Connor Franklin Rey
LIS 546: Data Curation II
Professor Nic Weber
26 May 2023
Final Protocol Project Report

Project Pitch

Continuing our final project from DCI, I am hoping to curate a directory of public libraries to investigate data tracking through public library websites in the United States. This dataset will be compiled to be run through The Markup's Blacklight Real-Time Web Privacy Inspector, with the aim of retrieving data on instances on library websites of third-party cookies, key logging, session recording, canvas fingerprinting, ad trackers, Facebook pixel, and Google remarketing analytics. The data will ideally be interoperable with IMLS Public Library Survey (PLS) data in order to research the prevalence of data tracking through public library websites, the types of libraries that utilize these tools, and their respective user populations.

Statement of Work

Topics of curation protocol:

Public libraries, data tracking, data privacy

Target audiences:

The intended audiences are researchers and policymakers interested in data privacy in public libraries. Additionally, concerned citizens may wish to access the completed dataset.

Project goals:

The aim of this project is to gain an understanding of the scope of data tracking on public library websites. While I was unable to achieve this goal in it's entirety within the timeframe of this class, I was able to make significant progress. What has been compiled is a directory of a majority of public library websites and online catalogs in the United States with accompanying demographic data from the PLS to aid in analysis.

Future researchers may wish to scrape information on the vendors providing online interfaces for US public library websites in order to gain a better understanding of their role data tracking. My unsuccessful attempts to scrape this information have been included in the R script as reference that can hopefully be of some user for future researchers.

Additionally, future researchers may wish to run the dataset through the Blacklight tool to collect information on data tracking on public library websites, as this was the initial goal of the project.

Examples of data:

- Compiled California public libraries directory dataset
    - This was our final project from DCI, in which we compiled a directory of California public libraries and joined it with key PLS data frames.
- Oregon public library directory
    - Directory of Oregon public libraries, available in a structured table
- Washington public library directory
    - Interactive directory of Washington public libraries. Unclear how I'm going to get structured data out of this one but I believe.

Users and User Stories

Potential users:

Personal data tracking on public library websites is a topic of potential interest for a wide audience including, but not limited to, academic researchers, policymakers, libraries, and anyone in the general public concerned about data privacy.

In a previous class, we compiled a directory of public libraries in California. This dataset combined datasets from the California Public Library Administrative Headquarters Directory 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.

User story:

Below is a user story envisioning the goals of a potential stakeholder to better understand how to design this repository to be of use to its audience:

As a California policymaker, I want a directory of public library websites, with information regarding their data tracking so that I understand the prevalence of data tracking on public websites both in-state, as well as out-of-state for the purposes of comparison.

Dataset usability:

This dataset provides users with a directory of all public library websites in the United States – at least all public libraries listed on librarytechnology.org. The dataset also includes an attribute that indicates its state, as well as some demographic information, which will allow for comparative analysis.

I was unable to capture information using the Blacklight tool, so users will not yet be able to do any analysis there yet. Nor was I able to scrape information regarding each library's vendors for their online catalogs and discovery interfaces to understand if certain companies are capturing more user data than others. Though, users may be able to scrape vendor information or utilize the Blacklight tool to incorporate this information later. Additionally, users may find other uses for the directory that will help our understanding of public libraries' presence online.

Collection Policy

Content	All submissions must be datasets concerning public library websites All submissions must pertain to the United States, at the municipal, county, district, or state level.
Data Type	Submissions may contain quantitative or qualitative data, so long as they meet the content criteria. Submissions must contain data recorded in alphanumeric characters.
Format	Electronic files of text or tabular data are accepted. Submissions must be in a non-proprietary, openly-documented format, such as: ● Text data - Plain Text (ASCII, UTF-8, or UTF-16 encoding), XML (with schema), JSON ● Tabular data - Character-Delimited Text (Tab-Separated Values, Comma-Separated Values, or other Delimited Text) ● All files must be unencrypted and uncompressed
Size	Individual submission files must be less than __GB
Preservation	We will undertake the following basic preservation actions for all

	repository files: ● persistent, permanent identifier ● bitstream maintenance ● preservation metadata ● backup copies ● regular virus and file corruption checks ● periodic refreshments to new storage media Due to capacity constraints, we cannot offer more than basic preservation measures.

Transformations and Data Quality

File / format:

Data should be stored in an open source format, ideally CSV. Data in proprietary format will not be accepted in the repository. Users who attempt to submit a proprietary format will be notified and referred to resources to convert to an open source format. All changes to data will be logged in the github repository, and the date of the last update will be recorded in the metadata.

Data values:

Tranformations will be made to normalize the data using an R script, so that users can understand how the data have been manipulated. Consistent variables will be recorded in a data dictionary. Every variable will have a singular type of value. Currently there is only one table, but variables of lib_tech_id and lib_name serve as reference to join with further librarytechnologies.org and PLS data respectively.

Data Dictionary

Variable	Variable Type	Allowed Values	Description
lib_name	string	capitalized string	Name of public library system
lib_website	string	active web address	Web address of public library system landing page
lib_catalog	string	active web address	Web address of public library system online catalog

city_county	string	string	Either the city or county of the public library system, depending on the information retrieved from librarytechnology.org
state	string	two letter string state postal code	State postal code
popu_lsa_2019	integer	integers greater than zero	PLS data included to illustrate the population of the Legal Service Area of library system
web_visits_2019	integer	integers greater than zero, -1, -3	PLS data included to indicate usage of library system website. Total visits (sessions) to library website -1–Missing -3–Not applicable (closed or temporarily closed administrative entity)

locale_mod	integer	11, 12, 13, 21, 22, 23, 31, 32, 33, 41, 42, 43	PLS data included to group library systems by size. Urban-centric locale code. The geographic location in terms of the size of the community in which it is located and the proximity of that community to urban and metropolitan areas. Assigned based on the modal locale code of associated stationary outlets (i.e., central and branch libraries). 11–City, Large: Territory inside an urbanized area and inside a principal city with population of 250,000 or more. 12–City, Mid-size: Territory inside an urbanized area and inside a principal city with a population less than 250,000 and greater than or equal to 100,000. 13–City, Small: Territory inside an urbanized area and inside a principal city with a population less than 100,000. 21–Suburb, Large: Territory outside a principal city and inside an urbanized area with population of 250,000 or more. 22–Suburb, Mid-size: Territory outside a principal city and inside an urbanized area with a population less than 250,000 and greater than or equal to 100,000. 23–Suburb, Small: Territory outside a principal city and inside an urbanized area with a population less than 100,000. 31–Town, Fringe: Territory inside an urban cluster that is less than or equal to 10 miles from an

			urbanized area. 32–Town, Distant: Territory inside an urban cluster that is more than 10 miles and less than or equal to 35 miles from an urbanized area. 33–Town, Remote: Territory inside an urban cluster that is more than 35 miles from an urbanized area. 41–Rural, Fringe: Census-defined rural territory that is less than or equal to 5 miles from an urbanized area, as well as rural territory that is less than or equal to 2.5 miles from an urban cluster. 42–Rural, Distant: Census-defined rural territory that is more than 5 miles but less than or equal to 25 miles from an urbanized area, as well as rural territory that is more than 2.5 miles but less than or equal to 10 miles from an urban cluster. 43–Rural, Remote: Census-defined rural territory that is more than 25 miles from an urbanized area and is also more than 10 miles from an urban cluster.
cen_tract	integer	integers formatted 0000.YY (YY=blank or numeric)	PLS data included to aid with potential interoperability with other demographic datasets. Census Tract code. 7 character - A small, relatively permanent statistical subdivision of a county or statistically equivalent entity delineated by local participants as part of the Census Bureau's Participant Statistical Areas Program. This field consists of four integers and two decimals, with an

			explicit decimal point.

Metadata Application Profile

What elements are necessary to meet your stakeholders' expectations?

Geographic and temporal scope are crucial pieces of context that must be included for stakeholders. Additionally the sources and any updates that have been made, must be recorded to allow for accuracy in any later research. Project Open Data serves as a adequate schema in meeting user needs. For the sake of simplicity, not changes will be made to the schema.

Namespace:

https://github.com/c-f-rey/public-library-website-directory/tree/main

Namespace definition:

This is the online directory hosted by Github. The name of the directory is meant to be as expressive as possible to aid with discoverability.

Status:

Mandatory

Metadata Schema:

Project Open Data

Example:

Attribute	Value
title	Directory of United States Public Library Websites and Descriptive Statistics
description	This dataset is being curated to investigate data tracking on public library websites in the United States. We have combined datasets from the Directory of Public Libraries in the United States 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	United States, public library, library, website, data, data tracking, data privacy
issued	05/24/2023
modified	03/26/2023
publisher	Connor Franklin Rey
contactPoint	Connor Franklin Rey, crey@uw.edu
accessLevel	public
license	https://opendatacommons.org/licenses/pddl/
spatial	United States
temporal	2019
describedBy	https://github.com/c-f-rey/public-library-website-directory/main/README_project_report.pdf
language	English
references	https://librarytechnology.org/libraries/uspublic/ , https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey
theme	Public Library Data Privacy
accessURL	https://github.com/c-f-rey/public-library-website-directory

format	CSV

You should also choose an encoding scheme for your examples (JSON or XML. Note: I don’t expect you to actually produce these records.)

Licensing

Open Data Commons Public Domain Dedication and License (PDDL)