**Data Profile**

The dataset I chose was sidewalk data from data.seattle.gov: <https://data-seattlecitygis.opendata.arcgis.com/maps/SeattleCityGIS::sidewalks-1>. The data captured is of sidewalks maintained by the Seattle Department of Transportation. Included in the dataset is the location of the sidewalk, curb type, condition, condition assessment date, etc. The data comes from the city of Seattle, and the key stakeholders are City of SeattleGIS Online, Seattle Department of Transportation, and the dataset owner, Claire Palay. Stakeholders can be inferred as well such as city planners and tourists or residents to plan walkable routes. There are six files available in the following formats: HTML, JSON, CSV, GeoJSON, Shapefile, and KML. No special software is needed to access the HTML or JSON files, however, special software GIS software or a third-party converter is needed to open the GeoJSON file, and Google Earth is needed for the KML file. Each file contains information for 46,361 sidewalk records. Additionally, the dataset is public and a custom usage license is linked to it that states that the City of Seattle cannot guarantee its accuracy regarding “labelling, dimensions, contours, property boundaries, or placement or location of any map feature.”

There is an opportunity for enrichment within the dataset regarding the attribute list. A lot of abbreviations are used, but when clicking into the attribute list there are no clear definitions. For example, there are attributes labelled MODDTTN, ASBLT, and MATL with no clear definition of what that means making it difficult for potential re-users to interpret the data. Attributes could be clearly defined so that users can properly understand the schema used Moreover, abbreviations are used for values as well, such as LSCP. So, controlled vocabularies for both attributes and values can be fully listed and defined to promote understanding. Additionally, when reviewing the CSV file, there are several attribute columns that are entirely empty. To keep such a large dataset as compact as possible, these attributes could be removed.

**Metadata**

The metadata included with the dataset is fairly comprehensive, although some pieces of schemas are missing values. When looking at the overview of the data before clicking into any of the files, metadata is structured using the Common Core standard. Views, downloads, data provided by, dataset owner, contact email, homepage, contact name, license info, last update, theme, unique identifier, geographic coverage, publisher, public access level, category, and tags are all included at the bottom of the page. This metadata is comprehensive and provides hyperlinks that enable cross-reference. For this metadata, it could be enriched by adding supplemental information in the schema. For example, the Geographic Coverage attribute lists a series of coordinates as its values, but it is implied that the user already knows that these numbers are coordinates. Instead, there could be something identifying them as such. Additionally, attributes such as Refresh Cycle could have additional information to clarify. The Refresh Cycle attribute is also lacking values.

Within the actual dataset, the metadata is provided in a supplemental document. This includes identification information, citation information, title, description information (such as abstract and purpose), keywords, theme (subcategories include Theme Key Word Thesaurus and Theme Keyword), access constraints, use constraints, data set credit, and metadata reference information (subcategories include are Metadata Standard Name, Version, and Metadata Time Convention). The metadata here is structured using FGDC Content Standard for Digital Geospatial Metadata. Opportunities for enrichment here could be combining the theme keywords into one line as there is currently a new line for each one.

Additionally, when opening the HTML file, a map opens in the browser that contains additional metadata located in a sidebar. Metadata here includes dataset, info updated and data update date, published, records, license details. One of the questions that comes up here is was feature layer means in relation to the dataset attribute field. This could potentially be explained in a supplemental document. Overall, the metadata for this dataset seems scattered. One way to enrich the metadata and promote reuse could be to combine it in a supplemental document to make each portion of metadata easily accessible and readable.

**Publication**

Currently, there are no publication listed within the dataset. I used Google and searched for “sdot sidewalk data.” A few news articles popped up regarding the current status of Seattle’s sidewalks and planned repair projects. An article by Guy Oron for *Real Change* cited the SDOT sidewalk dataset. The same query input into UW Libraries search resulting in finding a dissertation by Pedro Popoca regarding lack of sidewalks and pedestrian safety using the SDOT dataset.

**Repository Profile**

The repository I chose for this project is [GeoServer](https://geoserver.org/), which is a project maintain by OSGEO. I chose this repository because the repository is focused around geospatial data, which is exactly what the Seattle sidewalk dataset is comprised of (the dataset can be found [here](https://data-seattlecitygis.opendata.arcgis.com/datasets/SeattleCityGIS::sidewalks-1/about)). GeoServer allows users to search datasets by location, which would be advantageous when trying to review a specific neighborhood’s sidewalks within Seattle’s large dataset. Additionally, GeoServer advertises integration with programs such as ArcGIS, MapServer, and QGIS, which would all be beneficial when working with the sidewalk data.

Additionally, GeoServer is free and open source, allowing anyone to access and store data within the repository. The repository accepts various forms of vector and raster data. However, more types of data can be stored using appropriate extensions. For example, OSGeo has another project called [PostGIS](http://postgis.net/) that enables users to store various types of spatial data in both 2D and 3Ddata. Extensions such as PostGIS also enable integration with third party tools like ArcGIS and Tableau, which provides ease of access when working with geospatial data like what is provided in the sidewalk data set. The data stored by this repository is primarily focused on geospatial data, but otherwise there do not seem to be any restrictions on what is submitted. Moreover, the repository does not provide any guidance on what should be included in the submission information package nor is any human assistance provided to potential data submitters. However, there is the option to reach out via email with any questions. Although no metadata appears to be required when submitting data, metadata in the repository must be configured using the INSPIRE or ISO Geospatial metadata standards.

To download data, no login is required. Additionally, there are multiple access mechanisms provided. For example, when viewing a dataset on the hydrological basins in Africa, I have the option to download a zip file with the dataset, view it via and online link to the Food and Agricultural Organization of the United Nations, or the information directly to a map layer for viewing. After reviewing several datasets, the metadata provided is primarily in XML format and it adheres to the ISO Geospatial metadata standard. Finally, there is no identifiable information regarding the dissemination information package in this repository.

Finally, GeoServer does not archive data itself but allows for interoperability. When working with geospatial data, running GeoServer is beneficial as users can connect existing geospatial information from other published datasets. In addition to connecting information, GeoServer enables users to publish new information Google Earth and other web-based maps such as Open Layers. For Seattle sidewalk data, this is extremely beneficial as the data can be published directly to the ArcGIS map featured on the webpage for the data.

**Data Citation**

A persistent identifier is needed for this dataset. Currently, there is only an unreliable URL listed (<https://data.seattle.gov/dataset/Sidewalks/wgef-i7sp)>, which leads to a 404 error. Users then have to manually search for the sidewalk data on Seattle’s open data portal. Instead, a DOI should be created so that the dataset can be reliably accessed. Additionally, since the data is routinely updated, metadata including the date it was last updated should be at the top of the landing page when accessed using the DOI.

**Preservation**

The format present in the GitHub repository is a csv file, which is non-proprietary and can be opened using an array of different programs like Notepad, Excel, etc. Additional formats can be accessed from Seattle’s Open Data Portal. There are software requirements needed to open some of the other available formats. For example, GIS software or a third-party converter is needed to open the GeoJSON file, and Google Earth is needed for the KML file.

**Copyright License**

The data is available for public use and research. However, the City of Seattle does not guarantee the data’s accuracy at any given time. Due to the nature of the geospatial data, I do not recommend copyrighting the dataset as it is continually changing and it is being used to update other applications such as Open Street Map.

**Human Subject Considerations**

This data does not contain any human subjects and therefore human subject considerations are not applicable.