Connor Franklin Rey and Carly Stewart
LIS 545 Curation Protocol
2/21/2023

Abstract

Background Information

Libraries provide many online resources for patrons to utilize. Users can attend online community events, access databases and journal articles for research, check out reading materials, and learn about financial and language assistance. These free services are what make libraries unique pillars within communities. When patrons use online materials, however, they are given very little information as to how their website usage is being tracked, what data is being collected, and where their data goes after they are done pursuing their local library.

American Library Association has identified Confidentiality/Privacy as one of its core values, stating that “protecting user privacy and confidentiality is necessary for intellectual freedom and fundamental to the ethics and practice of librarianship,” (2020). The goal of this data is to determine the degree to which library systems are upholding the ALA values and protecting patrons when supplying free and accessible tools to users via their websites. This dataset will be utilized in a later investigation into the privacy and security of public library system websites and their data tracking within the state of California.

About the Data

This dataset was curated to investigate data tracking on public library websites in the state of California. We have combined datasets from the California Public Library Administrative Headquarters Directory (2023) and the IMLS 2019 Public Library Survey (PLS) from the IMLS to provide contextual information on the service population, website usage, and scale of the library systems being studied (2019). This dataset was made in preparation for an upstream curation project in which the data will be run through The Markup’s Blacklight Real-Time Web Privacy Inspector (Mattu, 2020). We aim to retrieve data on instances on library websites of third-party cookies, key logging, session recording, canvas fingerprinting, ad trackers, Facebook pixel, and Google remarketing analytics, as indicated by Blacklight's methodology (Mattu & Sankin, 2020). Additionally, we intend to curate and code a dataframe on each public library's privacy policy to investigate whether the potential presence of these trackers may be unlawful.

Intended Audience

In 2018, California passed the California Consumer Privacy Act (CCPA) with the intention of giving consumers more control over the personal information that is collected online (2023). CCPA gives California residents the right to know what personal information is collected by businesses and how it is used, opt out of sharing their personal data, and delete most of the personal information that was collected from them.

This dataset could provide information on the efficacy of the CCPA and inform lawmakers where they can improve data privacy. Furthermore, it can be used to determine how California libraries abide by the privacy values the ALA espouses. The intended audiences are researchers and California policymakers interested in data privacy in public libraries. Additionally, concerned citizens may wish to access the completed dataset.

Documentation

Metadata Using Project Open Data Schema

Attribute	Value
title	Directory of California Public Library Websites and Descriptive Statistics
description	This dataset is being curated to investigate data tracking on public library websites in the state of California. We will be combining datasets from the California Public Library Directory and the 2019 Public Library Survey from the IMLS to provide contextual information on the service population, website usage, and scale of the library systems being studied.
keyword	California, public library, library, website, data, data tracking, data privacy
issued	03/02/2023
modified	03/02/2023
publisher	Connor Franklin Rey , Carly Stewart
contactPoint	Connor Franklin Rey crey@uw.edu Carly Stewart cestew@uw.edu
accessLevel	public
license	https://opendatacommons.org/licenses/pddl/
spatial	California

temporal	2019
describedBy	https://github.com/c-f-rey/pl_web_tracking/blob/main/README.md
language	English
references	https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey , http://ca.countingopinions.com/pireports/report.php?ff5a767bd8ab7a953f295552669a26a5&live
theme	Public Library Data Privacy
accessURL	https://github.com/c-f-rey/pl_web_tracking
format	CSV

README.md File

A copy of our README file can be found on the project's Github repository (Franklin Rey & Stewart, 2023).

Reflection on Documentation Process

Given our dataset is compiled from government documents, we decided to use the Project Open Data schema – as it is the standard metadata schema for government data.

To normalize the datasets, we standardized variable titles to be consistently lower case, and include more expressive contextual information, such as the year when certain data points were collected. A great deal of information was removed from the original datasets, this was done for two reasons: Either to remove identifying information for library workers to maintain their privacy, or to remove PLS data that did not seem relevant for the purposes of this study. PLS data that has been included is to aid with eventual analysis.

Reflection

There were aspects of the curation process that posed some difficulties. Choosing a scope of our data (i.e. all public libraries, a specific state, a specific region) was a large task. We chose to focus on one state because it will allow us to test-run Blacklight. Also, California has a lot of structured data points with a variety of different-sized library systems and service populations. We thought it may be interesting to focus on California because of the California Consumer

Privacy Act that was passed in 2018, which could lead to exciting reuse potential. Similarly to the scope, the normalization process was a challenge because we had to determine how to structure the data in a way that would be useful for the eventual analysis. PLS documentation was quite large and difficult to navigate which was time-consuming, and because there was so much data, knowing what to include was hard to know; we were not sure what others might find useful. This is why we inevitably chose to include the original data and R script to allow reusers to analyze our curation decisions and make changes should they wish.

We did identify some potential ways to improve. The data is pretty expressive but this may prove less helpful when running through the Blacklight, so steps may need to be taken in the future to make it more tractable. Including unique identifiers for each row may help to improve machine readability and future reusability. Also, it may be helpful to standardize some kind of library data collection process. Because every library self-reports its data and there is not necessarily a method to check the accuracy, it is not very trustworthy and difficult to scale.

References

Admin. (2020, September 28). *Core values of librarianship*. Advocacy, Legislation & Issues. Retrieved February 20, 2023, from https://www.ala.org/advocacy/intfreedom/corevalues.

California Consumer Privacy Act (CCPA). State of California - Department of Justice - Office of the Attorney General. (2023, February 15). Retrieved February 24, 2023, from https://www.oag.ca.gov/privacy/ccpa.

California Public Library Administrative Headquarters Directory - Text Report. (2023). http://ca.countingopinions.com/pireports/report.php?ff5a767bd8ab7a953f295552669a26a5&live. Accessed 6 Mar. 2023.

Franklin Rey, C. & Stewart C. (2023). Directory of California Public Library Websites and their Descriptive Statistics 2019. https://github.com/c-f-rey/pl_web_tracking.

Mattu, S. (2020, September 22). Blacklight – the markup. https://themarkup.org/blacklight.

Mattu, S., & Sankin, A. (2020, September 22). How we built a real-time privacy inspector – the markup. https://themarkup.org/blacklight/2020/09/22/how-we-built-a-real-time-privacy-inspector.

Public Libraries Survey. (2019). http://www.imls.gov/research-evaluation/data-collection/public-libraries-survey. Accessed 6 Mar. 2023.