EPA IRIS Assessment Database Web Scrape for Use in Risk Assessment Fact Sheets

Ben Hildebrand, PhD

August 16, 2021

Abstract

Contents

List of Figures

1 Introduction

Web scraping is the process of extracting data from a website. These days, this process of web scraping is often done via the aid of a programming/scripting language. In particular, this work uses Python with the popular package BeautifulSoup to perform a web scrape of the Environmental Protection Agency's (EPA) online database of Integrated Risk Information System (IRIS) assessments.

At the request of Dr. Margaret MacDonell of Argonne National Laboratory (ANL), I was asked to perform a web scrape of the EPA's online, publicly-available dataset of chemicals that have IRIS assessments for the purpose of updating risk assessment fact sheets.

Motivation for the web scrape request is due to each IRIS assessment chemical containing a link to its respective information. Unfortunately, the fact that each chemical in the IRIS assessment database has an individual link makes manually accessing each chemical's key information inconvenient due to load times and page layout. To improve convenience and speed of accessing the IRIS assessment data, the web scrape performed in this work aims to agglomerate all of the key data for each chemical into one, single table to be viewed as a spreadsheet via MS Excel or similar software.

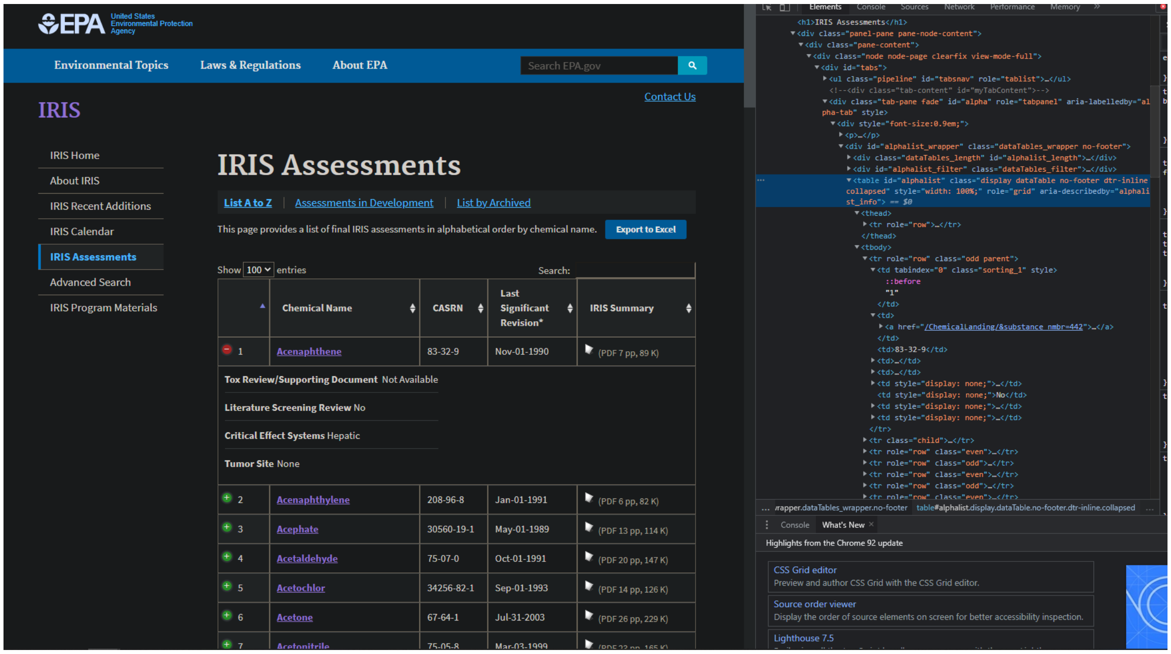
Before jumping in and scraping data from the webpage(s), a preliminary examination of what an IRIS assessment involves helps identify what important information is desired.

Part of the EPA's mission is to "protect human health and the environment" [1]. The EPA's IRIS Program aims to aid in achieving this goal by identifying and characterizing possible health hazards of chemicals found within the environment. IRIS assessments are performed for individual chemicals, a group of related chemicals, or a complex mixture of chemicals, and contain toxicity information that is used by many domestic and world-wide health agencies and organizations. IRIS assessments include the following toxicity values and information (see Basic Information about the Integrated Risk Information System [1] for definitions): Reference Concentration (RfC), Reference Dose (RfD), cancer descriptors, Oral Slope Factor (OSF), and Inhalation Unit Risk (IUR). Therefore, the main goal of this work is to collect all toxicity and assessment information for each chemical assessed in the IRIS assessments database into a single spreadsheet with chemical names listed as rows and toxicity information listed as columns. Sometimes there exist multiple toxicity values and/or descriptors for certain chemicals; this is reflected in the resulting tables by repeating the chemical name along with the respective toxicity information collected (e.g. some chemicals in the IRIS assessment database have multiple RfC and/or RfD values and are therefore listed over multiple rows with their differing values to reflect this fact).

2 Methodology

The Python programming language has a very useful package for web scraping called BeautifulSoup whose documentation is found here: BeautifulSoup Documentation [2].

To make use of the BeautifulSoup Python package, we first examine the EPA's IRIS assessment HTML page layout as well as the HTML page layout of some individual chemicals in order to get a sense of what relevant patterns and HTML tags exist in the html page contents [3] (see Figure 1). The IRIS Assessments main page contains a table listing the chemicals in the EPA's IRIS assessments database. Each chemical name contains a hyperlink to its respective assessment page (see Figure 2).



IRIS Assessments

This page provides a list of final IRIS assessments in alphabetical order by chemical name.

	Chemical Name	CASRN	Last Significant Revision*	IRIS Summary
1	Acenaphthene	83-32-9	Nov-01-1990	(PDF 7 pp, 89 K)
Tox Review/Supporting Document Not Available Literature Screening Review No Critical Effect Systems Hepatic Tumor Site None				
2	Acenaphthylene	208-96-8	Jan-01-1991	(PDF 6 pp, 82 K)
3	Acephate	30560-19-1	May-01-1989	(PDF 13 pp, 114 K)
4	Acetaldehyde	75-07-0	Oct-01-1991	(PDF 20 pp, 147 K)
5	Acetochlor	34256-82-1	Sep-01-1993	(PDF 14 pp, 126 K)
6	Acetone	67-64-1	Jul-31-2003	(PDF 26 pp, 229 K)
7	Acetonitrile	75-05-8	Mar-03-1999	(PDF 23 pp, 165 K)

```
<h1>IRIS Assessments</h1>
▼<div class="panel-pane pane-node-content">
  ▼<div class="pane-content">
    ▼<div class="node node-page clearfix view-mode-full">
      ▼<div id="tabs">
        ▶<ul class="pipeline" id="tabsnav" role="tablist">…</ul>
         <!--<div class="tab-content" id="myTabContent">-->
        ▼<div class="tab-pane fade" id="alpha" role="tabpanel" aria-labelledby="alpha-tab" style>
          ▼<div style="font-size:0.9em;">
            ▶<p>…</p>
            ▼<div id="alphalist_wrapper" class="dataTables_wrapper no-footer">
              ▶<div class="dataTables_length" id="alphalist_length">…</div>
              ▶<div id="alphalist_filter" class="dataTables_filter">…</div>
...           ▼<table id="alphalist" class="display dataTable no-footer dtr-inline collapsed" style="width: 100%;" role="grid" aria-describedby="alphalist_info"> == $0
                ▼<thead>
                  ▶<tr role="row">…</tr>
                  </thead>
                ▼<tbody>
                  ▼<tr role="row" class="odd parent">
                    ▼<td tabindex="0" class="sorting_1" style>
                        ::before
                        "1"
                      </td>
                    ▼<td>
                      ▶<a href="/ChemicalLanding/&substance_nmbr=442">…</a>
                      </td>
                      <td>83-32-9</td>
                    ▶<td>…</td>
                    ▶<td>…</td>
                    ▶<td style="display: none;">…</td>
                      <td style="display: none;">No</td>
                    ▶<td style="display: none;">…</td>
                    ▶<td style="display: none;">…</td>
                    </tr>
                  ▶<tr class="child">…</tr>
                  ▶<tr role="row" class="even">…</tr>
                  ▶<tr role="row" class="odd">…</tr>
                  ▶<tr role="row" class="even">…</tr>
                  ▶<tr role="row" class="odd">…</tr>
                  ▶<tr role="row" class="even">…</tr>
… wrapper.dataTables_wrapper.no-footer   table#alphalist.display.dataTable.no-footer.dtr-inline.collapsed …
```

Highlights from the Chrome 92 update

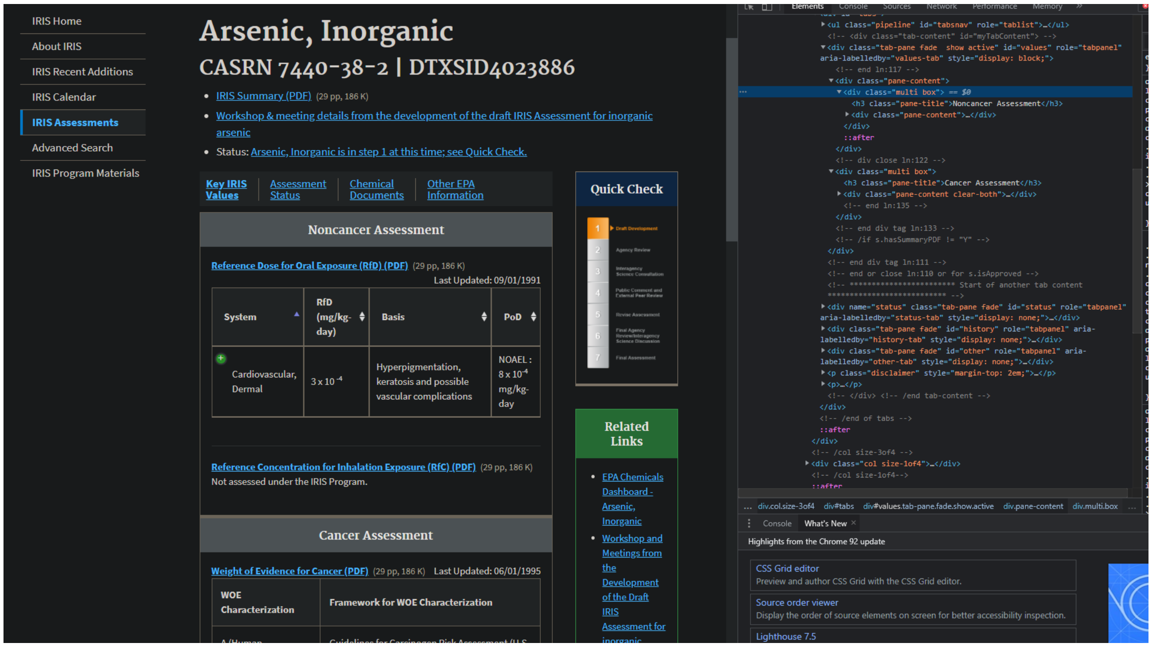
CSS Grid editor
Preview and author CSS Grid with the CSS Grid editor.

Source order viewer
Display the order of source elements on screen for better accessibility inspection.

Lighthouse 7.5

Figure 1:
IRIS Assessments main page with sample HTML layout.

Since the IRIS Assessments main page contains a table with specified ID, a scrape of the data from the table is quite standard via Beautiful-



Arsenic, Inorganic

CASRN 7440-38-2 | DTXSID4023886

- IRIS Summary (PDF) (29 pp, 186 K)
- Workshop & meeting details from the development of the draft IRIS Assessment for inorganic arsenic
- Status: Arsenic, Inorganic is in step 1 at this time; see Quick Check.

Key IRIS Values | Assessment Status | Chemical Documents | Other EPA Information

Noncancer Assessment

Reference Dose for Oral Exposure (RfD) (PDF) (29 pp, 186 K)

Last Updated: 09/01/1991

System	RfD (mg/kg-day)	Basis	PoD
Cardiovascular, Dermal	3×10^{-4}	Hyperpigmentation, keratosis and possible vascular complications	NOAEL : 8×10^{-4} mg/kg-day

Reference Concentration for Inhalation Exposure (RfC) (PDF) (29 pp, 186 K)

Not assessed under the IRIS Program.

Cancer Assessment

Weight of Evidence for Cancer (PDF) (29 pp, 186 K) Last Updated: 06/01/1995

WOE Characterization	Framework for WOE Characterization
A (Human	Guidelines for Carcinogen Risk Assessment (U.S.

Related Links

- EPA Chemicals Dashboard - Arsenic, Inorganic
- Workshop and Meetings from the Development of the Draft IRIS Assessment for inorganic

```
▶<ul class="pipeline" id="tabsnav" role="tablist">…</ul>
 <!-- <div class="tab-content" id="myTabContent"> -->
▼<div class="tab-pane fade  show active" id="values" role="tabpanel"
aria-labelledby="values-tab" style="display: block;">
   <!-- end ln:117 -->
  ▼<div class="pane-content">
...  ▼<div class="multi box"> == $0
       <h3 class="pane-title">Noncancer Assessment</h3>
      ▶<div class="pane-content">…</div>
     </div>
     ::after
   </div>
   <!-- div close ln:122 -->
  ▼<div class="multi box">
     <h3 class="pane-title">Cancer Assessment</h3>
    ▶<div class="pane-content clear-both">…</div>
     <!-- end ln:135 -->
   </div>
   <!-- end div tag ln:133 -->
   <!-- /if s.hasSummaryPDF != "Y" -->
 </div>
 <!-- end div tag ln:111 -->
 <!-- end or close ln:110 or for s.isApproved -->
 <!-- *********************** Start of another tab content
 ************************* -->
▶<div name="status" class="tab-pane fade" id="status" role="tabpanel"
aria-labelledby="status-tab" style="display: none;">…</div>
▶<div class="tab-pane fade" id="history" role="tabpanel" aria-
labelledby="history-tab" style="display: none;">…</div>
▶<div class="tab-pane fade" id="other" role="tabpanel" aria-
labelledby="other-tab" style="display: none;">…</div>
▶<p class="disclaimer" style="margin-top: 2em;">…</p>
▶<p>…</p>
 <!-- </div> <!-- /end tab-content -->
</div>
<!-- /end of tabs -->
::after
</div>
<!-- /col size-3of4 -->
▶<div class="col size-1of4">…</div>
<!-- /col size-1of4-->
::after
```

… div.col.size-3of4 div#tabs div#values.tab-pane.fade.show.active div.pane-content div.multi.box …

Highlights from the Chrome 92 update

CSS Grid editor
Preview and author CSS Grid with the CSS Grid editor.

Source order viewer
Display the order of source elements on screen for better accessibility inspection.

Lighthouse 7.5

Figure 2:
Arsenic, Inorganic assessment page accessed through Arsenic, Inorganic's link with sample HTML layout.

Soup. What is not-so-standard (at least in the experience of the author) is the scraping of the additional toxicity data that can be accessed through each link to each chemical's assessment due to the HTML structure of each page. Therefore, the general process is as follows. Each chemical assessment page contains two main sections: "Noncancer Assessment" and "Cancer Assessment" (Note: some chemicals have no assessments of one or both sections, but the Python script accounts for this). To breakdown the process of obtaining all relevant data, the web scrape is performed with the following procedure: (1) Scrape the main table of the IRIS Assessments main page (see Figure 1). (2) Access the link to each chemical assessment and scrape the data from the "Noncancer Assessment" section (see Figure 2). (3) Access the link to each chemical assessment and scrape the data from the "Cancer Assessment" section (see Figure 2). Note: The "Cancer Assessment" scrape is achieved in a two-stage web scrape: one scrape for Weight of Evidence (WOE) data, and one scrape for oral and inhalation data. (4) Merge each table obtained from each of the web scrapes mentioned in (1), (2), and (3). (5) Clean the data by columns for readability in MS Excel (or similar software). (6) Finally, drop all identical, duplicate rows (if any).

The details of how each of the above steps are carried out is not the subject of this work, but, for those who are interested, and for purposes of being thorough, the web scrape and data cleaning scripts are both available as two separate Jupyter notebooks which are available on the author's GitHub here: EPA IRIS Assessments Web Scrape. The Python script that performs the raw web scrape can be accessed directly here, and the Python script that performs the data cleaning (after running the web scrape) can be accessed directly here. Both Python scripts store the resulting data tables as CSV files (both CSV files are also on the author's GitHub).

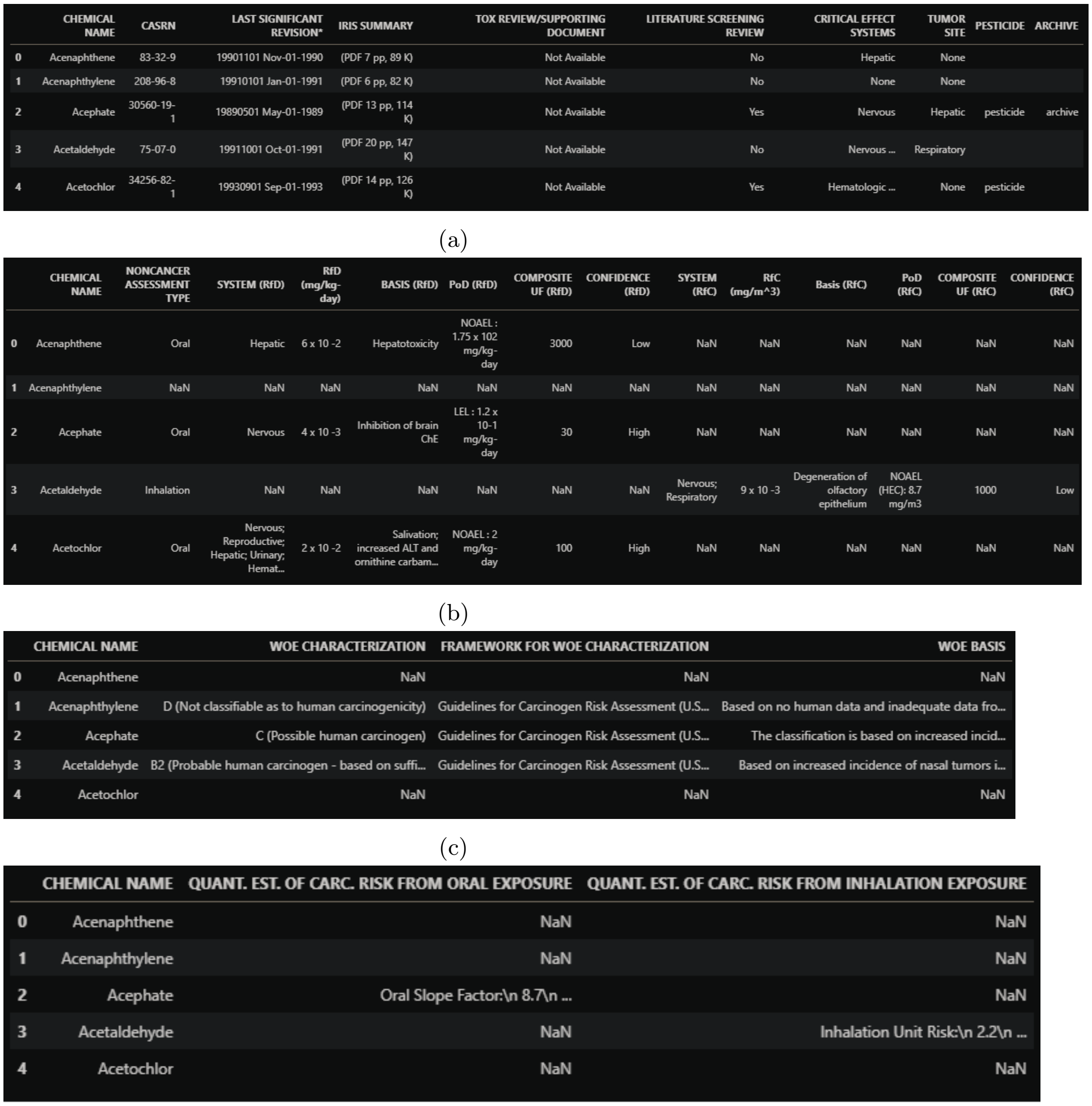
3 Results

The final table obtained from web scraping and merging contains 693 rows with 571 unique chemical names along with 28 columns. As was mentioned earlier, repeated rows result when a single chemical has additional assessment information for the same column(s) such as multiple RfC and/or RfD values. This means that some chemical names are repeated with different column information for at least *one* column. To view and/or download the entire data table see the author's GitHub: EPA IRIS Assessments clean data table.

See Figure 3 for snapshots of the resulting tables for each web scrape performed: initial web scrape of IRIS Assessments main table (3a); non-cancer assessment data (3b); cancer assessment data for WOE characterization, framework for WOE characterization, and basis (3c); and cancer assessments regarding oral and inhalation data (3d).

See Figure 4 for a snapshot of the table that results from performing each scrape, and then merging them all into a single table. Since the table has a total of 28 columns, it is too long to capture in a single screenshot. See Figure 4a for the first five rows and first 16 columns of the table and Figure 4b for the first five rows and last 12 columns of the table.

Lastly, see Figure 5 for the table that results after the data cleaning process (cleaning the data in the table from Figure 4). Again, the table is too long for a single screenshot. See Figure 5a for the first five rows and first 16 columns of the cleaned data table and Figure 5b for the first five rows and last 12 columns of the cleaned data table.



	CHEMICAL NAME	CASRN	LAST SIGNIFICANT REVISION*	IRIS SUMMARY	TOX REVIEW/SUPPORTING DOCUMENT	LITERATURE SCREENING REVIEW	CRITICAL EFFECT SYSTEMS	TUMOR SITE	PESTICIDE	ARCHIVE
0	Acenaphthene	83-32-9	19901101 Nov-01-1990	(PDF 7 pp, 89 K)	Not Available	No	Hepatic	None		
1	Acenaphthylene	208-96-8	19910101 Jan-01-1991	(PDF 6 pp, 82 K)	Not Available	No	None	None		
2	Acephate	30560-19-1	19890501 May-01-1989	(PDF 13 pp, 114 K)	Not Available	Yes	Nervous	Hepatic	pesticide	archive
3	Acetaldehyde	75-07-0	19911001 Oct-01-1991	(PDF 20 pp, 147 K)	Not Available	No	Nervous ...	Respiratory		
4	Acetochlor	34256-82-1	19930901 Sep-01-1993	(PDF 14 pp, 126 K)	Not Available	Yes	Hematologic ...	None	pesticide	

(a)

	CHEMICAL NAME	NONCANCER ASSESSMENT TYPE	SYSTEM (RfD)	RfD (mg/kg-day)	BASIS (RfD)	PoD (RfD)	COMPOSITE UF (RfD)	CONFIDENCE (RfD)	SYSTEM (RfC)	RfC (mg/m^3)	Basis (RfC)	PoD (RfC)	COMPOSITE UF (RfC)	CONFIDENCE (RfC)
0	Acenaphthene	Oral	Hepatic	6 x 10 -2	Hepatotoxicity	NOAEL : 1.75 x 102 mg/kg-day	3000	Low	NaN	NaN	NaN	NaN	NaN	NaN
1	Acenaphthylene	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Acephate	Oral	Nervous	4 x 10 -3	Inhibition of brain ChE	LEL : 1.2 x 10-1 mg/kg-day	30	High	NaN	NaN	NaN	NaN	NaN	NaN
3	Acetaldehyde	Inhalation	NaN	NaN	NaN	NaN	NaN	NaN	Nervous; Respiratory	9 x 10 -3	Degeneration of olfactory epithelium	NOAEL (HEC): 8.7 mg/m3	1000	Low
4	Acetochlor	Oral	Nervous; Reproductive; Hepatic; Urinary; Hemat...	2 x 10 -2	Salivation; increased ALT and ornithine carbam...	NOAEL : 2 mg/kg-day	100	High	NaN	NaN	NaN	NaN	NaN	NaN

(b)

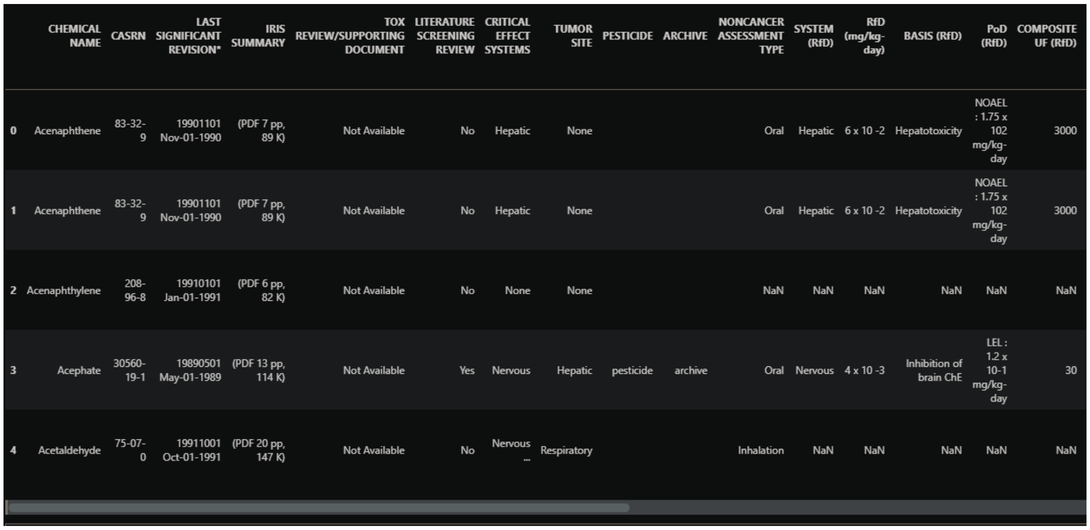
	CHEMICAL NAME	WOE CHARACTERIZATION	FRAMEWORK FOR WOE CHARACTERIZATION	WOE BASIS
0	Acenaphthene	NaN	NaN	NaN
1	Acenaphthylene	D (Not classifiable as to human carcinogenicity)	Guidelines for Carcinogen Risk Assessment (U.S...	Based on no human data and inadequate data fro...
2	Acephate	C (Possible human carcinogen)	Guidelines for Carcinogen Risk Assessment (U.S...	The classification is based on increased incid...
3	Acetaldehyde	B2 (Probable human carcinogen - based on suffi...	Guidelines for Carcinogen Risk Assessment (U.S...	Based on increased incidence of nasal tumors i...
4	Acetochlor	NaN	NaN	NaN

(c)

	CHEMICAL NAME	QUANT. EST. OF CARC. RISK FROM ORAL EXPOSURE	QUANT. EST. OF CARC. RISK FROM INHALATION EXPOSURE
0	Acenaphthene	NaN	NaN
1	Acenaphthylene	NaN	NaN
2	Acephate	Oral Slope Factor:\n 8.7\n ...	NaN
3	Acetaldehyde	NaN	Inhalation Unit Risk:\n 2.2\n ...
4	Acetochlor	NaN	NaN

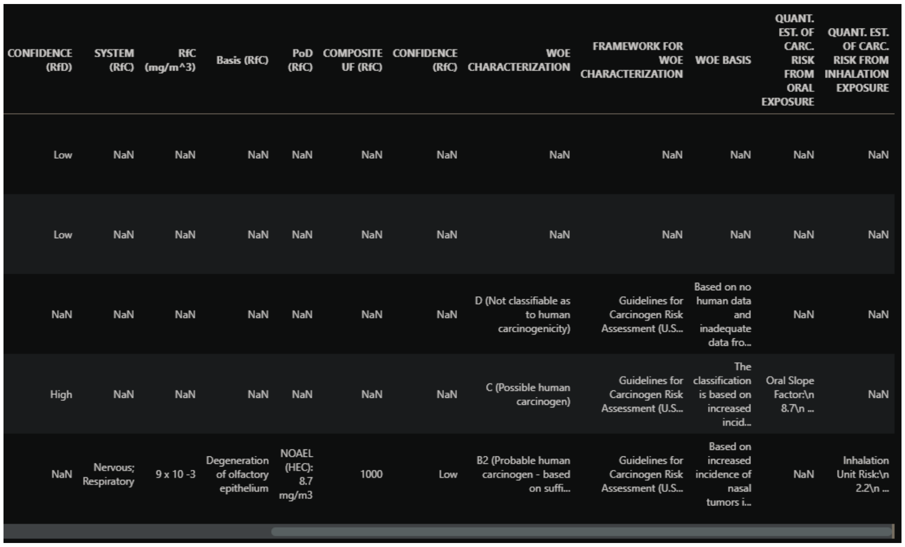
(d)

Figure 3: Snapshots of the resulting data tables from each web scrape described in the Methodology section. Each data table shows the first five rows with *all* of its columns. 3a: Initial web scrape of IRIS Assessments main page; 3b: Web scrape of "Noncancer Assessments" section of each chemical assessment; 3c: Web scrape of "Cancer Assessment" section of WOE table and basis for WOE characterization; 3d: Web scrape of "Cancer Assessment" section of oral and inhalation assessment data.



	CHEMICAL NAME	CASRN	LAST SIGNIFICANT REVISION*	IRIS SUMMARY	TOX REVIEW/SUPPORTING DOCUMENT	LITERATURE SCREENING REVIEW	CRITICAL EFFECT SYSTEMS	TUMOR SITE	PESTICIDE	ARCHIVE	NONCANCER ASSESSMENT TYPE	SYSTEM (RfD)	RfD (mg/kg-day)	BASIS (RfD)	PoD (RfD)	COMPOSITE UF (RfD)
0	Acenaphthene	83-32-9	19901101 Nov-01-1990	(PDF 7 pp, 89 K)	Not Available	No	Hepatic	None			Oral	Hepatic	6 x 10 -2	Hepatotoxicity	NOAEL : 1.75 x 102 mg/kg-day	3000
1	Acenaphthene	83-32-9	19901101 Nov-01-1990	(PDF 7 pp, 89 K)	Not Available	No	Hepatic	None			Oral	Hepatic	6 x 10 -2	Hepatotoxicity	NOAEL : 1.75 x 102 mg/kg-day	3000
2	Acenaphthylene	208-96-8	19910101 Jan-01-1991	(PDF 6 pp, 82 K)	Not Available	No	None	None			NaN	NaN	NaN	NaN	NaN	NaN
3	Acephate	30560-19-1	19890501 May-01-1989	(PDF 13 pp, 114 K)	Not Available	Yes	Nervous	Hepatic	pesticide	archive	Oral	Nervous	4 x 10 -3	Inhibition of brain ChE	LEL : 1.2 x 10-1 mg/kg-day	30
4	Acetaldehyde	75-07-0	19911001 Oct-01-1991	(PDF 20 pp, 147 K)	Not Available	No	Nervous ...	Respiratory			Inhalation	NaN	NaN	NaN	NaN	NaN

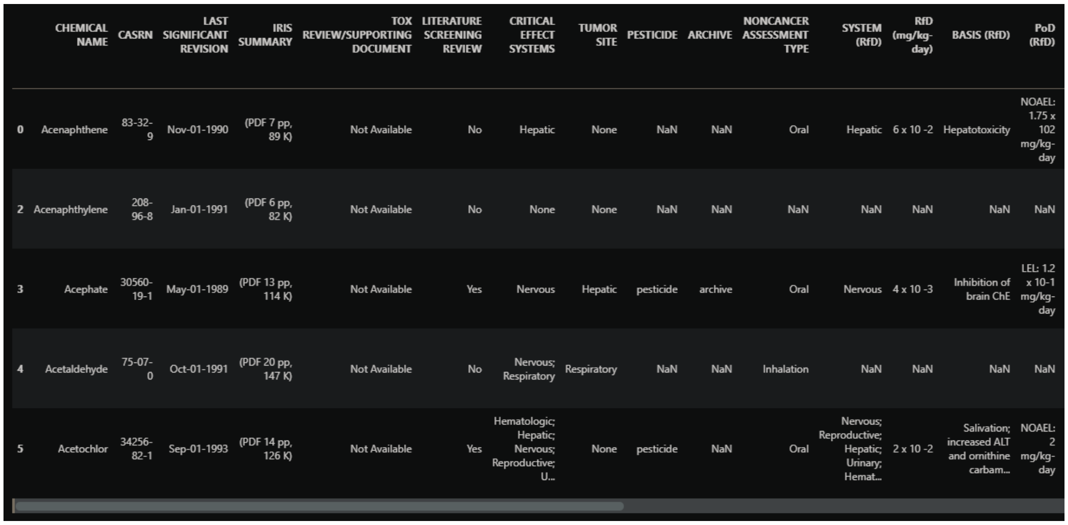
(a)



CONFIDENCE (RfD)	SYSTEM (RfC)	RfC (mg/m^3)	Basis (RfC)	PoD (RfC)	COMPOSITE UF (RfC)	CONFIDENCE (RfC)	WOE CHARACTERIZATION	FRAMEWORK FOR WOE CHARACTERIZATION	WOE BASIS	QUANT. EST. OF CARC. RISK FROM ORAL EXPOSURE	QUANT. EST. OF CARC. RISK FROM INHALATION EXPOSURE
Low	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Low	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	D (Not classifiable as to human carcinogenicity)	Guidelines for Carcinogen Risk Assessment (U.S...	Based on no human data and inadequate data fro...	NaN	NaN
High	NaN	NaN	NaN	NaN	NaN	NaN	C (Possible human carcinogen)	Guidelines for Carcinogen Risk Assessment (U.S...	The classification is based on increased incid...	Oral Slope Factor:\n 8.7\n ...	NaN
NaN	Nervous; Respiratory	9 x 10 -3	Degeneration of olfactory epithelium	NOAEL (HEC): 8.7 mg/m3	1000	Low	B2 (Probable human carcinogen - based on suffi...	Guidelines for Carcinogen Risk Assessment (U.S...	Based on increased incidence of nasal tumors i...	NaN	Inhalation Unit Risk:\n 2.2\n ...

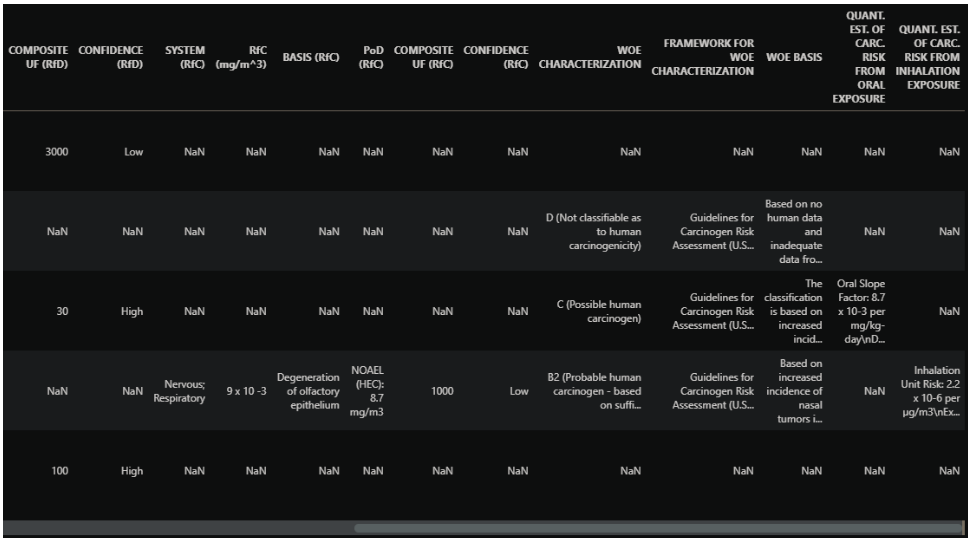
(b)

Figure 4: Resulting table from merging each of the tables from the individual web scrapes mentioned in the Methodology section without any data cleaning (raw web scrape, merged into one table). The table spans a total of 28 columns. 4a: First five rows and first 16 columns of resulting table. 4b: First five rows and last 12 columns of resulting table.



	CHEMICAL NAME	CASRN	LAST SIGNIFICANT REVISION	IRIS SUMMARY	TOX REVIEW/SUPPORTING DOCUMENT	LITERATURE SCREENING REVIEW	CRITICAL EFFECT SYSTEMS	TUMOR SITE	PESTICIDE	ARCHIVE	NONCANCER ASSESSMENT TYPE	SYSTEM (RfD)	RfD (mg/kg-day)	BASIS (RfD)	PoD (RfD)
0	Acenaphthene	83-32-9	Nov-01-1990	(PDF 7 pp, 89 K)	Not Available	No	Hepatic	None	NaN	NaN	Oral	Hepatic	6 x 10 -2	Hepatotoxicity	NOAEL: 1.75 x 102 mg/kg-day
2	Acenaphthylene	208-96-8	Jan-01-1991	(PDF 6 pp, 82 K)	Not Available	No	None	None	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Acephate	30560-19-1	May-01-1989	(PDF 13 pp, 114 K)	Not Available	Yes	Nervous	Hepatic	pesticide	archive	Oral	Nervous	4 x 10 -3	Inhibition of brain ChE	LEL: 1.2 x 10-1 mg/kg-day
4	Acetaldehyde	75-07-0	Oct-01-1991	(PDF 20 pp, 147 K)	Not Available	No	Nervous; Respiratory	Respiratory	NaN	NaN	Inhalation	NaN	NaN	NaN	NaN
5	Acetochlor	34256-82-1	Sep-01-1993	(PDF 14 pp, 126 K)	Not Available	Yes	Hematologic; Hepatic; Nervous; Reproductive; U...	None	pesticide	NaN	Oral	Nervous; Reproductive; Hepatic; Urinary; Hemat...	2 x 10 -2	Salivation; increased ALT and ornithine carbam...	NOAEL: 2 mg/kg-day

(a)



COMPOSITE UF (RfD)	CONFIDENCE (RfD)	SYSTEM (RfC)	RfC (mg/m^3)	BASIS (RfC)	PoD (RfC)	COMPOSITE UF (RfC)	CONFIDENCE (RfC)	WOE CHARACTERIZATION	FRAMEWORK FOR WOE CHARACTERIZATION	WOE BASIS	QUANT. EST. OF CARC. RISK FROM ORAL EXPOSURE	QUANT. EST. OF CARC. RISK FROM INHALATION EXPOSURE
3000	Low	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	D (Not classifiable as to human carcinogenicity)	Guidelines for Carcinogen Risk Assessment (U.S...	Based on no human data and inadequate data fro...	NaN	NaN
30	High	NaN	NaN	NaN	NaN	NaN	NaN	C (Possible human carcinogen)	Guidelines for Carcinogen Risk Assessment (U.S...	The classification is based on increased incid...	Oral Slope Factor: 8.7 x 10-3 per mg/kg-day\nD...	NaN
NaN	NaN	Nervous; Respiratory	9 x 10 -3	Degeneration of olfactory epithelium	NOAEL (HEC): 8.7 mg/m3	1000	Low	B2 (Probable human carcinogen - based on suffi...	Guidelines for Carcinogen Risk Assessment (U.S...	Based on increased incidence of nasal tumors i...	NaN	Inhalation Unit Risk: 2.2 x 10-6 per µg/m3\nEx...
100	High	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

(b)

Figure 5: Resulting table after cleaning the data from the merged table from Figure 4. The table again spans a total of 28 columns. 4a: First five rows and first 16 columns of resulting table. 4b: First five rows and last 12 columns of resulting table.

4 Discussion

The main item to discuss is data verification. The immediate question that comes to mind is: is the final data table obtained from the web scrape and merge correct?

To answer this question, we first note that the IRIS Assessments chemical database consists of 571 unique chemicals which can be quickly verified using the numbering at the left of the table on the IRIS Assessments webpage [3]. Checking the number of *unique* chemicals in the final resulting data table saved to the CSV also contains 571 *unique* chemicals. Therefore, it is reasonable to assume that each chemical from the original source appears at least one time. Furthermore, checking several rows of the data table with several chemical assessments from the original source have matching values. In particular, there are matching values for the first five rows of the data table; Arsenic, Inorganic also checks out; as well as Xylenes and Zinc and Compounds, both of which contain multiple rows due to multiple, differing assessment information.

5 Conclusion

Although the arguments regarding data correctness in the Discussions section does not provide a complete and thorough verification the data scraped and merged, it is reasonable to conclude that the final data table obtained contains all the necessary information from the IRIS Assessments database. Of course, for full-proof verification one would need to go through each and every chemical assessment and verify the data matches that which was collected. But naturally, this was precisely the process we were trying to avoid in the first place.

6 References

[1] U.S. Environmental Protection Agency. Basic information about the integrated risk information system. https://www.epa.gov/iris/basic-information-about-integrated-risk-information-system. Accessed: 2021-08-16.

[2] Leonard Richardson. Beautiful soup documentation - beautiful soup 4.9.0 documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/. Accessed: 2021=08-16.

[3] U.S. Environmental Protection Agency. Iris assessments. https://iris.epa.gov/AtoZ/?list_type=alpha. Accessed: 2021-08-16.