Digitization of data from published plots v1.0

Gustav Nilsonne[1], Love Ahnström[1]

[1]Karolinska Institutet

L Love Ahnström

APR 23, 2023

DOI:
dx.doi.org/10.17504/protocols.io.n2bvj8wnngk5/v1

Protocol Citation: Gustav Nilsonne, Love Ahnström 2023. Digitization of data from published plots v1.0. **protocols.io** https://dx.doi.org/10.17504/protocols.io.n2bvj8wnngk5/v1

Protocol status: Working
We use this protocol and it's working

Created: Apr 10, 2023

Last Modified: Apr 23, 2023

PROTOCOL integer ID: 80239

ABSTRACT

Purpose
This is a protocol for digitizing data that have been published in diagrams available as image files (such as .img or .png). The typical use-case is to extract data from published scientific papers for secondary analyses or for meta-analysis.

Protocol

1 Create a folder with a descriptive name to identify the data that are being digitized and the date,

for example the first author and year of the published paper, the year of publication, the figure number, and the date of creating the folder, for example: nilsonne_2022_fig1A_digitized_2023-04-22.

2 Add a ReadMe. The ReadMe should contain the full reference to the paper from which the figure is obtained, also to this protocol, and should say who is doing the digitizing.

3 Save a copy of the figure as an image file in the folder. Make sure to get the highest resolution that is available. If possible, download the image in the highest available resolution from the journal interface. If the image is only available in pdf format, do a screenshot after zooming in as far as the screen will allow.

4 Choose a suitable software for digitization, for example WebPlotDigitizer. Note in the ReadMe which version of the software was used.

5 Load the image into the software. Define axes as necessary in the software.

6 Identify the data points. Some softwares have automatic recognition of data points, but in many cases the data points must be identified manually.

7 Save an image where the axes and identified data points are overlaid on the original image, for documentation and to enable quality control.

8 Save the extracted data in a new file. Csv format is preferred. Column names should match the axis labels on the figure. Use the same naming convention for the file as for the folder (step 1 above).

9 If information is available about the expected number of data points, add it to the ReadMe. Note in the ReadMe whether the expected number of data points is the same as the detected number of data points.

10 Optional: Check if the same data are available in two different plots. For instance, there may be two scatterplots reporting the same variable on one axis and different variables on the other. In this case, digitize both.

11 Check and verify the accuracy of data extraction.

11.1 Optional: Plot data for comparison to the original plot. Load the extracted data into a statistical software and construct a plot of the same format as the digitized plot. Compare them side by side to identify any discrepancies. If this is done, save plot to folder and note in ReadMe whether the comparison was judged to confirm the accuracy of data digitization.

11.2 Optional: If any summary statistics were reported, such as means, medians, standard deviations etc, attempt to reproduce these numbers from the digitized data. Note any such checks and their results in the ReadMe.

11.3 Optional: If the granularity of data is known or can be inferred, this can be used to quantify the accuracy of digitization. For example, the variable “age” may have been recorded as integers. If most recorded numbers are close to an integer, the distance can be used to approximate the precision in digitization. If applicable, create a new column for the recorded variable, rounded to the appropriate precision. Then create another column and calculate the absolute difference between the recorded numbers and the rounded numbers. Add the mean, range, and standard deviation of the absolute difference to the ReadMe.

11.4 Optional: If step 10 was performed and two independent digitizations exist, check accuracy by sorting the values and plotting them against each other. A strict linear relationship will demonstrate that digitization was accurate. If possible, calculate the mean of the pair-wise sorted data points. This will average out error in data extraction. Note in the ReadMe whether this step was performed and what the results were.

12 Save the data folder in a suitable location such as an electronic laboratory notebook (ELN) and/or share it online through a suitable repository.