Searching Text Information in Reaxys

Learning Objectives

- Understand fundamental text search issues of truncation, single words versus phrases, and term proximity
- Understand how different text inputs in **Search Reaxys** affect the list of results options displayed
- Understand how to enter text terms into fields and subfields in **Query builder**

Background

Nearly all computer files are text-searchable, and behind the interface most have algorithms that broaden the search in various ways. Commonly, algorithms automatically search singulars and plurals, American and English spellings, as well as various forms of verbs and participles; less commonly, algorithms may additionally search synonyms.

An underlying principle of any search algorithm is to balance search *precision* (narrow, very targeted results) and *comprehension* (more comprehensive but therefore less targeted results). This balance may vary depending on the text and context searched. For example, when we go to **Edit** => **Find** in Microsoft Word and enter a few letters, Word automatically inserts what is known as left- and right-truncation. So, if we enter *meri*, we retrieve hits such as *merit*, *american* and *numeric*. This automatically applied left- and right-truncation is acceptable here because we are searching relatively small text files, but it may not be acceptable when we are searching trillions of words in large databases. In such instances, we may retrieve too many hits or too many irrelevant hits to manage.

So, developers of search engines for large databases need to consider very carefully how far algorithms should extend.

Another consideration for searches based on multiple terms is whether these terms should be searched individually or as part of a phrase. In general, scientific databases implement both options, though it depends on the database field searched. For example, terms like *Bob Belford* to be searched in the Author field should be searched as a phrase — or, at the very least, the terms should be near to each other (perhaps in either order).

Finally, developers need to consider whether all text entries should be algorithmically interpreted, or whether searchers should be given the option of "controlling the search" themselves. Some developers shy away from the latter, arguing that computers can handle all situations and that users shouldn't have to think about issues of truncation, single words versus phrases, or term proximity.

Reaxys *allows both* and users can easily choose either approach.

In fact, it is very easy to master these search issues, and when we do, we can construct innovative searches that algorithms could not handle as well.

Text searches in Quick search

In **Quick search**, we enter text terms into the **Search Reaxys** entry box, and click **Search**. Reaxys interprets the terms entered and generates a list of hitset options that may include Document, Substance, and/or Reaction Records.

The hitsets retrieved by Reaxys depend on what we entered. *In general,* results follow two basic rules:

1. If we enter only terms and prepositions, Reaxys goes to its dictionary which includes terms from Document, Substance (including property), and Reactions Records, and delivers a full list of options.

2. If we use Boolean (AND, NOT, OR) or proximity operators (e.g., NEAR), truncation — in Reaxys, the truncation symbol is * and may be applied before and/or after the word stem — or any special characters such as parentheses ( ) or quotes (''), Reaxys performs the search only in Document Records.

   Note: if performing a search like this, do NOT enter prepositions because they will be included as search terms and may restrict answers accordingly.

Reaxys uses several hundred basic algorithms to interpret terms entered in **Search Reaxys**, but the above is a good rule of thumb. If we do not get the hitset options we wanted, we can simply vary the terms entered a little.

As a rule, focus on one or two concepts (terms entered) in a text search and see what turns up. We can always combine search results later.

Some examples to try

Try some of the following examples to become familiar with some of the basic ways to enter search terms into **Search Reaxys**.

1. *Atenolol*

   This is a simple way to find information on a substance. We obtain Substance and Document Records; browse through some of them and try to understand the nature of the answers obtained.

2. *MgSO4*

   Formulas can be a good way to find simple inorganics in Substance Records, but may not be comprehensive search terms in Document Records.

3. *Magnesium sulfate*

   This gives many more Document Records – obviously!

4. *Magnesium sulfate or MgSO4*

   OK, the algorithms don't work as we want in this case. However, let's learn one way to "fix" this.

   Try: *(magnesium sulfat*) or MgSO4*. We get what we wanted because using the truncation completely turns off the algorithms and hence, the search is interpreted literally as we entered.

5. *Ritalin*

   This retrieves Substance Records all right, but not many Document Records. Here, this is because the most common Index Term (text) for the substance is *methylphenidate*.

   Try a search on *methylphenidate* and think about the implications.

6. *Global warming* with *Antarctica*

   Again, we need to "trick" the algorithm to get better results. We can use truncation like *global warm* antarctica,* or we can try a better search, like *(global warm* or climate chang*) and antarctica.*

   Note: we put parentheses around terms combined with the Boolean OR.

   It is difficult to write algorithms that cover every search option. Thus, at times it is preferable to have control of the search.

7. *Nanotub* batter**

   Note: Reaxys applies the AND operator between two terms by default. To see the effect of this operator, try entering *nanotub* NEAR batter** and compare the results. NEAR <u>restricts</u> answers (fewer hits) to those in which the terms entered are near to each other (i.e., separated by a few words).

8. *Electrical conductivity of copper*

   Note that this and related searches are examples in the resource "Finding property information in Reaxys", posted on the Explore Reaxys module as Reaxys_PropertiesSearch_200117. Try some of those examples if you have not done so already.

9. Find articles that describe the adsorption of phosphates from ground water by wood chips.

   There are 4 concepts here (with possible synonyms/search terms):

   a. Adsorption   (adsorption, adsorb => adsor*)
   b. Phosphates   (phosphate, phosphates => phosphate*)
   c. Ground water   (ground water(s), groundwater(s), wastewater(s)
                                 => *water*)
   d. Wood chips   (wood chip(s), woodchip(s), wood => *wood*)

   So, we could try *adsor* phosphate* *water* *wood** (the search may take a minute since there are lots of terms for Reaxys to consider).

   Browsing through the answers, we can look for further terms to expand our search, or to search for fewer terms. In complex situations like this, we may wish to break the search up into multiple smaller searches and then combine answers at the end. We will see how to do this in **Query builder** in a future resource that will be posted to the Explore Reaxys module.

Text searches in Query builder

A distinguishing feature of Reaxys is that text terms (words) can be searched in all Document, Substance, and Reaction Records. Arguably, there are as many searchable words in Substance and Reaction Records as there are in Document Records.

While the **Search Reaxys** entry box of **Quick search** taps into many of the fields in Substance and Reaction Records, it does not search *all* the text fields available. So, it is helpful to look at text searches through **Query builder**.

Access the **Query builder** by clicking the link in the top navigation bar or Reaxys. Click **Basic Indexes** in the right-hand panel. This reveals the 3 basic indexes:

- **Substance Basic Index** – searches single words in Substance Records
- **Reaction Basic Index** – searches single words in Reaction Records
- **Document Basic Index** – searches single words in Document Records. In fact, most of the search examples we described in the previous section operate through this index.

Drag the **Substance Basic Index** to the workspace.

Click the down arrow to the left of the search field, to display a menu of options to specify entries:

- **ends with** – this option adds left-truncation to the term entered
- i**s** – instructs Reaxys to search the term as entered
- **contains** – this option adds left- and right-truncation to the term entered
- **starts with** – this option adds right-truncation to the term entered

Use these options to have truncation entered automatically, or add the truncation directly to your search term with an asterisk (*).

Now click the **Lookup** icon on the right of the search field. A dialog box appears listing the first of over 1.4 million pages of terms that can be searched with the **Substance Basic Index**. Yes, there is a lot we can search!

Hover over the **Search** icon $\mathcal{Q}$ to make the search entry box appear and type in *groundwater*. Click the **Search** icon. Check the options shown below and then click **Transfer**.
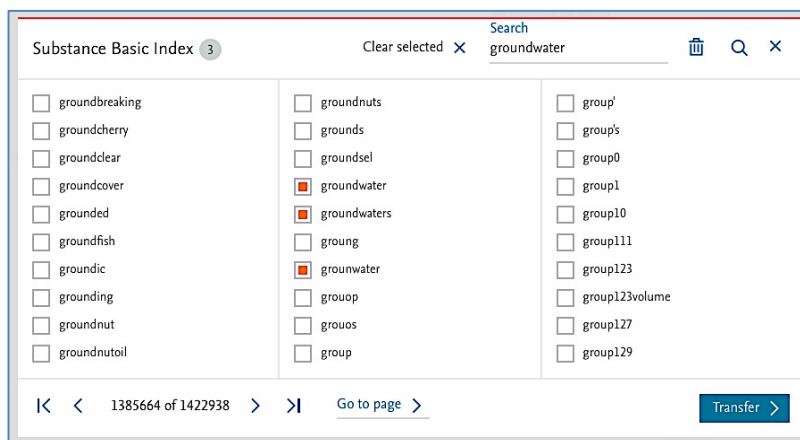
The terms we selected appear in the search field of the **Substance Basic Index**. Click **Search** at the top of the **Query builder** and select **Substances** from the resulting menu.

Our search retrieves almost 900 substances that contain the selected terms in one or more of the fields in their Substance Records.

By default, the substances are sorted by **No of References**. We are probably not interested in the first group of substances, but click on some of the **Hit Data** fields and browse through the entries to become familiar with how information is found in these fields.

We can find which of the almost 900 substances are "phosphates" following the steps below. For now, simply follow the instructions. We will soon create and post another resource document explaining the background to finding substances in Reaxys.

1. Go back to Query builder (use the button **Back to Query builder**).
2. Click **Molecular Formula** icon above the workspace, and enter *P?** in the **Molecular Formula** search field that appears in the workspace. This instructs Reaxys to search for substances with any number of phosphorus atoms (?) and any number of any other atoms (*).
3. Click **Search Substances**.

As a result, we obtain a few "phosphates". Click the **Hit Data** to display some information on them.

Starting with a text search in Substance Records, we found some phosphates that have been tested/analyzed in groundwater(s). Although we have obtained only a few substances, the example illustrates some principles for text searches in fields of Substance Records.

Some examples to try

1. Enter *adenocarcinoma** into the Substance Basic Index

   Our answer set contains over 47,000 substances that have been tested, used (and so forth) in studies on adenocarcinomas. Click some of the **Hit Data** to see precisely the information searched.

   There are many ways to narrow down the substance hits, which we will discuss in another resource.

2. We are interested in substances for which various Magnetic Data have been excerpted and added to Reaxys (this example also appears in the resource "Finding property information in Reaxys").

   Locate the terms in the **Description** subfield. Use **Search properties** in the **Query builder** to find the Magnetic Data field, drag it into the workspace and then click **Show fields**). Choose a couple of the terms and search them.

   The point of this exercise is to help you become familiar with the general types of data points that Reaxys has excerpted for this property. (The teaching point here is that we can't know all the types of information in the over 500 million data points in Reaxys – so at times we may want to perform a general search to see what is present and then, having found the information we want, construct a more targeted search accordingly).

3. Enter *parr near hydrogenator* into the Reaction Basic Index.

   Select Reaction Basic Index in the Query builder, drag the field into the workspace, enter the search terms, click **Search** and select **Reactions**.

Our hit set contains over 1,800 reactions that have been conducted using a Parr hydrogenator (a special apparatus for hydrogenation reactions). Most of the Hit Data can be found in the **Experimental Procedure** field.

4. Find reactions that have been catalyzed by lipase enzymes.

    Hint: follow the workflow in the example above, but instead enter *lipase* as the search term.

    Some of the Hit Data is located under **Conditions** while other Hit Data are found under **Experimental Procedures**.

    Clearly we need to narrow down such a large number of hits, but we'll learn how to do that in a future resource.

5. Find substances that have been isolated from *Grevillea* species.

    Hint: Search in the **Isolation from Natural Product** field in **Query builder**. The following screenshot is the final query.
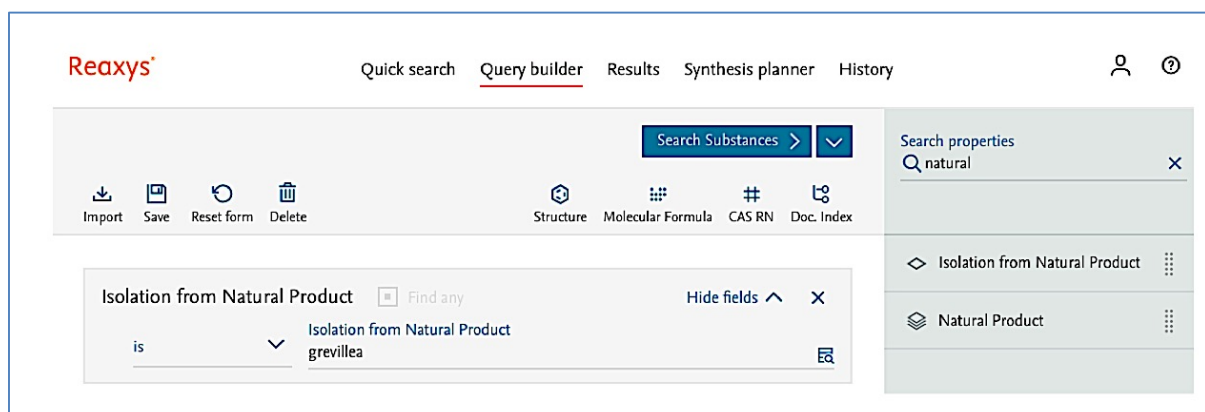
Summary

Text searches can be conducted in Document, Substance, and Reaction Records in Reaxys.

Text searches should consider synonyms, truncation (left- and right-truncation), Boolean and proximity operators, which are easily mastered. Synonyms can be difficult, but researchers generally know synonyms used in their field.

Generally, text searches in **Quick search** can best done by entering terms and prepositions only, or by using truncation and Boolean/proximity operators without any prepositions.

In depth text searches in all data fields of Reaxys can be performed in **Query builder**. The **Lookup** icon can help identify terms in the field.

Text searches can be combined with other search terms (e.g., by (sub)structure, formula or numeric terms).