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 | Harmfulness levels of major pesticides, water quality markers, and pollutant in the Water Quality Portal service. |
| Search keywords | Pesticides, water quality, pollutant |
| Data Dictionary? Y/N | Y |
| Does it contain PII? If yes – STOP HERE. | N |
| Use cases. Links to example uses or a few sentences for intended use. | Chemical pollution in Georgia |
| How to access the Raw Data Source. (Link to data source and access instructions.) | A website about water quality and health map, but I couldn’t find it. |
| Raw Data collector and trustworthiness of source. | A master student in UGA |
| Dynamic or static? (Does the analyst have to re-download and re-clean, or is the data set up-to-date) | Not up-to-date. The data can be only collected from the website. |
| How is the raw data maintained? | Not mentioned |
| Geographic and time coverage of data. | 2017-2021 |
| Caveats or limitations on data. | So many variables that makes the table hard to read |
| When was the raw data accessed for “cleaning”? | 2023 |
| How was the data verified? | Not sure |
| Link to code and instructions for “cleaning” | Look at the github repo |
| Completeness vs. sparsity | Great completeness |
| Link to any code and instructions for data analysis. | Not found yet |
| Link to cleaned dataset | Included in github repo |
| Link to GitHub Repository for READMEs, code, etc. | https://github.com/sci4ga/Health-and-WQ-2023.git |

# 