Samantha Parrish

1/27/23

LIS 545

Term Project - Data and Metadata Profile

The dataset studied in this profile is going to consist of the [Violent Crime Rate California 2000-2013](https://catalog.data.gov/dataset/violent-crime-rate-8340c/resource/83064e2e-3939-450c-bb8b-8702f2feba91), as located on [data.gov](https://data.gov/) (**NOTE: as of 2/15/23, this dataset has since been removed from data.gov. The dataset itself is still cited within this write-up and below and is thus available for review**). This data was published by the California Department of Health, though the stakeholders of this data themselves are much more varied. As stated on data.gov directly, “Crime and population data are from the Federal Bureau of Investigations, Uniform Crime Reports.” To that end, other stakeholders would be the numerous counties, cities, and other areas in California who contributed this data. It also should be noted that the citizens of the state of California are stakeholders because this information is pertinent to them as well.

The creators of this dataset have compiled a large data file in a single Excel spreadsheet, though that spreadsheet contains five tabs of metadata (all of which will be discussed below). As far as usage restrictions are concerned, there aren’t very many except it would behoove the user to move slowly through the data as it is quite a lot when taken at face value. Beyond that, all a user would need is a way to download the Excel file.

The metadata in this document is numerous, yet useful. As mentioned above, there are five tabs within the spreadsheet, which allow for additional data to be consumed. Those tabs (labeled: ViolentCrime, READ ME, DataDictionary, MPO\_CountyList, and DataFilteringInstructions, respectively) are provided to assist the user in maneuvering through the data.

They are simply the beginning of the provided metadata, however. In particular, READ ME and DataDictionary provide a user with useful metadata, such as descriptors for each available tab within the READ ME section and names, definitions, type, width/format, and coding comments within the DataDictionary. This type of metadata can prove invaluable simply because the data provided can otherwise be overwhelming without a sense of understanding of what one is reading.

In the section labeled Metadata within *Encyclopedia of Knowledge Organization*, the standard for most metadata as, “… commonly organized around a set of elements (such as “title”, “author”, “date”) that manifest as computer-readable documents in one of an alphabet-soup set of formats and mark-up languages, such as MARC, XML, JSON, and YAML.” In this way, the metadata within this dataset is not arranged in that way. However, it doesn’t necessarily need to be. This metadata is arranged in a way that arguably makes it easier for users who aren’t necessarily academics to understand it.

That said, while this is written for users who are primarily not academics, that doesn’t mean it’s easy to read for any laymen either. This is a document created by multiple government agencies for a government website, after all. However, not every user is going to be a government agent. If someone outside of government agencies wanted to understand the violent crime data of California from 2000-2013, they may struggle without a more generous breakdown provided within the DataDictionary. Just getting to the information may be a bit of a chore as well as the data.gov site definitely works better for users who already have a general idea of what they’re looking for. It’s not that difficult to find once you know what your query should be or what metadata to utilize on data.gov to narrow things down, but it still may take a bit to find what you’re seeking.

Adding to that, just using a general web search for this particular dataset provides even more datasets focusing on a similar subject. A search using a literature database like the UW library site provides much less pertinent results as the search engine essentially keys on words within the title of the dataset rather than looking for the use of the dataset itself. This is likely because the title of this dataset is broad. That’s not necessarily a bad thing, however, as it may aid users in further research. Just because the study itself isn’t directly linked or credited, it’s clear that the idea isn’t only represented in this single dataset on data.gov. Again, however, if one is indeed looking for this dataset alone, they will need to understand, even vaguely data.gov’s search capabilities.

In addition, there are no publications that are provided with the dataset that aren’t already a part of it. To clarify, there are links that will take you to the already listed metadata within the dataset (for example, the Data Dictionary has its own link on data.gov). This appears to be a feature of data.gov, rather than a bug, as clicking into other datasets just to explore didn’t yield any different results.

**Links to Repository + Dataset (just in case the embedded links don’t work)**

<https://data.gov/>

<https://catalog.data.gov/dataset/violent-crime-rate-8340c/resource/83064e2e-3939-450c-bb8b-8702f2feba91>

<https://data.chhs.ca.gov/dataset/99bc1fea-c55c-4377-bad8-f00832fd195d/resource/bc09f211-200c-4c4c-aa13-d2e89c0d5577/download/hci_crime_752_pl_co_re_ca_2000-2013_21oct15-ada.xlsx>