Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2017-52-AC1, 2021

Interactive comment on "The National Eutrophication Survey: lake characteristics and historical nutrient concentrations" *by* Jemma Stachelek et al.

Jemma Stachelek et al.

stachel2@msu.edu

Received and published: 13 November 2017

Reviewer Responses

0.1 Reviewer 1

- *This is an important data set and it is good to see that it is becoming available in electronic format. The authors appear to have used good methods to bring in a large dataset. I suspect that the manual entry error rate would be close to the*

lower end of their error rate.

- *There are multiple copies of these reports floating around in various libraries. Would it improve things to scan multiple copies of the same data and check the copies against each other?*

  This is a good suggestion. However, the effort to do this would be quite large. Also, since we manually verified the accuracy of every field, this task seems unnecessary. This may have been a more useful step at the beginning of our optical character recognition quality control step, affording us perhaps two other scan options, and the best of three could have been populated into our final repository. We would still suspect that a manual quality control step would have been necessary.

- *I am curious why the authors did not try to bring in the data that Stomp et al. (2011) digitized. They could then compare the two datasets. At a minimum they should try to combine the datasets, at least chlorophyll would be nice.*

  The R code provided in the article supplement could be extended to accomplish this task. It would require writing a parsing function to extract the values from the raw files produced as a result of the optical character recognition algorithm. However, this section has very complex formatting so we felt that doing so would be redundant. In addition, the Stomp et al. dataset is already archived and we felt that it was not good data management practice to duplicate their data in a second location.

- *I am not functional in R so could not assess that part of the data product.*

- *The data could use some quality checking. For example, there are points where phosphate exceeds total phosphorus. This is not possible.*

  We also agree that some data points reported through these four EPA documents did not make scientific sense. However, our goal was to repro-

duce the dataset exactly as it appears in the original documents, and not to judge the scientific accuracy of these data. We expect that a proper analysis would be conducted in cooperation with an original research project. We have added text to the manuscript pointing out this issue.

- *It also would be good if the all data were all quality checked. If it takes about 1 second per data point, I calculate it would take 3 hours each for the team of authors working in pairs to check the whole thing. That would lead to a cleaner data set as well as making the error rate certain.*

  We agree that manual quality checking is essential. We were able to calculate (and report) our error rate because we had already done this manual checking.

- *It would be nice to have retention all in one type of units (not years or days mixed)*

  We agree. However, if we converted the data in this way it would make it more difficult to check the data against the original documents.

- *Table 2. Could use the units*

  Good catch! We have updated the table.

0.2 Reviewer 2

- *Please use the full doi designation: https://doi.org/10.5063/f1kk98r5 ? The full designation allowed this reviewer to avoid a search through DataCite to access the KNB Site.*

  The ESSD author instructions say to report a DOI in the abbreviation form (10.5194/xyz). We suspect that this could be hyperlinked to the full designation on typesetting. For example, it appears that the partial doi code

provided in the abstract of the pdf manuscript has been hyperlinked to the full doi in the html abstract.

- *Data very well organized and easily accessible. Very good metadata. Spot checks (Montana, Illinois) showed believable locations and values, evidently quality control has worked reasonably well. Good product, potentially very useful as baseline for both chemical and hydrological / geomorphological purposes.*

- *No information about sampling date in the master .csv file? E.g. a reader gets reference to the report number (.pdf 475, published 1978) and to a page number (for Bloomington Lake, Maclean County, Illinois, actually on page 79 rather than 81 as in .csv file), but no reference to sampling dates. Bloomington Lake data shows nutrient and biological samples collected on 5/11/73, 8/9/73 and 10/17/73. For MacDonald Lake (.pdf 477, page 78 rather than 80 as in .csv) Montana, nutrient and biological sampling on 6/1/75 and 7/28/75. In text we read that sampling of geographic regions occurred by year (e.g. 1973 for southeastern including Illinois and 1975 for western) but the user does not see actual dates where available, or would need to extract those dates themselves? But apparently none of the raw files captured these sampling dates from the original .pdf?*

  The sampling dates you see are exclusive to the "Biological Characteristics" section, which we did not transcribe, because this has already been done in Stomp et al. (2011). In contrast, the data we report are annual means computed from monthly samples. The NES reports provide no further details about specific sampling dates. We have added clarifying text to the manuscript and metadata on this point.

- *Data from Illinois resides in two separate sections of .pdf 475 (page numbers 80 to 99 and 100 to 110 contiguous) but one needs to search by storet_code or state name to find all data per each state? This scattering arises from processing sequence?*

Yes

- *Page numbers in .csv file refer to page number of digitized .pdf, not to page numbers used within the individual reports? I did not see reference to this small discrepancy in the metadata.*

  We have appended additional text to the "pagenum" metadata field. It now reads: "page number of the pdf (not the report page number)".

- *Need specific clarification about the page number discrepancies and about whether the digitisation process captured the sampling date*

  See above.