Describing Collections with Datasheets for Datasets

Emily Maemura*[1] and Helena Byrne*[2]

[1]University of Illinois, Urbana-Champaign – United States
[2]British Library – United Kingdom

Abstract

This paper will report on the outcomes of several workshops conducted throughout Spring 2023 that explore the utility of a descriptive template for documenting web archives collections as datasets for research analysis. Past work in web archives scholarship has focused on addressing the description and provenance of collections and their data, including Dooley et al. (2018) who propose recommendations for descriptive metadata, and Maemura et al. (2018) who develop a framework for documenting elements of a collection's provenance. While recent efforts from researchers have provided a model for documenting their steps in generating a corpus from archived web data (e.g. Aasman et al., 2021; Fage-Butler, Ledderer & Brügger, 2022; Brügger, 2021), there is a need to both standardize documentation and strengthen communication between the curators collecting data and researchers who perform subsequent computational processing and analysis. Additionally, looking beyond libraries, archives, and cultural heritage settings, researchers in other fields are also addressing similar issues and developing their own approaches to the description of data. One approach to the challenge of describing large datasets comes from the field of machine learning where Gebru et al. (2018, 2021) propose developing "Datasheets for Datasets," a form of short document answering a standard set of questions arranged by stages of the data lifecycle.
Inspired by this work on Datasheets for Datasets, the authors conducted a series of workshops to explore how web archives collections can be described using that framework. The workflow presented by Gebru et al. includes a total of 57 questions to answer about a dataset, arranged into seven sections: Motivation; Composition; Collection Process; Preprocessing / Cleaning / Labeling; Use; Distribution; and, Maintenance. Through the workshops we aimed to understand how these questions can be adopted for the purposes of describing web archives datasets by asking participants to prioritize the questions and their information needs through a card-sorting exercise. Workshop participants considered how each question might be adapted and applied to describe datasets from the UK Web Archive curated collections. Each group evaluated a set of questions from the Datasheets framework and assessed them using the MoSCoW technique, sorting questions into categories of Must, Should, Can't, and Won't have. Through the discussion amongst participants, and across separate workshop venues, we aim to develop a set of emerging themes and core concerns for documentation from the card-sorting exercises.

The expected outcomes of these workshops are threefold. First, we will raise awareness of the Datasheets for Datasets framework in the web archiving community, providing potential connections and shared concerns of data documentation with fields like machine learning. Second, we generate a better understanding of the types of descriptive metadata web archive experts think should accompany web archive collections published as data. Third, we use

*Speaker

this discussion to promote stronger communication between web archivists and research users on priorities for documentation. The aim of these workshops and the presentation at RESAW is to generate a broader discussion of priorities and resources available for generating descriptive metadata and documentation for public web archives datasets.

References

Aasman, S., Bingham, N., Brügger, N., de Wild, K., Gebeil, S., & Schafer, V. (2021). Chicken and Egg: Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections. WARCnet Papers.

Brügger, N. (2021). Digital humanities and web archives: Possible new paths for combining datasets. International Journal of Digital Humanities. https://doi.org/10.1007/s42803-021-00038-z

Dooley, J., & Bowers, K. (2018). Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group (p.). OCLC Research. https://doi.org/10.25333/C3005C

Fage-Butler, A., Ledderer, L., & Brügger, N. (2022). Proposing methods to explore the evolution of the term 'mHealth' on the Danish Web archive. First Monday. https://doi.org/10.5210/fm.v27i1.11675

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for Datasets. ArXiv:1803.09010 (Cs).

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H.,, & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

Maemura, E., Worby, N., Milligan, I., & Becker, C. (2018). If These Crawls Could Talk: Studying and Documenting Web Archives Provenance. Journal of the Association for Information Science and Technology, 69(10), 1223–1233. https://doi.org/10.1002/asi.24048

Keywords: Web archives description, Data provenance, Data curation, Collections as data