USGS
science for a changing world

Duplicated Water Data - Causes, Implications, Solutions

Dorrie Gellenbeck and Jonathon Scott, U.S. Geological Survey

U.S. Department of the Interior
U.S. Geological Survey

National Water Quality Monitoring Council
April 2014

Introduction

Overview

- This is a story about :
  - How we broke the USGS hydrologic database through the best of intentions
  - The consequences of those intentions
  - How we are fixing it and the implications
  - What may lie ahead in the future

How we broke it.

- WATSTORE precursor to NWIS
  - Born in 1971
  - Disabled in 1997
- Followed technology changes
  - Mainframe->minicomputers.
- The WATSTORE database was moved into NWIS instances.
  - 48 separate NWIS databases

WATSTORE

U.S. GEOLOGICAL SURVEY'S NATIONAL WATER DATA STORAGE AND RETRIEVAL SYSTEM

The U.S. Geological Survey, through its Water Resources Division, investigates the occurrence, quantity, quality, distribution, and movement of the surface and underground waters that comprise the Nation's water resources. It is the principal Federal water-data agency and, as such, collects and disseminates about 70 percent of the water data currently being used by numerous State, local, private, and other Federal agencies to develop and manage our water resources. As part of the Geological Survey's program of releasing water data to the public, a large-scale computerized system has been developed for the storage and retrieval of water data collected through its activities.

2

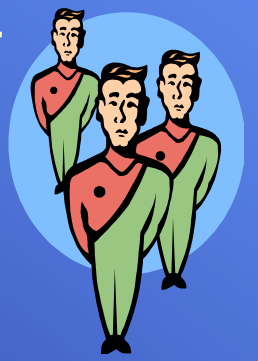
The National *Water* Data *Storage* and *Retrieval* System (WATSTØRE) was established in November 1971 to modernize the Geological Survey's existing water-data processing procedures and techniques and to provide for more effective and efficient management of its data-releasing activities. The system is operated and maintained on the central computer facilities of the Survey at its National Center in Reston, Virginia. As of 1975, data may be obtained from WATSTØRE through any of the Water Resources Division's 46 district offices listed at the end of this leaflet. General inquiries about WATSTØRE may be directed to:

Chief Hydrologist
U.S. Geological Survey
437 National Center
Reston, Virginia 22092

3

How we broke it - Best of intentions

- Given the huge time and expense of new data collection...
- Copying data was commonplace:
  - Within USGS between NWIS systems
  - Outside USGS with USGS data - STORET
- Coupled with large spatial and temporal variability
- It was natural to compile everyone's data



How we broke it - Best of intentions

- Congress identified difficulties with synthesizing data from multiple agencies
  - USGS copied data into STORET for years
- Warning signs: we observed multiple copies of USGS data in EPA STORET
  - ~2003 USGS data from STORET was removed
  - Triage step #1

Problems of duplicated data:

- Bias
  - Statistical analyses
  - Same data included more than once
- Incorrect analyses from stale copies
- Incomplete analyses
  - Copies between systems
  - Data updates in one version and not the other

How can USGS fix the problem?

- **Identify the problem**
  - ***A completely described problem is 50% solved ~Charles Kettering***
- **Gain support and resources**
  - ***Alone we can do so little; together we can do so much ~Helen Keller***
- **Identify the technical approach**
  - ***The devil is in the details ~Anonymous***

How we are fixing it.

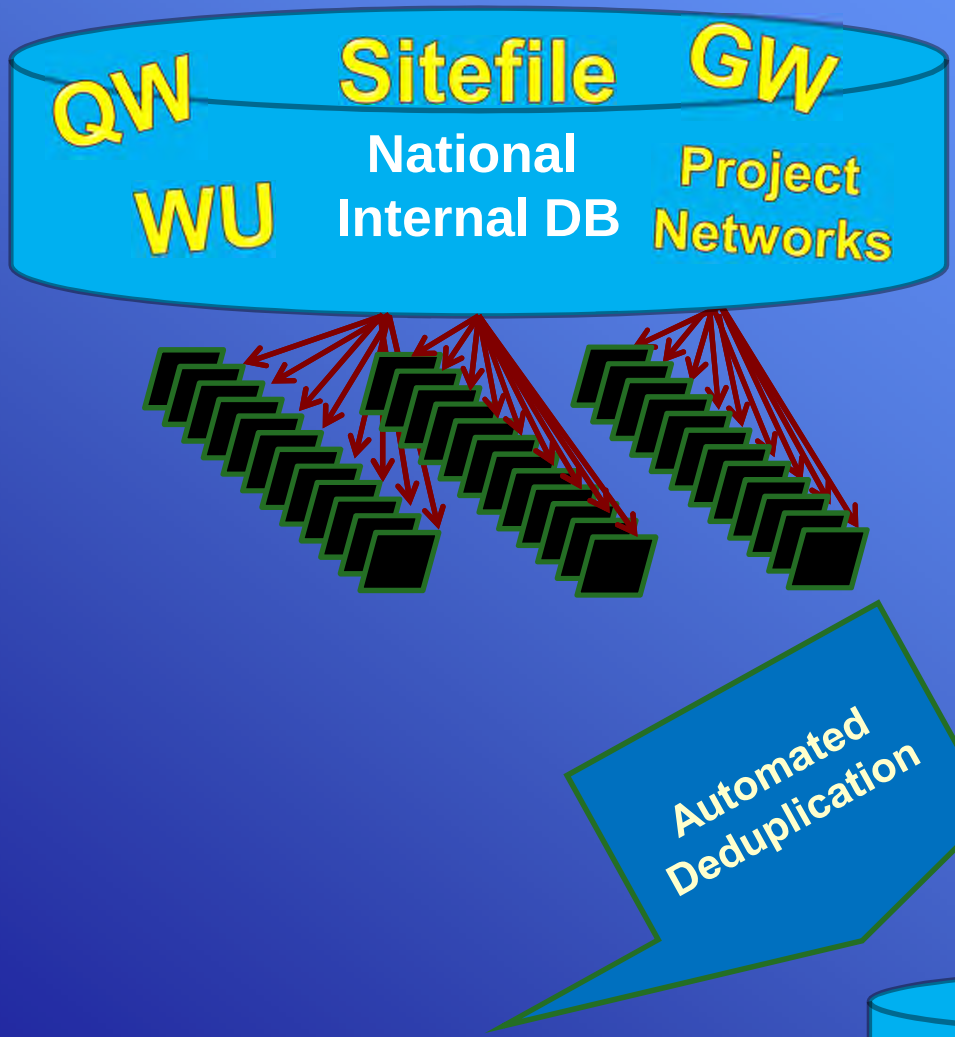
- Create aggregated national database
  - Includes all non-continuous data from separate NWIS systems
  - Water-quality data are one data type
  - Currently internal only
- Non-authoritative databases were identified
  - Not included in the aggregation
- Find & fix identical station ID for non-colocated stations
  - Where possible

How we are fixing it.

The short version

1. Group data into “Data Topics”
2. Develop algorithms to compare duplicates
3. Choose the “best” copy
4. Eventually, keep only the best record

Easy!! Right??



37 Data Topics

How we are fixing it.
The long version

- Divide DB into topics. For each topic:
  - Specify keys for exact duplicates
  - Specify keys for inexact duplicates
  - Computer automatically resolves exact duplicates
  - Subject matter experts formulate rules to resolve inexact duplicates
  - Devise scoring weights and tie breakers

Example: Subset of water-quality scores
Note: each of 37 data topics have unique rules

Score	Rule
+2	Most results
+1	Earliest result-creation date
+0.5	Most results with non-empty remark code
+0.5	Most results with non-empty value-qualifier codes
+1	Greatest number of non-null, non-mandatory fields
+0.1	Earliest sample creation date

Example: Subset of water-quality scores

- Scores are weighted
- More important characteristics get more points
- Determined by data experts
- Example:
  - Highest scoring rule (2.0 pts) – Most results
    - Explanation: ‘More is better’
  - Lowest scoring rule (0.1 pts) - Earliest sample creation date
    - Explanation: Indicates original sample record

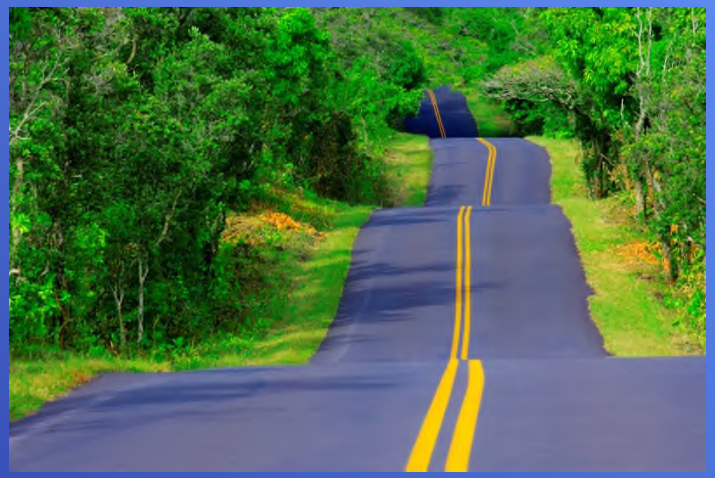
How we are fixing it.

- A water-quality example

Category	Arizona	score	New Mexico	score
Number of results	13	0	39	2
Earliest result creation date	no	0	yes	1
Most results with non-empty remark code	no	0	yes	0.5
Most results with non-empty value-qualifier codes	yes (tie)	0.5	yes (tie)	0.5
Greatest number of non null, non-mandatory fields	no	0	yes	1
Earliest sample creation date	no	0	yes	0.1
Total Score	**0.5**		***5.1***	

What's next for USGS?

- NWISWeb & QW Portal - Use ‘deduplicated’ aggregated database as single source for public display
- NWISWeb – expand data-types available
- Public may see some content changes:
  - Drainage area
  - Well depth
  - HUC



What's next for others?

Duplicates may exist from multiple agencies

- Agency code differences
- Site Identifier differences

- **USGS internal problems were just the tip of the iceberg.**

What's next for others?

- Employ similar approaches?
- Employ river reach?
- Track aliases for site identifiers among agencies
- Employ a common site numbering scheme
- **Stop COPYING!**

Questions?

Contacts

- Dorrie Gellenbeck, USGS
  - Denver, Colorado
  - djgell@usgs.gov; 303-236-1458
- Jon Scott, USGS Retired
  - jcscott@usgs.gov