Environmental Science & Technology

Viewpoint

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

pubs.acs.org/est

One Step toward Developing Knowledge from Numbers in Regional Analysis of Water Quality

Xianzeng Niu,*[,†] Tao Wen,[†] Zhenhui Li,[‡] and Susan L. Brantley[†]

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

[‡]College of Information Science and Technology, The Pennsylvania State University, University Park, Pennsylvania 16802, United States

Data define scientific research. While researchers in the past published data in print-only venues that were hard to find, nowadays data are mostly published online. In environmental science, the increase in data accessibility allows scientists to conduct regional water quality assessments using integrated data sets.[1,2] As data searching becomes easier, however, compiling data from multiple sources remains difficult because of issues related to data definition, structure, and integrity. Using our recent research as an example, this viewpoint clarifies challenges in integrating water quality data, and demonstrates the need to establish standards for data management. Rather than advocating "a complete fix" that may never happen, however, we advocate for simply one baby step to grow understanding of water quality. Each such small step will attract more water quality specialists to discover what can be learned with "big data".

We recently studied the impacts of shale gas extraction on water quality in Pennsylvania,[1] using four online databases: Water Quality Portal (WQP) sponsored by National Water Quality Monitoring Council (https://www.waterqualitydata.us/), the Susquehanna River Basin Commission (SRBC, http://mdw.srbc.net/waterqualityportal), ShaleNetwork (accessed through CUAHSI's HydroClient, http://data.cuahsi.org/), and Critical Zone Observatory (CZO, http://www.czo.psu.edu/). These innovative data portals have revolutionized water data discovery. As we learned, however, the easy discovery also highlighted other difficulties in data integration.

Data Collection and Screening. Mostly we were able to find the metadata that explains data structures and definitions (e.g., "User Guide" in WQP). However, some instructions were confusing or were unavailable, and we had to carefully inspect the data to make inferences. A metadata dictionary is essential to integrate data sets.

After data downloading, we screened and removed irrelevant data. For example, for the WQP data set, we determined that 23 out of 63 attributes (i.e., columns) were relevant to our question. We then selected 83 of 1556 chemical analytes of relevance to shale gas contamination. We also restricted the "sample media" to "streams/rivers" since we were interested in surface water only. The final resulting data set left ∼21% of the initially downloaded data.

Data Cleaning. This step included identification and correction of erroneous or implausible data. Data cleaning is often overlooked as compared to data analysis and interpretation, perhaps because it is often assumed that data are thoroughly reviewed by providers and can be used as "plug and play". This assumption is common, for example, when data come from government agencies. However, inconsistency in data definitions and mistakes happen everywhere.

The most common problem is inconsistency in variable names and units. We discovered 361 variable-unit combinations for 83 analytes. For nitrate alone, there were a total of 14 combinations (Table 1). If we had not standardized the data set, we would have lost data during querying due to unmatched variable names or our results would not have been trustworthy because of incorrect values (i.e., different units).

Other common problems included different codes for missing values (e.g., −9999, −8888) and mix of numerical and textural values (e.g., "NA", "No Data", "NULL" in a numerical field as "no data"). Issues of data integrity drove us to remove 42,525 values (about 1.3%).

Data cleaning also involved searching for redundancy. Redundancy becomes more problematic when data are compiled from multiple sources. We assumed records were duplicated when they were collected from the same location (latitude, longitude) at the same time with identical data values and units for the same analyte. A total of 194,864 redundant data values (6%) were removed.

Data Integration. After cleaning, we had to define a common terminology and data structure to integrate all data

Received: February 23, 2018

DOI: 10.1021/acs.est.8b01035
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

Table 1. Variable-Unit Combinations for Nitrate Found in Pennsylvania Water Quality Data Compiled from Multiple Data Sources (See Text for Details)

variable	unit	counts	new variable	new unit	data source
nitrate	μmol/L	45	Nitrate	μg Nitrate/L	CZO
nitrate	μM	406	Nitrate	μg Nitrate/L	CZO
Nitrate-N	mg/l	749	Nitrate	μg Nitrate/L	SRBC
Nitrate-N D	mg/l	1940	Nitrate	μg Nitrate/L	SRBC
Nitrate-N T	mg/l	5208	Nitrate	μg Nitrate/L	SRBC
Nitrate	mg/kg as N	40	Nitrate	μg Nitrate/L	WQP
Nitrate	mg/l	22413	Nitrate	μg Nitrate/L	WQP
Nitrate as N	mg/l	3262	Nitrate	μg Nitrate/L	WQP
Nitrate	mg/l as N	1105	Nitrate	μg Nitrate/L	WQP
Nitrate	mg/l as NO_3	21 098	Nitrate	μg Nitrate/L	WQP
Nitrate	ppb	6	Nitrate	μg Nitrate/L	WQP
Nitrate	ppm	123	Nitrate	μg Nitrate/L	WQP
Nitrate	μeq/L	24	Nitrate	μg Nitrate/L	WQP
Nitrate as N	μeq/L	31	Nitrate	μg Nitrate/L	WQP

sets into a single queryable database. Ontologies provide a way to do this.[3,4] In our project, we used a modified version of the Observations Data Model (ODM) from CAUHSI (https://www.cuahsi.org/). Accordingly, we also compiled a data dictionary to map our controlled vocabulary onto the terms used in original databases. For example, we assumed "filtered" water data should include data labeled "Dissolved", "Filterable", and "Filtered" in the original data sets. We likewise assumed that data could be labeled "Unfiltered" if they were originally labeled as "Total", "Total Recoverable", "Recoverable", or "Unfiltered". By combining data sets we were forced to lump some subtle differences in sample type.

Finally, the integrated and cleaned water quality data were stored as a relational database. All variables can now be easily queried and exported in a common format and used in programs for statistical analysis, data modeling and visualization. This experience leads us to advocate standardization in water data management to limit the time and effort needed for cleaning and integration.

A Way Forward. It would be efficient if we could agree to keep all water data "uniform" in all aspects of data from sampling, analysis, reporting, to storage. But uniformity will only be achieved in small steps. Even large data-sharing communities such as the worldwide network of Critical Zone Observatories do not agree upon standards for sampling and analysis. In fact, standardization is antithetical to the arc of chemical science where newly discovered analytical techniques are continually adopted.

Instead, we advocate an obvious baby step forward to "partial standardization". Each such step clarifies what we can do with bigger data sets and makes the next step easier. The first small step is to agree to report data in standard variable names and units with the same notation for censored or lacking data. It sounds easy but agreements among data providers are notoriously hard to reach. Achieving this step would lead to other small steps toward standard publishing formats such as ODM. We must stop trying to solve the entire data management problem in one big step. Rather, such small achievable steps must be pursued until more environmental chemists see the utility and power of "big data" and demand the harder agreements. Ten years ago, we lacked online data portals. As we discover data more easily nowadays, we clearly see how data and environmental scientists could work together to assess water quality in novel ways.[1,2,5] Surely, we believe that each baby step in "data standardization" would lead to a big leap toward developing knowledge from numbers in regional analysis of water quality.

AUTHOR INFORMATION

Corresponding Author

*E-mail: xzniu@psu.edu.

ORCID

Xianzeng Niu: 0000-0002-1702-5381

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Niu, X.; Wendt, A.; Li, Z.; Agarwal, A.; Xue, L.; Gonzales, M.; Brantley, S. L., Detecting the effects of coal mining, acid rain, and natural gas extraction in Appalachian basin streams in Pennsylvania (USA) through analysis of barium and sulfate concentrations. *Environ. Geochem. Health* **2017**, DOI: 10.1007/s10653-017-0031-6.

(2) Raymond, P. A.; Hartmann, J.; Lauerwald, R.; Sobek, S.; McDonald, C.; Hoover, M.; Butman, D.; Striegl, R.; Mayorga, E.; Humborg, C.; Kortelainen, P.; Dürr, H.; Meybeck, M.; Ciais, P.; Guth, P. Global carbon dioxide emissions from inland waters. *Nature* **2013**, *503*, 355.

(3) Earley, S. Really, Really Big Data: NASA at the Forefront of Analytics. *IT Professional* **2016**, *18* (1), 58−61.

(4) Niu, X. Z.; Williams, J. Z.; Miller, D.; Lehnert, K.; Bills, B.; Brantley, S. L. An Ontology Driven Relational Geochemical Database for the Earth's Critical Zone: CZchemDB. *Journal of Environmental Informatics* **2014**, *23* (2), 10−23.

(5) Brantley, S. L.; Vidic, R. D.; Brasier, K.; Yoxtheimer, D.; Pollak, J.; Wilderman, C.; Wen, T. Engaging over data on fracking and water quality. *Science* **2018**, *359* (6374), 395.

DOI: 10.1021/acs.est.8b01035
Environ. Sci. Technol. XXXX, XXX, XXX−XXX