UMD Open Dataset Research & Analysis

Google Dataset Search Critical Infrastructure User Guide

Terrabyte Solutions

INST490: Integrated Capstone for Information Science

May 2022

Table of Contents

1. Introduction

Purpose

The purpose of this guide is to support researchers within the College of Information Studies (iSchool), the Philip Merrill College of Journalism (JSchool), and other schools at the University of Maryland.

Background

This guide has been written by the project members of Terrabyte Solutions for the University of Maryland's Mary Ann Francis. This project is sponsored through the iConsultancy in partnership with the College of Information Studies' Integrated Capstone for Information Science (INST490).

There are many public datasets available that have limited documentation or shared information about how they can be best used and understood. Open source researchers and journalists need information on these open source datasets to effectively and efficiently use them. Projects like this would contribute to a growing repository of documentation about open source data sets that can be made available to a variety of stakeholders.

Scope

The scope of this project includes research into available information about the data source and content, an inventory and survey of the data sets included in the source, analysis of the capabilities of the data source, documentation regarding the datasets and capabilities, and technical data analysis examples of the type of research that could be done with the available data.

2. Describing the Data Source

About Google Dataset Search

Dataset Search is a search engine for datasets designed for a variety of users including but not limited to academic researchers, students, business, analytics, data scientists. It is a fairly new search engine that was first launched on September 5, 2018.

Using a keyword search, users can browse through and explore a variety of datasets hosted in thousands of repositories across the Internet. According to Google Research Scientist, Natasha Noy, Dataset Search has indexed "almost 25 million of these datasets, giving you a single place to search for datasets and find links to where the data is." Over the past few years, users have provided feedback after using the platform, allowing Dataset Search to officially come out of beta testing.

Dataset Search Key Features

Early adopters and users of Dataset Search have contributed to key features. Using on-board filters, you can now filter search results based on the types of datasets you want. The "Download format" selection filters datasets by tabular, document, image, text, archive, or other format. By selecting "Free", users can also choose to view datasets that are available free of charge from the provider. Users may also filter by "Usage rights" to search for results allowing for either commercial or noncommercial use. And the "Last Updated" tool filters datasets that have been last updated over the past month, year, or past 3 years.

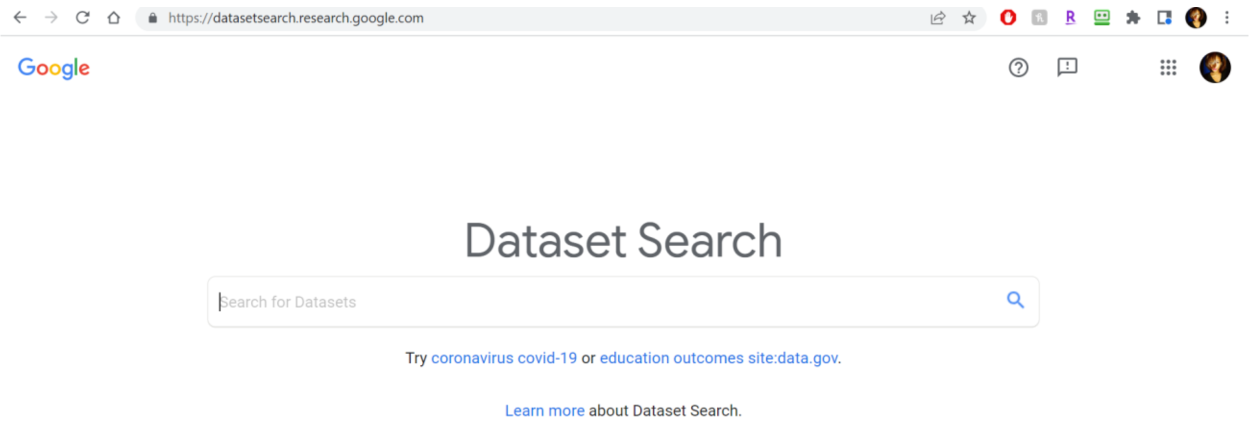
Dataset Search Snapshots Across the Web

According to Google Research's *The Keyword* blog post, the majority of topics that the datasets cover are from the geosciences, biology, and agriculture fields. With over 2 million collections, the United States is the front-runner in open government datasets available on the web. The most popular data format is **tables –** as of January 2020, Dataset Search has indexed over 6 million of them.

Dataset Search continues to grow as future developers describe their datasets with schema.org, an open standard. If you found a dataset that is not yet available on Dataset Search, users can ask the data providers to add the schema.org descriptions so that others can find and learn about their dataset too.

Walkthrough Tutorial of Dataset Search Engine

1. Navigating to the https://datasetsearch.research.google.com/ brings you to the Dataset Search Engine's front page.



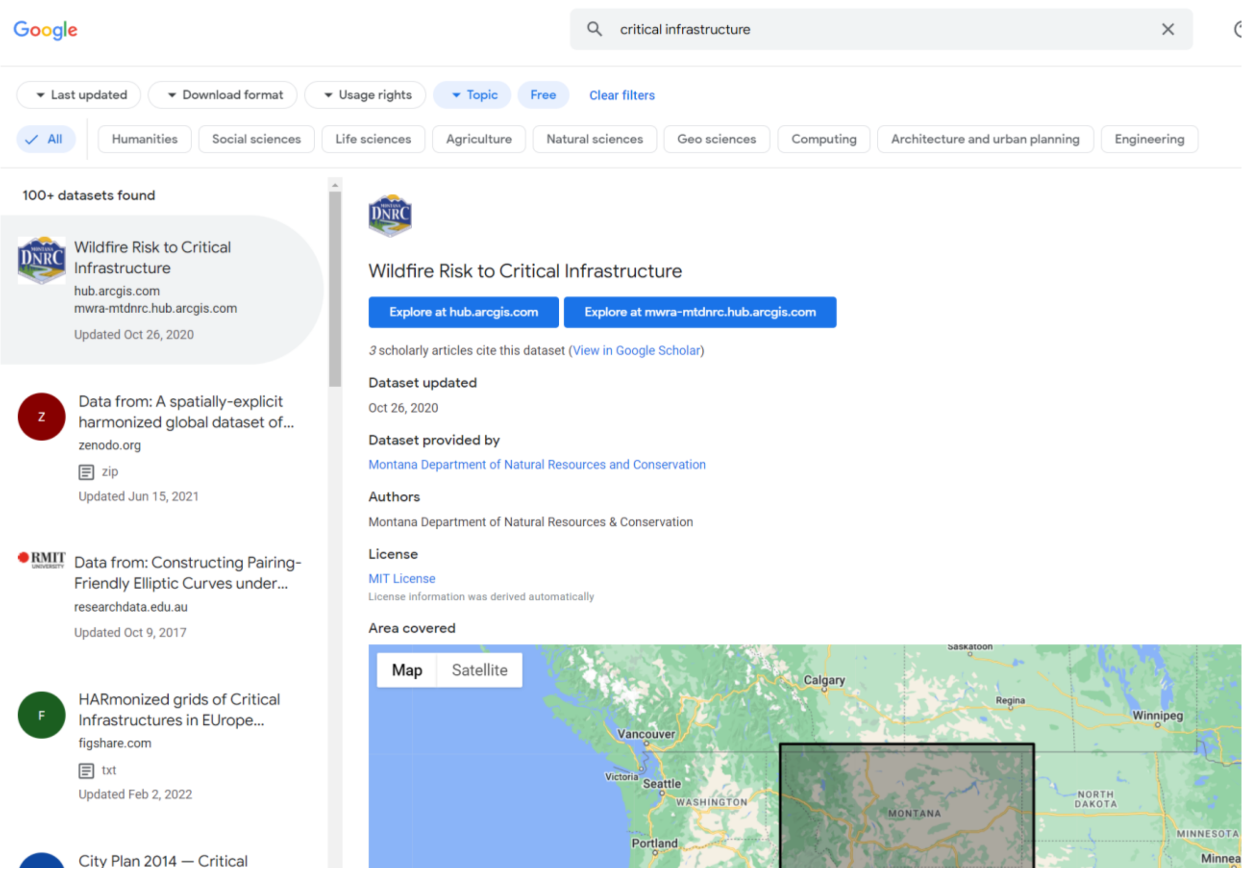
Dataset Search

Search for Datasets

Try coronavirus covid-19 or education outcomes site:data.gov.

Learn more about Dataset Search.

2. After querying “critical infrastructure’ and filtering on free datasets, these are the returned dataset collections.



100+ datasets found

Wildfire Risk to Critical Infrastructure
hub.arcgis.com
mwra-mtdnrc.hub.arcgis.com
Updated Oct 26, 2020

Data from: A spatially-explicit harmonized global dataset of...
zenodo.org
zip
Updated Jun 15, 2021

Data from: Constructing Pairing-Friendly Elliptic Curves under...
researchdata.edu.au
Updated Oct 9, 2017

HARmonized grids of Critical Infrastructures in EUrope...
figshare.com
txt
Updated Feb 2, 2022

City Plan 2014 — Critical

Wildfire Risk to Critical Infrastructure

Explore at hub.arcgis.com

Explore at mwra-mtdnrc.hub.arcgis.com

3 scholarly articles cite this dataset (View in Google Scholar)

Dataset updated
Oct 26, 2020

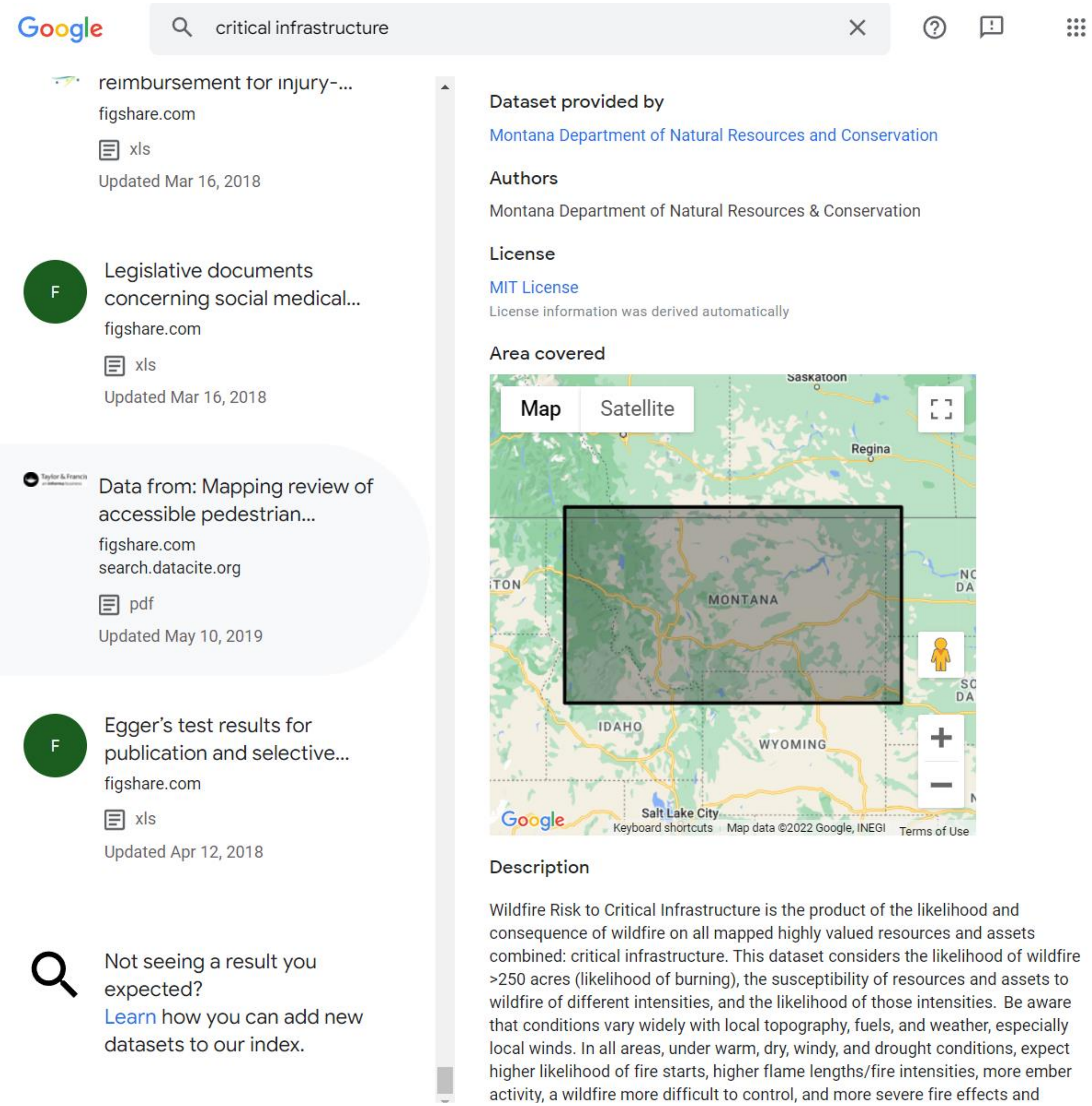
Dataset provided by
Montana Department of Natural Resources and Conservation

Authors
Montana Department of Natural Resources & Conservation

License
MIT License
License information was derived automatically

Area covered

3. Scroll all the way down on the scrollable panel on the left side of the page to fulling access to all datasets in terms of searching.



reimbursement for injury-...
figshare.com
xls
Updated Mar 16, 2018

Legislative documents concerning social medical...
figshare.com
xls
Updated Mar 16, 2018

Data from: Mapping review of accessible pedestrian...
figshare.com
search.datacite.org
pdf
Updated May 10, 2019

Egger's test results for publication and selective...
figshare.com
xls
Updated Apr 12, 2018

Not seeing a result you expected?
Learn how you can add new datasets to our index.

Dataset provided by
Montana Department of Natural Resources and Conservation

Authors
Montana Department of Natural Resources & Conservation

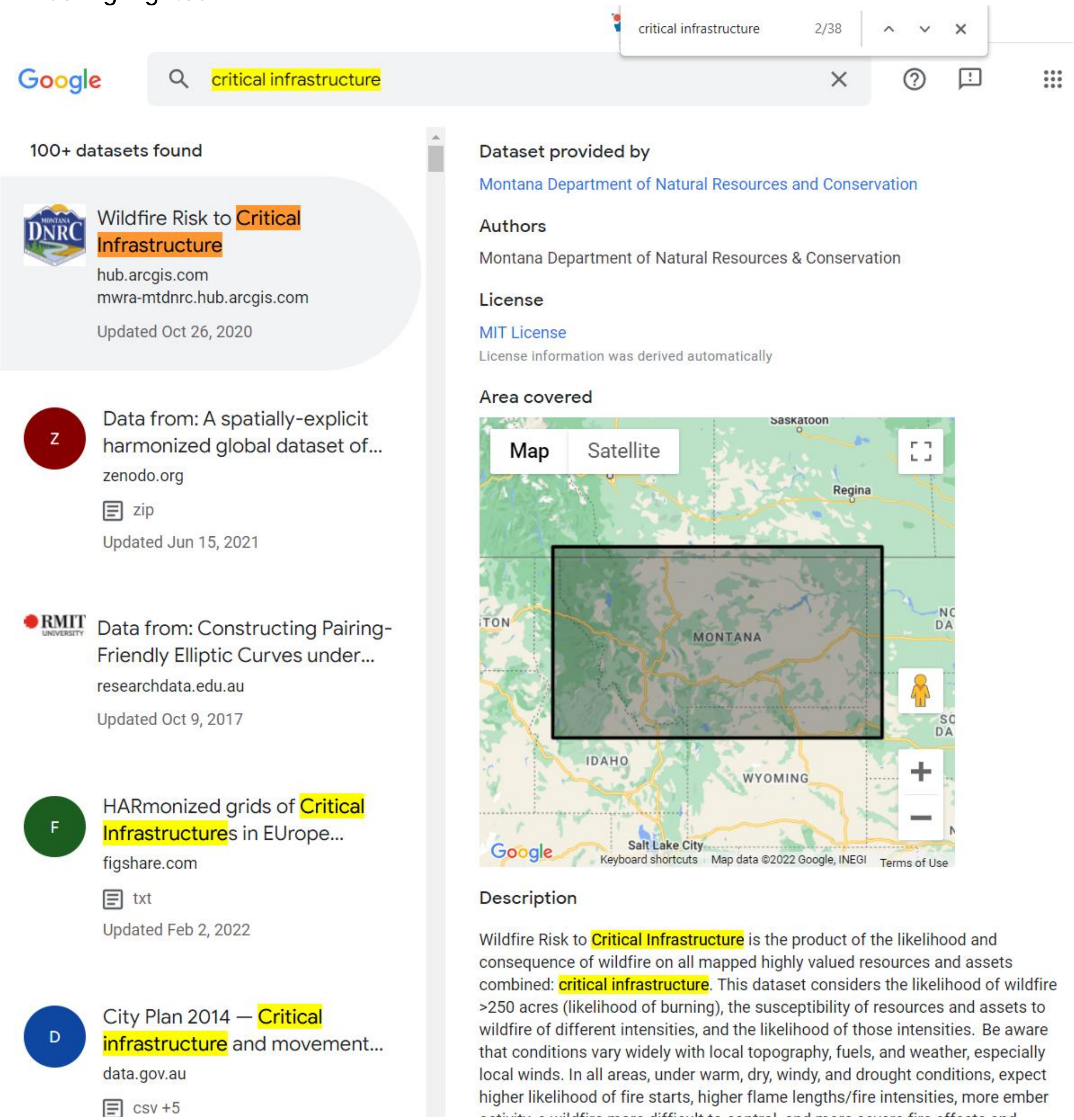
License
MIT License
License information was derived automatically

Area covered

Description

Wildfire Risk to Critical Infrastructure is the product of the likelihood and consequence of wildfire on all mapped highly valued resources and assets combined: critical infrastructure. This dataset considers the likelihood of wildfire >250 acres (likelihood of burning), the susceptibility of resources and assets to wildfire of different intensities, and the likelihood of those intensities. Be aware that conditions vary widely with local topography, fuels, and weather, especially local winds. In all areas, under warm, dry, windy, and drought conditions, expect higher likelihood of fire starts, higher flame lengths/fire intensities, more ember activity, a wildfire more difficult to control, and more severe fire effects and

4. Press Ctrl + F on the keyboard to search the specific dataset. The related datasets will be highlighted.



100+ datasets found

Wildfire Risk to Critical Infrastructure
hub.arcgis.com
mwra-mtdnrc.hub.arcgis.com
Updated Oct 26, 2020

Data from: A spatially-explicit harmonized global dataset of...
zenodo.org
zip
Updated Jun 15, 2021

Data from: Constructing Pairing-Friendly Elliptic Curves under...
researchdata.edu.au
Updated Oct 9, 2017

HARmonized grids of Critical Infrastructures in EUrope...
figshare.com
txt
Updated Feb 2, 2022

City Plan 2014 — Critical infrastructure and movement...
data.gov.au
csv +5

Dataset provided by

Montana Department of Natural Resources and Conservation

Authors

Montana Department of Natural Resources & Conservation

License

MIT License

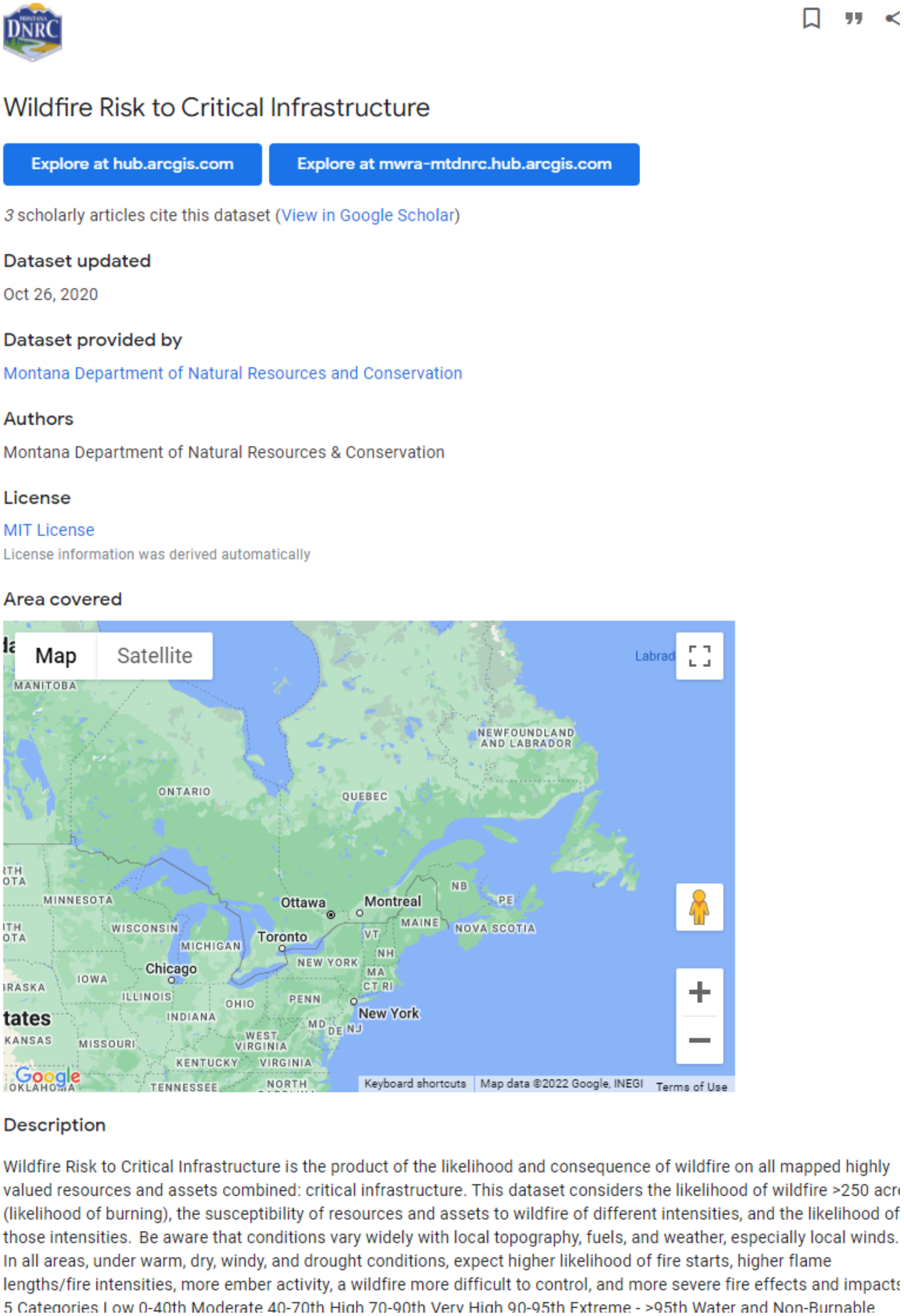
License information was derived automatically

Area covered

Description

Wildfire Risk to Critical Infrastructure is the product of the likelihood and consequence of wildfire on all mapped highly valued resources and assets combined: critical infrastructure. This dataset considers the likelihood of wildfire >250 acres (likelihood of burning), the susceptibility of resources and assets to wildfire of different intensities, and the likelihood of those intensities. Be aware that conditions vary widely with local topography, fuels, and weather, especially local winds. In all areas, under warm, dry, windy, and drought conditions, expect higher likelihood of fire starts, higher flame lengths/fire intensities, more ember activity, a wildfire more difficult to control, and more severe fire effects and

5. Click on the selected dataset on the left side of the webpage, the basic information which included: title, article link, dataset updated date, dataset provider, authors, license, the country/area of article, and the description of article.



Wildfire Risk to Critical Infrastructure

Explore at hub.arcgis.com

Explore at mwra-mtdnrc.hub.arcgis.com

3 scholarly articles cite this dataset (View in Google Scholar)

Dataset updated

Oct 26, 2020

Dataset provided by

Montana Department of Natural Resources and Conservation

Authors

Montana Department of Natural Resources & Conservation

License

MIT License

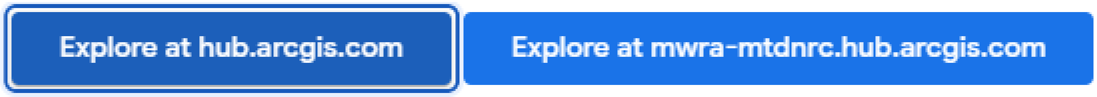
License information was derived automatically

Area covered

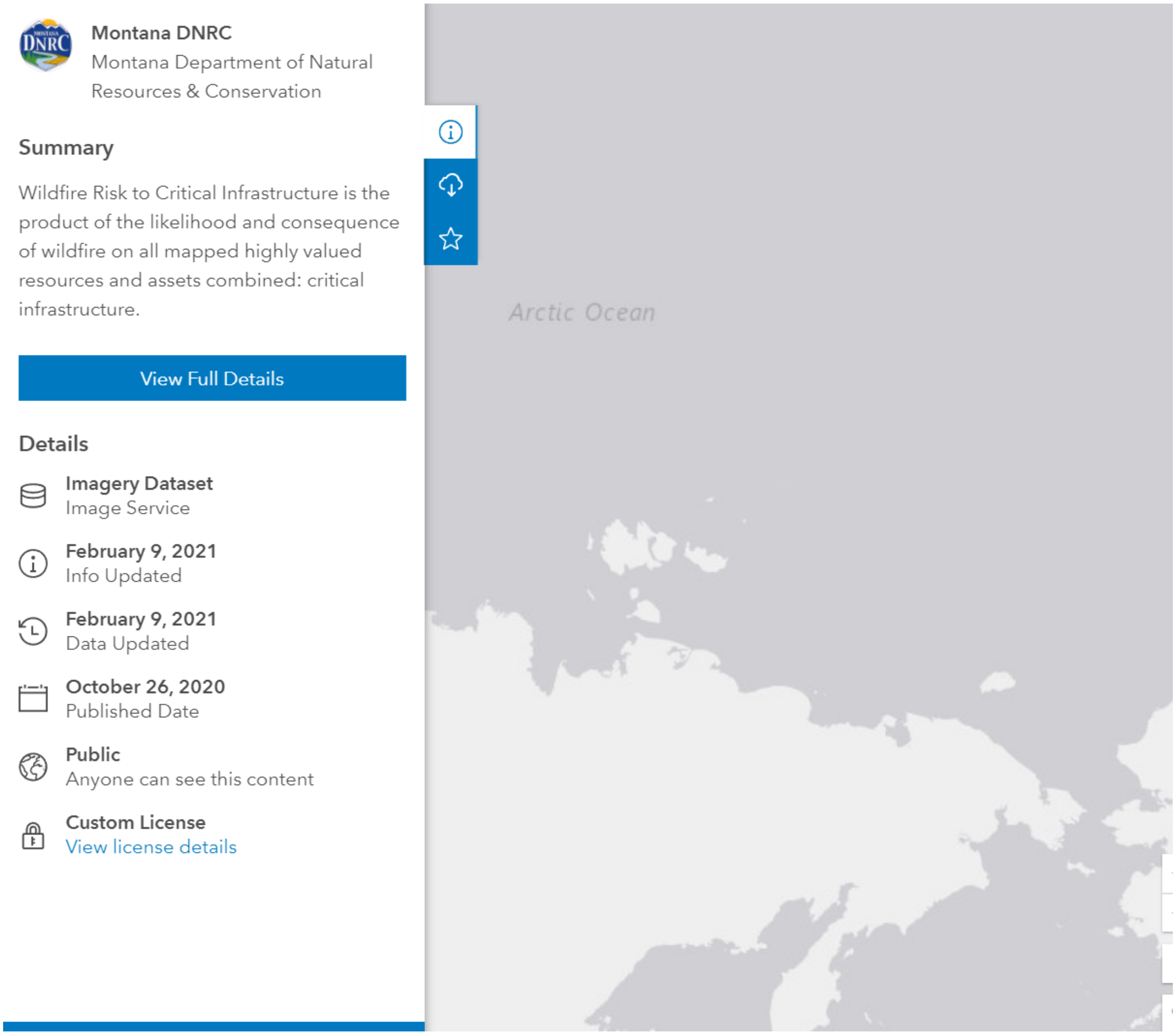
Description

Wildfire Risk to Critical Infrastructure is the product of the likelihood and consequence of wildfire on all mapped highly valued resources and assets combined: critical infrastructure. This dataset considers the likelihood of wildfire >250 acr (likelihood of burning), the susceptibility of resources and assets to wildfire of different intensities, and the likelihood of those intensities. Be aware that conditions vary widely with local topography, fuels, and weather, especially local winds. In all areas, under warm, dry, windy, and drought conditions, expect higher likelihood of fire starts, higher flame lengths/fire intensities, more ember activity, a wildfire more difficult to control, and more severe fire effects and impact 5 Categories Low 0-40th Moderate 40-70th High 70-90th Very High 90-95th Extreme - >95th Water and Non-Burnable

6. Click on the “Explore at” links to be redirected to more details regarding the dataset source, and for access to direct file downloads.


Explore at hub.arcgis.com
Explore at mwra-mtdnrc.hub.arcgis.com

7. This is an example of the “Explore at” link redirecting the user to ArcGIS Hub platform where the raw data is accessible and available for download.



Montana DNRC
Montana Department of Natural Resources & Conservation

Summary

Wildfire Risk to Critical Infrastructure is the product of the likelihood and consequence of wildfire on all mapped highly valued resources and assets combined: critical infrastructure.

View Full Details

Details

Imagery Dataset
Image Service

February 9, 2021
Info Updated

February 9, 2021
Data Updated

October 26, 2020
Published Date

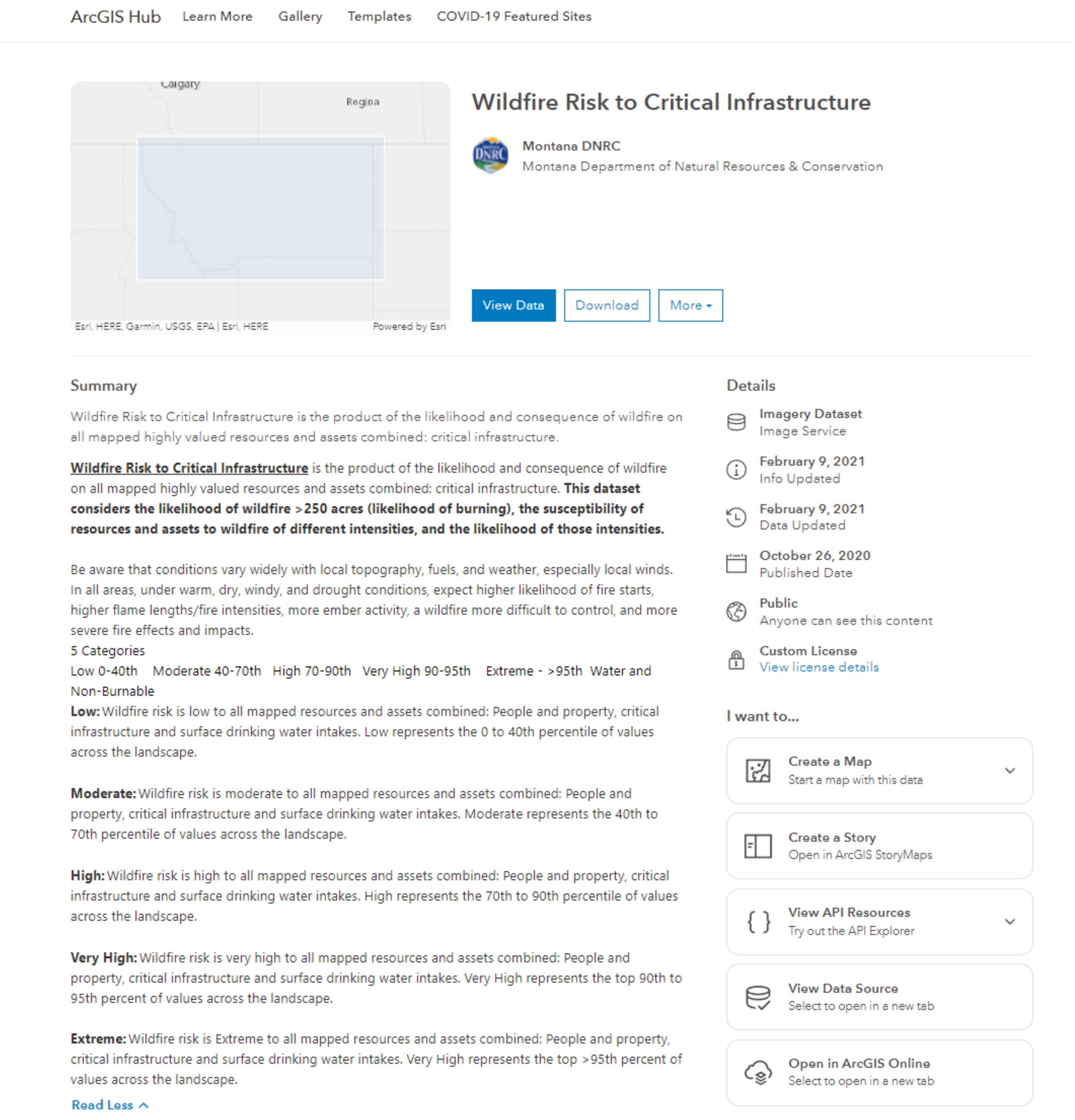
Public
Anyone can see this content

Custom License
View license details

8. Click on View Full Details.

9. Here, more details about the above dataset are available below through the “Explore

at" links. Users will have access to any additional information they were not able to retrieve directly from the Dataset Search Engine results preview page.



Wildfire Risk to Critical Infrastructure

Montana DNRC
Montana Department of Natural Resources & Conservation

Summary

Wildfire Risk to Critical Infrastructure is the product of the likelihood and consequence of wildfire on all mapped highly valued resources and assets combined: critical infrastructure.

Wildfire Risk to Critical Infrastructure is the product of the likelihood and consequence of wildfire on all mapped highly valued resources and assets combined: critical infrastructure. **This dataset considers the likelihood of wildfire >250 acres (likelihood of burning), the susceptibility of resources and assets to wildfire of different intensities, and the likelihood of those intensities.**

Be aware that conditions vary widely with local topography, fuels, and weather, especially local winds. In all areas, under warm, dry, windy, and drought conditions, expect higher likelihood of fire starts, higher flame lengths/fire intensities, more ember activity, a wildfire more difficult to control, and more severe fire effects and impacts.
5 Categories
Low 0-40th Moderate 40-70th High 70-90th Very High 90-95th Extreme - >95th Water and Non-Burnable
Low: Wildfire risk is low to all mapped resources and assets combined: People and property, critical infrastructure and surface drinking water intakes. Low represents the 0 to 40th percentile of values across the landscape.

Moderate: Wildfire risk is moderate to all mapped resources and assets combined: People and property, critical infrastructure and surface drinking water intakes. Moderate represents the 40th to 70th percentile of values across the landscape.

High: Wildfire risk is high to all mapped resources and assets combined: People and property, critical infrastructure and surface drinking water intakes. High represents the 70th to 90th percentile of values across the landscape.

Very High: Wildfire risk is very high to all mapped resources and assets combined: People and property, critical infrastructure and surface drinking water intakes. Very High represents the top 90th to 95th percent of values across the landscape.

Extreme: Wildfire risk is Extreme to all mapped resources and assets combined: People and property, critical infrastructure and surface drinking water intakes. Very High represents the top >95th percent of values across the landscape.

Read Less ^

Details

Imagery Dataset
Image Service

February 9, 2021
Info Updated

February 9, 2021
Data Updated

October 26, 2020
Published Date

Public
Anyone can see this content

Custom License
View license details

I want to...

Create a Map
Start a map with this data

Create a Story
Open in ArcGIS StoryMaps

View API Resources
Try out the API Explorer

View Data Source
Select to open in a new tab

Open in ArcGIS Online
Select to open in a new tab

3. Critical Infrastructure Web Scraping Spreadsheet

General Information

After researching available information about Dataset Search and its content, we surveyed and assessed the ***critical infrastructure*** related datasets. This work product serves as a snapshot of the available dataset results as of March 23, 2022. The dataset results contained a total of 174 entries.

Data Collection Process

The Developer team of Terrabyte Solutions used the Beautiful Soup library in Python to build a script to scrape and parse data from the Dataset Search results page for more streamlined analysis and processing.

Although we were able to automate many of the resulting datasets' title, dataset provider, published date, and website links, we noticed that Google Dataset often reported an outdated last "Dataset updated" date, which was only accessible after directly accessing the "Explore at" links.

To maximize data completeness and accuracy, as a team, we manually verified the web scraped data and added several additional fields (keyword tags, doi, geographic context, etc.) to provide a more enhanced context into the capabilities of the data source and the resulting "critical infrastructure" related datasets. Our collection is available in the Google Sheets cloud-based application, which can be easily accessed and exported into several file formats based on user needs.

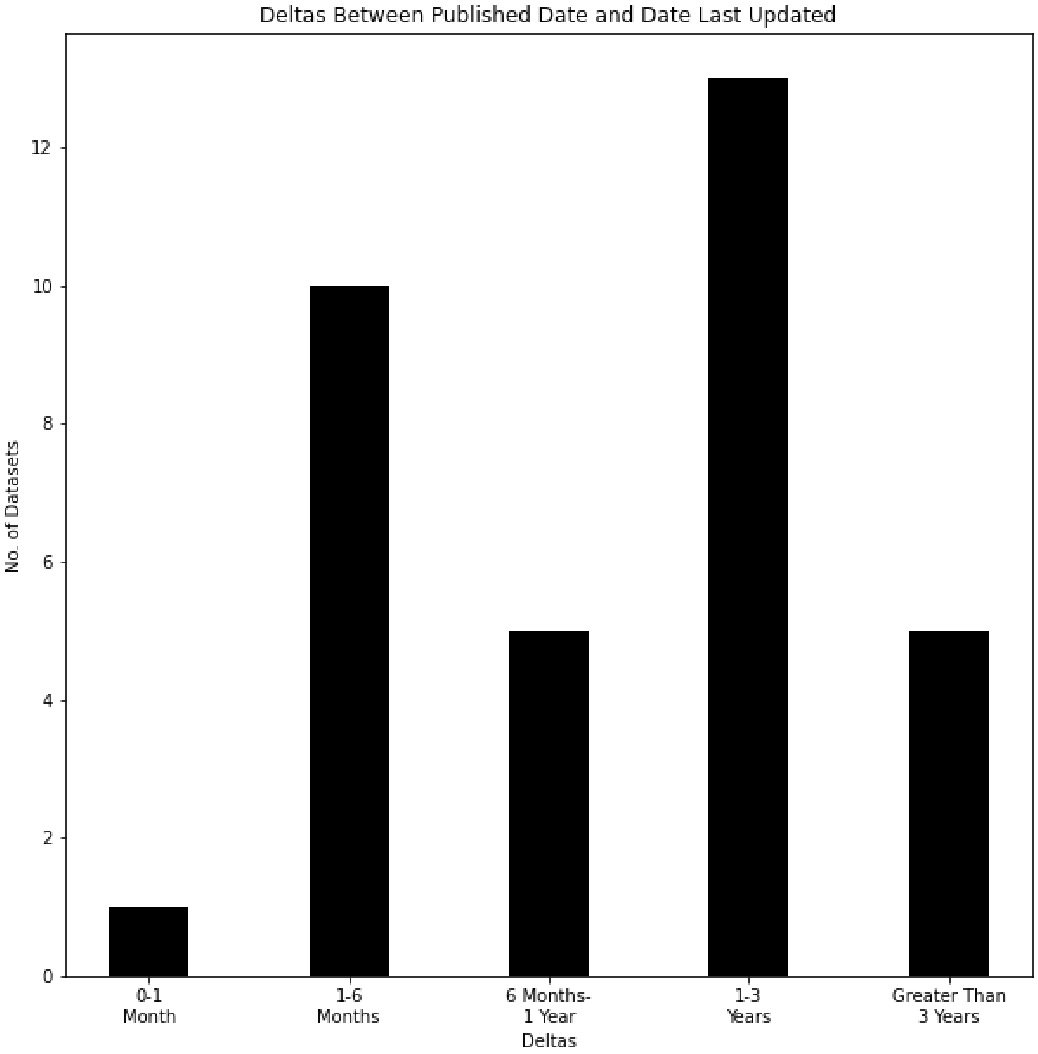
How is This Spreadsheet Organized?

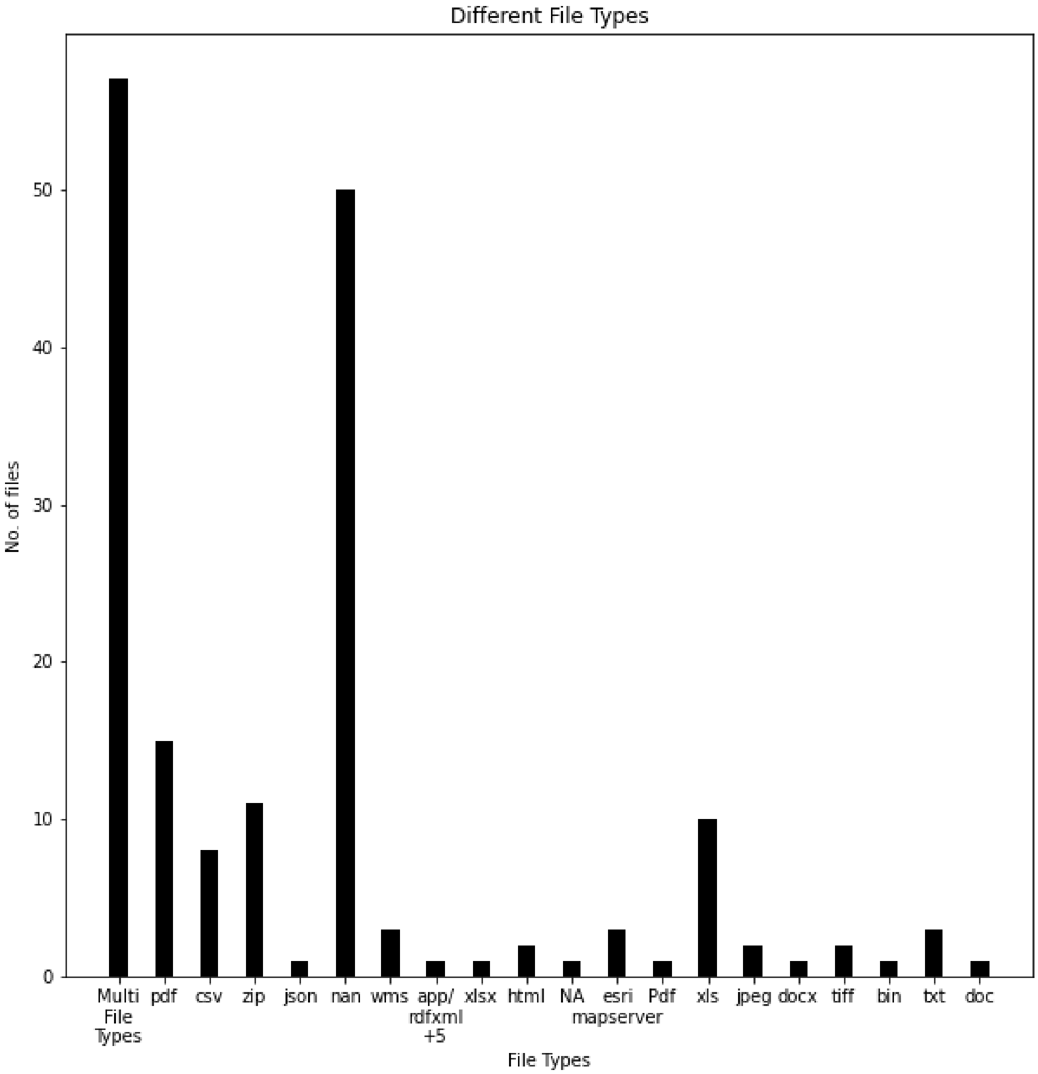
The spreadsheet document variables are explained in the following table:

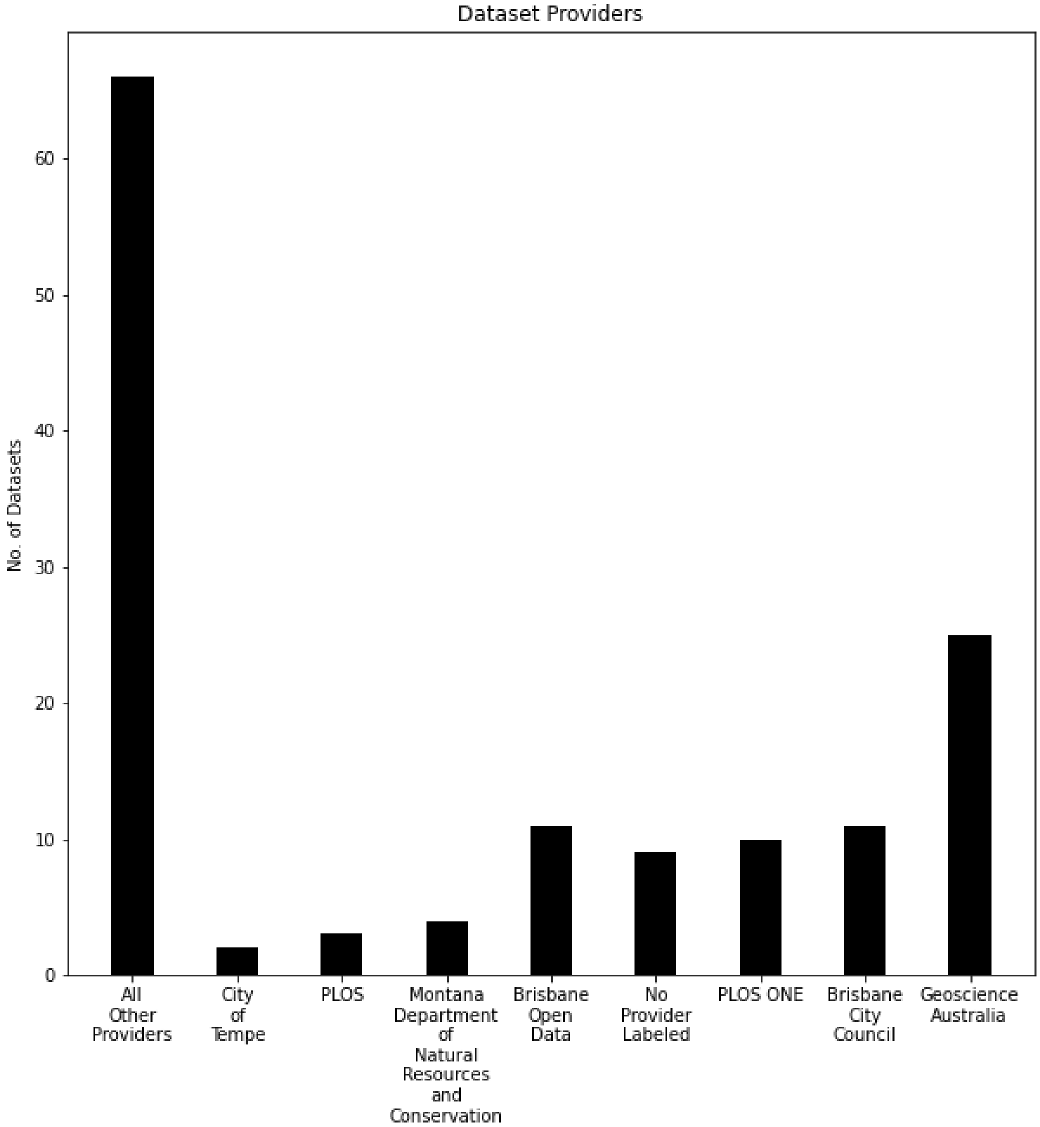
Variable	**Operational Definition**
Title	Title of the dataset.
Provider	Identifies those who provided data to the client dataset and/or external data store. Data providers are often organizations that provide internal data to external parties and companies.
Authors	Identifies the authors of the dataset – if applicable.
Country of Dataset Focus	Indicates the geographic area that a dataset focuses on.
License	Specifies the license type for each dataset. Indicates permissions and rights for those who seek to build upon, distribute, tweak, or remix a creator's work.
DOI (digital object identifier)	If applicable, provides the unique, permanent string that is assigned to online journals, academic articles, books, and other works.
Keyword Tags	Tags placed by the creator of the content to describe what the content is and what it relates to.
Published Date	The initial date that the dataset was published.
Date Last Updated	Most recent date that the dataset was last updated.
Website Link	Lists the dataset's website domain and is hyperlinked directly to the site for direct access by users.
Data Status	Describes whether or not the original data is currently accessible and available for

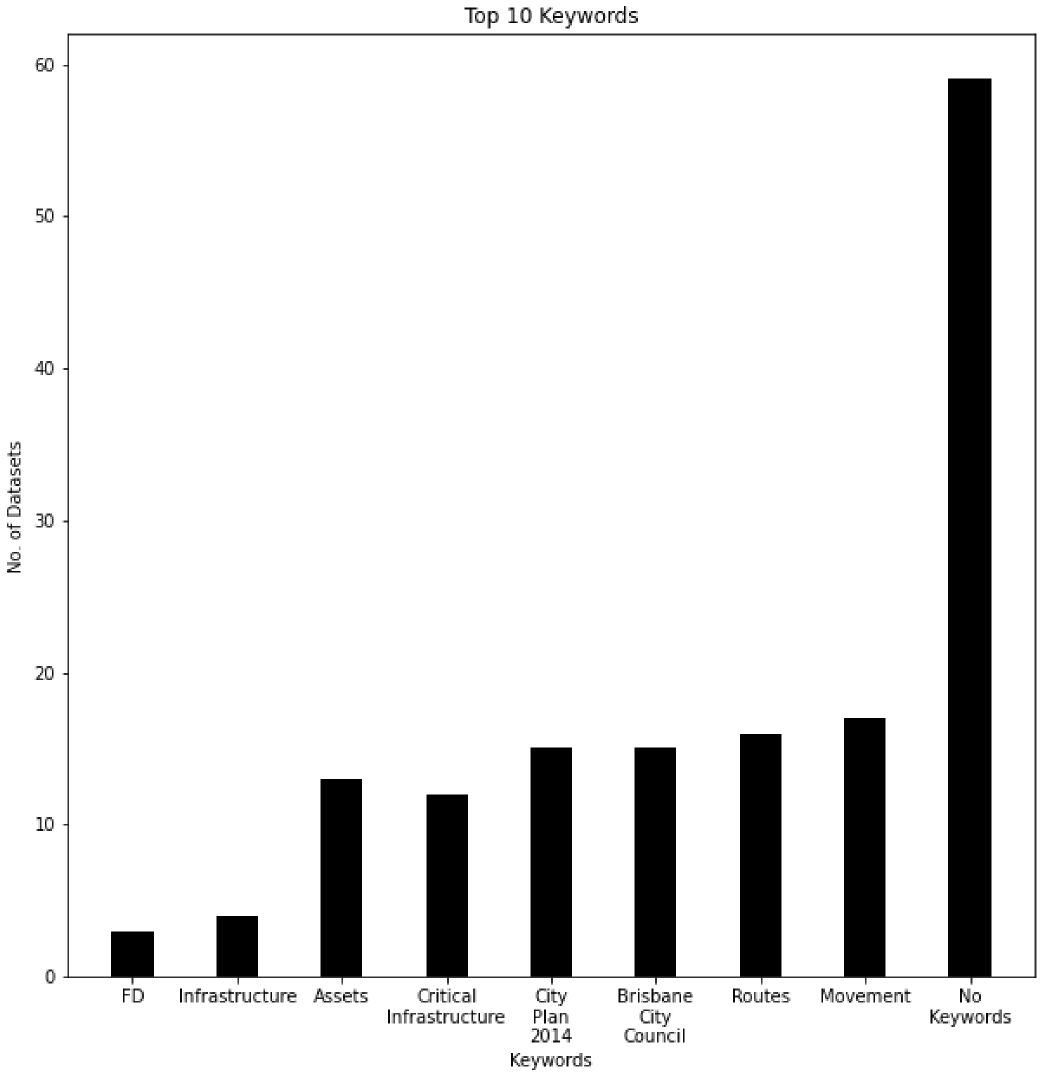
	public and immediate download.
Filetype	Provides a list of the various file types included in the original dataset(s).
Description	A brief description and summary of the dataset, sourced from the authors/dataset providers. These descriptions offer insight into the source’s goals, assumptions, core ideas, and implications.

4. Sample Data Analysis Examples




Different File Types
No. of files
0
10
20
30
40
50
Multi File Types
pdf
csv
zip
json
nan
wms
app/ rdfxml +5
xlsx
html
NA
esri mapserver
Pdf
xls
jpeg
docx
tiff
bin
txt
doc
File Types


Dataset Providers
No. of Datasets
0
10
20
30
40
50
60
All Other Providers
City of Tempe
PLOS
Montana Department of Natural Resources and Conservation
Brisbane Open Data
No Provider Labeled
PLOS ONE
Brisbane City Council
Geoscience Australia


Top 10 Keywords
No. of Datasets
0
10
20
30
40
50
60
FD
Infrastructure
Assets
Critical Infrastructure
City Plan 2014
Brisbane City Council
Routes
Movement
No Keywords
Keywords