Lily Yan
LIS 545
Term Project - Additional Information

Citing datasets recognizes their significance in advancing research, promoting transparency, reproducibility, and proper attribution within the scholarly community. Furthermore, data citations should facilitate proper attribution to the creators of the dataset, referencing metadata such as provenance, title, and their unique and persistent identifiers. Ball and Duke (2011) emphasizes the importance of citing datasets, recognizing that data plays a valuable part in research, outlining a set of data attribution practice guidelines. These citation practices should also be applied to the *Water Potability* dataset, a synthetically generated dataset, that falls under the umbrella of a legitimate, citable product of research. In the absence of a Digital Object Identifier (DOI), utilizing a URL as an identifier for this dataset is a practical approach to ensure proper attribution and accessibility. While DOIs are commonly used for citing datasets due to their persistence and ability to locate resources, a URL can serve as an alternative identifier, especially for datasets in this case that may not have been assigned a DOI. That being said, here is a recommended citation:

Kadiwal, Aditya. (2021). "Water Quality." Kaggle.
https://www.kaggle.com/datasets/adityakadiwal/water-potability.

It is also important to consider responsible data citation for authors as well, in which it would be advisable to have a *data access statement* that articulates information about data access and access restrictions. Although this dataset is licensed under CC0 Creative Commons Zero and open to the public domain. This means that the dataset can be freely used, modified, and redistributed by anyone, for any purpose, without restrictions. No attribution is required, however, having a data access statement that communicates this information increases transparency, supporting reproducibility through the provision of clear access instructions.

Upon download, the dataset is in a compressed .zip format. In order to extract .zip files, special softwares designed to work with compressed files is needed. There is no documentation or indication as to why this smaller dataset may need to be in a .zip file, however, .zip allows for efficient download, organization, and data-sharing. This dataset itself is provided to users in the format of a .csv file, an easy-usable open format, allowing for interoperability, re-use, and to facilitate the preservation through mitigating the loss of data during format conversions (Hart et al., 2016). This format is a widely used file format, and can be accessed without special softwares. Many applications and programs are also available to read .csv files such as Notepad, Microsoft Excel, Google Sheets, as well as programming languages Python, and R. To further support usability of this data, a JSON representation of this data allows for metadata inclusion such as provenance, creator, copyright, and other descriptive metadata. Centralizing this

metadata into this JSON file will ensure that this information stays associated with our dataset throughout its lifecycle.

The files I have provided consistently bring up the synthetic nature of the dataset. Wynholds (2011) discusses the importance of capturing the identities of digital objects, including datasets, leveraging this knowledge in a way that preserves the identity of digital objects across different contexts and technological change. Given four functional requirements that contribute to the identity of digital objects, our dataset would likely fall under the second function requirement, which is that the identity of the dataset is embedded, inherent and/or inseparable. The challenge presented here is that there is a lack of detailed metadata attached to indicate the provenance, and the creation of data. This leaves us with the only piece of information on the origins of the data being synthetically generated, in which this identity needs to be preserved across technologies and platforms to ensure ethical and informed use of this data in research and reference.

The dataset contains synthetically generated data on variables and their impact on water potability, void of any personally identifiable data about people.

References

Ball, A. & Duke, M. (2011). How to Cite Datasets and Link to Publications. DCC How - to Guides. Edinburgh: Digital Curation Centre. https://www.dcc.ac.uk/guidance/how-guides/cite-datasets.

Hart, E., et al. (2016). Ten Simple Rules for Digital Data Storage. PeerJ Preprints, 4: e1448v2. https://doi.org/10.7287/peerj.preprints.1448v2.

Wynholds, L. (2011). Linking to Scientific Data: Identity Problems of Unruly and Poorly Bounded Digital Objects. International Journal of Digital Curation, 6(1). https://doi.org/10.2218/ijdc.v6i1.183.