# READ ME Template

|  |  |  |
| --- | --- | --- |
| **Version** | **Date** | **Description** |
| V1 | 10-16-2024 | Initial Commit- data/file navigation/format/ Carter and Mike contributors added |
| V2 | 12-4-2024 | Added Software Explanation and Flowchart |
| V3 |  |  |

**TITLE** Superfund NPL Scraper

## BACKGROUND

**STUDY RATIONALE** In the background describe how they study came about, the rationale for the study, the study design, and study partners. Any additional notes (e.g., sampling at a facility and the facility shut down shortly after the study concluded) should also be included.

**POPULATION OF INTEREST/ MODEL ORGANISM/STUDY LOCATION** As appropriate, describe the population of interest (e.g., environmental justice community in West Eugene, OR), the animal or cell model, or the study location (e.g., for environmental exposure studies describe where the study took place).

**RESEARCH AND COMMUNITY GOALS/HYPOTHESIS** This section should be modified based on the study design.

**IRB/ACUP DETERMNATION**. Describe if the project was reviewed by IRB or IACUC, and the final determination. Example for an IRB-approved study: “The study was reviewed and approved by the OSU Institutional Review Board, #IRB-2021-1087. All data has been de-identified for analysis. The codebook containing participant identities and addresses is accessible only to Rohlman and Germano.”

**STUDY RESOURCES** List any helpful resources that describe the context for the study, or would be useful context for someone new to the study who is trying to get acquainted with the study. Do not just bullet out resources, describe each resource.

* Research Data Management Basics. George mason University. This is geared predominantly towards data dictionaries and codebooks, but also includes good information on how to handle sensitive information and personal identifiers. <https://infoguides.gmu.edu/data-management/document>
* How to write a good README file for your GitHub project. This resource (2021) is geared for README files for programming files, but contains excellent advice for developing a README that will give users a detailed description of the project. <https://www.freecodecamp.org/news/how-to-write-a-good-readme-file/>
* The ultimate guide to writing a great README.md for your project. Written by Medium, and designed for users of Markdown, this guide nonetheless details the benefits of a README. <https://www.freecodecamp.org/news/how-to-write-a-good-readme-file/>
* How to write a good Readme for your Data Science project on GitHub. Again written for data scientists that are writing and documenting code, this piece highlights the importance of documenting your data sources (descriptions of the data, pre-processing steps), as well as a results and evaluation section for the document. <https://medium.datadriveninvestor.com/how-to-write-a-good-readme-for-your-data-science-project-on-github-ebb023d4a50e>

## STUDY/PROJECT TEAM

Here, list the major project team members and their contributions to the project, along with contact information, as appropriate so that the person using the README can reach out for more information as needed. To help with this section, consider using the CRediT statements to first identify how team members contributed, and expand as needed. These roles were developed for writing author contributions for manuscripts, so all roles may not be applicable in this situation. Delete the CRediT table when done.

PARTNERSHIPS

SOW or MOU or PARTNERSHIP AGREEMENT: (link to document)

|  |  |  |
| --- | --- | --- |
| **Name** | **Project Role** | **Contact Information** |
| Carter Deal | Software: Designed and developed the program to the specifications given | dealc@oregonstate.edu |
| Mike Barton | Supervision: Oversaw development of the program and provided direction of development | Micheal.barton@oregonstate.edu |

**Credit: CONTRIBUTOR ROLE TAXONOMY** | <https://credit.niso.org/>

|  |  |
| --- | --- |
| **Contributor Role** | **Description** |
| Conceptualization | Ideas; formulation or evolution of overarching research goals and aims. |
| Data curation | Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later re-use. |
| Formal analysis | Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data. |
| Funding acquisition | Acquisition of the financial support for the project leading to this publication. |
| Investigation | Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection. |
| Methodology | Development or design of methodology; creation of models. |
| Project administration | Management and coordination responsibility for the research activity planning and execution. |
| Resources | Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools. |
| Software | Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components. |
| Supervision | Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team. |
| Validation | Verification, whether as a part of the activity or separate, of the overall replication/reproducibility of results/experiments and other research outputs. |
| Visualization | Preparation, creation and/or presentation of the published work, specifically visualization/data presentation. |
| Writing – original draft | Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation). |
| Writing – review and editing | Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or post-publication stages. |

**FUNDING ACKNOWLEDGEMENTS**

Here, list any funding that contributed to the project. If funding is tied to a specific person (e.g., training grant), indicate that here as well.

## DOCUMENTS

In this section, describe where files can be found, e.g., a shared Box or Google Drive folder. Documents that are restricted due to containing personally identifiable information (PII) may be in different locations. This should be described, and conditions of access should also be described (e.g., only individuals listed on the IRB may be able to access those files). Use the following table to describe the necessary documents. For folders with restricted access (e.g., Box, Google Drive, Microsoft OneNote) include the name of the person that owns the folder.

If there is an IRB or ACUP, include those approved files in this table as well (recruitment materials, consent form, questionnaires, etc.)

|  |  |  |  |
| --- | --- | --- | --- |
| **Format** | **Title** | **Description** | **Location** |
| .py  Python | Superfund.py | Python file containing the execution script for the program.  There are two versions of this file for the two different versions of superfund programs, one that only utilizes the xlsx file given by the user, and one that uses the EPA list site given by the user and/or the xlsx file | Github repository (Hosted by Mike Barton)  Link to xlsx only version:  <https://github.com/bartonmike/superfund-npl-scraper/tree/main/superfund_xlsx>  Link to xlsx and/or site version:  <https://github.com/bartonmike/superfund-npl-scraper/tree/main/superfund_site_and_xlsxv> |
| .txt  Text file | Requirements.txt | Txt file that contains the required python packages for executing the superfund.py scripts. There are two versions of this file for the different superfund.py versions. | Github repository (Hosted by Mike Barton)  Link to xlsx only version:  <https://github.com/bartonmike/superfund-npl-scraper/tree/main/superfund_xlsx>  Link to xlsx and/or site version:  <https://github.com/bartonmike/superfund-npl-scraper/tree/main/superfund_site_and_xlsxv> |

**FILE NAVIGATION**

All programs and files can be found on the github repository accessed through this link: <https://github.com/bartonmike/superfund-npl-scraper>

The files in the repository can be broken down into three separate sections:

* Folder containing the superfund.py and requirements.txt needed for the site and xlsx version of the code
* Folder containing the superfund.py and requirements.txt needed for the xlsx only version of the code
* The README file containing the instructions on how to run the code and what it does

Outline of how the files appear in the repository:

* superfund\_site\_and\_xlsx (folder)
  + superfund.py
  + requirements.txt
* superfund\_xlsx (folder)
  + superfund.py
  + requirements.txt
* README.md

Describe where information in files can be found. This should include the names of main files, as well as information about specific sheets or sub-sections within the file. For example, here you may describe where specific graphs and visualizations were developed, and where the underlying dataset was pulled from. Tables or flow charts may be helpful. A flow chart section is provided at the end of the document as an example.

For large projects, consider using a Data Management Plan tracking document. This Excel document (see description and link below) provides an overview of all data collected in the project, the status of the data collection and data analysis, and recommended metadata and nomenclature conventions.

* Data Management Plan template. Tabs cover different elements of the study, and examples are provided, specifically the tab ‘DMP Klamath’.
  + This template is located in BOX: <https://oregonstate.app.box.com/file/1609802550199>
    - BOX > Rohlman Lab > Data Documentation and READ ME Templates > Data Management Plan Template

## STUDY DESIGN

Briefly describe how the study will be designed (e.g., single time point, longitudinal).

**RECRUITMENT**

As necessary, describe how recruitment will be conducted.

**CONSENT**

As necessary, describe how consent will be sought (written, verbal, remote) and how informed consent will be determined.

**SAMPLING DESIGN**

If needed, describe the sampling design, e.g., samples will be collected in concentric circles located at 0.25, 0.5, and 1.0 miles from the point source.

## FIELD WORK

In studies where field work was conducted, describe when and where field work was conducted, and for how long (e.g., sampling duration). Include any relevant notes (e.g., heavy wildfire smoke on day 2 of sampling).

## DATA PROCESSING AND ANALYSIS

This section will have multiple sub-sections. Use only those that are applicable, and add as appropriate.

If specific protocols were used for processing or analysis, reference those in the appropriate sub-sections.

**DATA COLLECTION TOOLS**

Python

Selenium

In this section, list and briefly describe the data collection tools that were used. This may include air monitors, surveys, focus groups, etc. For a literature review, describe the search platforms that were used.

**DESCRIPTION OF DATA**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Site Name | City | State | EPA ID | Search ID | Score | NAI | NAI Entity |

Above are the values that the xlsx files will always have (the NAI and NAI entity will not be available in the xlsx and site version of the program if the user chooses to only read the site)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Initial Assessment Completed | Proposed to the National Priorities List | Finalized on the National Priorities List | Remedial Investigation Started | Final Remedy Selected | Final Remedial Action Started | Remedy Selected | Remedial Action Started | Construction Completed | Deleted from National Priorities List | Most Recent Five-Year Review | Achieved Sitewide Ready for Anticipated Reuse |

Above are potential values that will be dynamically added after the first values, depending on what each entry has in its table

This section may be useful to describe the different types of data that were collected. For example in a study that used passive stationary samplers, silicone wristband samplers, survey data, and PM2.5 monitors, the data may need to be described in terms of how many of each sampler were deployed, the analytes they monitor, the survey platform used, etc. This is also a good place to identify any outliers in the data, or influential data points. In some cases, a table may be helpful to visualize the different data types, data units, and more.

Here, you should also note any issues that arose with the data (e.g., incorrect units were reported, there were issues with how samples were numbered, etc.)

If additional existing data is being integrated, for example data from the NOAA Hazard Mapping System, this should also be described.

For a literature review, include the search keywords, instructions or code for API requests or database queries, and search criteria. Consider using the PRISMA checklist and flowchart here (<https://www.prisma-statement.org/>).

### SOFTWARE EXPLANATION

As appropriate, describe the software that is necessary for data processing and/or data analysis. For example, ArcGIS, R Studio, Qualtrics, etc. Include the version number and any other information that would be necessary for publication (e.g., City, State).

#### General Explanation

This program is a python source file (ending with .py). It is coded in the python programming language. It differs from traditional .exe (executable) files as it needs to be ran on the windows or Linux command line through a command written by the user, instead of just being clicked and opened from the windows file explorer.

The entire program logic is written and executed in the .py source file, but it does need an additional file to be able to execute. This additional file is a .xlsx (Microsoft excel) file that contains the information acquired through an advanced search of NPL sites on the epa website.

There are two different versions of this program, one that uses the .xlsx file exclusively for its information, and another that can use either only a website version of the information, both the website and .xlsx file.

No matter what version is chosen, there will be the same type of output. The program outputs a .csv file (a general table file similar to excel or gsheets). This file contains a table with a series of headings detailed in the “Data Processing and Analysis” section of this document, and an entry for each Superfund site on that table. There are minor changes in the headers depending on which version is chosen.

#### Explanation of versions:

#### XLSX and site version:

This version takes information from the xlsx file and provided epa website page that has a table of superfund sites. What the program is looking for in a site is one with the same layout of information as the one in this link: <https://www.epa.gov/superfund/national-priorities-list-npl-sites-state>

What specifically it is scanning for is the dynamic table that is loaded on the page. It is pulling the state names, amount of sites, site names, cities, epa id, score, and site progress profile. Even if there are other tables that can provide that data, the program is hard coded to only be able to read tables with the exact order and format of all the data and headers as the one in the provided link, even if the certain headers or data is not being read by the program, as it relies on not what the data is, but where it is on the table. For example, the program is looking for the city in the second column of the table, if something else was there instead it would still store the data as the city. A very specific thing it is also looking for is the linked Site Progress Profile being the second bullet point on the 7th column of the table.

This version can completely rely on just the provided website link, but it can’t pull the NAI or NAI Entity from the website, so to be able to utilize that is must have a provided xlsx file with the same sites as the provided website to reference, so that it can find and pull the NAI and Entity data from the relative sites.

#### Xlsx only version:

This version only uses the xlsx file provided for information. This version outputs the NAI and NAI Entity, as well as an additional column labeled “NPL Status” that details whether the NPL site is deleted, ongoing, etc. This column is not in the site and xlsx version because that status is usually given through the provided website.

Unlike the site version, the xlsx only version is searching for the headers by name and not location (this is also true for the NAI and NAI Entity, if the xlsx is provided, in the site version), therefore the xlsx file format could be varied, as long as the correct headers exist, which are the following (case sensitive):

* “EPA ID”
* “Native American Interest (NAI)”
* “Indian Entity (NAI Status)”
* “Site Name”
* “City”
* “HRS Score”
* “State”
* “Superfund Site Profile Page URL”
* “NPL Status”

Keep in mind that just because this version only utilizes xlsx file does not mean it is an “offline” version. It still needs internet access and a version of chrome installed to be able to access the link and scrape the provided superfund site profiles given in the xlsx file.

#### How to download the xlsx file:

To download the xlsx file the user must go to the advanced search page for superfund site information using this link: <https://cumulis.epa.gov/supercpad/CurSites/srchsites.cfm>

The user can then select whatever advanced criteria they want, then click search at the bottom of the page

Keep in mind when using the xlsx and site version, there could be mixed results with the if the criteria in the .xlsx doesn’t match the criteria of the given page.

After clicking search, there will be a link on the page labeled “Download Excel file containing values for all search criteria”. Clicking this link will download the .xlsx file.

#### How To Download and Run Program:

1. The machine must have Python 3 installed, please go to the python website to download and install the latest version
2. Go to the github site for the program, provided using this link: <https://github.com/bartonmike/superfund-npl-scraper.git>
3. Click on the “carrot” on the green button labeled “code”, a dropdown will appear
4. A screenshot of a computer

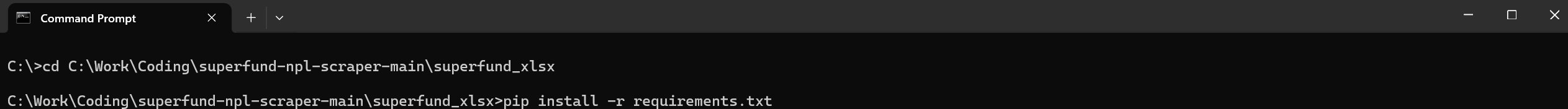
   Description automatically generatedA screen shot of a computer

   Description automatically generatedClick “Download ZIP”, a zip file will then download.
5. Put the ZIP in the desired folder on your computer, then extract it, Windows 11 has unzipping built in, so if you are using Windows 11 extracting is as easy as right clicking on the file in the file explorer and clicking “extract all”
6. Double click into the file that was created by the extraction, labeled “superfund-npl-scraper-main”, then double click into the folder of the desired version that you would want to run, either “superfund\_site\_and\_xlsx” or “superfund\_xlsx”.
7. Click the address bar on the top of the file explorer, then copy the contents
8. Press the windows key, then search for “Command prompt”, press enter, a command prompt window should open

A black screen with a black background

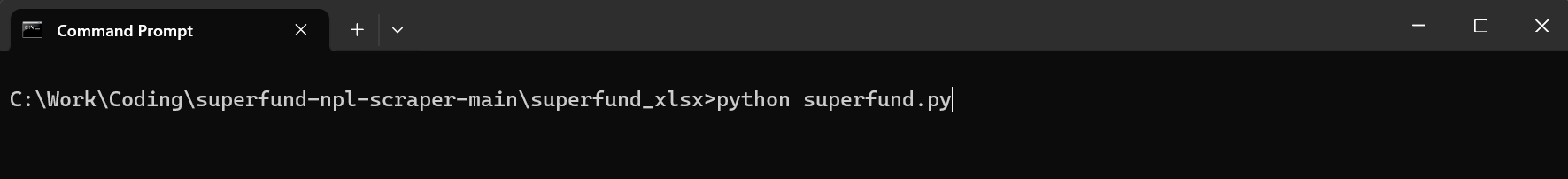
Description automatically generatedA screenshot of a computer program

Description automatically generated

1. Type “cd”, space, then paste and press enter (if ctrl+v does not work in command prompt, paste using right click). The Command prompt should then display the address on the left, indicating that command prompt has now moved to the location of the program and knows where it is so that you can run it, if there is an error, type “cd ..” + enter a few times and try again.
2. A screenshot of a computer

   Description automatically generatedA screenshot of a computer

   Description automatically generatedThe Program needs to download some libraries and additional content to run from the python database, to install that content you can use the provided requirements.txt file, just type the command “pip install -r requirements.txt” and it will install all the content needed. This command needs to be run only once per machine, but each version of the program has slightly different requirements, so running one install doesn’t equate to running both, switching to the other version would require running the install command if not done so already.
3. After installing everything, we can finally put in the command to run the program “python superfund.py”



Example of how the program starts:

A black screen with white text

Description automatically generated

There may be times when the program encounters an error, and when it does it will stop and write what it has done so far and then exit.

The method that this program uses to process the data is pretty simple. It systematically goes through all the superfund sites listed on either the website or xlsx (depending on the version), looks for the data mentioned in “Data Processing and Analysis” section of this document and then saves it, it then adds the superfund site with its data to an internal array (or “list”) of sites.

Once the program has gone through all the sites, it should then have a an internal “list” of every site, with every site having its own data. It then takes that list and systematically writes it to a csv file. Each row on the csv file is a unique site, and each column is the category of data that the program was looking for in the sites. The final result should be a csv filled with individual site entries and the data associated with them organized as a table of values.

### DATA PROCESSING

Describe how samples were processed. You may need sub-sections here to describe processing steps for different data types, e.g., processing silicone wristband data versus survey data. An example introductory section may include the following:

All samples were analyzed by the Food Science and Environmental Stewardship laboratory directed by Dr. Kim Anderson, at Oregon State University. Each sample was analyzed for 64 PAHs using a targeted analyte method. Raw data was received in an Excel document. Prior to analysis, the data was processed under the following conditions:

1. Complete background corrections as needed (wristband dataset only)
2. Filter out analytes that were detected in less than 25% of the dataset
3. Replace below-limit-of-detection values with BLOD/2

### DATA ANALYSIS

Describe how the data was analyzed. This may include how data were stratified, the statistical analysis’ conducted, etc. Sub-sections may be needed to describe how different data types were analyzed.

## PUBLICATIONS & PRESS

In this optional section, include press releases relevant to the research and/or study location.

## FILE NAVIGATION FLOW CHART

For projects with multiple documents, a simple flow chart may be helpful to identify where files are, and how they are organized in a folder or folders.

