LIS 545 Data Curation
Curation Protocol
Ruizhe Wang
03/07/2022

Project Abstract

Background & Objectives

The repository "LIS545 Curation Protocol" is owned by Ruizhe Wang (Github: rachelwww224).

Homelessness has been a serious and prevalent issue in Seattle and other surrounding areas in Washington state. As a complex social problem, many underlying factors have affected or worsened homelessness, such as poverty, mental health issues, drug and alcohol abuse, housing affordability, etc. This project aims to assist policymakers and researchers in learning more about the homelessness crisis in Washington state over the past few years. Containing homogenous homelessness data, the resource helps them quantify the scale of homelessness from 2017 to 2020 by different regions across the state. The resource can serve multiple purposes for different audiences. It provides policymakers with a better understanding of the homelessness crisis to make informed public policy decisions; researchers can use or reference this resource for their research projects.

About the Data

The dataset is sourced from the report "Seattle/King County Point-in-Time Count of Individuals Experiencing Homelessness" published in 2020. The data was collected by All Home, the Seattle/King County Continuum of Care (CoC) lead. Ranging from 2017 to 2020, the data provides a point-in-time count of the sheltered and unsheltered homeless populations across different regions in Washington state. For example, the 2020 unsheltered homeless population was counted by volunteers on the night of January 24, 2020. We assume that this one-time counting can be a good estimation of the actual homelessness situation.

The data was initially collected through multiple sources, including a street count for unsheltered homelessness and information retrieval from Homeless Management Information System (HMIS) Data for sheltered homelessness.

I encountered some challenges in extracting the data from the report in PDF format. I tried to use Tabula, but the output file format was incorrect. Thus, I manually typed the data into the Excel spreadsheet. The data in the original report is in a wide format, with years being the columns. I transformed and reshaped the data into a long format to facilitate the data documentation process and improve the data usability for future analy1tics.

Potential Audiences & Users

Besides policymakers and researchers, fellow curators are the primarily targeted audiences for this protocol. It helps them understand how and why the data was curated and the values for reuse. They may reuse this data in their curation processes or reuse it for future analytics.

Documentation

1) Metadata Using Project Open Data Schema

Attribute	Value
Title	Washington State Homelessness Count (2017-2020)
Description	The dataset includes the annual number of homeless individuals across regions in the Washington state from 2017 to 2020. The intended audience is policymakers or reserchers. This dataset was curated for a course taught at the University of Washington.
Keyword	“Homelessness”, “Washington”
Issued	March 3, 2022
Modified	March 3, 2022
Publisher	Ruizhe Wang University of Washington
ContactPoint	Ruizhe Wang rwang37@uw.edu
AccessLevel	Public
Rights	The data is freely available to the public.
Spatial	U.S. Washington State
Temporal	2017 - 2020
Format	CSV
MediaType	CSV
Language	English
References	https://kcrha.org/wp-content/uploads/2020/07/Count-Us-In-2020-Final.pdf

2) Readme.txt File

3) Documentation Recap

I selected Project Open Data as the metadata schema since government agencies also use it for their datasets and APIs. Since the primary audiences for this project are researchers and policymakers, they may have the experience of working with this metadata schema before. Thus, it would be easier for them to understand the metadata and further explore the reuse values. According to the Project Open Data schema, I included most of the attributes that are always required in the guideline, such as title, description, keyword, etc. Besides, I also included attributes that are not necessary but applicable to this project, such as rights, spatial, temporal, etc. In the readme file, I clearly explain the data normalization process, the file naming conventions, and I also include a data dictionary that provides complete documentation for the dataset.

Reflection

The curation process for the Washington state homelessness data is pretty simple. The primary data source is the report published by AllHome in 2020, which provides homogenous count data across different regions in Washington state. Since the report is in PDF format, I encountered some difficulty extracting the data. Due to the formatting issues, Tabula did not work well in producing a well-formatted CSV file. Thus, I manually typed the data into the Excel spreadsheet. Besides, the data table in the report was in wide format. I converted it into a long format in the curation process. Saving data in the long format can facilitate data management and data reuse. Fellow curators and data management personnel can easily update the data; researchers and policymakers can directly use this data file in their analytics without worrying about data transformation. After the data had been extracted, I created a data dictionary for documentation and metadata using Project Open Data schema. I provide complete documentation of the dataset and its critical components in the data dictionary and metadata, such as variables, description, naming convention, etc. By reading the documentation, users and fellow curators can learn more about the data and help them to update the data or improve the curation process if needed. I provided two versions of the data, raw and normalized. They can cater to different needs of the data users or curators. The normalization process is pretty simple. I used built-in functions in Excel to conduct normalization. The data is saved in.CSV format, which can be accepted by most software on the market.

My main concern about the curation is the data quality issue. Data quality plays a significant role in data management and data reuse. According to critiques of the report, the unsheltered homelessness count might be inaccurate. First, they pointed out that the number of volunteers is positively correlated with the number of unsheltered homeless being found, and the number of volunteers was not consistent over the years (SCC Insight, 2020). In 2020, the number of volunteers had sharply decreased due to multiple reasons. Second, the report uses the “point-in-count” method, where they assume that counting in a specific day/night can be a good estimation of the homelessness situation for that year. The critiques pointed out that on January 24, 2020, a heavy rainstorm in Seattle made many unsheltered homeless people move into sheltered places on that night (SCC Insight, 2020). Thus, the counting might be biased for 2020. Therefore, I mentioned the potential data quality issues in the documentation so the users can be mindful of the issue in their analytics.

I use Github as the data repository since it is stable, safe, and freely available to the public. Users can easily view the files and download the data from the Github interface. Besides, Github allows for easy updates for the files, which ensures proper data management and preservation process in the future.

References

AllHome. (2020). *2020 Seattle/King County Point-in-Time Count of Individuals Experiencing Homelessness.* **https://kcrha.org/wp-content/uploads/2020/07/Count-Us-In-2020-Final.pdf**

Seattle City Council Insights. (2020, July 12). *Annual homeless count report released, and it's still a mess*. SCC Insight. **https://sccinsight.com/2020/07/12/annual-homeless-count-report-released-and-its-still-a-mess/**