Environmental Science & Technology

Viewpoint

Cite This: Environ. Sci. Technol. XXXX, XXX, XXX−XXX

pubs.acs.org/est

Three Principles to Use in Streamlining Water Quality Research through Data Uniformity

Andrew R. Shaughnessy,*[,†] Tao Wen,[§] Xianzeng Niu,[§] and Susan L. Brantley[†,§]

[†]Department of Geosciences, The Pennsylvania State University, University Park, Pennsylvania 16802, United States

[§]Earth & Environmental Systems Institute, The Pennsylvania State University, University Park, Pennsylvania 16802, United States

Integrating data collected from government agencies, the private sector, and academia is a difficult but important task in sustaining water research.[1] Easy access to data through online repositories has created new opportunities in research using data analytics that were difficult in the past. For example, Kaushal et al. (2018) compiled data from United States Geological Survey (USGS) sites to investigate a phenomenon that they call "freshwater salinization syndrome".[2] But to really facilitate such studies, problems in reporting convention should be addressed. Niu et al. (2018) called for standardization in water quality reporting, and after that paper, several authors wrote with questions as to how to decide such a framework.[3] We looked at data in the largest U.S. water quality data repository, that is, the Water Quality Portal (WQP, www.waterqualitydata.us), to suggest a next step.

As of 29 August 2019, the WQP contains data from 990 462 surface water and 1 508 639 groundwater sites globally. Before implementation of the WQP in 2012, the data from >400 federal, state, and local agencies were published in different locations; now researchers can find these data sets in one central location.[4] The WQP allows researchers to easily find and access data; however, data compiled from different providers highlights a new issue, data consistency.

Among the most inconsistently reported water quality parameters are some of the most important: the nutrients, including species of nitrogen (e.g., nitrate, nitrite, ammonia) and phosphorus (e.g., orthophosphate, total phosphorus). Understanding nutrient dynamics is important for mitigating harmful algal blooms and informing best management practices, among others. One issue with anion data in general is that they are commonly reported in both polyatomic and elemental mass concentrations. Such inconsistencies decelerate research through the time-consuming task of data cleaning.

For example, in the WQP, nitrate has 32 name-unit combinations, which includes 3 772 511 samples from 274 764 sites. Of these samples, 93% report nitrate as "Nitrate". Within the "Nitrate" category, 38% of samples have missing or ambiguous units (Table 1). Reporting nitrate as "mg/L" without specifying "as N" or "as NO_3", is ambiguous. Utilizing such data without going through methodologies or using specialized domain knowledge is difficult to impossible. Combing through method codes may be impractical depending on the volume of data or the background knowledge of the investigator.

An additional issue related to the inconsistency in units is double reporting. Today, some samples are reported twice in the WQP, both as "mg/L as element" and "mg/L as polyatomic." To retain the maximum number of samples, users need to convert all data to the same units. Duplicates must then be removed before analysis, which is time-consuming. Some may never remove them. Alternately, scientists might simply use data listed for one set of units, which leaves out some unique observations since not all data are double reported. Thus, inconsistency in reporting can cause double use or can lower the volume of data for analysis.

From the point of view of fundamental chemistry, molar units would avoid all issues of ambiguity and repetition; however, molar units are rarely reported in the WQP (<0.01%). Instead, we recommend adopting the most commonly used unit-name combinations already in the WQP. We adopt this "majority-rules" approach over the more fundamental approach so that change is more manageable and fewer data providers need to update the reporting standard. We thus advocate to use "mg/L" as the concentration unit.

Using the majority-rules principle to choose between "as N" and "as NO_3" is problematic because the proportions of use are roughly equally split. Here, we invoke another principle based on safety. Our "safety-first" approach is evident when misinterpretation is considered. The EPA drinking water standard for nitrate is 10 mg/L as N. If the nitrate concentration in a sample is 20 mg/L as NO_3 but is misinterpreted as 20 mg/L as N, then one might incorrectly

Received: October 24, 2019

DOI: 10.1021/acs.est.9b06406
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Table 1. Number of Samples of Nitrate and Phosphate in the WQP by Sample Name and Unit

name	mg/L as N or P	mg/L	mg/L asNO_3 or asPO_4	no unit	other units
Nitrate	1 165 165	1 120 103	833 907	284 019	88 244
Nitrate as N	0	145 135	0	46 857	3186
Nitrate-N	0	72 532	0	0	0
Nitrate-Nitrogen	0	13 287	0	0	0
Nitrate-nitrogen	0	73	0	0	0
Orthophosphate	714 369	1 882 526	616 578	586 763	51 876
Phosphate-phosphorus	0	1 698 200	0	96 802	52 721
Phosphate-phosphorus as P	0	482 418	0	21 228	238 719
Orthophosphate as P	0	385 506	0	58 391	12 385
Orthophosphate as PO4	0	30 203	0	4633	2380
Phosphate-phosphorus as PO4	0	18 422	0	667	0
Phosphate	0	21 966	0	0	0
Ortho-Phosphate-Phosphorus	0	2322	0	0	0

conclude that the sample is unsafe to drink. However, if the sample was 20 mg/L as N and misinterpreted as 20 mg/L as NO_3, then one might incorrectly conclude that the water is safe to drink. The latter is more dangerous to humans than the former; therefore, we recommend the units of "mg/L as NO_3" on the principle of "safety-first".

Additionally, reporting nitrate as nitrate is less ambiguous than reporting nitrate as nitrogen. This emphasizes a third principle, namely "elimination of ambiguity". Some analytical methods measure just nitrate, while others measure mixed forms of nitrogen. Data providers should report "Nitrate" when nitrate is being measured and "Inorganic Nitrogen" when other nitrogen forms are lumped together in an analysis.

We can apply these principles to P as well. Phosphate in the WQP contains 56 different name-unit combinations, which includes 6 979 075 samples from 301 625 sites. The majority of samples are listed in units of "mg/L" and we recommend reporting phosphate as "Orthophosphate" in units of "mg/L as PO_4". Like Nitrate as a descriptor, Orthophosphate unambiguously reveals the speciation of the analyte. Using that nomenclature eliminates the problem pointed out by Sprague et al. (2017): the name "phosphate-phosphorus" is used variously to mean total phosphorus, mixed forms of phosphorus, or orthophosphate.[5] Although we have not gone through all the other species in the WQP, we think that the three principles of majority-rules, safety-first, and eliminating-ambiguity would be helpful for names and units of other analytes as well.

As pointed out in Niu et al. (2018), small and achievable steps are needed to promote efficiency in important research on water quality data, and these steps can be taken by each researcher or agency.[3] Such steps may not be made as easily by the major government agencies that provide data (i.e., USGS and USEPA) because they might have to change long-standing traditions. Perhaps individuals should implement the "majority-rules", "safety-first", and "elimination of ambiguity" principles with nitrate and phosphate as a first achievable step forward. Although change is often resisted, most data providers want their data to be used and maintained, and they want to continue data collection: why not move toward reducing ambiguity and increasing uniformity in data sets to make data utilization easier and improve data reusability?

AUTHOR INFORMATION

Corresponding Author

*E-mail: ars637@psu.edu.

ORCID

Andrew R. Shaughnessy: 0000-0001-9574-5053

Tao Wen: 0000-0002-6113-7532

Xianzeng Niu: 0000-0002-1702-5381

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Patterson, L.; Doyle, M.; King, K.; Monsma, D. *Internet of Water: Sharing and Integrating Water Data for Sustainability*; The Aspen Institute: Washington, DC. 2017.

(2) Kaushal, S. S.; Likens, G. E.; Pace, M. L.; Utz, R. M.; Haq, S.; Gorman, J.; Grese, M. Freshwater salinization syndrome on a continental scale. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (4), E574−E583.

(3) Niu, X.; Wen, T.; Li, Z.; Brantley, S. L. One Step toward Developing Knowledge from Numbers in Regional Analysis of Water Quality. *Environ. Sci. Technol.* **2018**, *52* (6), 3342−3343.

(4) Read, E. K.; Carr, L.; De Cicco, L.; Dugan, H. A.; Hanson, P. C.; Hart, J. A.; Kreft, J.; Read, J. S.; Winslow, L. A. Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resour. Res.* **2017**, *53* (2), 1735−1745.

(5) Sprague, L. A.; Oelsner, G. P.; Argue, D. M. Challenges with secondary use of multi-source water-quality data in the United States. *Water Res.* **2017**, *110*, 252−261.

DOI: 10.1021/acs.est.9b06406
Environ. Sci. Technol. XXXX, XXX, XXX−XXX