Data Catalogue Requirements

Created Jan 31, 2024

# Purpose:

A “cleaned” dataset is one that had been accessed and readied for analysis. This is a valuable resource for others. A data catalogue enables a data analyst to utilize “cleaned” data by understanding:

1. what “cleaned” data is available for use;
2. what the data can be used for;
3. how to find both the “processed” and “raw” data.

# Intended user:

The data catalogue is written for a person who would like to access readily available data for use in analyzing data to answer questions.

# Instructions for a complete Data Catalogue entry:

A complete data catalogue entry consists of the following things:

1. Filled in README Content Table (see below – cut and paste that into a separate document titled “<data set> Catalogue README” – either a Word or Google Doc
2. Updated table entry in Data Catalogue.xlsx

<https://sci4ga.sharepoint.com/:x:/s/programming_committee/ETziw497gWFAserShs76-PYB27AYWmDNRzc6UPeXE_Mt6w?e=o48J5u>

rows to fill in:

1. Cleaned Data File Link (already filled in – double check this is right)
2. Analysis Code GitHub Repo Directory Link (already filled in – double check this is right)
3. High Level Overview
4. Search Keywords

# What you need to do:

1. Find the “cleaned” data set.
2. Find associated cleaning / analysis code.
3. Create a README Catalogue File. (Turn the file into Amy and Michael – we will upload in the associated data directory on Sci4Ga’s sharepoint)
4. Make sure there is a GitHub Directory for that **individual dataset code and analysis and related files**
   1. If not, make one. For example – in ej-bucket-exploration/DataThink – all the eight categories are in the one main directory. Make a separate one for “Climate Change” “Energy” etc and sort the files into there
5. Put a link to that individual GitHub Directory in the Data Catalogue Spreadsheet in “Analysis Code GitHub Repo Directory Link”
6. Ensure there is a link to the **cleaned data** in the “Cleaned Data File Link” column
   1. If that cleaned data is in GitHub let Amy and Michael know – we need to get it out of GitHub and into Sci4Ga sharepoint for space limitations.
7. Add “High Level Overview” and “Search Keywords” into the Data Catalogue Spreadsheet

# Data Catalog README Content Table

(see next page)

|  |  |  |
| --- | --- | --- |
| Category | Information | |
| High level, 2 sentence overview | | The dataset contains variables that describe the whole picture of water quality by showing the characteristic of water pollutant and its result and measurement. |
| Search keywords | | UGA Spring Internship 2023, Water Quality, data |
| Data Dictionary? Y/N | | Yes, <https://sci4ga.sharepoint.com/:x:/r/sites/programming_committee/_layouts/15/Doc.aspx?sourcedoc=%7BA28EEFFB-87D2-4484-A7C6-A962C701918B%7D&file=Potential%20Harmful%20Chemicals%20from%20WQP.xlsx&action=default&mobileredirect=true> |
| Does it contain PII? If yes – STOP HERE. | | No |
| Use cases. Links to example uses or a few sentences for intended use. | | | This dataset named ‘waterqualitydatafinal’ is used for the application of the identification of potential sites of major pollution as well as the correlation between water quality and health outcome. |
| How to access the Raw Data Source. (Link to data source and access instructions.) | | <https://sci4ga.sharepoint.com/:x:/r/sites/programming_committee/_layouts/15/Doc.aspx?sourcedoc=%7B97A122BB-FF14-48A9-B3D6-B9D48450B7AA%7D&file=waterqualitydatafinal.csv&action=default&mobileredirect=true> |
| Raw Data collector and trustworthiness of source. | | Not mentioned. |
| Dynamic or static? (Does the analyst have to re-download and re-clean, or is the data set up-to-date) | | Dynamic |
| How is the raw data maintained? | | Stored in GitHub, the old data set can be found in data/archive. |
| Geographic and time coverage of data. | | From 2017 to 2021 |
| Caveats or limitations on data. | | Not seen at this moment. |
| When was the raw data accessed for “cleaning”? | | April 24th, 2023 |
| How was the data verified? | | Not mentioned. |
| Link to code and instructions for “cleaning” | | <https://sci4ga.sharepoint.com/sites/programming_committee/Shared%20Documents/Forms/AllItems.aspx?ga=1&id=%2Fsites%2Fprogramming%5Fcommittee%2FShared%20Documents%2FData%20Focused%20Programs%2FDavid%27s%20Files%2FUGASpringInternship2023RProject%2Fcode%2Fcodeforcleaningdata%2ER&viewid=5cdfb080%2D5731%2D46bd%2D993a%2D197bb56ae965&parent=%2Fsites%2Fprogramming%5Fcommittee%2FShared%20Documents%2FData%20Focused%20Programs%2FDavid%27s%20Files%2FUGASpringInternship2023RProject%2Fcode> |
| Completeness vs. sparsity | | Rather complete |
| Link to any code and instructions for data analysis. | | <https://sci4ga.sharepoint.com/sites/programming_committee/Shared%20Documents/Forms/AllItems.aspx?ga=1&id=%2Fsites%2Fprogramming%5Fcommittee%2FShared%20Documents%2FData%20Focused%20Programs%2FDavid%27s%20Files%2FUGASpringInternship2023RProject%2Fcode%2Farchive%2FWaterQualityDataMarkdown%2ERmd&viewid=5cdfb080%2D5731%2D46bd%2D993a%2D197bb56ae965&parent=%2Fsites%2Fprogramming%5Fcommittee%2FShared%20Documents%2FData%20Focused%20Programs%2FDavid%27s%20Files%2FUGASpringInternship2023RProject%2Fcode%2Farchive> |
| Link to cleaned dataset | | <https://sci4ga.sharepoint.com/:x:/r/sites/programming_committee/_layouts/15/Doc.aspx?sourcedoc=%7B97A122BB-FF14-48A9-B3D6-B9D48450B7AA%7D&file=waterqualitydatafinal.csv&action=default&mobileredirect=true> |
| Link to GitHub Repository for READMEs, code, etc. | |  |

# 