Abstract Sifter user guide

Nancy C. Baker, Thomas Knudsen, Antony Williams

Availability: The Abstract Sifter and documentation is freely available for download at ftp://newftp.epa.gov/COMPTOX/Sustainable_Chemistry_Data/Chemistry_Dashboard/Abstract_Sifter/ .

Contact: baker.nancy@epa.gov

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

Abstract Sifter user guide

This user guide describes the functionality of the PubMed Abstract Sifter version 1.0. The reader is invited to download the tool from the freely accessible ftp site and follow along:

ftp://newftp.epa.gov/COMPTOX/Sustainable_Chemistry_Data/Chemistry_Dashboard/Abstract_Sifter/

This document will illustrate the use of the tool through a series of screen shots walking through a few simple tasks. First open the Abstract Sifter file AbstractSifter_v1.xlsm. A security warning may appear. If so, be sure to enable content as shown in Figure 1.

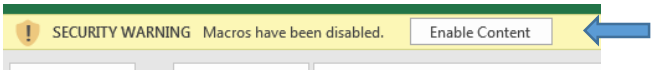


Figure 1. Enable macros upon opening

The Abstract Sifter Excel file consists of seven sheets. We'll start this documentation on the Main sheet. The Main sheet is where the basic functionality takes place, including functions we call "sifting". To use the Abstract Sifter, the end-user clicks on the *Query PubMed* button at the top of the screen. A form is displayed in which the end-user types a PubMed query of interest (Figure 2). In the example, we are showing a very simple query: "chlorpyrifos", but these queries can be more complex. The text that the user enters into the box is sent to PubMed, so all PubMed syntax rules apply.

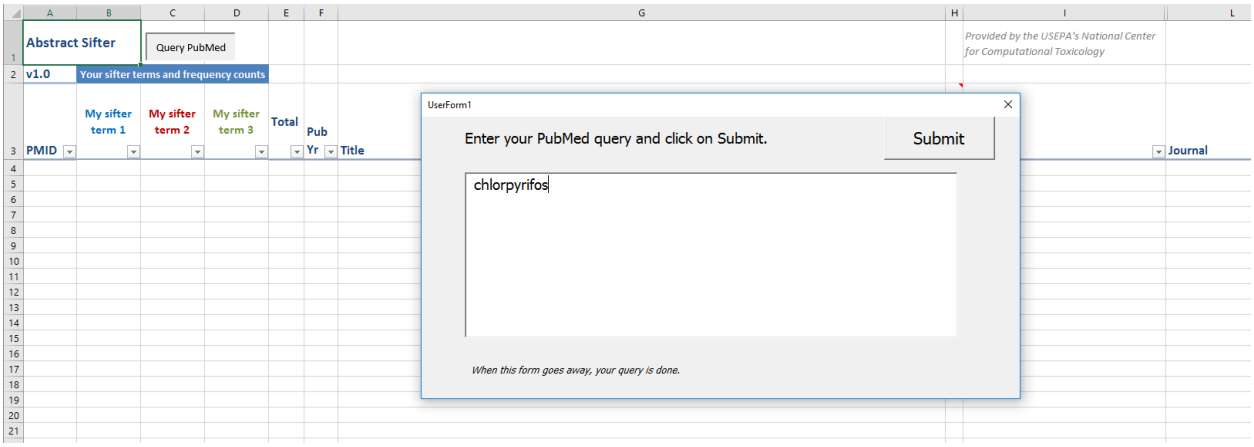


Figure 2. Running a PubMed query

When finished entering the query, the user clicks on *Submit* and the query is packaged by VBA into an e-utility command that is passed to the NCBI (National Center for Biotechnology Information) web services. (Note that using Sifter Query PubMed capability requires internet access.) The first response returned by the utility is the number of articles found. (Figure 3) This number is displayed, and the user is asked if he/she want to continue. If the number of articles is over 5,000, the query will not be run and the user is encouraged to refine the query to return fewer records.

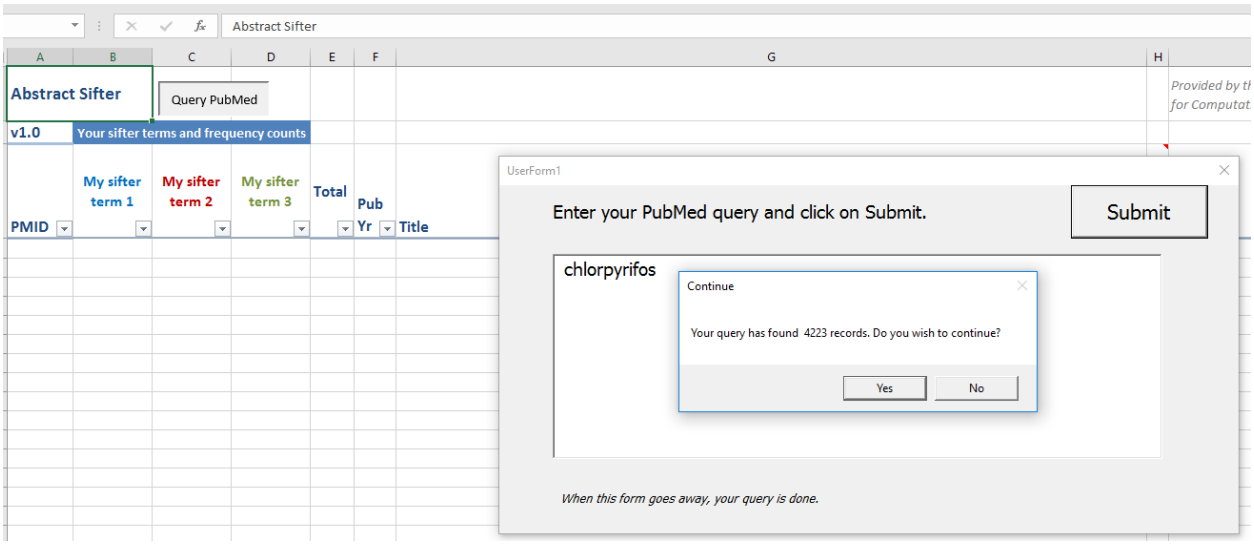


Figure 3. Responding to PubMed

If the returned results are fewer than 5,000 and the user indicates he/she wants to continue, the articles are downloaded from NCBI by Excel, and regular expressions are used to parse the citations for title, abstract, authors, publication year, journal, and PubMed identifier. Each record returned is inserted into a row in the Main sheet. Any rows in the Main sheet from a previous query are deleted.

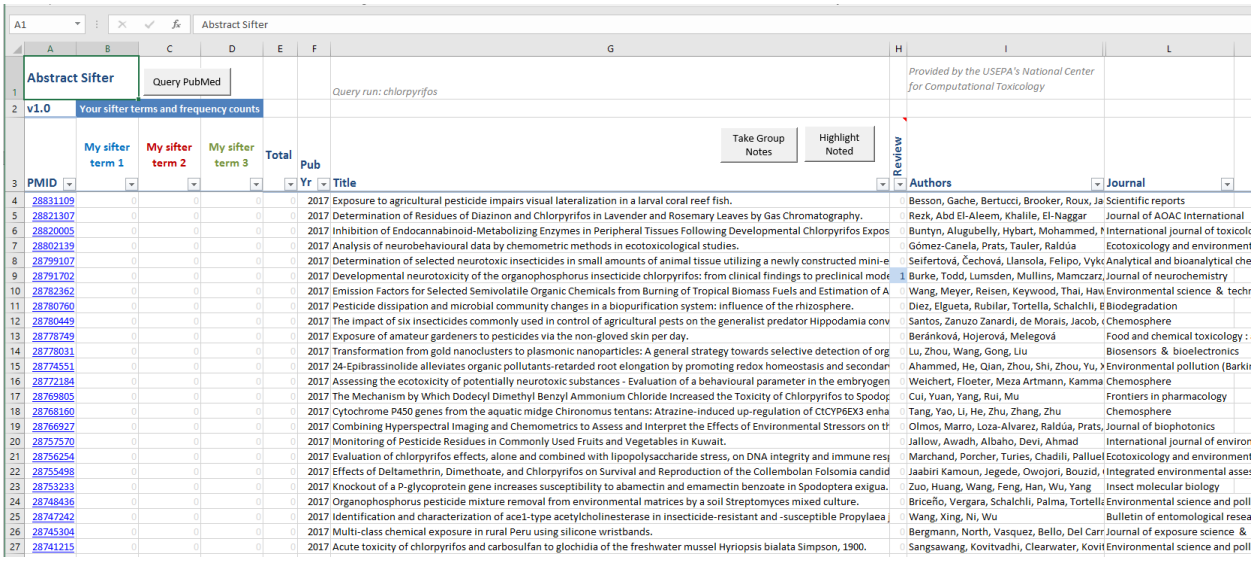


Figure 4.Results from PubMed query - before sifting

At this point the results of the query are stored in the Main sheet and can be browsed like any other data in a spreadsheet (Figure 4); however, the most effective way to find an articles of interest is to use the innovative sifter functionality. To demonstrate this functionality, we will continue to use our example of chlorpyrifos.

Let us suppose at this point that we are looking for dose studies of chlorpyrifos in rats. We type the term "mg/kg" in cell B3, "rat" in C3, and "brain" in D3. As we finish typing and move to the next cell, the Abstract Sifter will count the occurrences of the terms in the title and abstract combined. The citations can be sorted by these counts, either individually or by the total. Figure 5 shows what the Sifter looks like when these terms have been entered into B3, C3, and D3 and then the entries sorted by occurrence counts of "mg/kg" in descending order. PubMed 16472551 has 12 occurrences of "mg/kg", 22 of "rat", and two of "brain". This article indeed describes a dose-response study in rats.

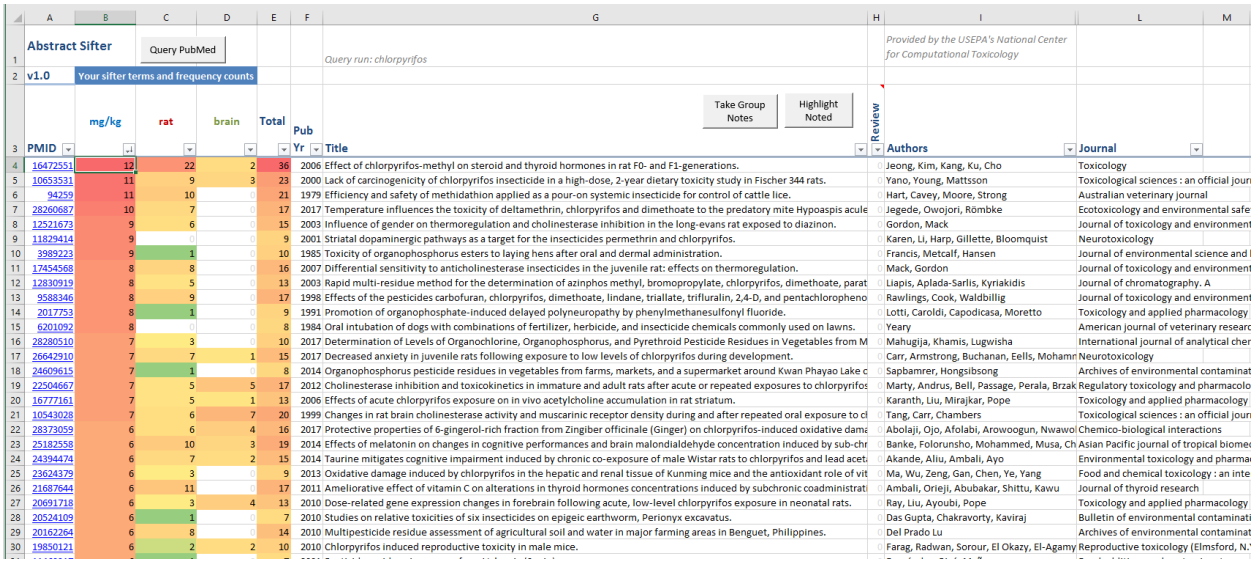


Figure 5. After sifter terms entered into cells B3, C3, D3 and sorting on B3

To see the abstract, we can either click on the PubMed ID hyperlink to be taken to PubMed, or we can double-click on any other cell in the row for this article. This action brings us to the Abstract sheet where the abstract is displayed along with other article meta-data like title (Figure 6).

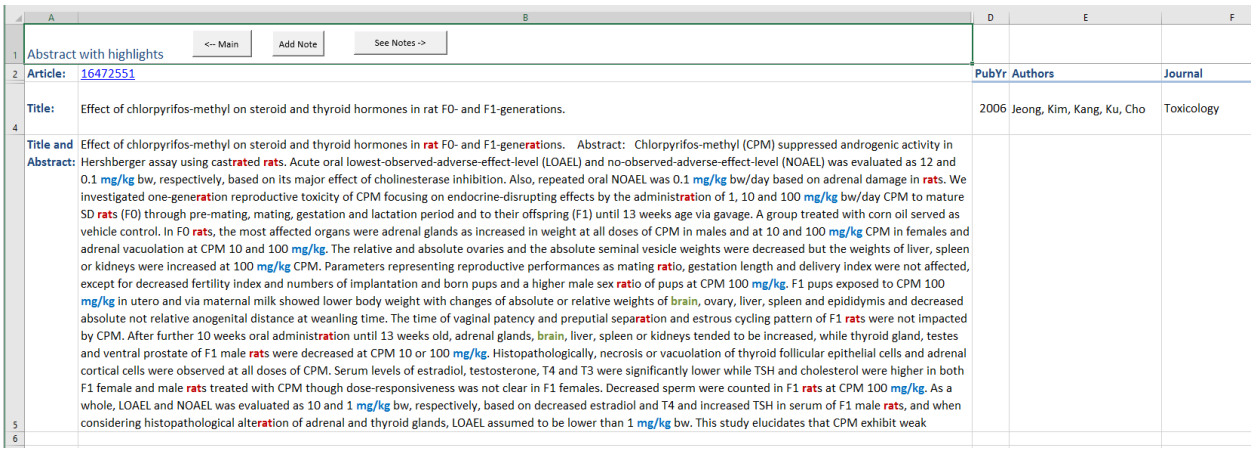


Figure 6. An abstract with highlighted sifter terms.

There are several aspects of the Abstract sheet that are important to note. First, the sifter terms are highlighted. The font colors reflect the colors of the fonts in cells B3, C3, and D3. This highlighting makes the reading process easier and draws attention to sections of the abstract that might be of most

interest. It is also interesting to note that the counts and highlighting for “rat” also picked up “administration”. Putting a space before (“ rat”) would eliminate some of these occurrences.

Sifting the results through specifying sifter terms in B3, C3, and D3 can be repeated as many times as the user wishes. Similarly, new PubMed queries can be run, altered, rerun. There are no restrictions on either of these activities other than the 5000 record return limit. Go for it!

Taking notes

Particularly given the dynamic nature of the sifter, many users find it helpful to be able to make notes on articles that they want to keep track of. There are two ways using the Sifter to take notes, one way through the Main page and the other way starting at the Abstract page. To return to our case study, let us say that we have found a set of articles on the Main page that we know we need to read in depth or to ask the library to get. We can select these articles and then click on the *Take Group Notes* button. A form appears where we can enter information into fields called Tag and Notes. These elements are self-defined. We can also click on *yes*, *no*, or *maybe*. This set of a variable is a quick way to associate articles with a note. Notice that these choices each come with a color (yes-green, no-red, and maybe-yellow). Entering any of these fields is optional. (Figures 7 and 8.) When we click on the *OK* button, each article selected will be inserted into the Notes page with the corresponding information. (Figure 9)

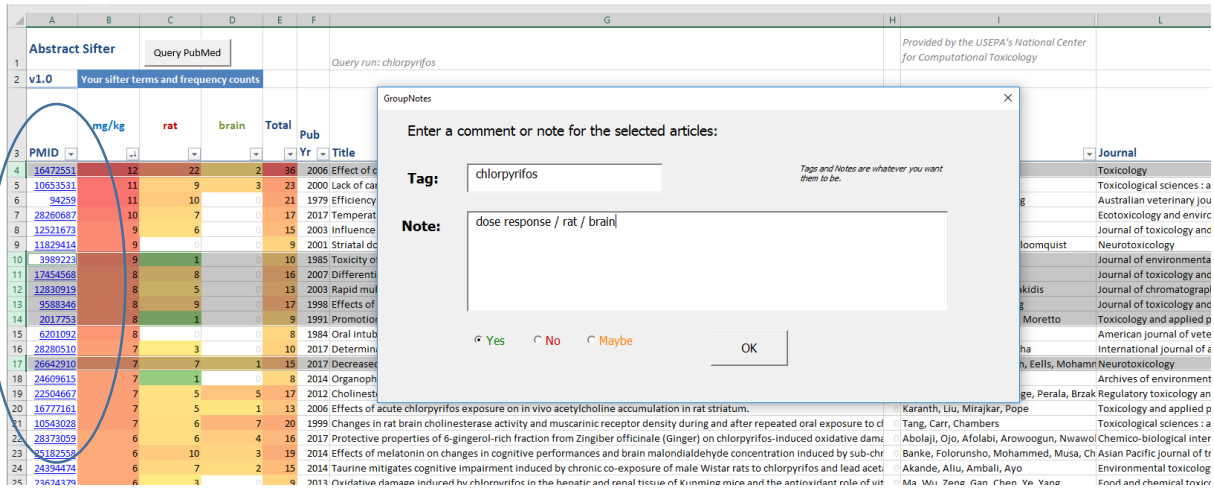


Figure 7. Taking group notes.

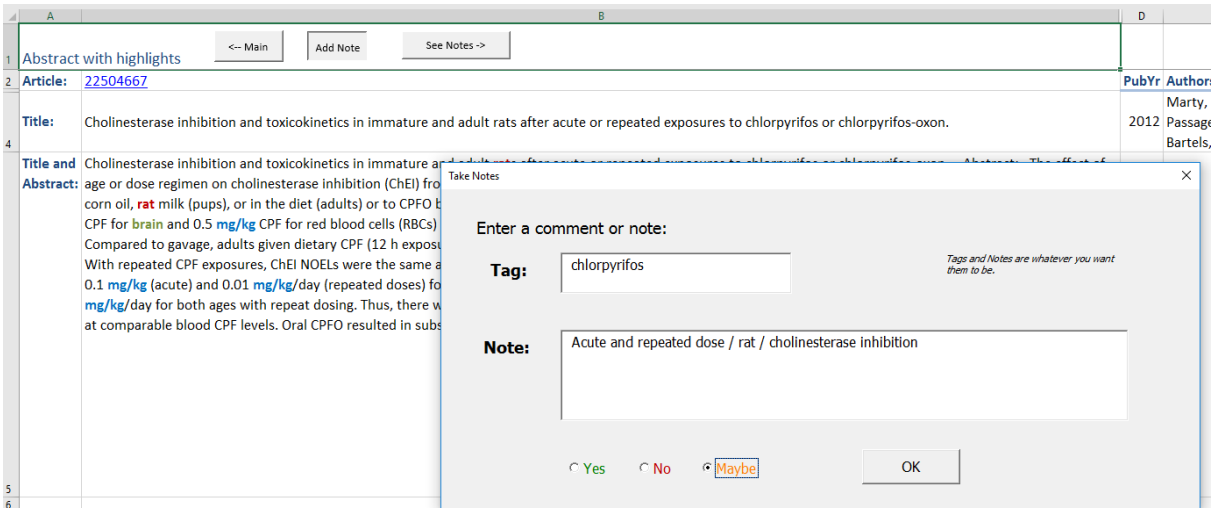


Figure 8. Taking group notes - another example.

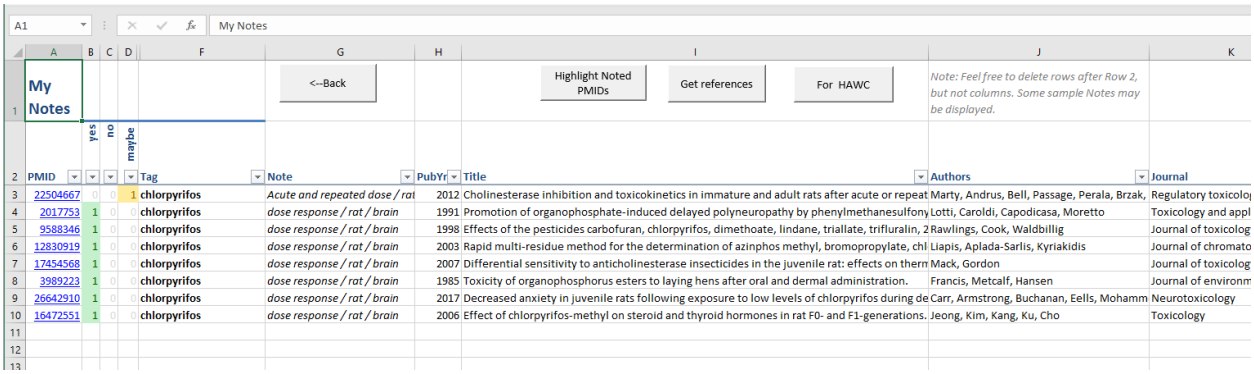


Figure 9. The Notes page. Remember to save your workbook.

The note-taking process can be used to help keep track of which citations have been read and evaluated and which have not. By clicking on the *Highlight Noted* button on the Main page or on the Notes page, the color of each noted PubMed ID will be modified as shown in Figure 10.

The user is welcome to make changes on the Notes sheet and add or delete rows after Row 2.

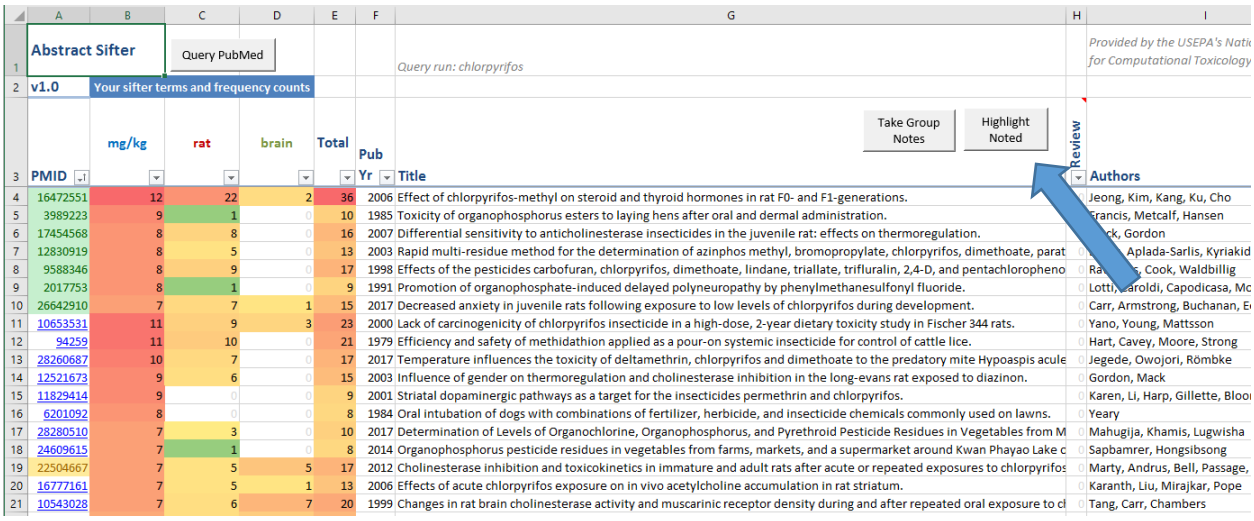


Figure 10. After clicking on Highlight Notes and then sorting by color

The Log sheet

The Log sheet keeps track of the queries you have run. The Abstract Sifter routines insert a row into the sheet every time you complete a query. These queries can be viewed and rerun. To rerun a query, simply double-click on it. (Figure 11.) Delete rows after Row 2 if you want to clear old entries.

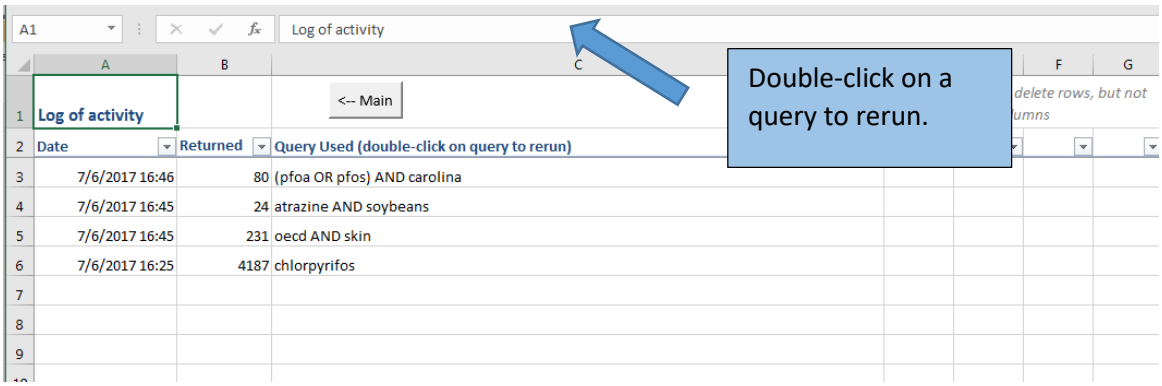


Figure 11. View of the Log sheet

The Landscape sheet

The Landscape sheet allows the user to get an overview of the literature for a set of entities like chemicals. Figure 12 shows an example of a Landscape sheet built by a researcher interested in the toxicity of a particular set of chemicals. Let's take a look at that first. Queries designed to find the chemicals of interest are entered into Column C and in this case, a short version of the chemical is in

Column B. The queries in Row 3 are typical ones used in searching for articles about different kinds of chemical toxicity. We will refer to these queries as subject matter queries.

The premise behind the design of the Landscape sheet is very simple: PubMed queries will be built by taking the values in Column C (in this example chemical names and corresponding CAS numbers) and appending this query text to the subject matter query text in Row 3 with an " AND " in between the two query parts.

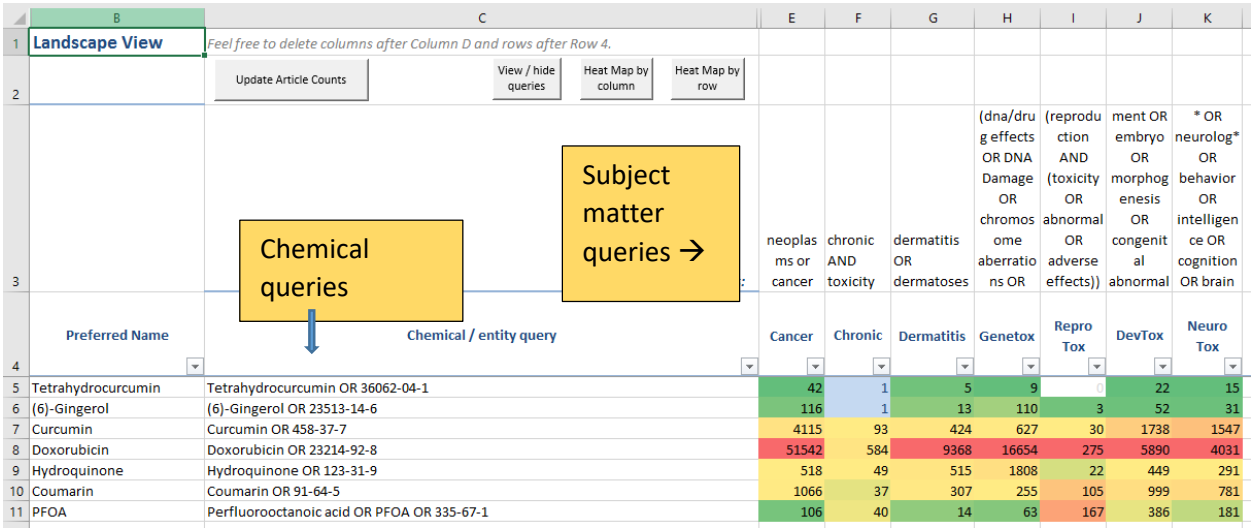
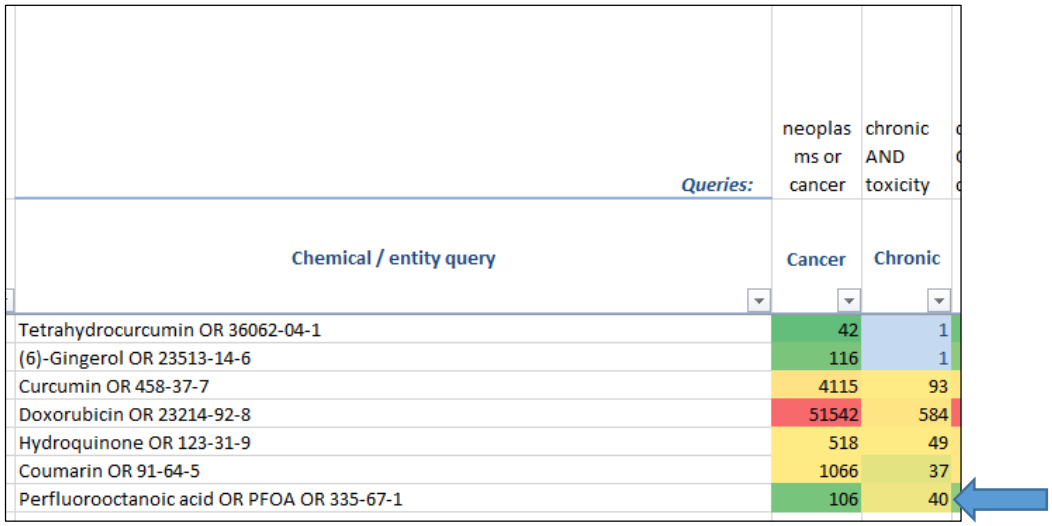


Figure 12. Example of Landscape sheet use

To illustrate, we will double-click on the cell highlighted in Figure 13. When we double-click on this cell this tells the Abstract Sifter to take the query text in Column C about Perfluorooctanoic Acid and append it to (chronic AND toxicity). Figure 14 shows the constructed query. We can then click on *Submit* and the query gets sent to PubMed and we can then see the results on the Main page. The number of articles retrieved from PubMed is 40. That count is placed in the corresponding Landscape cell that we just clicked on.



Queries:	neoplasms or cancer	chronic AND toxicity
Chemical / entity query	Cancer	Chronic
Tetrahydrocurcumin OR 36062-04-1	42	1
(6)-Gingerol OR 23513-14-6	116	1
Curcumin OR 458-37-7	4115	93
Doxorubicin OR 23214-92-8	51542	584
Hydroquinone OR 123-31-9	518	49
Coumarin OR 91-64-5	1066	37
Perfluorooctanoic acid OR PFOA OR 335-67-1	106	40

Figure 13. Double-click on article count cells

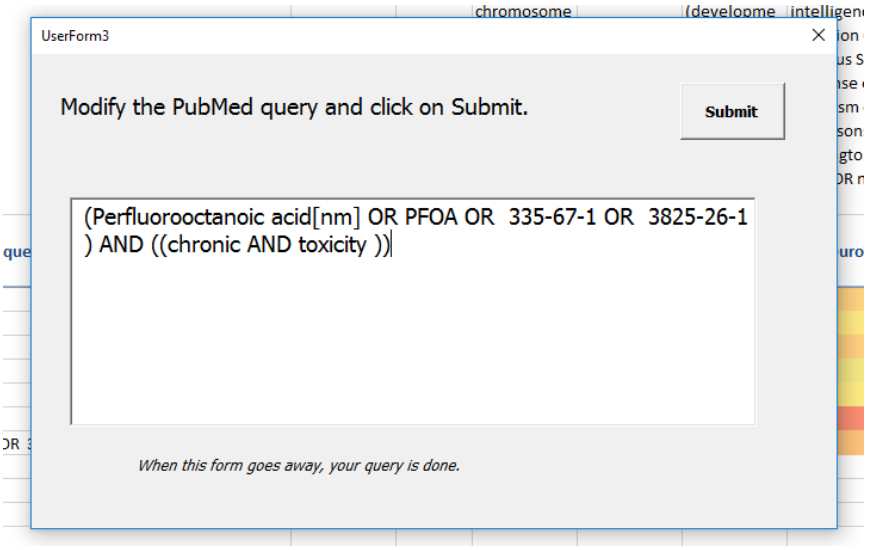


Figure 14. Constructed query

Now let's add to the Landscape sheet. Figure 15 shows how we added a new chemical to the list: atrazine. To find out the article counts for atrazine, select empty cells on the same row as atrazine, then click on *Update Article Counts* button. Excel will build each query from the atrazine part and the subject matter part and send each query to PubMed to find out how many citations satisfy the query. The

article counts are placed in the corresponding cells. To run the query and retrieve the results, just double-click on any of the article count cells.

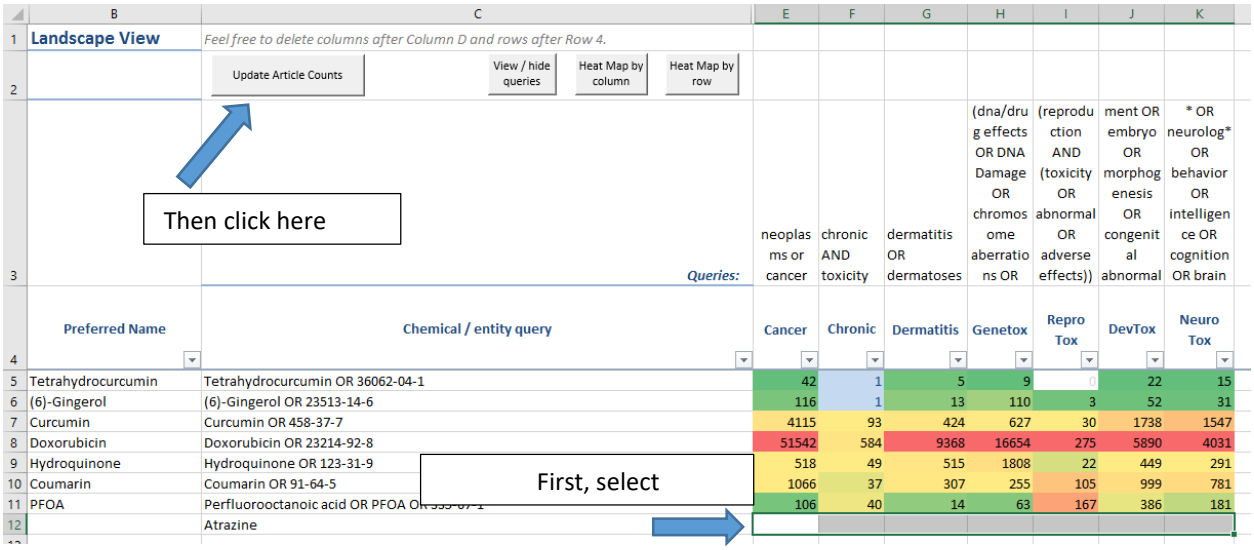


Figure 15. Adding rows to the Landscape sheet

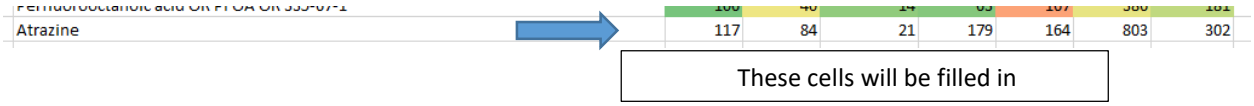


Figure 16. After clicking on Update Article Counts

New subject matter queries can be entered as well. The query part goes into Row 3 and a heading (your choosing) goes into Row 4. See the example below where the PubMed query part: skin OR dermatitis is entered with the heading skin. Next highlight the cells underneath and click on the *Update Article Counts* button. The counts of articles satisfying the queries are placed in the cells. What's happening behind the scenes? For each cell, a query is being built by the Abstract Sifter and sent to PubMed to retrieve a record count. That record count is then inserted into the corresponding cell. (Figure 17.)

Keep in mind that our examples revolve around chemicals, but that does not mean anyone is limited to chemicals. The entries in Column C and in Row 3 can be whatever you the end-user want them to be: genes, proteins, diseases, authors …

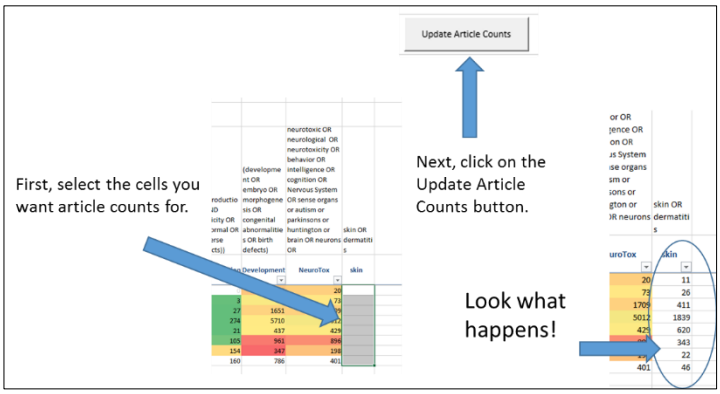


Figure 17. Steps for retrieving counts

Making things look good

The Landscape page has three buttons that make formatting easy (Figure 18). You can choose to hide the query row or show it. This button simply automates hiding and unhiding Row 3. The heat map buttons will quickly apply heat map coloring to the cells with article counts either by column or by row. Try them out!

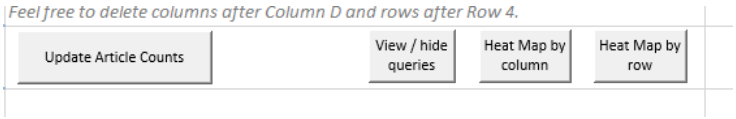
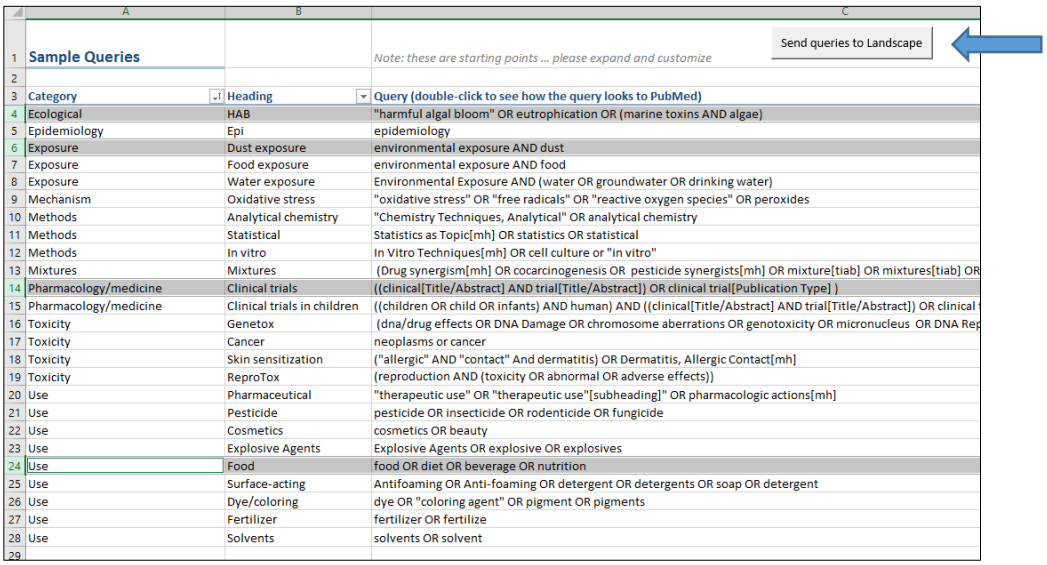


Figure 18. Buttons on the Landscape page include formatting actions

Using Sample Queries

The SampleQueries sheet contains a number of sample subject matter queries that the end user can use as a starting point for building a Landscape view of a set of entities. Let's see how. First, we will clean off the old subject matter queries by deleting columns E-J. Next, on the SampleQueries sheet we will

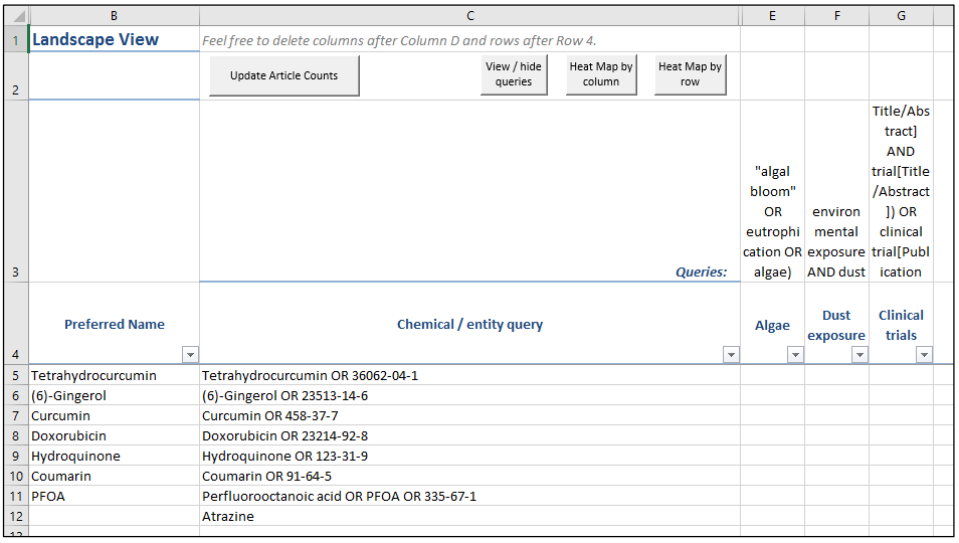
select rows with queries of interest then we click on the button *Send Queries to Landscape* (Figure 19). A form will ask for confirmation; to which we reply *OK*.



	A	B	C
1	**Sample Queries**		*Note: these are starting points ... please expand and customize* — Send queries to Landscape
2			
3	**Category**	**Heading**	**Query (double-click to see how the query looks to PubMed)**
4	Ecological	HAB	"harmful algal bloom" OR eutrophication OR (marine toxins AND algae)
5	Epidemiology	Epi	epidemiology
6	Exposure	Dust exposure	environmental exposure AND dust
7	Exposure	Food exposure	environmental exposure AND food
8	Exposure	Water exposure	Environmental Exposure AND (water OR groundwater OR drinking water)
9	Mechanism	Oxidative stress	"oxidative stress" OR "free radicals" OR "reactive oxygen species" OR peroxides
10	Methods	Analytical chemistry	"Chemistry Techniques, Analytical" OR analytical chemistry
11	Methods	Statistical	Statistics as Topic[mh] OR statistics OR statistical
12	Methods	In vitro	In Vitro Techniques[mh] OR cell culture or "in vitro"
13	Mixtures	Mixtures	(Drug synergism[mh] OR cocarcinogenesis OR pesticide synergists[mh] OR mixture[tiab] OR mixtures[tiab] OR
14	Pharmacology/medicine	Clinical trials	((clinical[Title/Abstract] AND trial[Title/Abstract]) OR clinical trial[Publication Type])
15	Pharmacology/medicine	Clinical trials in children	((children OR child OR infants) AND human) AND ((clinical[Title/Abstract] AND trial[Title/Abstract]) OR clinical
16	Toxicity	Genetox	(dna/drug effects OR DNA Damage OR chromosome aberrations OR genotoxicity OR micronucleus OR DNA Rep
17	Toxicity	Cancer	neoplasms or cancer
18	Toxicity	Skin sensitization	("allergic" AND "contact" And dermatitis) OR Dermatitis, Allergic Contact[mh]
19	Toxicity	ReproTox	(reproduction AND (toxicity OR abnormal OR adverse effects))
20	Use	Pharmaceutical	"therapeutic use" OR "therapeutic use"[subheading] OR pharmacologic actions[mh]
21	Use	Pesticide	pesticide OR insecticide OR rodenticide OR fungicide
22	Use	Cosmetics	cosmetics OR beauty
23	Use	Explosive Agents	Explosive Agents OR explosive OR explosives
24	Use	Food	food OR diet OR beverage OR nutrition
25	Use	Surface-acting	Antifoaming OR Anti-foaming OR detergent OR detergents OR soap OR detergent
26	Use	Dye/coloring	dye OR "coloring agent" OR pigment OR pigments
27	Use	Fertilizer	fertilizer OR fertilize
28	Use	Solvents	solvents OR solvent
29			

Figure 19. Selecting rows with queries of interest

Our Landscape sheet then looks like Figure 20.



	B	C	E	F	G
1	**Landscape View**	*Feel free to delete columns after Column D and rows after Row 4.*			
2		Update Article Counts — View / hide queries — Heat Map by column — Heat Map by row			
3		***Queries:***	"algal bloom" OR eutrophication OR algae)	environmental exposure AND dust	Title/Abstract] AND trial[Title/Abstract]) OR clinical trial[Publication
4	**Preferred Name**	**Chemical / entity query**	**Algae**	**Dust exposure**	**Clinical trials**
5	Tetrahydrocurcumin	Tetrahydrocurcumin OR 36062-04-1			
6	(6)-Gingerol	(6)-Gingerol OR 23513-14-6			
7	Curcumin	Curcumin OR 458-37-7			
8	Doxorubicin	Doxorubicin OR 23214-92-8			
9	Hydroquinone	Hydroquinone OR 123-31-9			
10	Coumarin	Coumarin OR 91-64-5			
11	PFOA	Perfluorooctanoic acid OR PFOA OR 335-67-1			
12		Atrazine			

Figure 20. New queries on Landscape sheet

Next, we select the article count area and then click on *Update Article Counts*.

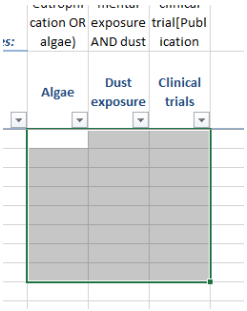


Figure 21. Selecting the cells for article counts

Once the article counts are populated, we click on Heat Map by Row and then on Hide queries. Our resulting Landscape view looks like Figure 22. To run the query and retrieve the results, just double-click on any of the article count cells.

	B	C	E	F	G
1	**Landscape View**	*Feel free to delete columns after Column D and rows after Row 4.*			
2		Update Article Counts / View / hide queries / Heat Map by column / Heat Map by row			
4	**Preferred Name**	**Chemical / entity query**	**Algae**	**Dust exposure**	**Clinical trials**
5	Tetrahydrocurcumin	Tetrahydrocurcumin OR 36062-04-1	0	0	3
6	(6)-Gingerol	(6)-Gingerol OR 23513-14-6	0	0	5
7	Curcumin	Curcumin OR 458-37-7	4	0	258
8	Doxorubicin	Doxorubicin OR 23214-92-8	6	1	9206
9	Hydroquinone	Hydroquinone OR 123-31-9	12	1	130
10	Coumarin	Coumarin OR 91-64-5	5	1	162
11	PFOA	Perfluorooctanoic acid OR PFOA OR 335-67-1	7	52	5
12		Atrazine	90	14	3

Figure 22. New Landscape view

Exporting to other applications from the Notes sheet

The Abstract Sifter allows the user to export articles from the Notes sheet to outside applications. On the Notes sheet there are two buttons labeled *Get References* and *For HAWC*. By clicking on the *Get References* button, the form in Figure 23 appears. The box on the left contains the PMIDs for each of the rows in the Notes sheet. Clicking on Send to PubMed will send the PMIDs to PubMed for retrieval. In PubMed, the citation records will be displayed. At the top of the screen, the user can click on *Send to* (Figure 25). This starts a dialog leading the user through the steps of downloading the citations to a

reference management software. (Figure 26). The user can download all the PubMed IDs on the Notes page, or a selected set of rows through the All or Selected radio buttons shown in Figure 24.

Figure 23. Buttons on Notes sheet that allow export

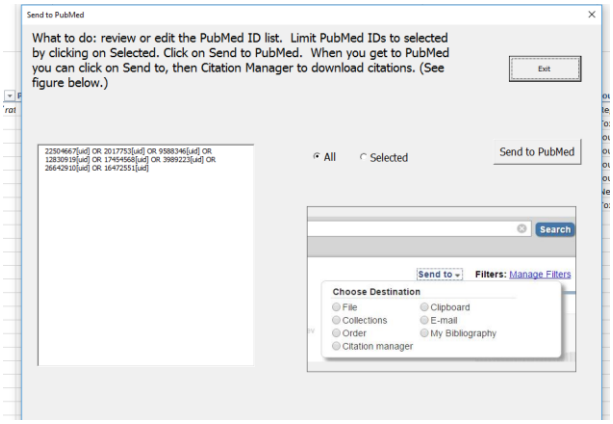


Figure 24. Form that appears after clicking on Get references button

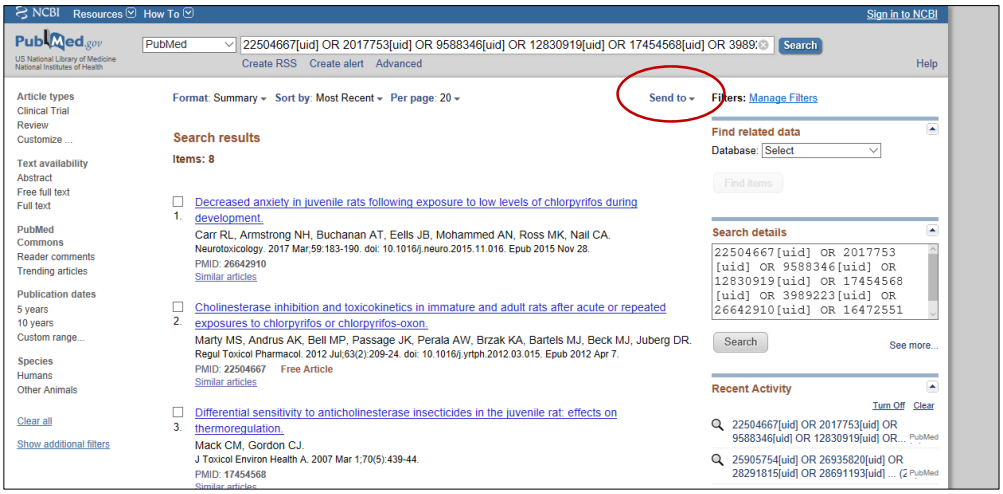


Figure 25. In PubMed, click on Send to

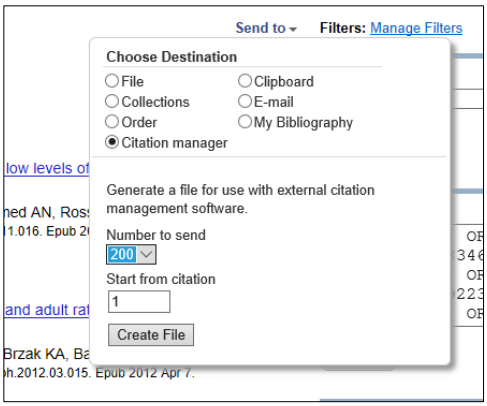


Figure 26. Dialog box for exporting to citation manager

When the user clicks on the Create File button, a file is created and downloaded in nbib format and can be imported into any reference manager software.

Similarly, the PMIDs can be formatted for input into Health Assessment Workspace Collaborative, also known as HAWC (Figure 27). HAWC tutorials are available to demonstrate the import process so it will not be covered here.

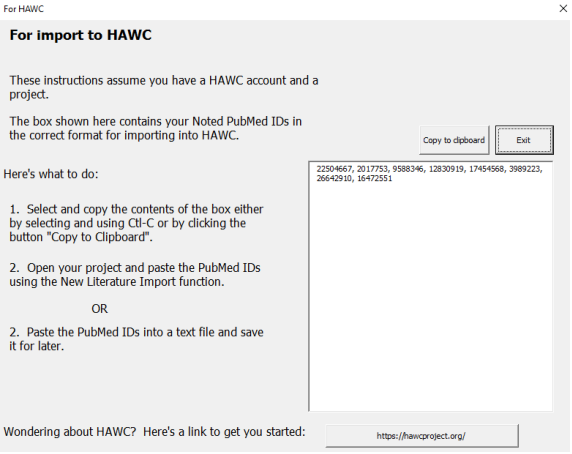


Figure 27. Form to send PMIDs to HAWC

Helpful Tips

Tip 1

The Landscape sheet is a great way to explore a set of chemicals but some chemical names are long, complex, and a challenge to PubMed. If you get unexpected results from a chemical query, it's a good idea to check it in PubMed. You can take any query generated by the Abstract Sifter and copy and paste it into PubMed using Ctl-C to copy and Ctl-V to paste. For example, the query in the box shown in Figure 28 is selected and copied (with Ctl-C). Then in PubMed the query is pasted into the query line at the top as shown in Figure 29. After we click on search we see that PubMed brings back 51 records. On the right side of the page is a box entitled *Search Details*. Click on the *See More* … link to expand this box. (Figure 30) Figure 31 shows the information provided by PubMed about how it expands the query. If you need to learn more about PubMed queries, click on *Help* on the PubMed home page.

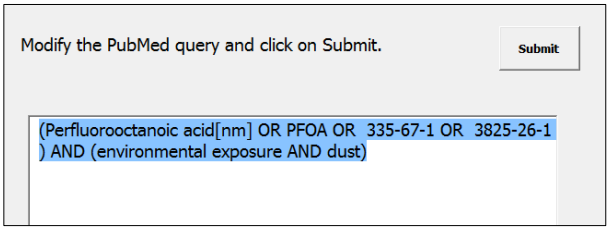


Figure 28. Select and Ctl-C to copy

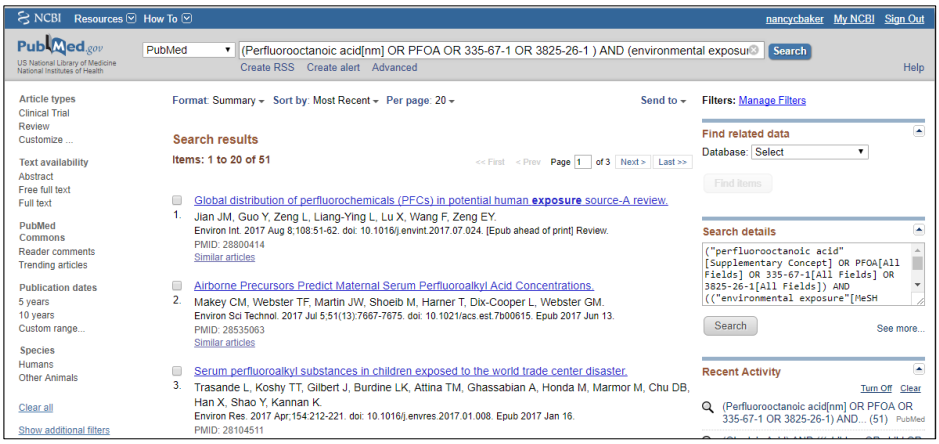


Figure 29. Ctl-V to paste in PubMed then search

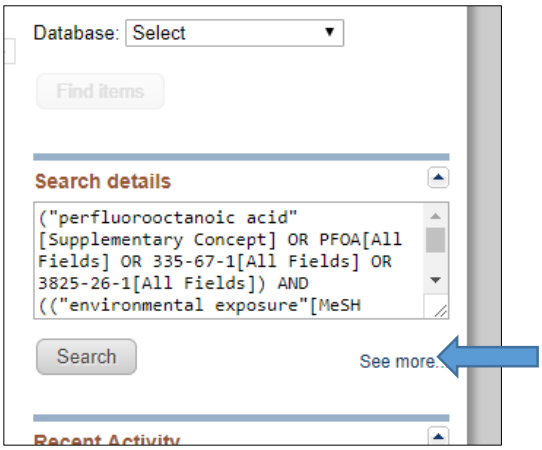


Figure 30. See what PubMed does to expand your search

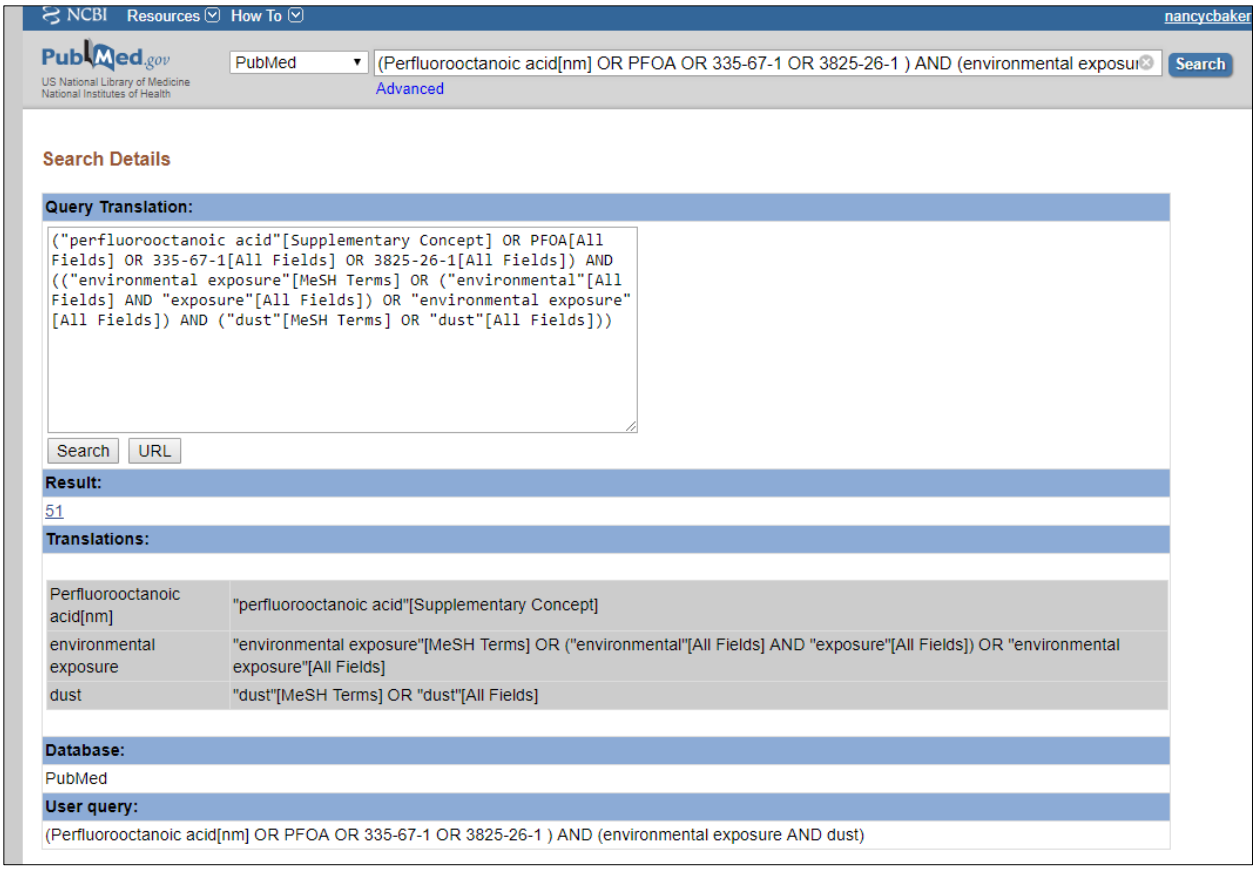


Figure 31. PubMed query breakdown and expansion

Tip 2

It can also be very helpful in chemical research to include the chemical name in the sifting process. This is because a chemical can be mentioned in an abstract even in cases where the article is not really about the chemical and will be retrieved in the PubMed query (depending on how the query is worded). Counting the occurrences of the chemical name in the abstract through the sifting process can help the user discriminate between articles mentioning a chemical or those actually about a chemical.