Please see the following manuscript for detailed methods used to harmonize the datasets in LAGOS-NE; in particular, many of the specific methods and procedures are found in the online “Additional Files” available at: https://gigascience.biomedcentral.com/articles/10.1186/s13742-015-0067-4

Soranno, P.A., E.G. Bissell, K.S. Cheruvelil, S.T. Christel, S.M. Collins, C.E. Fergus, C.T. Filstrup, J.F. Lapierre, N.R. Lottig, S.K. Oliver, C.E. Scott, N.J. Smith, S. Stopyak, S. Yuan, M.T. Bremigan, J.A. Downing, C. Gries, E.N. Henry, N.K. Skaff, E.H. Stanley, C.A. Stow, P.-N. Tan, T. Wagner, K.E. Webster. 2015. Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science and data reuse. GigaScience 4:28 doi:10.1186/s13742-015-0067-4

We acquired water quality datasets for LAGOS-NE-LIMNO v1.087.1 by contacting individuals at each of the 17 state and 5 tribal agencies. These contacts helped us to identify the state-agency collected dataset required by the Clean Water Act and which is most likely to be in the public domain. In this way, we were able to acquire at least one (and typically more) dataset from each of the 17 states. Because states vary in sampling approach and intensity (see below for details), we sought to supplement these datasets with other known sources of water quality data, including university researchers, federal agencies, and non-profit groups. The full list of data sources acquired is in Soranno et al. (2015) in “Additional File 17”. Based on resources and effort required to include them, we incorporated a subset of these datasets in LAGOS-NE-LIMNO v1.087.1.

All methods to create this module are described in Soranno et al. (2015). Briefly, for each dataset acquired, we authored LAGOS-NE metadata in EML to aid in data provenance. We also incorporated key metadata features (e.g., methods used, censor codes (if applicable)), and sampling program information) into the database so that future users could easily identify these important attributes. Because each dataset was unique in structure, file format, and naming conventions, we manually processed each dataset and its metadata so that they could be translated into the standard LAGOS-NE vocabulary and data model. Although labor-intensive, we created customized R scripts to process and load each dataset separately, which are available in LAGOS-NE-RAWDATA data module.