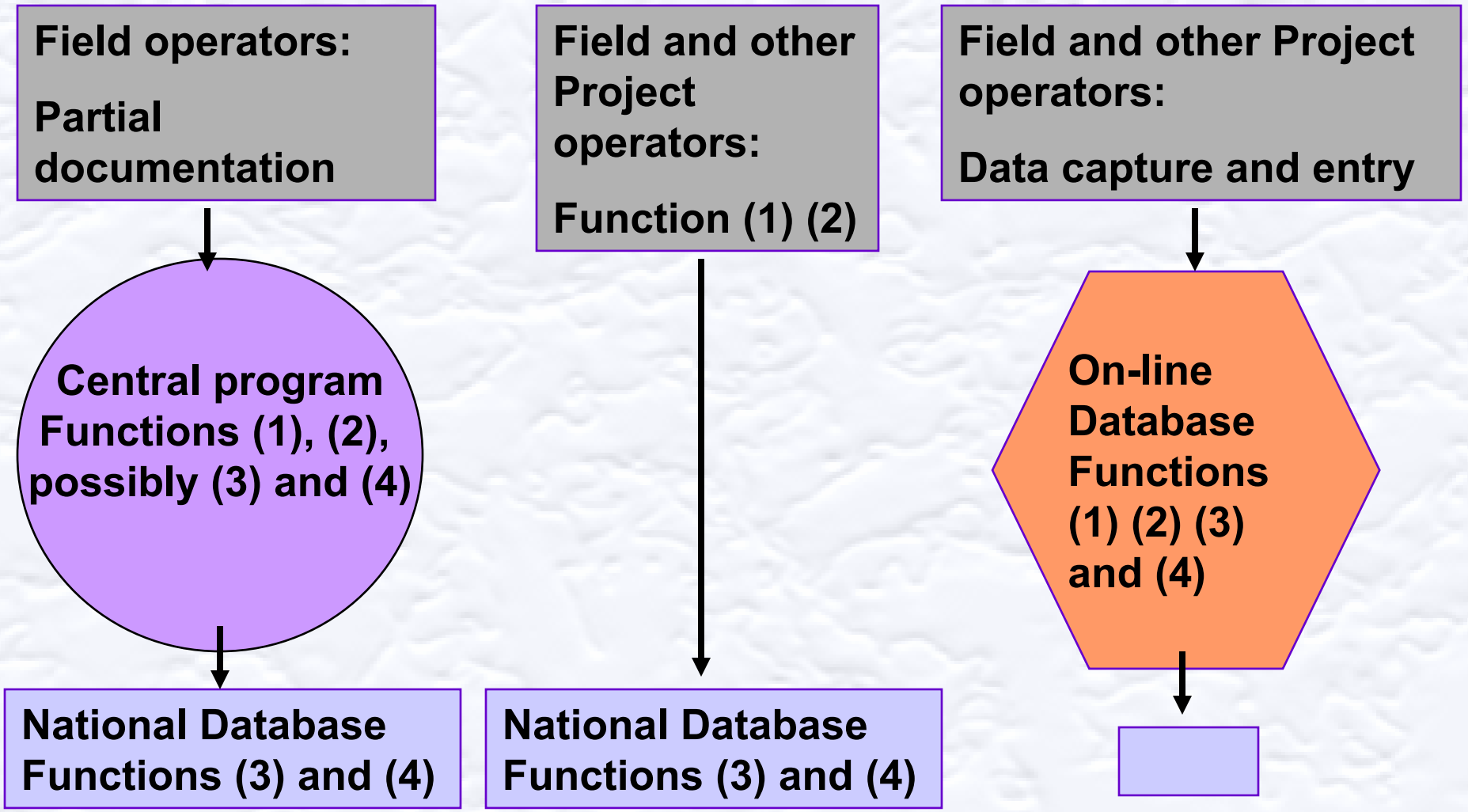

I am no less than 600 uS, no more than 700 uS
EC=650 uS


I come with a cumulative error range of 50% to 100%…
NH_3=5 mg/l

Examples: Projects and Programs

- Field data sheets in drawer (too many folks)
- Excel spreadsheets – home made
- Excel spreadsheet templates and data transfer tools
- Excel regional database with web and data transfer interfaces
- Access database for Project – home made
- Regional Access database
- “Program central” – Access or Oracle centralized database

Models of data management systems



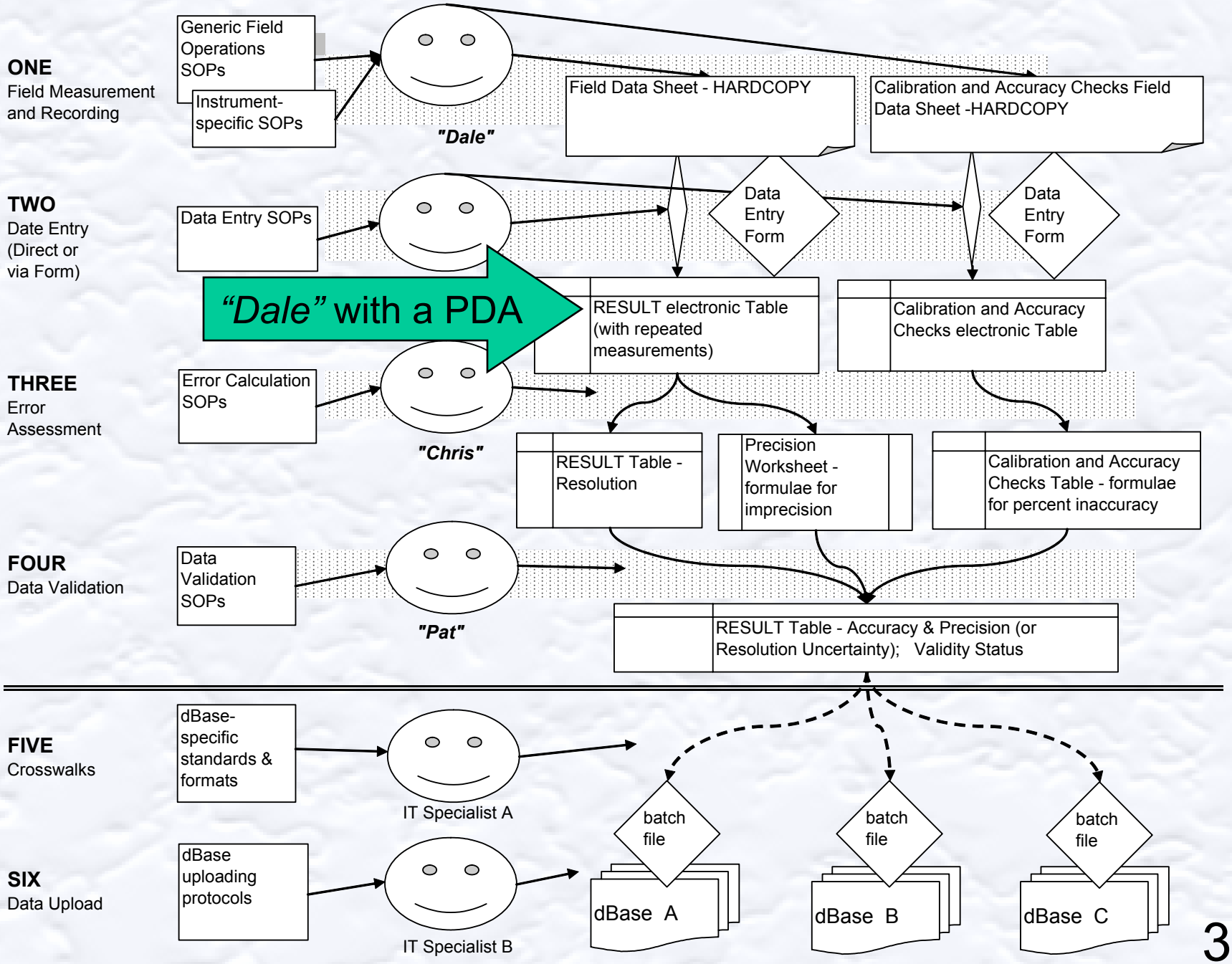
Functions: (1) documentation & QA/QC; (2) storage & sharing; (3) retrieval; (4) presentation.

Web hosting

If you want to create your own web-based database, even just for for function 3 (retrieval), check out Web Hosting opportunities: For $10-20 per month you can have

- **Your own domain**
- **MySQL database with several GB of storage**
- **Periodic backup of your data**

But you will be the one designing the database with all its tools, setting it up, uploading data, and updating the data.


ONE
Field Measurement and Recording
Generic Field Operations SOPs
Instrument-specific SOPs
"Dale"
Field Data Sheet - HARDCOPY
Calibration and Accuracy Checks Field Data Sheet -HARDCOPY
TWO
Date Entry (Direct or via Form)
Data Entry SOPs
Data Entry Form
Data Entry Form
"Dale" with a PDA
RESULT electronic Table (with repeated measurements)
Calibration and Accuracy Checks electronic Table
THREE
Error Assessment
Error Calculation SOPs
"Chris"
RESULT Table - Resolution
Precision Worksheet - formulae for imprecision
Calibration and Accuracy Checks Table - formulae for percent inaccuracy
FOUR
Data Validation
Data Validation SOPs
"Pat"
RESULT Table - Accuracy & Precision (or Resolution Uncertainty); Validity Status
FIVE
Crosswalks
dBase-specific standards & formats
IT Specialist A
batch file
batch file
batch file
dBase A
dBase B
dBase C
SIX
Data Upload
dBase uploading protocols
IT Specialist B

Excel versus Access: Advantages

Advantages of Excel:

- Small files, easy to e-mail, easy to exchange
- Intuitive, easy to learn, transparent, easy to see your data
- Supports drop-down menus to reduce data entry errors
- Easy to sort and filter data
- Good for calculations and graphing

Advantages of Access:

- More practical for large databases
- Supports sophisticated queries and security features
- Can generate data reports & tables in various formats
- Controlled data entry, and less hands-on manipulation of data

Excel versus Access: Disadvantages

Disadvantages of Excel:

- File gets cumbersome with large data sets
- Requires a lot of hands-on manipulation
- Potential for human error when manipulating data
- No capabilities for complex queries

Disadvantages of Access:

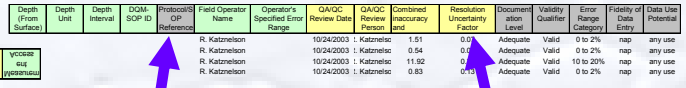
- Harder to learn, takes dedication and experience
- Large files, 20 or 30MB -- harder to exchange
- Cannot do calculations or graphs (but data are easily exported to Excel for that)

Use capture tools for all Water Quality Data Elements (WQDEs)

What

How Good?

(worksheets)



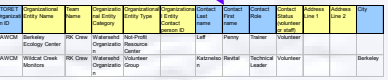
Who

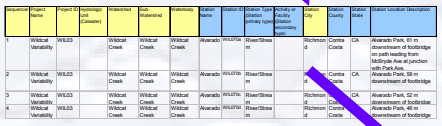
Where

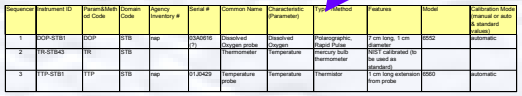
When

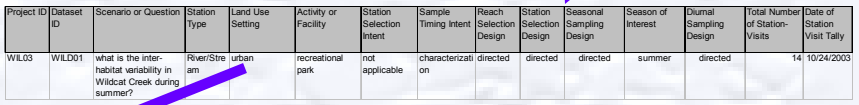
How

Why



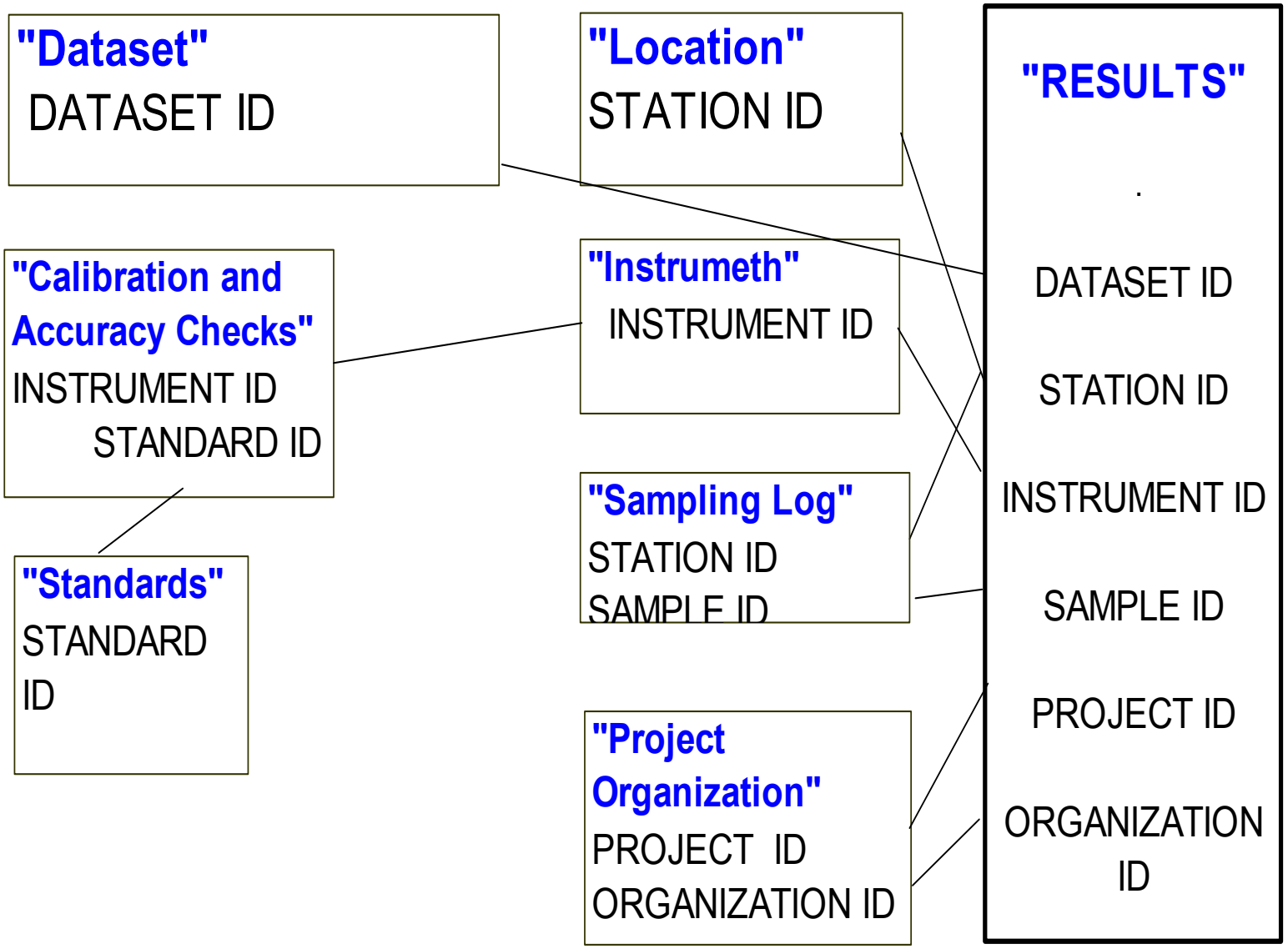




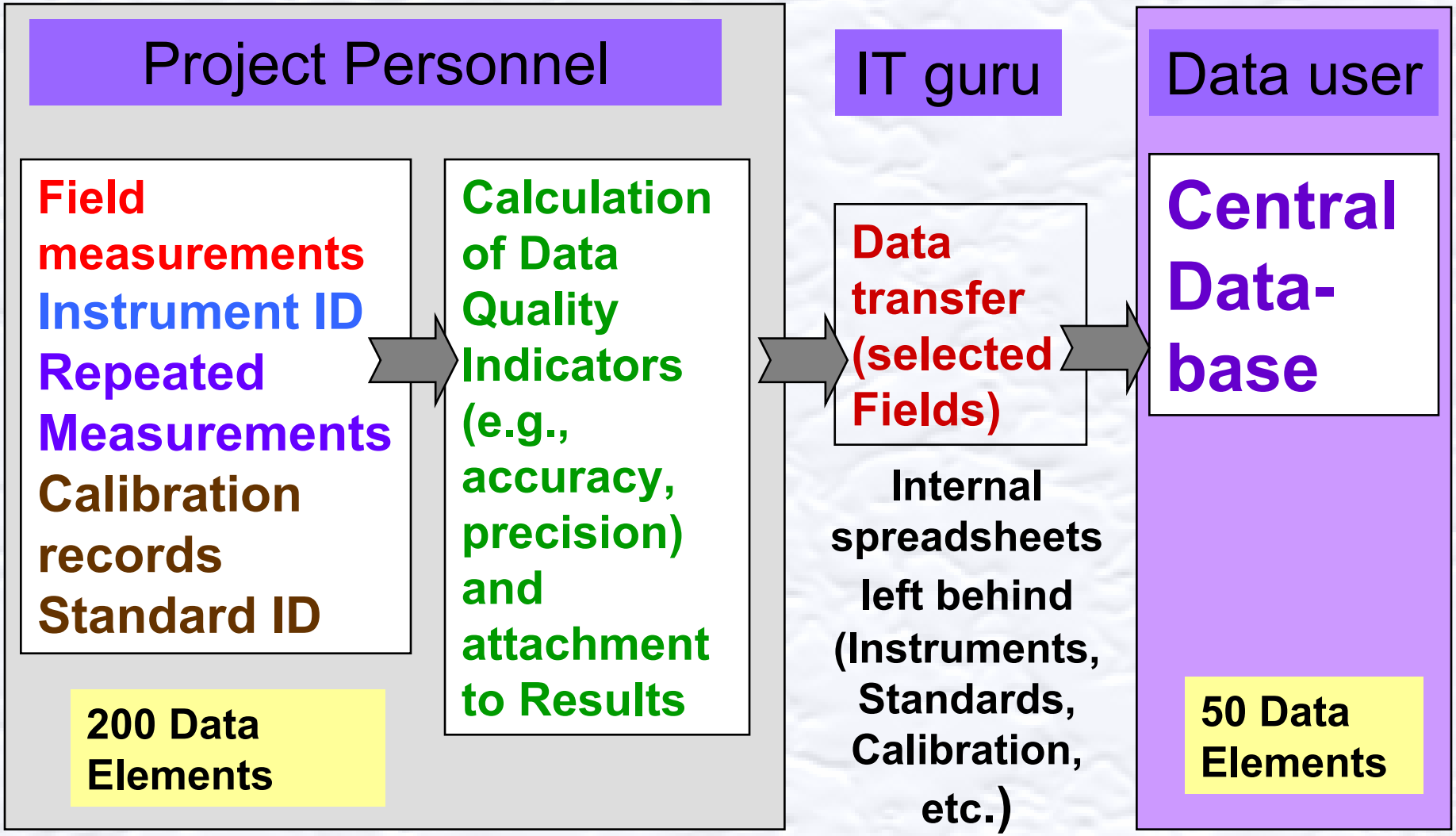


What does it represent?

You can package it all in the Project File…

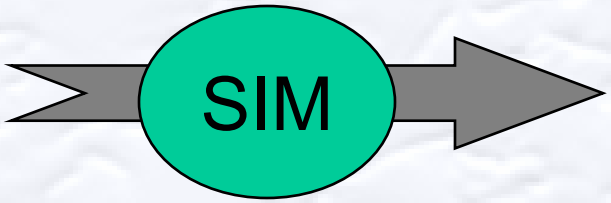


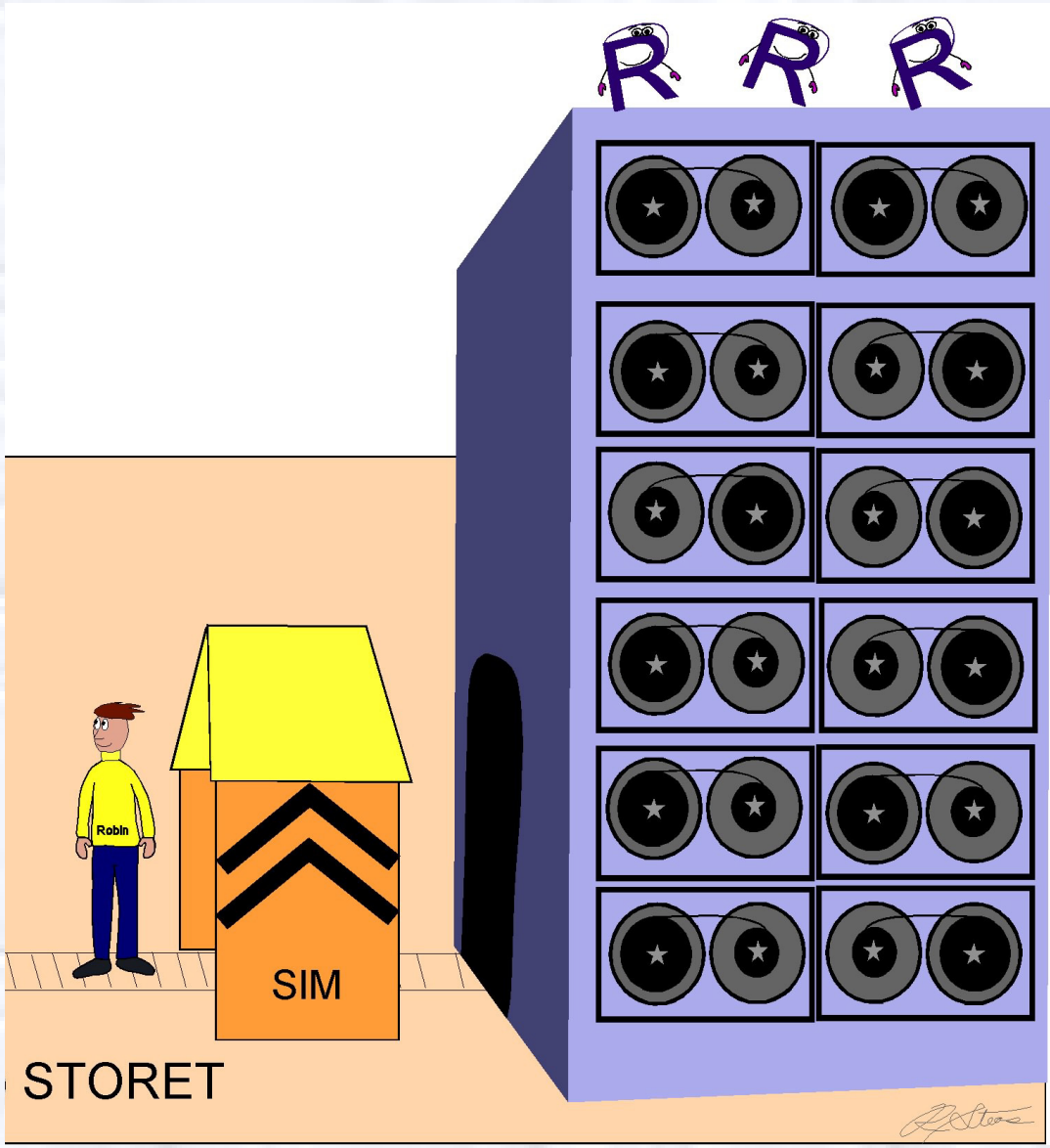
You can have Project personnel document and manipulate the data; Then transfer only selected elements to the Central Database



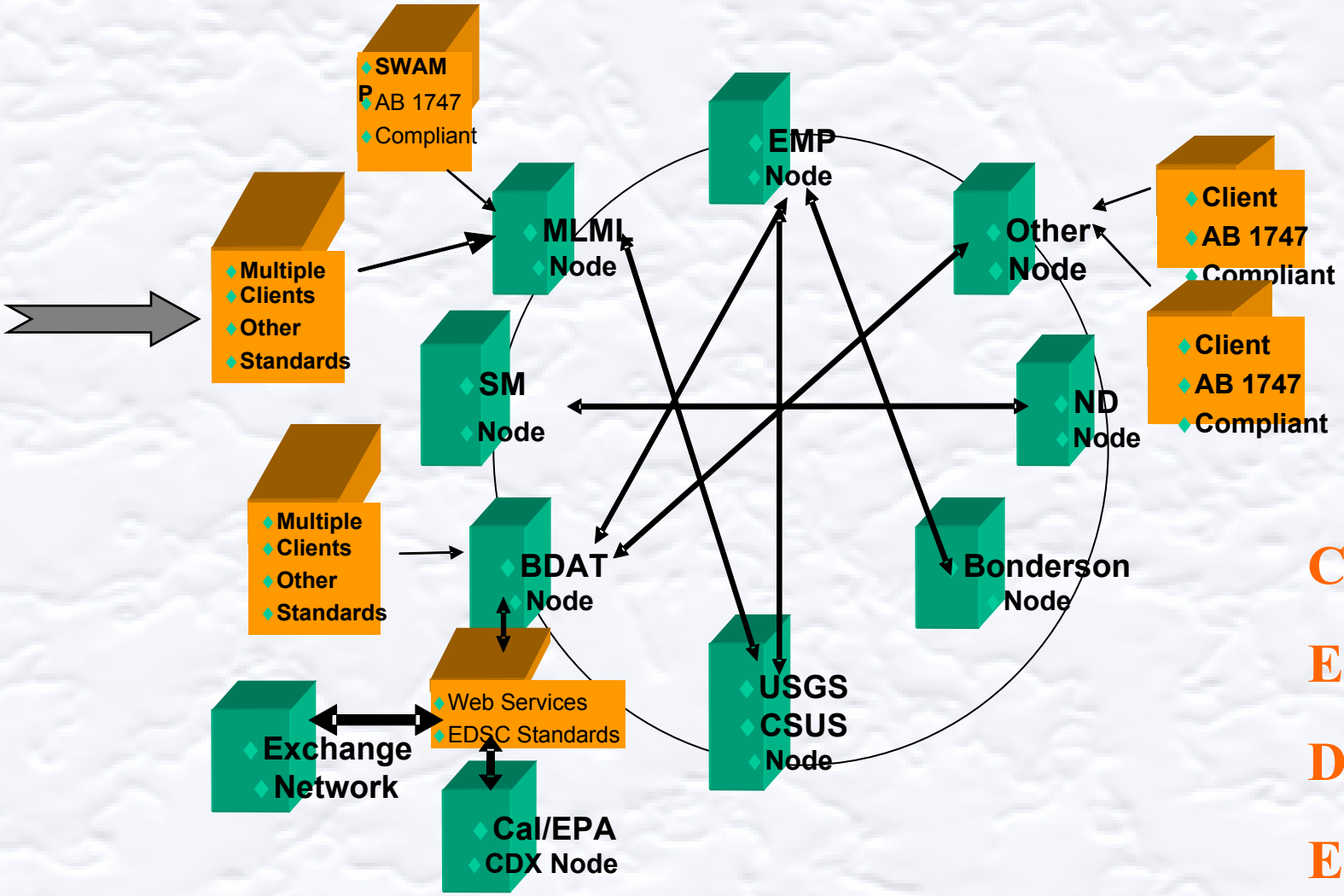
The Central Database can be…

STORET





….Or a node in the California Cooperative Data Management System



California

Environmental

Data

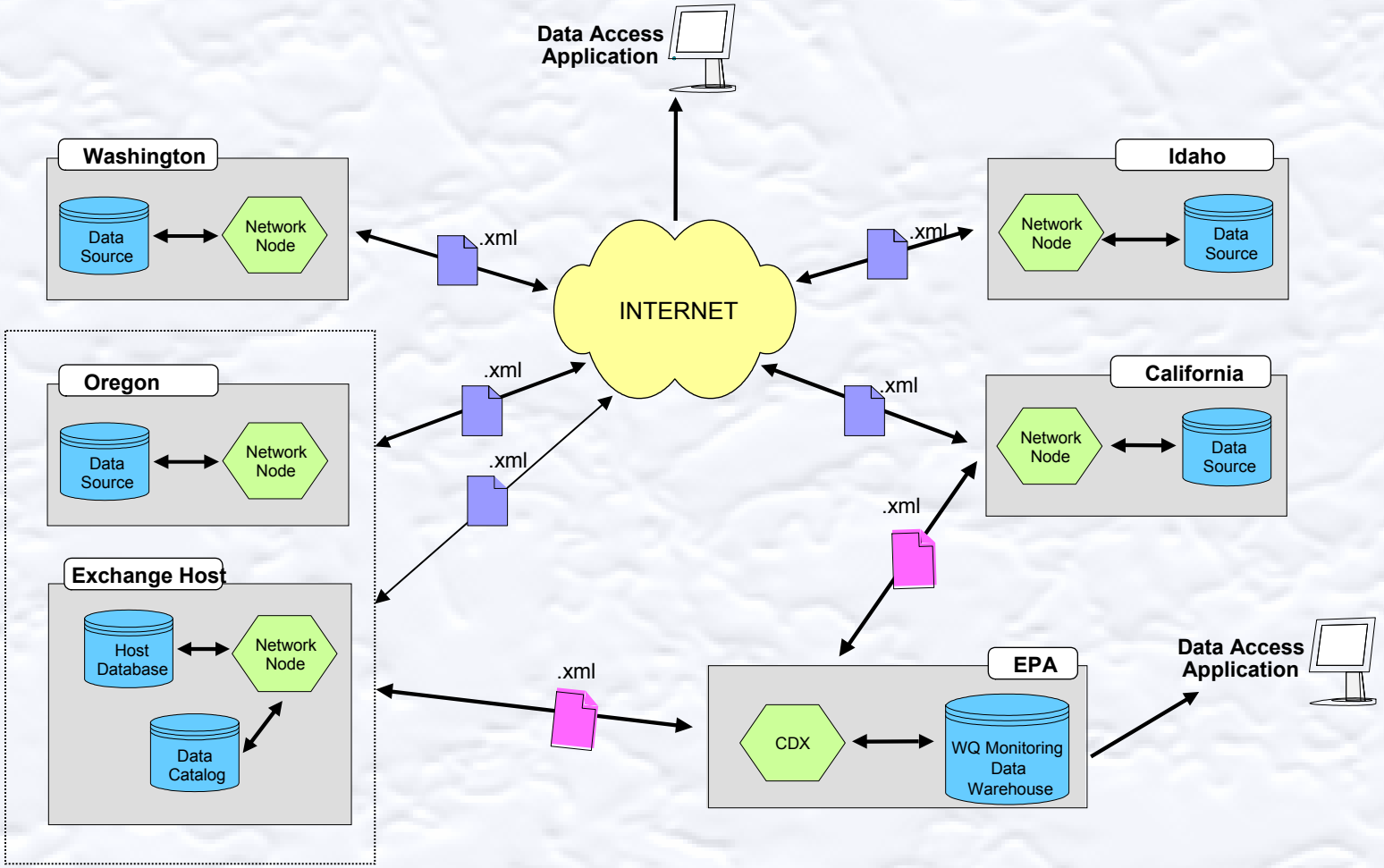
Exchange

Network

Source: Karl Jacob, DWR

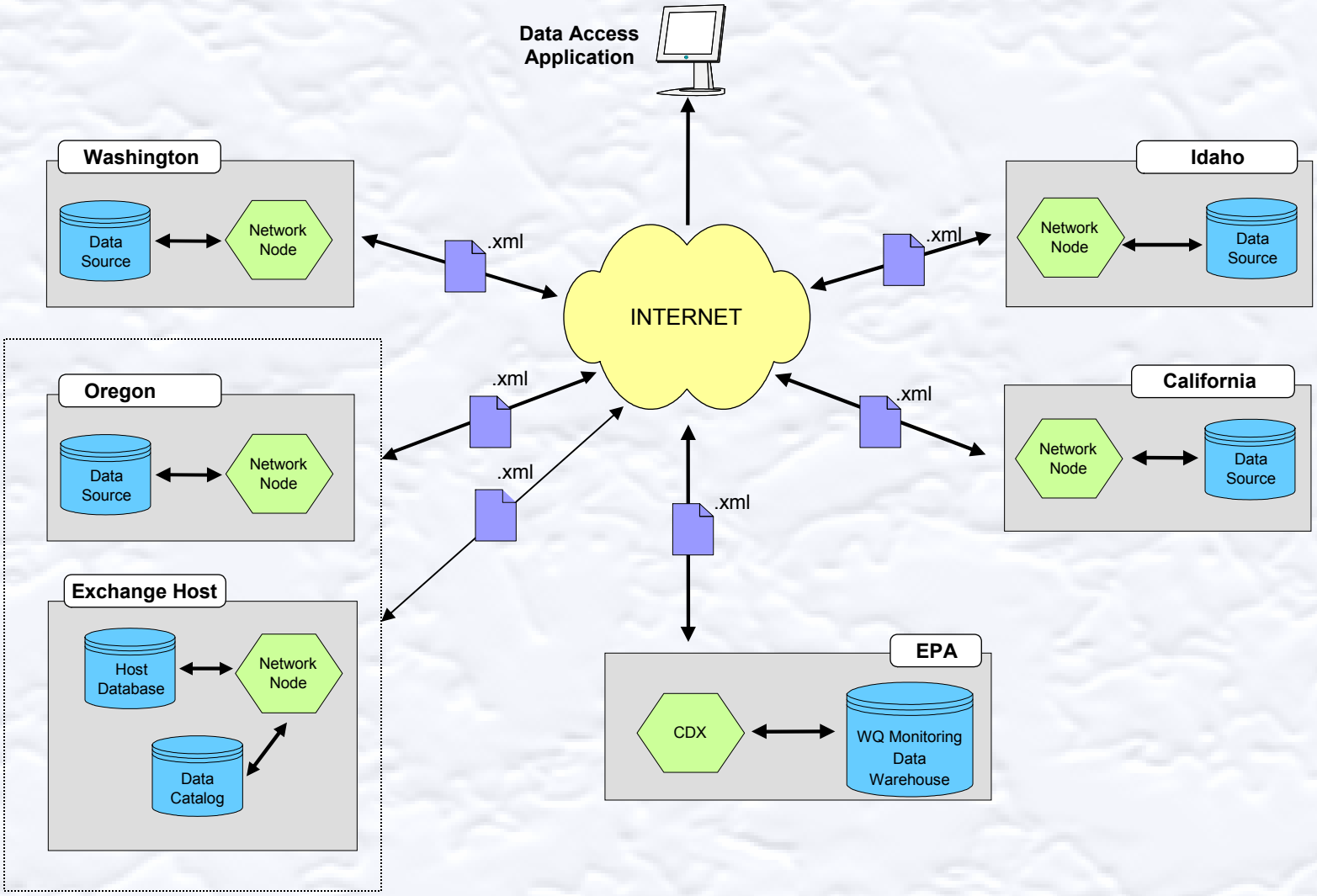
…. or the National data exchange network!

A. Nearer-Term Vision for the Data Flow



Source: Curtis Cude, Pacific Northwest Water Quality Data Exchange

B. Long-Term Vision for the Data Flow



Source: Curtis Cude, Pacific Northwest Water Quality Data Exchange

Ready to transfer your data?

Find out the about the restrictions (business rules, formats, permitted values),

Identify the data flow pathways, and

Decide if you want to use the updatable or the non-updatable mode in your target central database.

XML Schema

```
<?xml version="1.0" encoding="utf-8"?>
<xsd:schema targetNamespace="urn:us:net:exchangenetwork" xmlns:xsd="http://www.w3.org/2001/XMLSchema" xmlns:pnwwqx="urn:us:net:exchangenetw
    <xsd:annotation>
        <xsd:documentation>
        Schema Name :                PNWWQX_ProjectDetailsType_v.1.3.xsd
        Current Version Available At :
        Description :                This schema defines the data elements to be shared through the Pacific Northwest Water Quality Data Exchange relat
and assess the water quality.
        Application :                Pacific Northwest Water Quality Data Exchange
        Developed by :               Pacific Northwest Exchange States; Windsor Solutions, Inc
        Point of Contact :           Curtis Cude (cude.curtis@deq.state.or.us)
                                     Kevin Jeffery (kevin_jeffery@windsorsolutions.com)
        </xsd:documentation>
    </xsd:annotation>
    <xsd:complexType name="ProjectDetailsType">
        <xsd:sequence>
            <xsd:element ref="pnwwqx:ProjectIdentifier"/>
            <xsd:element ref="pnwwqx:ProjectName"/>
            <xsd:element ref="pnwwqx:ProjectDescription"/>
            <xsd:element ref="pnwwqx:ProjectQAPPIndicator"/>
            <xsd:element ref="pnwwqx:ProjectQAPPDescription" minOccurs="0"/>
            <xsd:element ref="pnwwqx:ProjectStartDate"/>
            <xsd:element ref="pnwwqx:ProjectEndDate" minOccurs="0"/>
            <xsd:element ref="pnwwqx:ProjectAreaDescription" minOccurs="0"/>
        </xsd:sequence>
    </xsd:complexType>
    <xsd:element name="ProjectIdentifier" type="xsd:string"/>
    <xsd:element name="ProjectName" type="xsd:string"/>
    <xsd:element name="ProjectDescription" type="xsd:string"/>
    <xsd:element name="ProjectQAPPIndicator" type="xsd:boolean"/>
    <xsd:element name="ProjectQAPPDescription" type="xsd:string"/>
    <xsd:element name="ProjectStartDate" type="xsd:date"/>
    <xsd:element name="ProjectEndDate" type="xsd:date"/>
    <xsd:element name="ProjectAreaDescription" type="xsd:string"/>
</xsd:schema>
```

Source: Curtis Cude, Pacific Northwest Water Quality Data Exchange

Summary

Actions for capture, quality management, and storage of monitoring data involve many tasks, employs many roles, and require many tools

The two extremes are a totally centralized system (Region or State) versus a local database at the Project level

Centralized data management options require lots of resources and IT support

The choice of tools and platforms are not always yours, but when it is – plan ahead