Abstract Sifter User Guide, Version 1.0

Nancy C. Baker¹, Thomas Knudsen², Antony J. Williams²

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

¹Leidos, Research Triangle Park, NC, 27711, USA and ²National Center for Computational Toxicology, U.S. Environmental Protection Agency, Research Triangle Park, NC, 27711, USA.

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 – Version 1.0

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 provides guidance on the use of the Abstract Sifter through a series of screen shots showing the most common tasks in the tool followed by some helpful tips.

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.

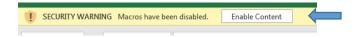


Figure 1. Enable macros upon opening

Once open you will see that the Abstract Sifter Excel file consists of seven sheets. Each sheet is described briefly in the table below.

| Sheet name | Sheet Function |
|---------------|--|
| ReadMe | Basic information on the sifter with links to more documentation |
| Main | Starting point for PubMed queries and for sifting |
| Abstract | The sheet where the citation abstract is shown |
| Notes | Notes and tags are inserted here |
| Log | Log of every query run on Main sheet |
| SampleQueries | Repository of sample queries to use as starting points |
| Landscape | High level view of literature for entities |

Main Sheet

The Main sheet is where the basic functionality of Abstract Shifter occurs, including functions we call "sifting". To begin using the Abstract Sifter, the end-user clicks on the *Query PubMed* button at the top of the screen in the Main sheet. 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. (For a review of this syntax, visit

https://www.ncbi.nlm.nih.gov/books/NBK3827/

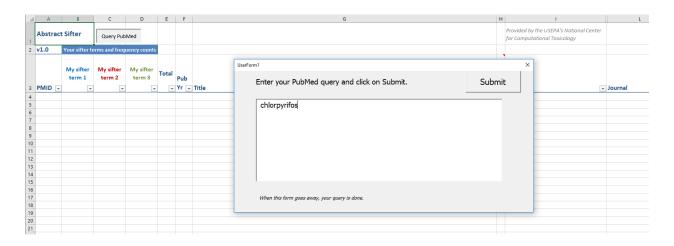


Figure 2. Running a PubMed query

When finished entering the query, the user clicks on *Submit* and the query is packaged by a Visual Basic Application (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.

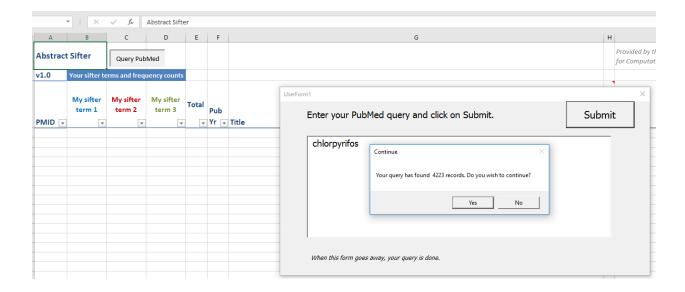


Figure 3. Responding to PubMed

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.

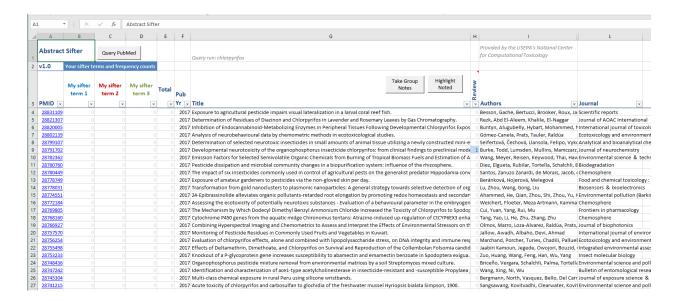


Figure 4.Results from PubMed query - before sifting

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 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-response data for chlorpyrifos in rat studies that focused on brain effects. 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 then 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 cells 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.

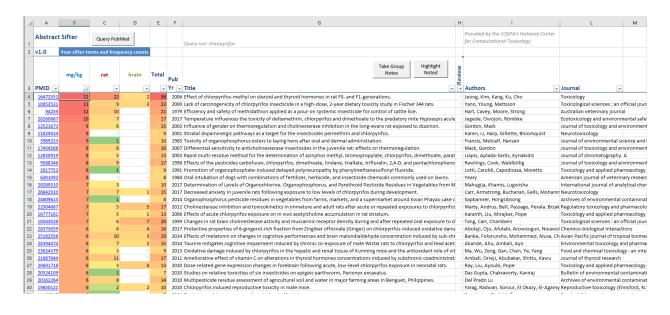


Figure 5. After sifter terms were entered into cells B3, C3, D3 and sorting on B3

Abstract Sheet

To see the abstract for any of the retrieved articles, 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 and authors (Figure 6).

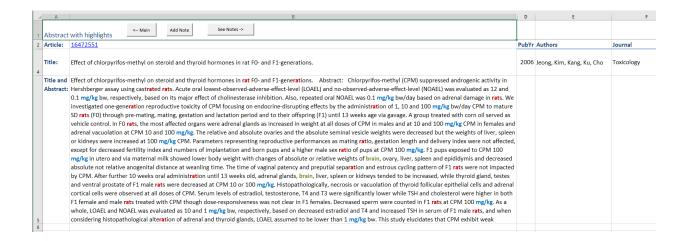


Figure 6. An abstract with highlighted sifter terms.

There are several aspects of the Abstract sheet that are important to note. First, the sifter terms in the abstract are highlighted. The font colors reflect the colors of the fonts in cells B3, C3, and D3 entered into the Main Sheet. This highlighting makes the reading the abstract easier by drawing attention to sentences that might be of more interest. It is also interesting to note that the counts and highlighting for "rat" also picked up "administration". Putting a spaces before or after search terms (e.g., " 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.

Notes Sheet:

Given the dynamic nature of the sifter, many users find it helpful to be able to make notes on articles that they want to track. There are two ways using the Sifter to take notes: one way is through the Main page, and the other way starts with 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. 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 variables 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)

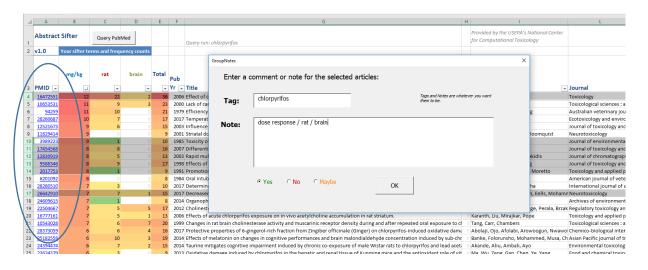


Figure 7. Taking group notes.

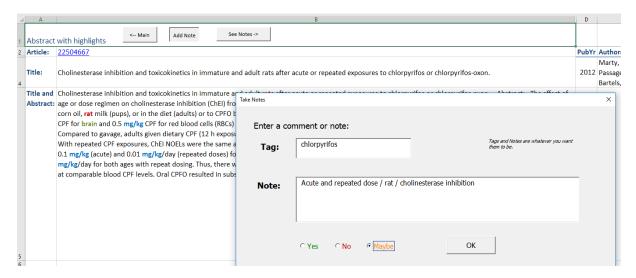


Figure 8. Taking group notes - another example.

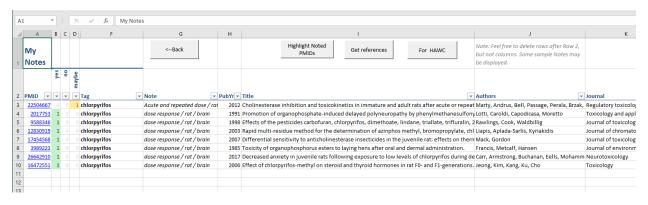


Figure 9. The Notes page. Remember to save your workbook.

The second option for note taking starts with the Abstract Sheet. The "Add Note" button in the top row allows notes to be inserted into the Notes Sheet using the same form used to add Group Notes described above.

The note-taking 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 can make changes to the notes sheet by editing, adding or deleting rows below row 2.

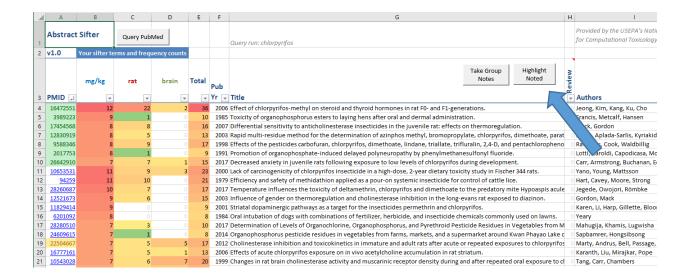


Figure 10. After clicking on Highlight Notes and then sorting by color

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 any or all rows after Row 2 if you want to clear old entries.

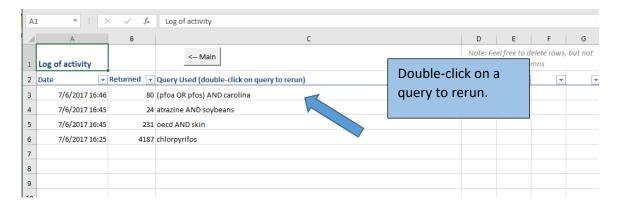


Figure 11. View of the Log sheet

Landscape sheet

The Landscape sheet provides an overview of the literature to the user for a set of entities, for example, a list of chemicals or genes. 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 name is in Column A. 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.

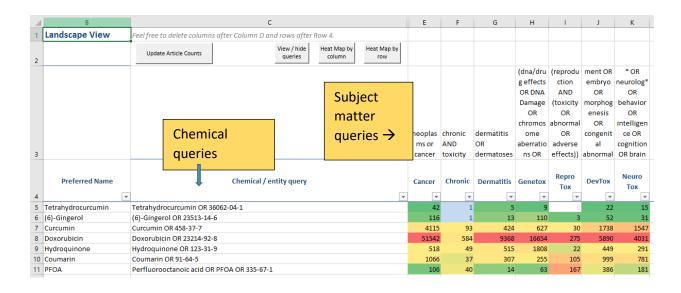


Figure 12. Example of Landscape sheet use

To illustrate, we will double-click on the cell with the arrow pointer 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.

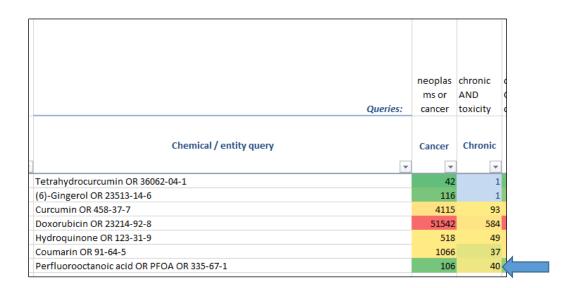


Figure 13. Double-click on article count cells

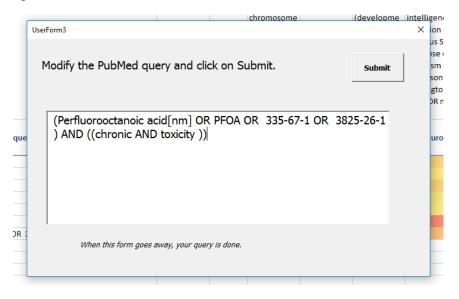


Figure 14. Constructed query

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.

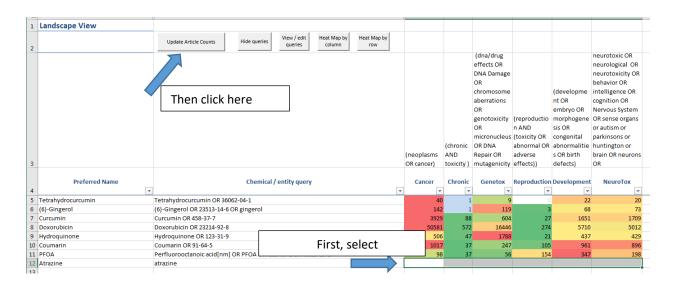


Figure 15. Adding rows to the Landscape sheet

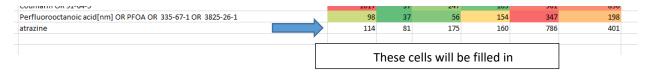


Figure 16. After clicking on Update Article Counts

New subject matter queries can be entered as well. The query part goes into Row 3 and a heading (of your choice) 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 ...

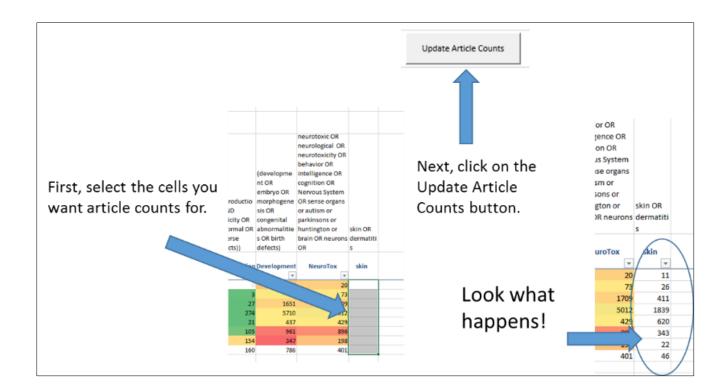


Figure 17. Steps for retrieving counts

Making things look good

The Landscape sheet has three buttons that make formatting easy (Figure 18). You can choose to hide the query row or show it. 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

SampleQueries Sheet

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 on the Landscape sheet. 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*.

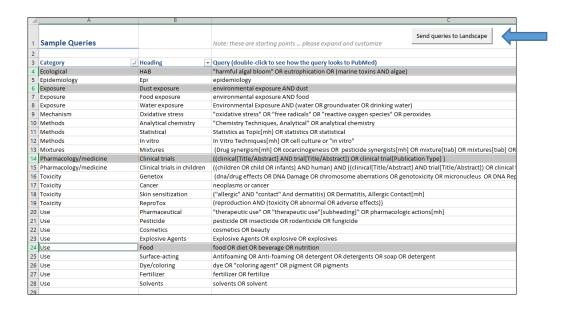


Figure 19. Selecting rows with queries of interest

Our Landscape sheet then looks like Figure 20.

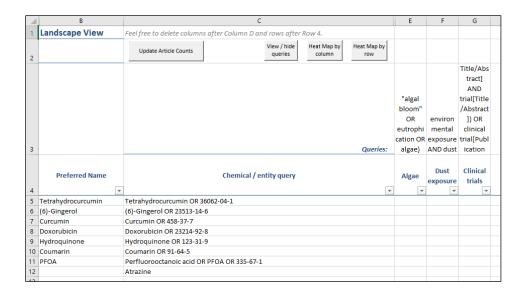


Figure 20. New queries on Landscape sheet

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

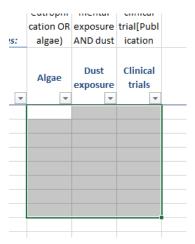


Figure 21. Selecting the cells for article counts

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.

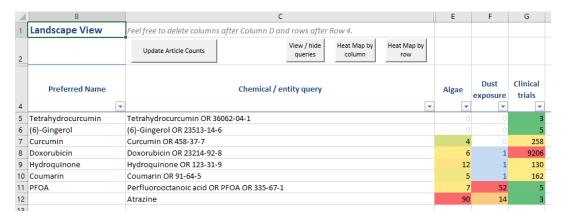


Figure 22. New Landscape view

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 24 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

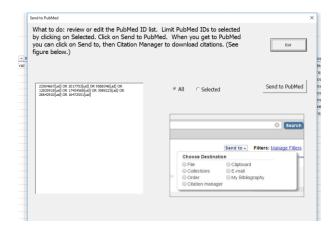


Figure 24. Form that appears after clicking on Get references button

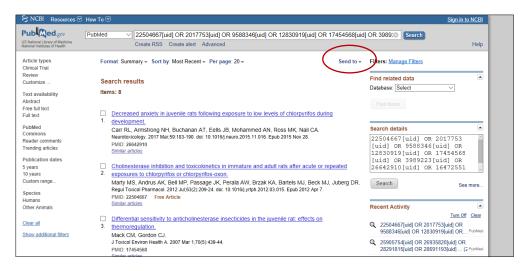


Figure 25. In PubMed, click on Send to

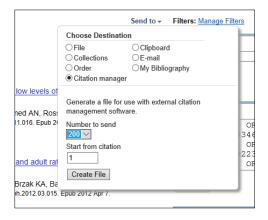


Figure 26. Dialog box for exporting to citation manager

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

Similarly, the PMIDs can be formatted for input into HAWC (Health Assessment Workspace Collaborative) (Figure 27). HAWC tutorials are available at https://hawcproject.org/about/. The import process is demonstrated there, so it will not be covered here.

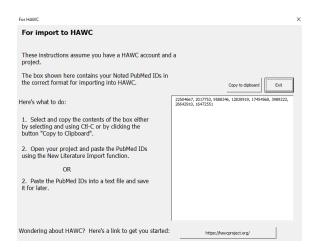


Figure 27. Form to send PMIDs to HAWC

18

Helpful Tips and Guidelines

Tip 1 – checking quality of results

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.

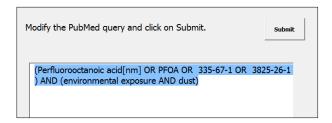


Figure 28. Select and Ctl-C to copy



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

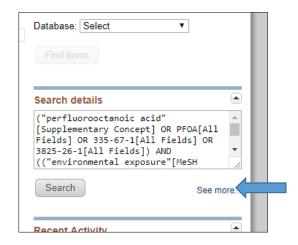


Figure 30. See what PubMed does to expand your search

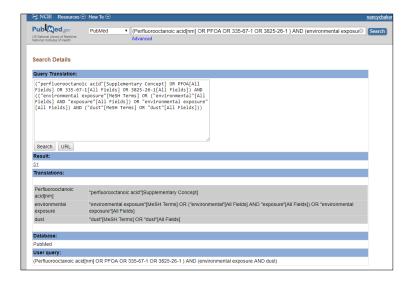


Figure 31. PubMed query breakdown and expansion

Tip 2 – Sifting the chemical literature

It can 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 that are actually about the chemical.

Tip 3 – cleanup and customization

The Abstract Sifter can be cleaned up by deleting rows and columns from previous work, but the Abstract Sifter programming requires certain columns and rows to be in certain places. To learn how to clean up your sifter without disrupting the behind-the-scenes coding, consult the table below.

| Sheet name | Sheet Function |
|---------------|---|
| Main | Do not add columns. Rows will be added and deleted by the Sifter. |
| Abstract | Do not add or delete rows or columns. The Sifter software updates this sheet. |
| Notes | Delete any unwanted rows after Row 2. Do not add or delete columns. |
| Log | Delete any unwanted rows after Row 2. Do not add or delete columns. |
| SampleQueries | Delete unwanted rows after Row 3. Modify and add rows as desired, following the pattern of current rows. (That is, keep the heading in column B and the query text in column C.) Hint: use this sheet to keep queries important to your organization. |
| Landscape | Delete or modify rows after Row 4 and columns after Column D. |

Keep in mind that the Abstract Sifter is an Excel file. You can rename it, mail it, and of course, if you want to keep your Log, Notes, and Landscape entries, you should save it. The SampleQueries sheet

provides an opportunity for you and your organization to start collecting and organizing queries that you have found useful.

Tip 4 – Connections to the EPA Chemistry Dashboard

The Environmental Protection Agency's Chemistry Dashboard is a great place to find chemical information to enhance your chemical search queries with synonyms and CAS numbers. Future releases of the Dashboard will offer opportunities to download a list of chemicals formatted for easy insertion into the Landscape sheet. You'll find the Chemistry Dashboard here:

https://comptox.epa.gov/dashboard.

The EPA Chemical Dashboard also contains its own (slightly different) version of the Abstract Sifter. It works on the same basic premise as the Excel version, but has some interesting differences. To see it, start with a chemical search. Let's look at the chemical fipronil by entering the name in the search box and clicking on the search icon (magnifying glass) (Figure 32).

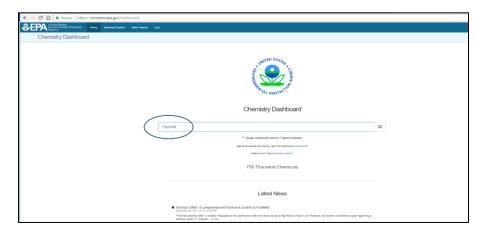


Figure 32. Searching for fipronil on the EPA's Chemistry Dashboard entry form

The main page for fipronil is displayed with the structure diagram and a selection of tabs below that lead to other information about the chemical. Click on the Literature tab as shown in Figure 33.

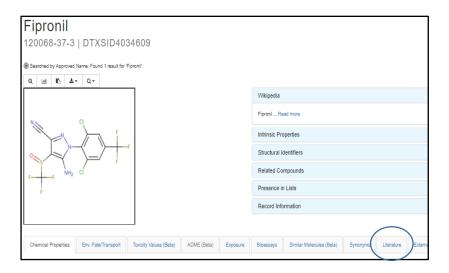


Figure 33. Select the Literature tab then on PubMed Abstract Sifter (see below).

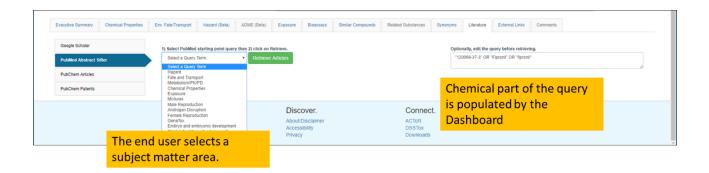


Figure 34. How to select prepared queries.

Select PubMed Abstract Sifter on the left set of buttons. The Dashboard helps you to build queries for this chemical. The chemical identifier part of the query is prepopulated on the right with name and CAS number. The subject matter part of the query is determined by selecting a topic area in the pull-down box in the center of the form. The user has several pre-composed queries to choose from. When one of them is chosen, the query is modified by appending the subject matter text. Figure 35 shows that when Metabolism/PK/PD is chosen, the text (metabolism OR metabolite OR tissue distribution OR pharmacokinetics OR pharmacodynamics) is appended to the chemical identifiers. The query can be modified manually as well. When ready, the user clicks on Retrieve Articles.

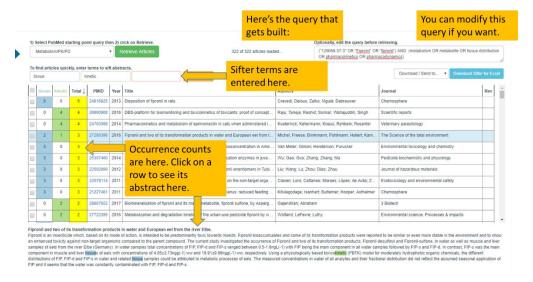


Figure 35. Sifting on the EPA Chemistry Dashboard's PubMed Abstract Sifter.

After the user clicks on Retrieve Articles, the article information is retrieved from PubMed and inserted into the results table. The articles can be sifted by entering terms into the boxes shown. In the example in Figure 35, the user has entered tissue into one box and kinetic into the other. The occurrences of these terms are counted for each PubMed citation and displayed. The table can be sorted on these values. Clicking on a row tells the Dashboard Sifter to display the title and abstract below the table with the sifter terms highlighted.

Check back periodically with the EPA's Chemistry Dashboard to learn about enhancements to PubMed Abstract Sifter.

Contact:

Contact Nancy Baker at baker.nancy@epa.gov with issues, ideas, and feedback.

24