Water Research 110 (2017) 252–261

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Challenges with secondary use of multi-source water-quality data in the United States

Lori A. Sprague [a, *], Gretchen P. Oelsner [b], Denise M. Argue [c]

[a] U.S. Geological Survey, Denver Federal Center, Box 25046, Bldg. 53, MS 415, Denver, CO, 80225, United States
[b] U.S. Geological Survey, 6700 Edith Blvd. NE, Suite B, Albuquerque, NM, 87113, United States
[c] U.S. Geological Survey, 331 Commerce Way, Pembroke, NH 03275, United States

ARTICLE INFO

Article history:
Received 29 August 2016
Received in revised form
28 November 2016
Accepted 15 December 2016
Available online 22 December 2016

Keywords:
Rivers
Water quality
Nutrients
Data
Metadata

ABSTRACT

Combining water-quality data from multiple sources can help counterbalance diminishing resources for stream monitoring in the United States and lead to important regional and national insights that would not otherwise be possible. Individual monitoring organizations understand their own data very well, but issues can arise when their data are combined with data from other organizations that have used different methods for reporting the same common metadata elements. Such use of multi-source data is termed "secondary use"—the use of data beyond the original intent determined by the organization that collected the data. In this study, we surveyed more than 25 million nutrient records collected by 488 organizations in the United States since 1899 to identify major inconsistencies in metadata elements that limit the secondary use of multi-source data. Nearly 14.5 million of these records had missing or ambiguous information for one or more key metadata elements, including (in decreasing order of records affected) sample fraction, chemical form, parameter name, units of measurement, precise numerical value, and remark codes. As a result, metadata harmonization to make secondary use of these multi-source data will be 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. The value of these ambiguous data is estimated at $US12 billion, a substantial collective investment by water-resource organizations in the United States. By comparison, the value of unambiguous data is estimated at $US8.2 billion. The ambiguous data could be preserved for uses beyond the original intent by developing and implementing standardized metadata practices for future and legacy water-quality data throughout the United States.

1. Introduction

Substantial investments have been made in water-quality monitoring in the United States by Federal, State, and local governments, Tribes, water utilities, universities, and many others since the inception of water-quality monitoring in the early 20th century (Myers, 2015), but funds for monitoring have slowly been eroding over time (National Research Council, 2004). For example, the U.S. Geological Survey (USGS) began monitoring a network of 505 stream sites throughout the United States in 1991 through the National Water-Quality Assessment (NAWQA) project of its National Water-Quality Program. Since that time, reductions in funding have led to a reduction in the NAWQA network to 117 sites, limiting the spatial and temporal resolution of key findings from the program (Rowe et al., 2013). Two other USGS national monitoring networks—the Hydrologic Benchmark Network (focused on undeveloped streams) and the National Stream-Quality Accounting Network (focused on large rivers)—were reduced from 54 to 15 sites and from 518 to 39 sites, respectively, between the 1970s and the 1990s (Hooper et al., 2001). Funding decreases have affected other organizations as well. For example, the Temporally Integrated Monitoring of the Environment (TIME) and Long Term Monitoring (LTM) program—a collaborative program managed by the U.S. Environmental Protection Agency (USEPA)—was initiated in the 1980s to examine trends in surface-water chemistry in response to changing air emissions and acid deposition. The number of lakes and streams monitored by the integrated program has dropped

* Corresponding author.
E-mail addresses: lsprague@usgs.gov (L.A. Sprague), goelsner@usgs.gov (G.P. Oelsner), dmargue@usgs.gov (D.M. Argue).

http://dx.doi.org/10.1016/j.watres.2016.12.024

over time due to funding cuts, leading to a loss of sites in Colorado, the Upper Midwest, and Vermont (U.S. Environmental Protection Agency, 2009). An evaluation of historical nutrient data collected by multiple organizations that reported data in the USGS National Water Information System (NWIS) database and the USEPA Storage and Retrieval (STORET) database found that the number of monitoring sites nationwide having at least 2 years of nutrient data and at least 20 samples increased throughout the 1970s, but widespread decreases were occurring by the early- to mid-1990s. The timing of the peak number of sites ranged from the mid- to late-1970s (in the Midwest and Pacific Northwest) to the early- to mid-1990s (in the South) (Saad et al., 2011).

As the collective resources for water-quality monitoring in the United States have decreased, it has become increasingly critical to leverage information from the disparate monitoring networks to address regional and national water-resource issues. Major monitoring organizations have made substantial progress in making their data publicly available (Myers, 2015), but fragmented data storage practices (including nomenclature, use of metadata, and data storage and dissemination platforms) continue to pose challenges when combining data from multiple organizations. Individual monitoring organizations understand their own data very well and are able to make use of those data locally. But problems can arise when their data are combined with data from one or more other organizations using a different dissemination platform and(or) using a different approach to describe the same common metadata elements.

In this study, we surveyed available nutrient data and metadata from 488 organizations that have conducted monitoring in U.S. streams at any point since 1899 (Supplementary Table 1). The survey included ambient monitoring data that were publicly accessible from Federal, State, regional, and local government agencies and non-governmental organizations, and included more than 25 million water-quality records from 321,927 sites. Nutrients are the focus of this survey because they were recently found to be the most widespread chemical stressor in U.S. streams—46% of U.S. streams have high levels of nitrogen, phosphorus, or both (U.S. Environmental Protection Agency, 2016a). Nutrients also are among the most commonly monitored water-quality parameters in the United States, and as a result, they provide a window into common challenges encountered for a wide variety of parameters when using data from multiple sources. The ultimate objective of this study was to identify major inconsistencies in metadata practices that limit the secondary use of multi-source nutrient data. The term “secondary use” is defined herein as distinct from the term “primary use”. “Primary use” refers to use of data for the original intent determined by the organization that collected the data. “Secondary use” refers to the use of the same data for other purposes. Data users making secondary use of the data are defined herein as “secondary data users”.

2. Material and methods

All available nutrient concentration records for streams and rivers were retrieved from NWIS and STORET, the two largest national water-quality databases in the United States. STORET serves water-quality data from the USEPA and multiple submitting organizations; NWIS primarily serves data from the USGS, but also contains a small amount of data collected by other organizations. Data were retrieved from NWIS in May 2013 and from STORET in October 2013; more recent data from both data bases were obtained from the Water-Quality Portal (which serves data from NWIS and STORET) between January and March 2015. Additional nutrient concentration records from major water-resource agencies in each State also were included in the evaluation. These data were available from local agency databases or directly from staff within the agency. These data were obtained between June 2010 and May 2015; some organizations provided data on more than one date. Overall, approximately 70% of the final nutrient records came from NWIS and STORET.

The 488 organizations across the United States included in the survey comprised 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 nongovernmental organizations; 34 volunteer organizations; and 50 private organizations (Table 1). One-half of the data came from State water, natural resources, or environmental protection agencies; another one-third of the data came from Federal agencies.

Variations in nomenclature, terminology, and jargon in the water-resources community have long been problematic when aggregating water-quality data from multiple sources. In particular with nutrients, there are different chemical and physical forms in natural waters (notably, partitioning between organic and inorganic forms and dissolved and particulate phases) and different methods of field collection and laboratory analysis. Defined standards and universally accepted nomenclature for nutrients in their various forms have been unavailable or have not been widely adopted. Ambiguous and(or) inconsistent terminology, ambiguous and(or) incomplete metadata, and clearly incorrect data and metadata were all encountered in our survey. To evaluate these issues, we identified the key result-level metadata elements needed to unambiguously interpret each value. Result-level metadata applies to an individual numerical value from a sample, and includes metadata elements such as units of measurement. Often, there are multiple results in a sample (for example, nitrate, ammonia, and orthophosphate results together in a single sample). Sample-level metadata applies to all results in the sample, and includes elements such as date, time, site name, site location, and sampler type. This study evaluates only result-level metadata.

The key result-level metadata elements in our evaluation included parameter name, sample fraction (filtration status), chemical form (molecular or elemental), numerical value of the analysis, units of measurement, and remark codes indicating either poor quality or values detected below the laboratory reporting limit. The number of records that could be unambiguously interpreted were tallied, along with the number of records that had missing or ambiguous information for each of the key metadata elements.

3. Results and discussion

3.1. Records with missing or ambiguous metadata

Nearly 14.5 million of the original 25 million records surveyed—or over one half of the original records—had missing or ambiguous information for one or more of the key metadata elements. To make use of any of these 14.5 million records, a secondary data user would need to make some assumptions about the value.

3.1.1. Parameter name

Of the many different nutrient parameters reported by organizations throughout the United States, ten are most commonly used to characterize nutrient concentrations in streams and are most frequently reported in water-quality databases. These common parameters included 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,