Data Set URL: [Chicago Green Roof Data](https://www.kaggle.com/chicago/chicago-green-roofs)

The data in this file is a compilation of the different buildings within Chicago, Illinois that appear to have vegetation on the top of the building. Vegetation on top of building is a seen as a clear effort to introduce appeal for building residents, community gardens, upscaling a building and curb appeal, and introducing biodiverse plant species in places that otherwise would not have been able to sustain life. Green Roofs is also an initiative to promote sustainability within an urban landscape or city. They have been shown to reduce dust and smog production within a city and can assist with drainage issues within the city (About Green Roofs). The data set consists of address/location, total area of the green space and vegetation on the building, and total area of the roof. This data has come from the City of Chicago. The information in this data set is derived from satellite imaging technology. The job was outsourced to a third party, but the City of Chicago made clear the boundaries in which to collect the data. The data was collected in August, September, and October 2010. The data file has been recently updated on the Kaggle with an edit in 2020. It is unclear what has been updated about the files.

The key stakeholders for this data can include city planners, other building planners and owners, as well as environmental researchers within urban landscapes. There also may be interest in this data for those who live and work within the city of Chicago as they could possibly utilize or visit these sites included in the data. It could also be beneficial for building planners that are looking to implement vegetation on rooftops which are interested in studying irrigation setup or foundation of roof gardens. Looking at the City of Chicago website, there is a section on Chicago Green Roofs which has taken this data and made it available in map format on their website. The way in which each location is placed within the map matches format in which the data appears in the .csv file. For example, any address or location written in the data is not one full address; it is separated. 1140 N North Branch ST is separated by each identity. *1140* is the house number, *N* is the direction, *North Branch* is the street name, and *ST* is the street type. The data appears separated in the data file as well as the website where the data is visualized on a map. It is plausible that the data is in this format for organization of data as well as to easily import data to a map visualization (Chicago Green Roofs).

Within the folder there are two files. One is the .csv file which contains the data. There is also a .json file which is the metadata profile. This data is considered part of the open data initiative and exists in the public domain. The .csv file can be opened with any spreadsheet processor—Microsoft Excel or Numbers. The metadata file is stored as a .json file. It can be opened with any simple text processing application that comes preinstalled on computers—Mac is TextEdit and Windows being Notepad. It is worth noting that this file, when viewed in TextEdit, is unformatted with no/ hierarchal formatting. I downloaded a Chrome Web App to read the file and display it in the formatted version to make it organized and easier to read. The. json formatting tool can be found with a web search in the [Chrome Web App store.](https://chrome.google.com/webstore/detail/json-formatter/bcjindcccaagfpapjjmafapmmgkkhgoa) It appears to be written in a JavaScript coding language (Noman, N.). The metadata file seems to reiterate what is seen in the .csv file. Some notable and interesting features of the metadata include counts of repeated street names on line 312. There is also an occurrence of counts of street directions within the data. For example, the direction including “W,” which is west, is counted 117 times in this document in reference to street direction. This may become helpful for city planners or developers looking to increase green spaces where there is less prevalence of such space direction wise.

The metadata is also full of information which allows users to discover where sources for the data were derived from. This can be especially important in considering the authority of the data. The metadata does grant some additional information regarding the creation of the data. It features tags in which this information might be useful. The metadata could benefit from being enriched. It would benefit from adding more keywords to allow more access to the data. Adding words like: “green, eco, environmental, city, urban, landscapes, gardens, biodiversity…” could all be added and would allow the data to be more easily found when searched by keyword. There does not appear to be any publications associated with this data. A keyword search including related phrases and words to this dataset was completed. There did not appear to be published scholarly works associated with this data. Additionally, a Google Search with the same keywords did not show published materials associated with this data. As mentioned previously, this data has been used internally by the City of Chicago to publish a visualization of the data on a map. The data has a public domain license meaning it can be redistributed and used by the general public. This data is visualized on the City of Chicago website, which is the only place I have been able to find it.

**Repository Chosen**: [Texas Data Repository](https://data.tdl.org/)

Selection of a repository was time consuming for this dataset. The dataset can be seen in many different lenses of research. It is a simple dataset which could be interpreted in many ways. For example, university students could use this dataset to find areas within Chicago where successful Green Roofs are implemented and use this data to craft better new and innovative architecture plans for their cumulative assignments. City planners of major cities could use this information to understand ways in which the city can create new greenspaces in their city and use this data to visit the sites. It could also be flipped to natural science research where this data could be used as a starting point in finding ways in which human interaction and intervention help further or reduce effects of climate change. Since it seems this data could be understood in more than one lense discipline wise, it was important to be able to store this data in a general repository.

A general repository allows for cross-disciplinary data to be added and searched within one space. This can be beneficial and possibly disadvantageous to the ease in finding this data set for target users. It is beneficial as it is less specific and constrained in uploading material requirements than discipline specific repositories. Many discipline specific repositories require certain data schemas, accompanying files, or only allow data to be uploaded when notable scholarly articles are published as a result of said data. It is understandable that these constraints are in place for discipline specific data repositories as it continues standard use and organization. This data set seems to be simpler in that it is one .csv file (concerning location data and measurements) and one .json metadata file.

While the Chicago Green Roof data is not groundbreaking research with scholarly articles published from it, there is still use in it. The general repository which allows more types of data to be uploaded seems to be a perfect fit for this dataset. Uploading to a general repository could also be disadvantageous to the data in terms of visibility. The data could get lost among the sea of other datasets and projects within the general repository, which houses numerous amounts of other datasets. To combat this possible gap, it will be important to include keywords and relevant metadata tags when uploading to the repository to make it linked to more disciplines and topics when keyword searched in the repository, making it more relevant to the user.

This repository is called the Texas Data Repository, which is also a unit of a larger Dataverse Project ran by Harvard University (Dataverse Project About). This repository can be accessed by members of the Texas Digital Library (TDL) institutions (Texas Data Repository Dataverse). The institutions include University of Texas School System, Baylor University, and Texas Woman’s University, just to name a few. This data repository does require affiliation within selected institutions for access to upload datasets. Users who are outside of this affiliation are allowed to search, view, and download all data files and projects on this website. Anyone can search and view on the repository website, however in order to download, the user will need to log-in through the third-party authentication service using a Google Gmail account.

To upload data to this repository there is an affiliation needed with listed institutions on the website, most of which are higher education institutions within Texas. This repository will allow multiple types of data to be submitted. “Spreadsheets, sensor and instrument data, surveys, GIS data, imagery…codebooks or data dictionaries…” (Texas Digital Library Wiki) can be included as accepted data for upload in this repository. These are accepted in any file type. Additionally, any individual file uploaded to this repository has to be less than 4GB, which means you can separate one (large, more than 4GB) file to many files as long as they are less than 4GB each, if needed. Lastly, this repository only accepts data which includes a CC0 license, or public domain. Only affiliated users with the approved institutions can upload a submission of data. The Chicago Green Roofs data package may be beneficial for students, researcher, or any other user (who does not have to be affiliated) of the Texas Data Repository.

Any discipline can be represented in this data repository making it multidisciplinary and a general repository. For someone interested in uploading their data, there are mainly two components required in the submission package. This would be the data files and the descriptive metadata. Data files can also be interchanged with documentation and/or code if applicable. Those who are looking to upload can fill out a support ticket concerning any questions or assistance in uploading data to the repository. The dissemination information package appears as a folder bundled with the files requested. It appears as a direct download to the user, however as previously mentioned, authentication is required to download. Files will include documentation, code, dataset(s), and metadata. Selection of what files are downloaded can be chosen by the user instead of downloading the whole package. Lastly, this repository follows OAIS structure and has ISO certification which ensures organization and quality in archival methods established by this repository (Digital Preservation Resources).

This GitHub has five total files within the ZIP download. The first will be the original data file in csv which was found on Kaggle. The second file will be the original metadata that came with the dataset file in JSON format. This original metadata file has information which may be pertinent to keep for some users. Additionally, the JSON file used a proprietary schema. This proprietary schema risks becoming obsolete as not many users may have access to the exact schema protocol to understand the elements it uses. With ambiguous elements and classification, the user can possibly loose what the intended meaning of some entries are because of undefined elements. It is easy to open the file as a text formatted file, usually what the computer will default to if no application can open JSON files. In order to open the JSON file in a hierarchical view, it is important to have a JSON format viewer which can be downloaded from application stores on the internet. The third file will be this document which attempts to help explain the documents, origin, and data within the three other files in the ZIP. The fourth file will be a newly made metadata file in XML format attempting to establish more visibility of the data. The final fifth file will be the project description (README file) which attempts to organize and frame the data.

It is using DataCite schema, which is an interoperable schema which allows for more access to a wider scope of users. The new metadata file has pertinent information as well as additional subject elements to aid in making this data more relevant or discovered in a keyword search for data in a repository. These files are copyrighted CC0 meaning it is in the public domain allowed for download and redistribution. These files do not contain any information consisting of human subjects. The data has not needed to be anonymized as it contains unsensitive data.

Works Cited:

About Green Roofs. (n.d.). Retrieved January 28, 2021, from https://greenroofs.org/about-green-roofs

Chicago Green Roofs. (2013). Retrieved January 28, 2021, from <https://www.chicago.gov/city/en/depts/dcd/supp_info/chicago_green_roofs.html>

Dataverse Project: About. (n.d.). Retrieved February 11, 2021, from <https://dataverse.org/about>

Digital Preservation Resources. (n.d.). Retrieved February 11, 2021, from https://texasdigitallibrary.atlassian.net/wiki/spaces/DPS/pages/751960128/Digital+Preservation+Resources

Noman, N. (2020, August 24). JSON File Extension: How to open. Retrieved January 28, 2021, from <https://openfileextension.com/open-json-file/>

Texas Data Repository Dataverse. (n.d.). Retrieved February 11, 2021, from <https://dataverse.tdl.org/>

Texas Digital Library Wiki. (n.d.). Retrieved February 11, 2021, from https://texasdigitallibrary.atlassian.net/wiki/spaces/TDRUD/pages/289144946/About