Challenges with the Secondary Use of Multi-Source Water-Quality Data in the United States

Lori Sprague
U.S. Geological Survey
National Water-Quality Program
National Water-Quality Assessment Project
Surface Water Trend Studies Coordinator

USGS
science for a changing world

NAWQA Trend Study

Scope

- Stream and river trends in nutrients, pesticides, sediment, carbon, salinity, and aquatic ecology
- NWIS, STORET, and other Federal, State, and local databases
- 147 million water-quality records from 468,000 sites and 518 organizations

Survey of multi-source data

- 25 million nutrient records from 322,000 sites and 488 organizations
- NWIS, STORET, and other Federal, State, and local databases

  19 Federal agencies
  6 regional (multi-State) organizations
  100 State water, natural resources, or environmental protection agencies
  130 tribal organizations
  108 County or subcounty organizations
  24 academic organizations
  17 non-governmental organizations
  34 volunteer organizations
  50 private organizations

Survey of nutrient data

These issues occur with secondary use of data

- Primary use – the use of data for the original intent determined by the organization that collected the data
- Secondary use – the use of the same data for other purposes

Individual monitoring organizations understand their own data very well

- Issues arise when their data are combined with data from other organizations using different reporting methods

Key metadata elements (result level)

Metadata needed to unambiguously identify a result value

- Parameter name
- Sample fraction (filtration status)
- Chemical form (molecular or elemental)
- Numerical value
- Units
- Remark codes

Parameter name

10 most commonly reported nutrient parameters

- Ammonia
- Kjeldahl nitrogen (ammonia and organic nitrogen)
- Nitrite
- Nitrate
- Nitrite plus nitrate
- Nitrogen (mixed forms, including nitrite, nitrate, ammonia, and organic nitrogen)
- Organic nitrogen
- Organic phosphorus
- Orthophosphate
- Phosphorus (mixed forms, including orthophosphate, polyphosphates, and organic phosphorus)

Parameter name

- 1,046 unique variations on those 10 parameter names
  - 115 could not be unambiguously mapped
  - Mapping had to be done manually for the remaining 931
- Greatest number of unique variations
  - Orthophosphate -- 147
  - Ammonia -- 141
  - Phosphorus, mixed form -- 119

Parameter name

Reported Parameter Name	Harmonized Parameter Name
Inorganic nitrogen, calculated as $NH_3+NO_2+NO_3$	Inorganic Nitrogen $NH_3+NO_2+NO_3$
Inorganic nitrogen (nitrate and nitrite) as N	Nitrite + Nitrate
Inorganic Nitrogen	**?**
Nitrogen, Inorganic, Total	?
Nitrogen,Inorganic||Nitrogen,inorganic,total(ug/LasN)	?
Nitrogen,Inorganic||Nitrogen,inorganicasN	?
NitrogenInorganicTotal||InorganicNitrogen	?
NitrogenInorganicTotal||TotalInorganicNitrogen,asN	?

Parameter name

"Phosphate": Inorganic Phosphorus or Phosphorus (mixed forms)?

- Some use "phosphate" to mean phosphorus (mixed forms) (based on the field perspective)
- Some use "phosphate" to mean ortho- or total phosphate (based on the laboratory perspective)
- In many cases, it is not clear

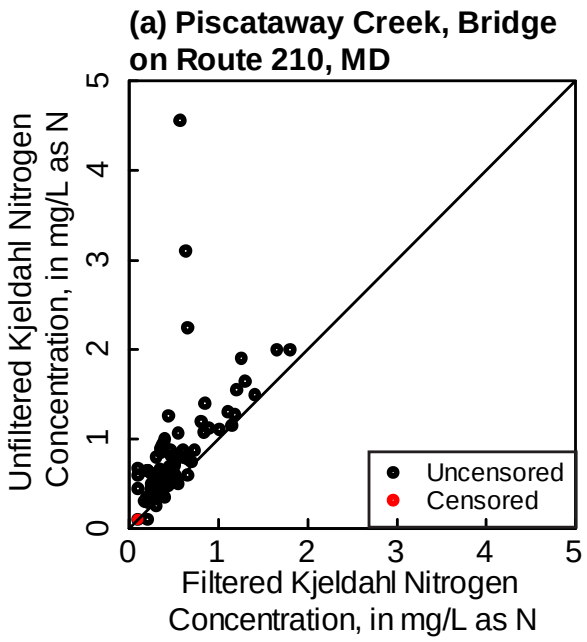
- In STORET, "phosphate-phosphorus" has been used in the past to represent phosphorus (mixed forms)
- Laboratory analytical method can be helpful
  - Not always populated or meaningfully populated (e.g. "Method 1")

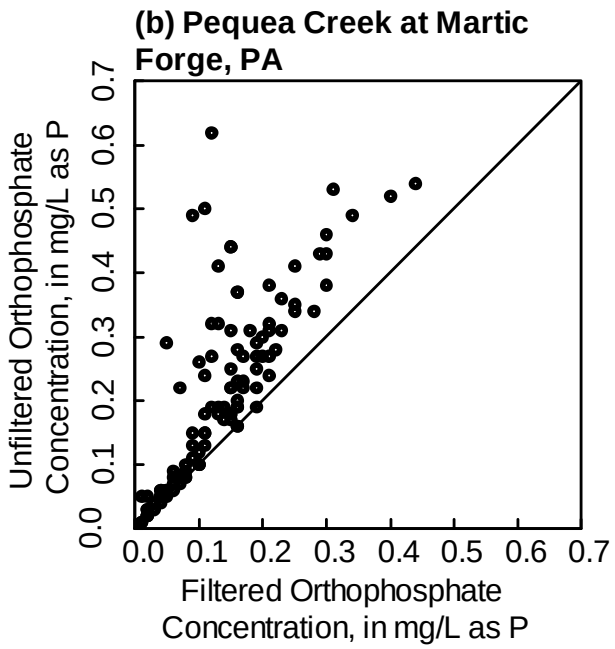
Filtration Status

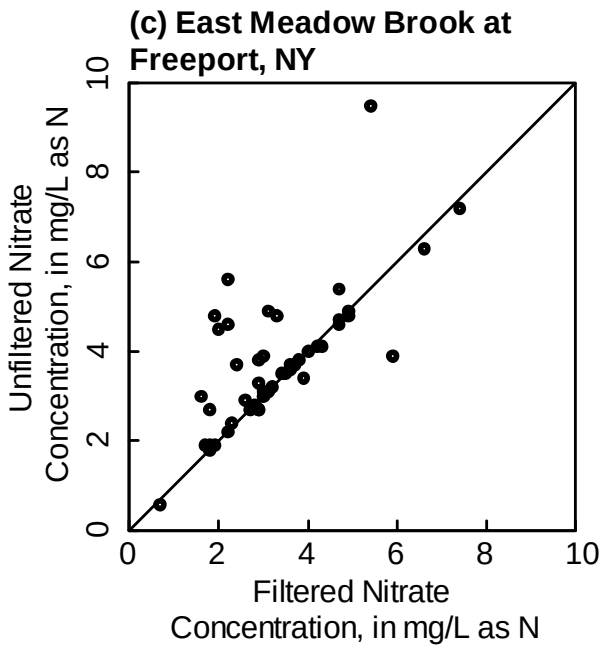
Filtration is the physical process used to separate the particulate and aqueous fractions of a water sample

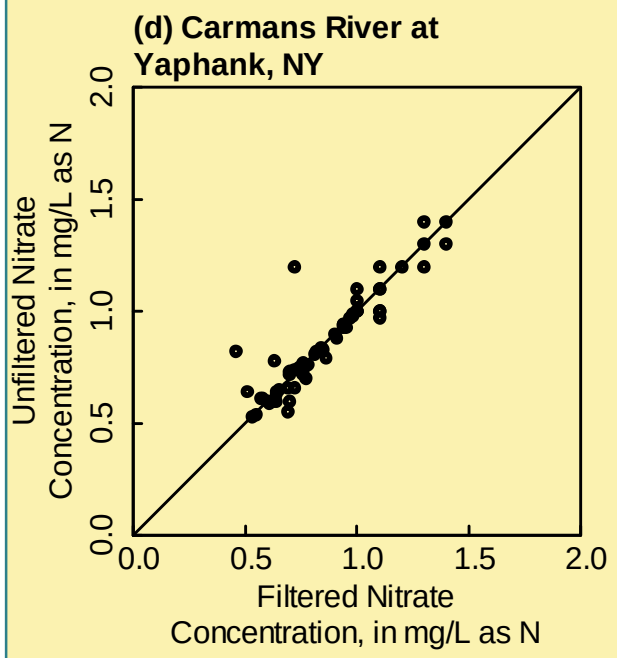
- In a single stream sample, it often is possible to determine both unfiltered and filtered variations of the same chemical
  - For example, total nitrogen and total dissolved nitrogen
  - The resulting values may be very different from one another
  - Not always tied to the laboratory method

Filtration Status









USGS
science for a changing world

Filtration Status/Parameter Name

Use of the word "TOTAL"

Used two ways

1. summation of multiple species (NH_3 + Organic N = Kjeldahl N)
2. an unfiltered sample

Unambiguous
Total Kjeldahl Nitrogen, Unfiltered
Total Kjeldahl Nitrogen, Filtered
Dissolved Total Kjeldahl Nitrogen

Ambiguous
Total Kjeldahl Nitrogen
Kjeldahl Nitrogen

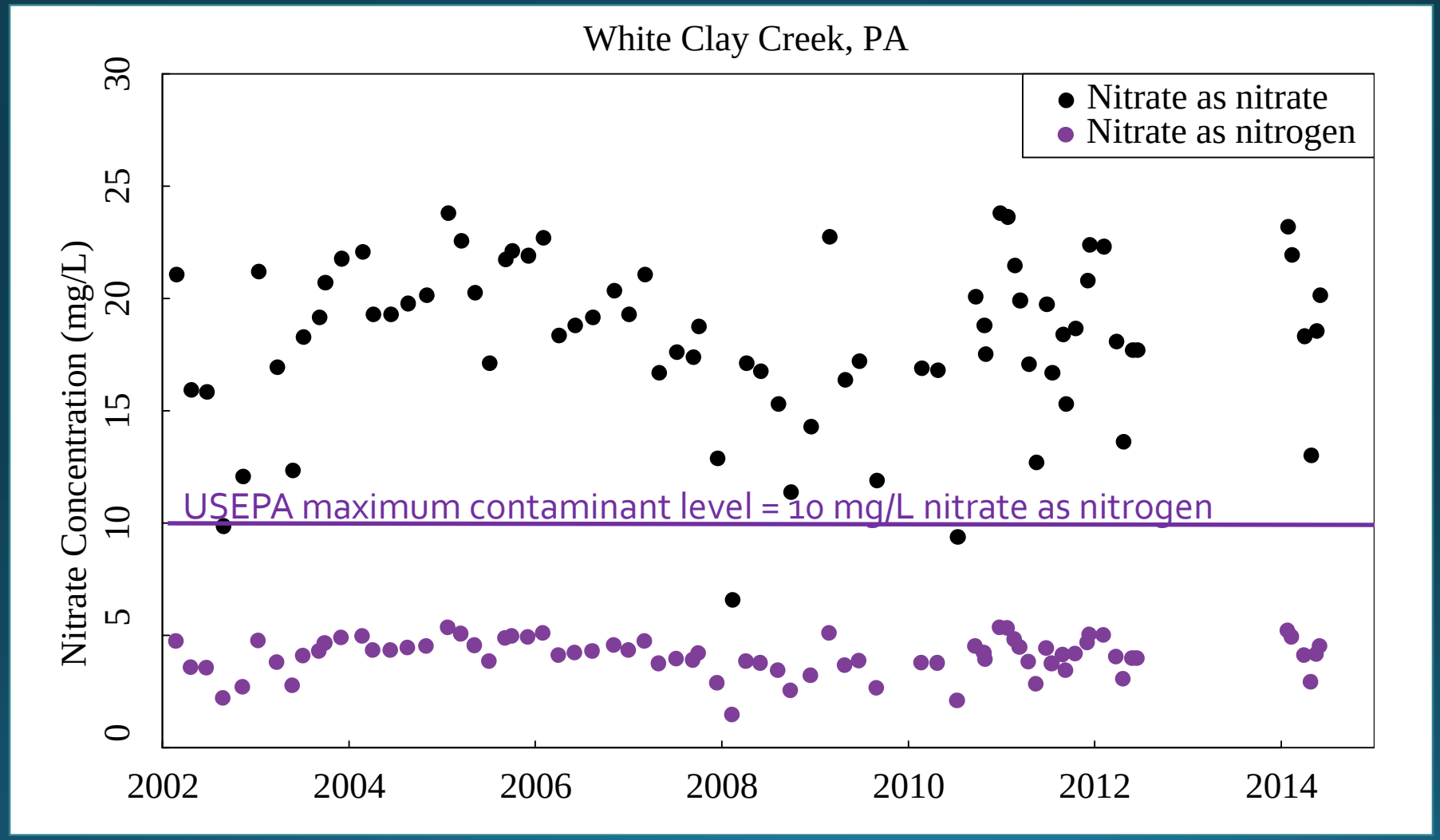
Chemical Form

Some water-quality results can be reported in two chemical forms—molecular or elemental

- For example, the same nitrate value can be reported two ways:
  1. "as nitrate" (the molecular form, which includes the full set of nitrogen and oxygen elements in the nitrate molecule)
  2. "as nitrogen" (the elemental form, with includes just the nitrogen element)

  The two differ by a factor of 4.5
- Similar reporting options for ammonia, nitrite, and orthophosphate

Chemical Form



USGS
science for a changing world

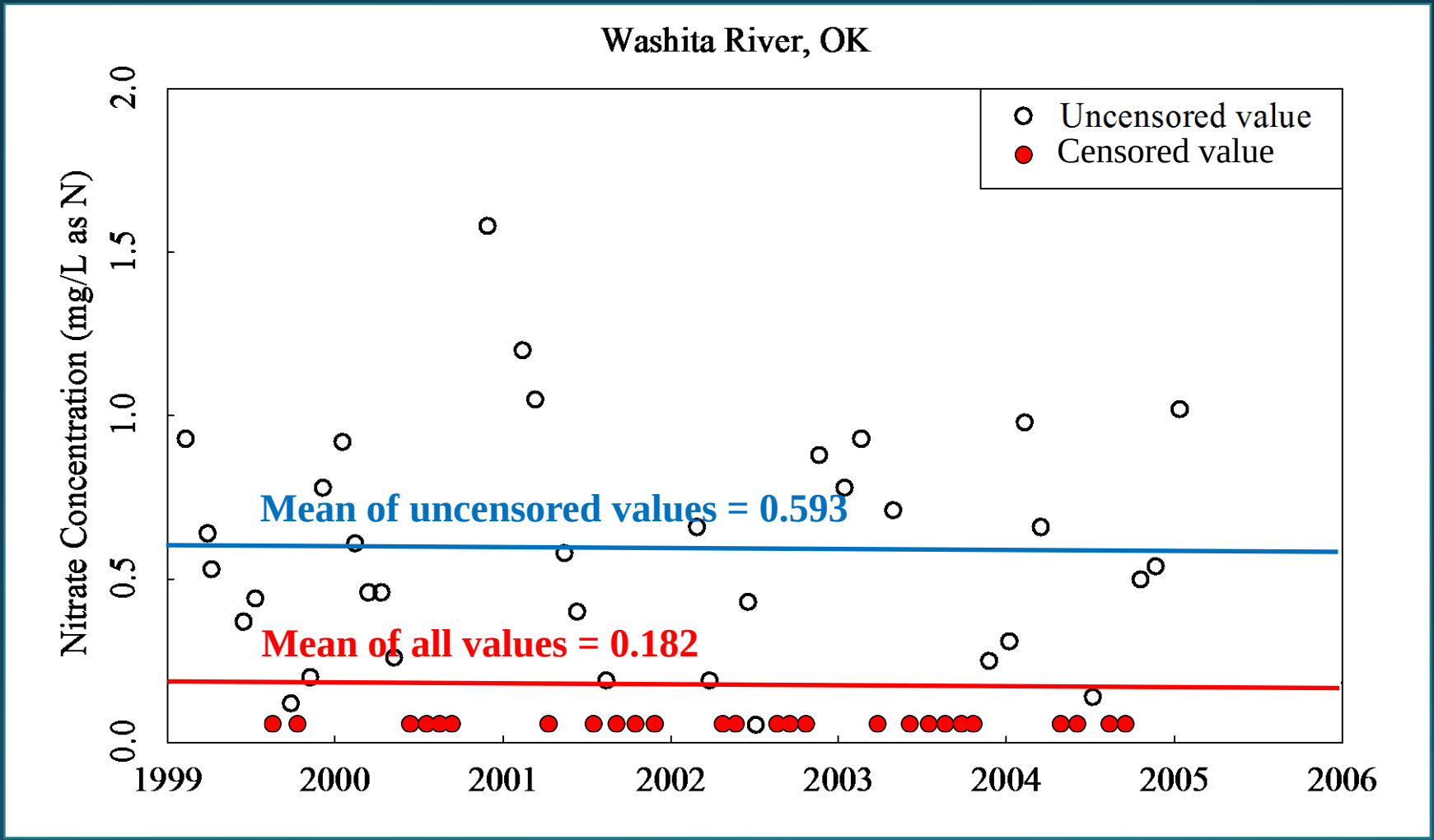
Units

- Many data values were reported without units (mg/L, µg/L, etc.)
- Other values were reported with units that were clearly inappropriate
  - For example, nephelometric turbidity units (NTU) with a parameter like ammonia

Remark Codes

- 587 unique remark codes
  - Many of these were not defined or were ambiguously defined
  - Not stored consistently (remark field versus comment field)
- 63 unique remark codes indicated laboratory censoring
  - Had to be manually identified
- Mis-identification of censored data can lead to substantial bias in data analyses
- Of the 488 sources surveyed, 118 did not provide any records at all with censored remark codes

Remark Codes



USGS
science for a changing world

Zero, negative, and missing values

- Zero values
  - Cannot have a nutrient concentration of zero
  - Often used to represent laboratory censoring
- Negative values
  - Laboratory censoring?
- Missing values with censored remark code
  - Was the sample not analyzed?
  - Laboratory censoring?

Count of affected records

- Of the 25,125,379 original records, 14,453,492 had missing or ambiguous information for one or more of the key metadata elements

Starting records	Affected records					
	Parameter name	Filtration status	Chemical form	Units	Remark codes	Zero, negative, missing
25,125,379	3,557,821	11,946,455	4,265,615	1,311,096	124,523	636,454

Count of affected records

- These counts do not include the unknown number of censored records that may have been deliberately withheld

Starting records	Affected records					
	Parameter name	Filtration status	Chemical form	Units	**Remark codes**	Zero, negative, missing
25,125,379	3,557,821	11,946,455	4,265,615	1,311,096	**124,523**	636,454

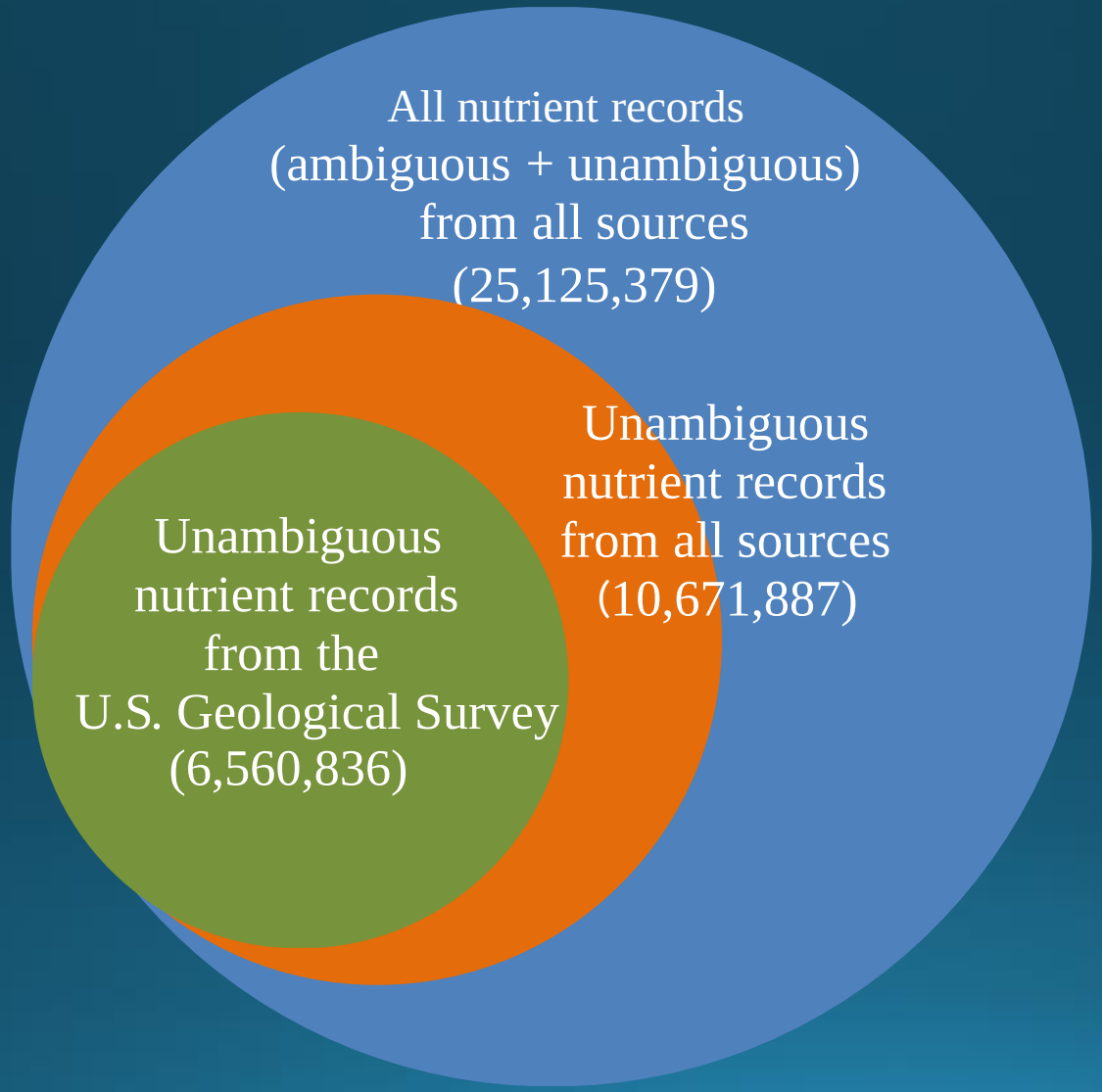
Increasing data usability

1. Use consistent parameter naming conventions—such as those curated by the Open Geospatial Consortium Naming Authority
2. Use “filtered”, “unfiltered”, and “particulate” to describe filtration status and avoid “total”
3. Report filtration status, chemical form (when relevant), and units of measurement
4. Use consistent remark code conventions, encode remark codes with data values (no text narratives), and fully report all censored data with detection limits
5. Discontinue the use of zero or negative values to denote censored data

Why does this all matter?

- Combining water-quality data from multiple sources can help counterbalance diminishing resources for stream monitoring
  - It can also lead to important regional and national insights that would not otherwise be possible
- Currently, metadata harmonization to make use of these multi-source data is time consuming, expensive, and inexact
- Different data users may make different assumptions about the same ambiguous data, potentially resulting in different conclusions about important environmental issues

Value of the affected records



Value of the affected records

- Cost to collect a stream-quality sample
  - Salary, travel, supplies, equipment, laboratory analysis, administrative support, database support, and quality control and quality assurance management costs
- Published estimates
  - Betanzo et al. (2015): $3,022 - $6,148 (adjusted for inflation)
  - Horowitz (2013) : $4,108 - $6,162 (adjusted for inflation)
  - Herrera Environmental Consultants and Aspect Consulting (2010): $2,179 (adjusted for inflation)

Value of the affected records

- Cost to collect a stream-quality sample
  - Maximum estimate $6,162
  - Minimum estimate $2,179
  - Average estimate $3,788 (average of $3,022, $6,162, and $2,179)

Value of the affected records

- 14,453,492 affected RECORDS
  - With multiple affected records in a given sample, that translates to 3,928,774 unique SAMPLES
- Estimated 20% of the samples are duplicated
  - 3,143,019 unique samples
- 3,143,019 x $3,788 = $12 billion
- Range= $6.8 billion to 19 billion

Value of the UNaffected records

- 2,706,136 unique SAMPLES = $8.2 billion
- Range= $4.7 billion to 13 billion

Value of the affected records

$12 billion represents a substantial collective investment by monitoring organizations in the United States

- This investment can be protected by implementing standardized metadata
- Standardized metadata can also help increase the use and value of legacy and future data beyond their original intent

Challenges of implementing standardized metadata

The full cost to implement standardized metadata is unknown, but will not be trivial

- Could require changes to local data processing, data bases, and web interfaces
- Funding is not readily available
- Metadata issues may be easier to address in recent and future data than in legacy data

Progress is underway

1. Water-Quality Portal
   - Developed by USEPA and USGS to be a single point of access for water quality data
   - Uses certain standardized metadata elements formatted according to the Water Quality Exchange (WQX) Outbound XML schema
2. Water-Quality eXchange (WQX) Nutrient Best Practices Guide
   - Recently created to promote consistency when submitting data through WQX
   - Produced through a collaborative effort between USEPA, USGS, and several State monitoring organizations
3. National Water-Quality Monitoring Council has published detailed metadata recommendations

QUESTIONS?

lsprague@usgs.gov

Publication

Water Research (DOI number 10.1016/j.watres.2016.12.024)

Acknowledgments

Gretchen Oelsner and Denise Argue

USGS
science for a changing world