



Text Analytics User Guide: Industry

Legal Notice

The information in this document does not constitute legal advice. Usage of the information remains under the sole responsibility of the user. The European Chemicals Agency does not accept any liability with regard to the use that may be made of the information contained in this document.

Title: Text Analytics User Guide: Industry

Issue date: July 2023

Language: en

Text Analytics is developed by the European Chemicals Agency.

IUCLID 6 is developed by the European Chemicals Agency in association with the OECD.

© European Chemicals Agency, 2023

Reproduction is authorised provided the source is fully acknowledged in the form

“Source: European Chemicals Agency, <http://echa.europa.eu/>”, and provided written notification is given to the ECHA Communication Unit (publications@echa.europa.eu).

If you have questions or comments in relation to this document, please send them to ECHA via the information request form at the address below, quoting the reference and issue date given above:

<http://echa.europa.eu/contact>

European Chemicals Agency

Mailing address: P.O. Box 400, FI-00121 Helsinki, Finland

Visiting address: Telakkakatu 6, FI-00150 Helsinki, Finland

Changes to this document

Date	Modification
07/07/2023	TA version 3.11.3. For a summary of changes, see the release notes on the IUCLID website.
13/03/2023	TA version 3.10.1. For a summary of changes, see the release notes on the IUCLID website.
02/02/2022	TA version 3.8.0.
13/12/2021	TA version 3.7.2.
19/08/2020	TA version 3.4.0.
17/06/2020	TA version 3.3.0.
18/12/2019	First official version for industry. TA version 3.2.1.

Table of Contents

Changes to this document	i
Table of Contents	i
Table of Figures	iii
Table of Tables	iv
1. Introduction	1
1.1. Ways to search in Text Analytics	1
2. Search for a phrase – level 1	2
3. Technical aspects of searching	4
3.1. Indexing	4
3.1.1. Whitespace/standard analyser	4
3.1.2. Language analyser	4
3.1.2.1. Case insensitivity	4
3.1.2.2. Language detection	5
3.1.2.3. Stemming	5
3.1.2.4. Stop words	6
3.1.2.5. Characters that are neither letters nor numbers	6
3.1.2.6. Numbers	6
3.2. IUCLID field	7
3.3. Attachment	7
3.4. Searching for words where word order matters	8
3.5. Search options	8
3.5.1. Search scope	9
3.5.2. IUCLID User groups	9

3.5.3. Highlight	10
3.5.3.1. Total highlights	11
3.5.3.2. Highlights length	11
3.5.4. Only latest dossier submissions	11
3.5.5. Collapse optional metadata	12
4. Search results	12
4.1. Relevance	12
4.2. Search results – The flat view (default)	12
4.3. Search results - More options	15
4.3.1. Search results – More options - Hierarchy	16
4.3.2. Search results – More options - Document	17
4.3.3. Search results – More options - Dossier metadata	17
4.3.4. Search results – More options - Download (text)	17
4.3.5. Search results – More options - Download (original)	17
4.4. Search results – The aggregated view	17
4.4.1.1. Show flat view for a single dossier	20
4.4.2. Search options for the aggregated view	21
4.4.2.1. Aggregate by	21
4.4.2.2. Dossier metadata	21
4.5. Hyperlinks to dossiers in IUCLID	21
4.6. Saving and sharing search criteria	22
4.7. Export	22
5. Search using controls applied to the search terms – level 2	24
5.1. Requiring/Excluding terms	24
5.2. Boolean operators	24
5.3. Regular expression	25
6. Search using the query language of Text analytics – level 3	26
6.1. Searchable properties of fields, entities, and attachments	26
6.1.1. Searching properties with no language analysis applied	28
6.1.1.1. Case sensitive search	28
6.1.1.2. Search for a specific form of a word	28
6.1.1.3. Wildcards	28
6.2. joinLevel	29
6.2.1. joinLevel:parent	29
6.2.1.1. joinLevel:parent{entity.entity_type}	29
6.2.1.2. joinLevel:parent{entity.hierarchy}	29
6.2.2. joinLevel:relative	30
6.2.2.1. joinLevel:relative(desc:<N>)	30
6.2.2.2. joinLevel:relative(asc:<N>)	31
6.2.3. joinLevel:dossier	31
6.3. Searching with no stemming	31
6.4. Query builder	32
6.4.1. The property Label and its autocomplete function	35

6.4.2. Fields / Attachments	38
6.4.3. Containing document.....	38
6.4.4. Containing dossier.....	39
6.5. Search	41
6.5.1. Search: Automatic syntax suggestions.....	41
6.5.2. Search: Traffic light syntax indicator.....	43
6.5.3. Search: Error messages	46
7. Worked examples of searches.....	47
7.1. Simple keyword search - level 1	47
7.2. Intermediate searching - level 2.....	47
7.3. Search with the query language of text analytics - level 3	48
7.3.1. Example using only the Query builder	48
7.3.2. Example of manual refinement of a query generated in Query builder	51
8. Getting help	52
Appendix A. Quick training exercise on query syntax	53
Appendix B. What to do if the search results are unexpected	55
Appendix C. Advanced topics	55
Appendix D. Queries obsolete from v3.8.0 onwards.....	56

Table of Figures

Figure 1: Example of a simple search of level (1)	2
Figure 2: Accessing an attachment from an endpoint study record.....	7
Figure 3: The search options under the menu when the type of view is the default "Flat"	9
Figure 4: IUCLID User groups – Instance Based Security (IBS) off (left), on (right)	10
Figure 5: Your search has been performed on user groups different than the default ones.....	10
Figure 6: Highlights of a search result.....	11
Figure 7: Example of properties shown for a search result in the default (flat) view for a IUCLID attachment.....	13
Figure 8: More options in the search result.....	15
Figure 9: Example output of the option Hierarchy for a search result	16
Figure 10: Selecting either the flat or aggregated views for search results.....	18
Figure 11: Search results in the aggregate view, aggregated by dossier UUID	19
Figure 12: Switch to the flat view for a single dossier.....	20
Figure 13: Go back to the aggregate view after viewing flat results for a single dossier	20
Figure 14: Aggregated view - options.....	21
Figure 15: Export query results.....	22
Figure 16: Controlling what is exported.....	23
Figure 17: Inserting NOT from the suggestion of functions.....	25
Figure 18: Opening and closing the list of Searchable properties	26
Figure 19: Example of where to find a searchable property – inventory description.....	27
Figure 20: Open Query builder	32
Figure 21: Managing criteria in Query builder.....	33
Figure 22: Example of creation of Boolean logic: A AND (B OR C).....	34
Figure 23: The Query window in Query builder.....	34

Figure 24: Search in a drop-down menu	35
Figure 25: The in-line help for properties.....	36
Figure 26: Selecting the label of a field	37
Figure 27: An example of a criterion defined under Field / Attachment	38
Figure 28: An example of a criterion defined under Contained document	39
Figure 29: An example of a criterion defined under Containing dossier	40
Figure 30: Filter for all REACH registrations.....	40
Figure 31: Open the tab for Search.....	41
Figure 32: Building a search criterion from the automatic suggestions in the Search.....	42
Figure 33: Search for an entity by the Working context (submission_type) of the containing dossier	43
Figure 34: A red indicator of incorrect syntax, and where the break in the syntax is located	43
Figure 35: A green indicator of correct syntax	44
Figure 36: Traffic light indicator in combination with the auto-suggest feature.....	45
Figure 37: Error message when searching through a red traffic light	46
Figure 38: Using autosuggest to correct queries	47
Figure 39: Location of the properties: value, other_text, and remarks_text in the web interface of IUCLID.....	48
Figure 40: Example of Query builder	50
Figure 41: Example of refinement of a query using a regular expression in the label field	51
Figure 42: How to direct a question to the IUCLID helpdesk	52

Table of Tables

Table 1: Detectable languages	5
Table 2: Language analysis for numbers	6
Table 3: Supported file types of IUCLID attachments	7

1. Introduction

Text analytics (TA) is a search engine that searches the database of an installation of IUCLID. TA provides complementary search functionality to that available via the standard IUCLID user interfaces. A search can read nearly all data that a user can enter in to IUCLID. This includes data in IUCLID fields, and the content of attachments. The search results state where within the instance of IUCLID a match was found, and contain hyperlinks that open a user-interface of IUCLID at the search result. TA has been designed to return search results quickly enough to allow users to browse the data, honing search criteria as they go along. The search criteria accepted by the TA range in complexity from simple keywords, right up the use of the TA's own query language which is used when highly specific information needs to be selected from large databases.

The user interface of *Text analytics* is viewed using a web-browser. The default address is shown below:

`http://<address of IUCLID>/ta`

If you do not know the address of IUCLID, ask your local administrator.

1.1. Ways to search in Text Analytics

There are three main levels of complexity of searches, as summarised below, with the simplest first:

1. Enter words that are related to the intended search results, just as you would in an internet search engine such as Google. This approach is used in the example shown below;
2. Enter a search term in text, but also include controls in the search criteria such as wildcards, Boolean logic, and/or groupings;
3. Use the custom query language of *Text Analytics*. Under the tab *Query builder* there are menus to help build a query.

If you need to do only a simple search, or you are not that familiar with the terminology and data structure of IUCLID, start with method (1). If you have more specific information about what you are looking for, for example you want to exclude certain search results, try (2). In that case it is recommended that you check the syntax described later in this manual. Finally, if you are looking for something highly specific, or you want to control exactly what type of data is found, you should create your own query using the query language of Text Analytics (3). Bear in mind that the query language is specific to *Text Analytics*, and requires the ability to construct queries manually by referring to the documentation in this manual. This process may be aided by the menus under the tab *Query builder* which can be used to create a query, or parts of a query, which can then be refined, and/or combined with each other.

The methods defined above (1-3) are described in detail in subsequent sections of this document.

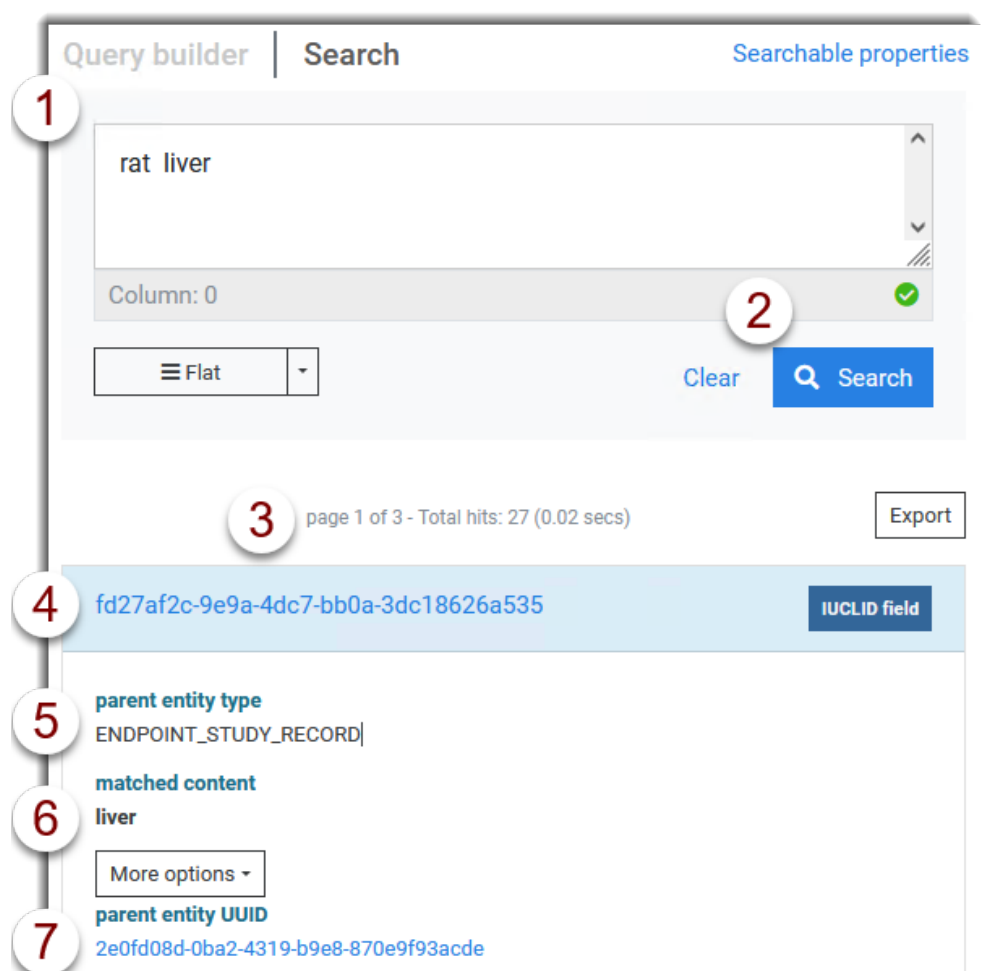
It is recommended to read at least the first three chapters of this manual. It will save you time compared to clicking around in the interface unaided. If you are only ever going to do simple level 1 searches, after reading section three, you can go straight to the examples in the section *7 Worked examples of searches*.

2. Search for a phrase – level 1

To get you started and to provide some context for reading the rest of the manual, this section contains an example of the simplest type of search (level 1). It also shows how to view search results in the user-interface of IUCLID.

First, enter a search term or terms into the box, and then click on the magnifying glass icon. Search results contain links that allow the IUCLID user-interface to be opened at the dossier or document in which a match was found. In the example shown below, the search terms are: `rat` `liver`. This will find either **rat** or **liver** in any order.

Figure 1: Example of a simple search of level (1)



Legend for Figure 1:

1. Enter search term(s) into the box under the tab *Advanced search*;
2. Click on the *Search* button;
3. The number of search results (hits) and the duration of the search;
4. The UUID of the *Dossier* that contains the search result. This is a link to the dossier in the web interface of IUCLID;
5. The type of document of the parent document of the search result. In this example it is an *Endpoint study record*.

6. The text that matched the search. If relevant there is a link to other matches in the same search result.
7. The UUID of the parent document of the search result. The UUID is a link to the parent document in the web interface of IUCLID.

3. Technical aspects of searching

Text analytics searches within only IUCLID dossiers in a IUCLID database and what they contain, such as attachments, entities, and documents. Therefore, for example, a *Substance dataset* entity that is in a IUCLID database but outside a dossier is not searched. However, a read-only copy of that same *Substance dataset* within a dossier is searched.

A search can be conducted either within the values of IUCLID fields, or within the text of attachments. The default is to search both, but this can be changed from the search options, as shown in Figure 3. Aspects of the two types of searching are explained in the subsections below.

3.1. Indexing

Text analytics does not search the IUCLID database directly, but searches an index. This is done to improve the efficiency, and therefore the speed of searching. Indexing is done periodically in the background. Changes in the IUCLID database take effect in search results only after the TA has been re-indexed. Whilst indexing is in progress, searches can take longer, and the results can change, reflecting changes in the IUCLID database. The date and time of the most recent indexing is stated in the user interface.

Text analytics uses analysers to create two indices. The first is created by a whitespace/standard analyser, whilst the second is a language analyser which carries out various transformations as described below. The index used in a search depends on how the search query is written and the type of data being searched. To find out how this works it is best to continue reading this document in order, and to consider the examples given later on.

3.1.1. Whitespace/standard analyser

Where the whitespace/standard analyser finds a space character, it cuts the text in to a substring, and then places it in the index. No other changes are made to the text. In a sentence such as this one, all the sub-strings would be words.

3.1.2. Language analyser

The language analyser also cuts text at space characters, but it then carries out various transformations designed to make search queries simpler and more intuitive.

3.1.2.1. Case insensitivity

Upper case characters are converted to lower case characters in the index. In searches where the language analysed index is used, this is also done to the search terms. This makes searching insensitive to case.

For example, if the original text contained the word "Toxicity", the index contains the form "toxicity". If the search term is either "Toxicity" or "toxicity", what is actually searched for is the lower-case form "toxicity", which matches the index.

3.1.2.2. Language detection

The language analyser tries to automatically detect the language in the indexed text using a set of rules that describe each language it knows. The detectable languages and their searchable codes are stated in the table below. If none of the languages in the table is detected, English (EN) is set by default. More than one language can be detected.

The codes of the languages detected in a document are stored in the index, and therefore constitute a searchable property of the document.

Table 1: Detectable languages

Language	Searchable code for the language
English	EN
German	DE
Spanish	ES
Danish	DA
French	FR
Swedish	SV
Finnish	FI
Greek	EL
Hungarian	HU
Italian	IT
Norwegian	NO
Portuguese	PT
Romanian	RO
Russian	RU

3.1.2.3. Stemming

The language analyser takes each word from the whitespace/standard analysed index and determines the base word from which it is derived, known as the stem. This process is known as *stemming*, and therefore the index is said to have been *stemmed*. This requires knowledge of the detected language, see above, and the lexical rules of that language, which are built into TA.

In searches where the language analysed index is used, this is also done to the search terms. This makes a search match all the forms of a particular word, whilst entering only one form into the search term.

For example, the stem of "rats" is "rat", so the search term "rats" matches "rat" and "rats". The search term "ratio" will not match "rat" because the language analyser knows that the detected language is English, and in English "rat" is not the stem of "ratio".

For example, the stem of "assessment" is "assess", so the search term "assessment" matches "assess", "assessment", "assessed" etc.

3.1.2.4. Stop words

The language analyser leaves certain words out of the index because they are considered as not containing information that adds to the usefulness of the search. These are referred to as *stop words*. In English they include some, but not all, conjunctions and prepositions. For example, and, "at", "to" are stop words, but "over" is not.

The stop words depend on the detected language. A stop word in one language is not necessarily a stop word in another language. For example, the word "to" is a stop word in English, but not in either German or Spanish.

For example, where the detected language is English, the search term `biodegradable water soluble` gives the same search results as `biodegradable` and `water soluble`.

3.1.2.5. Characters that are neither letters nor numbers

The standard analyser cuts the text at whitespaces, but then the language analyser goes further and removes characters that are neither letters nor numbers. This means that language analysed text cannot be searched for most non-alphanumeric characters. For example, searching for `15% saline` returns the same hits as `15 saline`.

There is an exception for full-stop and comma, as described below.

3.1.2.6. Numbers

Numbers are split at the white space character in the same way as letters. The characters that represent a decimal point, full-stop and comma, are not removed if they are bounded by a number on both sides. If the comma has been used as a thousand separator, it is left in place and must be searched for separately. For example, to find all instances of standard atmospheric pressure would require the following:

`101325 101,325`

The behaviour is summarised in the table below.

Table 2: Language analysis for numbers

Search term	Word after language analysis
<code>3.14</code>	<code>3.14</code>
<code>3,14</code>	<code>3,14</code>
<code>3.</code>	<code>3</code>
<code>0.3</code>	<code>0.3</code>
<code>.3</code>	<code>3</code>
<code>3.0</code>	<code>3.0</code>
<code>101,325</code>	<code>101,325</code>

3.2. IUCLID field

IUCLID fields are the places in the IUCLID user-interface into which users can enter data. For example, they include free text fields, pick lists, numbers, checkboxes, dates, and numeric ranges.

The index created by the language analyser in Text analytics classifies IUCLID fields as being one of the following three types:

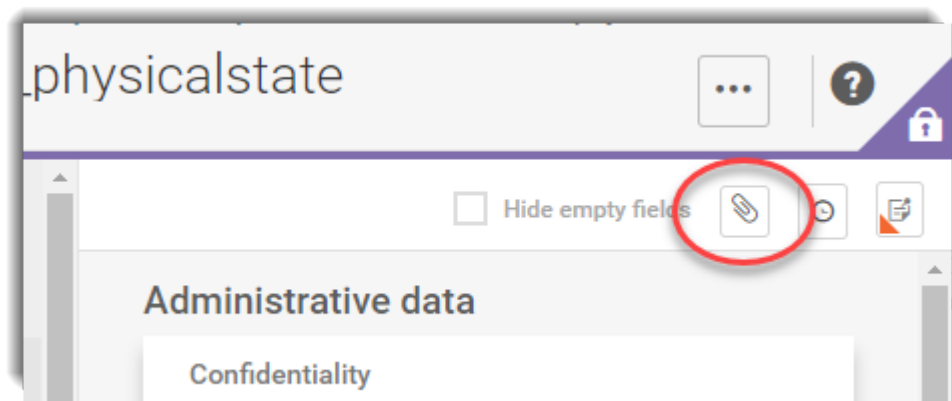
1. Free text
2. Pick list
3. Range, date, number, checkbox etc.

Language detection is done for only the first two types of fields. For the third type, the language English is set by default. For fields of type range, the characters in the field are concatenated into a single string.

3.3. Attachment

A IUCLID attachment is a file that is attached to a IUCLID entity/document. An example is shown below of an attachment that is attached to an endpoint study record. The attachment(s) for the record are accessed by clicking on the icon ringed in red.

Figure 2: Accessing an attachment from an endpoint study record



Text analytics extracts text from IUCLID attachments, and then indexes it using the language analyser. The types of files that can be processed are listed in the table below.

Table 3: Supported file types of IUCLID attachments

File type or application	File extension
PDF	pdf
Microsoft Word	doc, docx
Microsoft Excel	xls,xlsx
HTML (web page)	htm, html
RTF	rtf

In the case of PDF files, if the file contains a scan, Text analytics extracts the text using optical character recognition (OCR). The scan must have a resolution of least 300 dpi.

The OCR text, and the original attached file, can be downloaded from a search result via the menu More options.

3.4. Searching for words where word order matters

If a phrase is entered, any of the words can match, and in any order. To search for words in a particular order, place the whole phrase in double quotes.

3.5. Search options

The menu for search options is accessible via the black arrowhead icon indicated in the figure below. These options affect all searches.

Figure 3: The search options under the menu when the type of view is the default "Flat"

The screenshot shows the 'Search' tab in the 'Query builder' interface. A dropdown menu is open for the 'Flat' view, showing various search options. The options include:

- Search scope**
 - ☒ IUCLID fields
 - ☒ IUCLID attachments
- IUCLID User groups**
 - Common
- ☒ External content
- Total highlights**
 - 3
- Highlight length**
 - 50
- ☒ Only latest successful file
- ☒ Asset status active
- ☐ Collapse optional properties

Buttons at the bottom of the menu: OK, Cancel, Reset.

3.5.1. Search scope

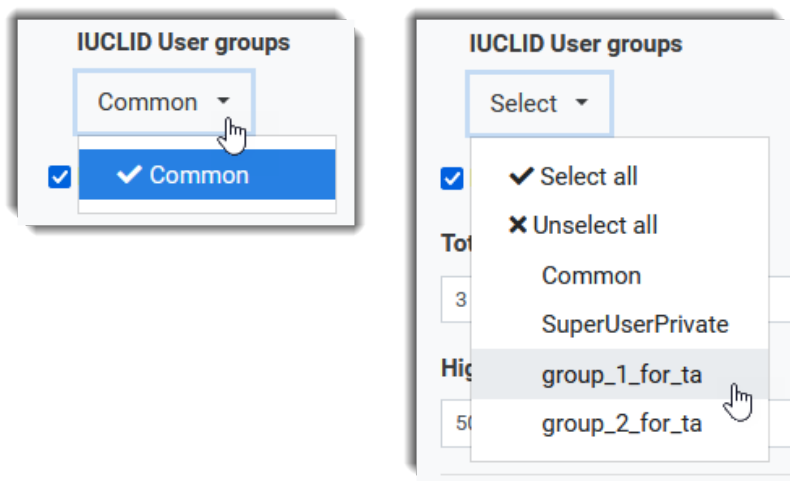
Select whether the search looks for matches inside fields, attachments, or both at the same time.

3.5.2. IUCLID User groups

This feature has an effect only if security groups are in use in the instance of IUCLID to which *Text Analytics* is connected. This feature of IUCLID is known as *Instance Based Security (IBS)*. In TA, it limits search hits to those within dossiers that have been shared with read access to the selected security groups. For management of security groups, contact your system administrator.

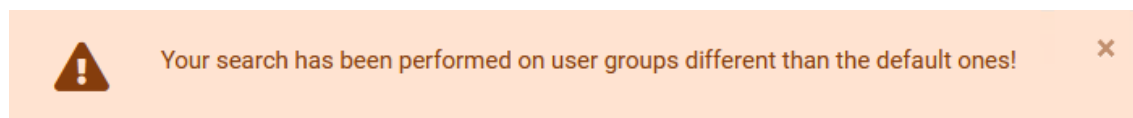
If security groups are *not* in use in IUCLID, there is still a default group named *Common*, which can be ignored. In the examples below, on the left IBS is *not* in use, and on the right, it is in use where the user of TA has access to the groups: *Common*, its private group, and two other groups that are defined in IUCLID.

Figure 4: IUCLID User groups – Instance Based Security (IBS) off (left), on (right)



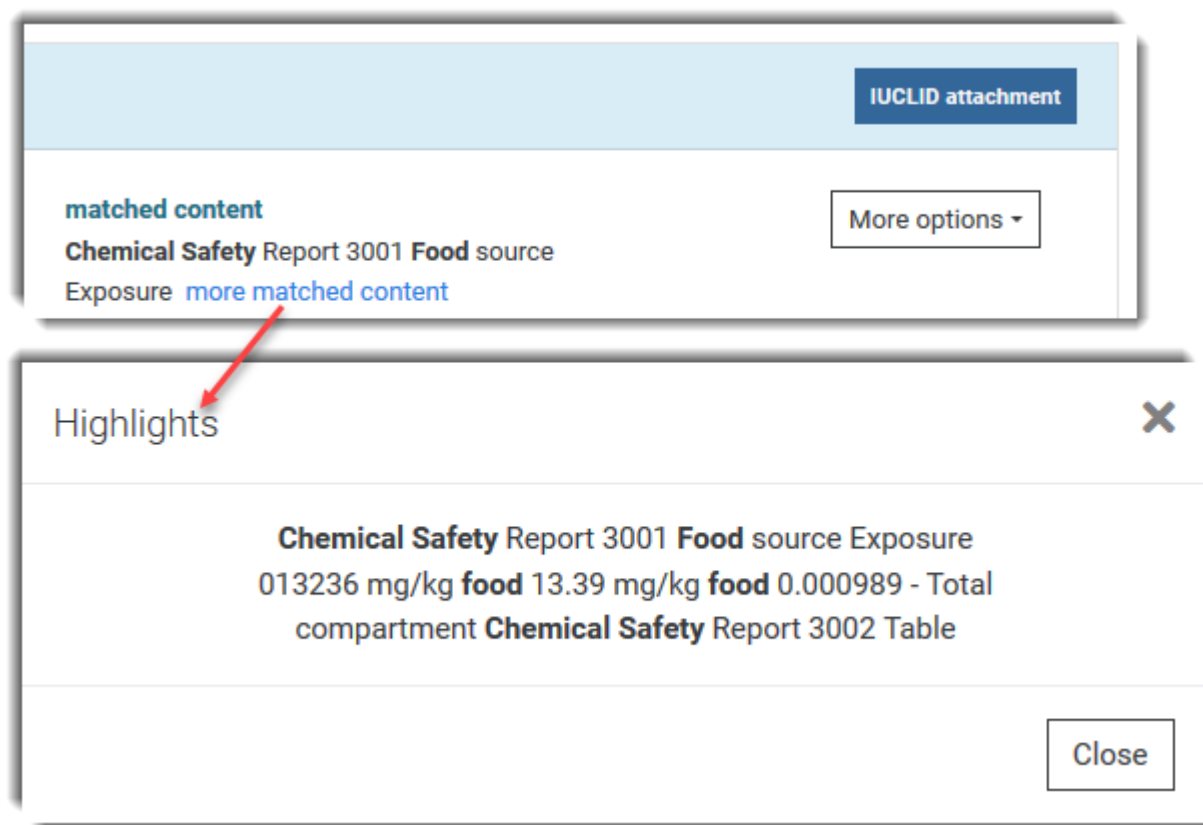
If the following message is displayed, contact the administrator of the installation of TA. This can happen even where IBS is not in use. There are instructions in the installation manual on how to prevent it from appearing.

Figure 5: Your search has been performed on user groups different than the default ones



3.5.3. Highlight

A highlight is a short piece of text that surrounds the matching text of a search. Matching words are shown in bold. A search can be set up to return one or more highlights per search result. The first highlight is shown in the search result in the field *matched content*. If more than one highlight is available, the complete set can be viewed by clicking on the link *more matched content*. An example is shown below in which a simple search was carried out in attachments for the phrase *Appraisal of the Safety of Chemicals in Foods*.

Figure 6: Highlights of a search result

The first highlight is " **Chemical Safety** Report 3001 **Food** source ". Clicking on the link *more matched content* shows all the highlights.

Note that in this example the highlights contain instances of the words "Chemical" and "Food" with no "s" on the end, even though the search phrase contained only the plural versions. This happens because, the search engine searches for the stems of the words in the search phrase. For more information about that behaviour, see section 3.1.1.

3.5.3.1. Total highlights

The maximum number of highlights that are displayed (0-10). A value of zero displays all the text within a search result. In the example shown above in Figure 6, the value is 3.

3.5.3.2. Highlights length

The number of characters per highlight (20-500). In the example shown above in Figure 6, the value is 30.

3.5.4. Only latest dossier submissions

This can be ignored by industry users because it has no effect on either the searching, or the display of search results. The value is set to "true" for all dossiers.

3.5.5. Collapse optional metadata

Use this to save space in the display of the search results, by hiding some of the properties that are shown. If required, the values of the properties can easily be unhidden from within the search results.

4. Search results

A search result is created for every unique match between the search criteria, and either the field or attachment that matches. Therefore, there can be more than one search result per dossier, or document. Search results are shown in a paginated list, ordered by relevance. A brief description of relevance is given below. Each entry in the list is surrounded by a blue border, and has a blue header that contains the UUID of the matching dossier, which is linked to the instance of IUCLID searched by *Text analytics*. There are two formats in which search results can be displayed:

- 1) **Flat (default):** An individual search result is shown for every match between the search criteria and a dossier. If there is more than one match for a particular dossier, the dossier appears in multiple entries in the list.
- 2) **Aggregated:** Search results are grouped by dossier, so that there is only one entry in the list per dossier.

It is recommended to first familiarize yourself with the *flat* view before progressing to the *aggregated* view.

4.1. Relevance

Results are returned in descending order of relevance. Each search result is given a numerical score to express how relevant it is to the search query. The scoring takes the following factors into account:

1. Term frequency: How often does the term appear in the text? The more often, the more relevant. A text containing five mentions of the same term is more likely to be relevant than a text containing just one mention.
2. Inverse document frequency: How often does each term appear in the index? The more often, the less relevant. Terms that appear in many documents have a lower weight than less common terms.
3. Field-length norm: How long is the text? The longer it is, the less likely it is that words in the text will be relevant. A term appearing in a short title text carries more weight than the same term appearing in a long content.

4.2. Search results – The flat view (default)

Every search result is shown in the user interface in its own box with a pale blue border and header. The various properties displayed for a search result are described in an annotated example below in Figure 7. The name of a property is shown in a pale grey bold font beneath its value. The width of the interface affects the location of some fields. To see the layout as it is shown below in Figure 7, set the width of the browser page to match that of the figure.

Figure 7: Example of properties shown for a search result in the default (flat) view for a IUCLID attachment

1 6a6c6cb2-1941-4c49-a6ec-15677b228091 2 IUCLID attachment

3 **parent entity type**
FLEXIBLE_RECORD

4 **matched content**
Chemical Safety Report 3001 Food source Exposure [more matched content](#)

5 More options ▾

6 **parent entity UUID**
dc4031fa-e729-4beb-a90e-11f50df7d54d

7 **filename**
577-11-7_CSR_updateFeb2017_complete_3001-end.pdf

8 **label**
REACH Registration above 1000 tonnes->13 Assessment reports->13.1
Chemical Safety Report (CSR).Chemical Safety Report (CSR).Chemical safety
report (CSR)

9 **size**
10.2 MB

10 **Detected language**
EN

11 **order in section**
Not available

12 **IUCLID type**
Single file attachment

13 **IUCLID path**
FLEXIBLE_RECORD.ChemicalSafetyReport.CSR.CSRDocument

14 **file type (MIME)**
application/pdf

15 **optical character recognition**
false

16 Optional metadata ^

17 **All languages**
EN(0.86),DE(0.14)

Key to Figure 7:

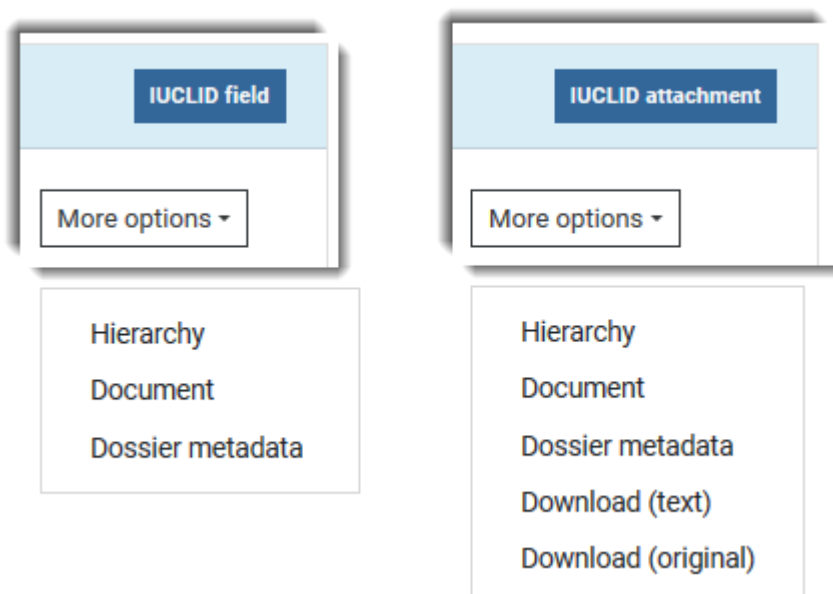
No.	Name	Remark
1	Dossier UUID	This is the UUID of the dossier that contains the IUCLID parent document of the search result. It is a hyperlink to the instance of IUCLID that contains the document. Following the link opens a IUCLID client window that displays the dossier.
2	Type of data source	The value of this is <i>IUCLID field</i> for a field in IUCLID, or <i>IUCLID attachment</i> for a file that is attached anywhere within a dossier. For <i>IUCLID attachment</i> , specific properties are shown in the search result, as indicated below by an asterisk*.
3	parent entity type	The type of IUCLID entity that contains the data source in (2). In the example above it is a <i>Substance</i> . For an attachment, if the parent entity type is DOSSIER, the attachment is attached to the <i>Dossier</i> header.
4	matched content / show value	A highlight of the text, or the value of the field, that matched in the search. The link <i>More matched content</i> points to the highlights of the search result. Highlights are explained in section 3.5.3.
5	More options	These are options for how to display the search result, and related document(s). The options are described below in section 4.3 <i>Search results - More options</i> .
6	parent entity UUID	This is the UUID of the IUCLID parent document of the search result.
7	filename*	The filename of the attachment.
8	label	The name of the IUCLID field that contains the data which matched the search criteria. If the match is an attachment that is not in a field but is attached to a whole document, the value is "Attachments". That is the case in the example shown above. For another example, see section 3.3.
9	size*	The file size of the attached file.
10	Detected language	Text Analytics tries to automatically detect the language in the text being searched using a set of rules that describe each language it knows. This is the most prevalent language detected in the document that contains the search result. See entry 15 below.
11	Order in section	Some flexible records in IUCLID can contain more than one data block. <i>Order in section</i> is the position of the particular matching data block down a list of several blocks. The third one down would have a value <i>Order in section</i> = 3.

12	IUCLID type	The data type in IUCLID for the matching value. Examples include: "Text area", "Open list", "Half-bounded with closed list decimal".
13	IUCLID path	The path in IUCLID to the field that contains the matching value. The IUCLID path can be displayed within the classic IUCLID interface by pressing the function key F12. It is not available in the current web interface.
14	file type (MIME)*	The media type (MIME) of the attached file.
15	optical character recognition*	This indicates whether Optical Character Recognition (OCR) was used on an attachment.
16	Optional metadata	This toggles the display of properties shown optionally at the bottom of the search result. The default can be set under the search options described in section 3.5 <i>Search options</i> .
17	All languages	This property is shown only for search results that are a <i>IUCLID field</i> . This is a list of the languages detected in the search result, with the fraction of the total stated in parentheses after the language code. For example, half English and half Finnish is EN(0.5)FI(0.5).

4.3. Search results - More options

Under the button labelled *More options* there are additional options for how to display the search result, and the related document(s). Some options are presented only when the search result is an attachment, as shown in the figure below. The box with the blue background at the top right of a search result indicates whether the search result is a *IUCLID field*, or a *IUCLID attachment*.

Figure 8: More options in the search result

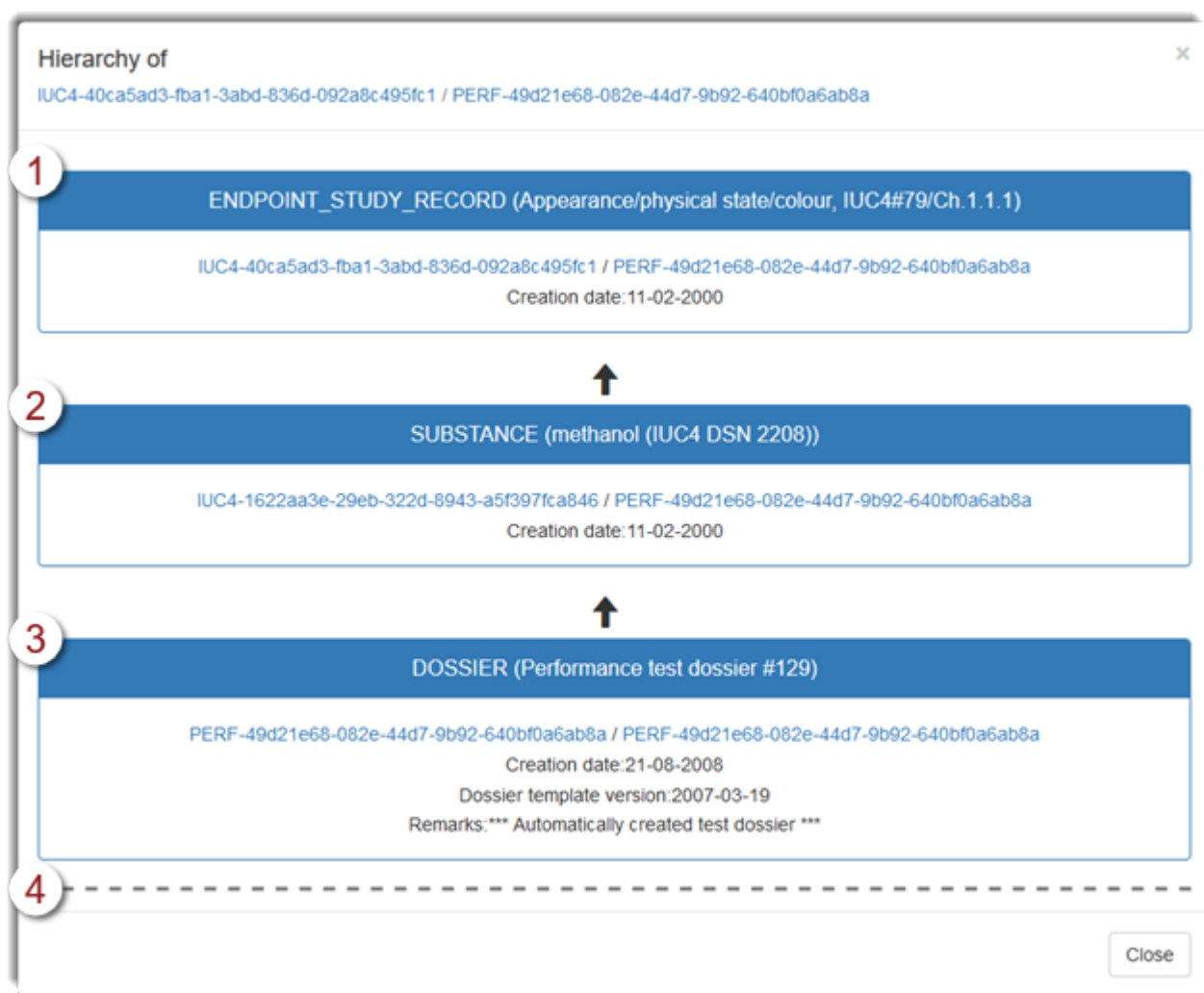


4.3.1. Search results – More options - Hierarchy

This option shows where the parent document of the search result resides in the IUCLID document hierarchy. It shows the documents that refer to it; even if indirectly. The hierarchy is presented with the search result at the top of the window. The referring documents are shown progressing down the window, which is going up the IUCLID hierarchy. The highest level in the hierarchy is the dossier. Each unique path through the IUCLID hierarchy to the search result is shown. Multiple unique paths are separated by a dashed horizontal line. An example is shown in the figure below for a search result whose parent document is an *Endpoint study record*. The hierarchical structure is:

Dossier > Substance > Endpoint study record

Figure 9: Example output of the option Hierarchy for a search result



Key to Figure 9:

1. The parent document of the search result. This is a document that contained text in a field that was matched by the search criteria.
2. A IUCLID document that refers directly to the parent document. In this example it is a Substance referring to a Flexible record in IUCLID section 1.2 Composition.

3. At the foot of the item is the dossier that contains the search result.
4. A dashed line is shown at the foot of each unique path through the IUCLID hierarchy to the parent document.

The UUIDs of the documents are clickable hyperlinks to the instance of IUCLID that contains the documents. Following a link opens a IUCLID client window that then displays the document.

4.3.2. Search results – More options - Document

Displays the IUCLID parent document of the search result. The text of the IUCLID document is presented as a web page. Links in the document to another IUCLID document are presented as hyperlinks that when followed, open the referred document as a webpage.

4.3.3. Search results – More options - Dossier metadata

This is for the presentation of dossier metadata obtained from systems outside IUCLID. Only users of the ECHA production system have access to such data.

4.3.4. Search results – More options - Download (text)

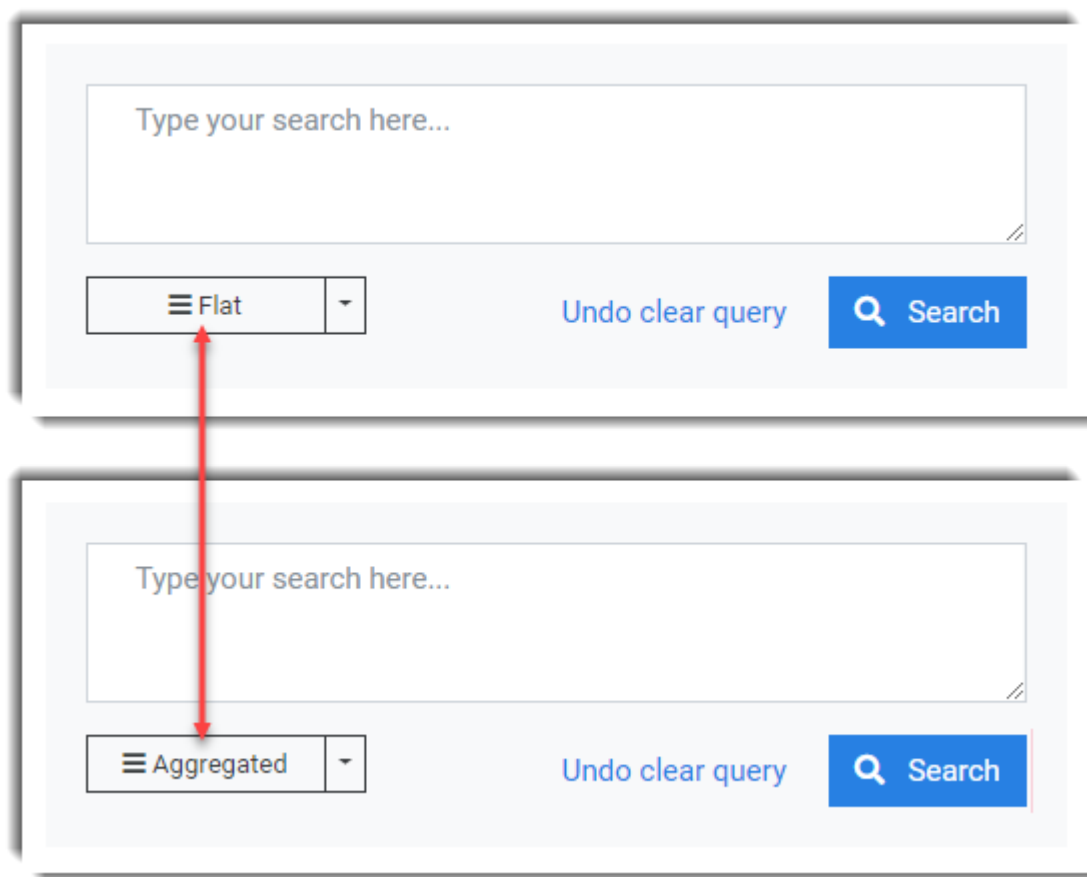
This applies only to a search result that is an attachment. It is the text that was extracted from the attachment either directly, or using OCR. If the text came from OCR, as indicated by the OCR property for the search result, this text may be used to check the accuracy of the OCR. For example, mistakes in the OCR may explain why the attachment does not match the search criteria it is expected to match.

4.3.5. Search results – More options - Download (original)

This applies only to a search result that is an attachment. It is the original of the attached file.

4.4. Search results – The aggregated view

The *Aggregated* view shows the search results grouped by the dossier that contains them. To switch from the default *Flat* view to the *Aggregate* view, click on the button indicated by the red arrow in the figure below. This button can also be used to switch back from the *Aggregate* to the *Flat* view.

Figure 10: Selecting either the flat or aggregated views for search results

In the *Aggregated* view, search results that occurred in the same dossier are shown in a box that contains the dossier UUID in the header, in blue. The UUID is a link to IUCLID. The search results are shown grouped within the IUCLID hierarchy as follows:

dossier > parent document (entity) > search result (matching field or attachment)

For each dossier, the number of each document and search result displayed within it is limited to two. An example of the aggregate view is shown below.

Figure 11: Search results in the aggregate view, aggregated by dossier UUID

1 [IUC5-37ee64ac-9f59-4139-ae73-7e54047e1a53](#) 2 Flat view

3 Dossier metadata

Latest submission flag
true

4 FLEXIBLE_RECORD 5 00fd0ffc-20b3-3c8c-84f7-613866a4ddbc

6 label matched content

REACH Registration above 1000 tonnes->2
Classification & Labelling and PBT
assessment->2.2 DSD - DPD.Labelling.Risk
phrases.Risk phrases

R62 - Possible risk of impaired fertility

7

8 FLEXIBLE_RECORD 02dc3748-80da-3f61-aa74-4e9ea713fceb

9 label matched content

REACH Registration above 1000 tonnes->2
Classification & Labelling and PBT
assessment->2.2 DSD - DPD.Labelling.Risk
phrases.Risk phrases

R62 - Possible risk of impaired fertility

10 [IUC5-abba9f61-8f37-47f9-a567-7296ce72cded](#) Flat view

Key to Figure 11:

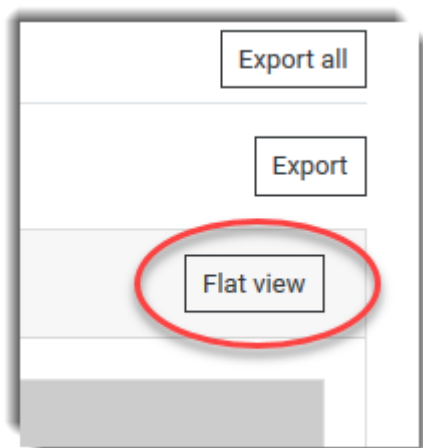
1. The dossier UUID. This is a hyperlink to IUCLID.
2. Switch to a flat view for only this dossier. For more information see section 4.4.1.1 below.

3. The dossier metadata. The fields shown here are determined by the search options described below in section 4.4.2.2 *Dossier metadata*.
4. The first document in the first dossier that contains a match. A maximum of two documents are shown per dossier.
5. The UUID of the IUCLID document. It is not a hyperlink.
6. The first matching field in the first document. A maximum of two fields are shown per document.
7. The second field that matches, in the first document.
8. The second document that contains a match.
9. The second field that matches, in the first document.
10. The second dossier that contains a match.

4.4.1.1. Show flat view for a single dossier

All the search results within a particular dossier can be shown in the flat view by clicking on the button labelled *Flat view*, in the right-hand side of the header for the dossier, as can be seen in the figure below.

Figure 12: Switch to the flat view for a single dossier



To go back to the aggregate view for all the dossiers in the original search, click on the link *Back to aggregate view*, which is at the upper left of the interface, as shown below.

Figure 13: Go back to the aggregate view after viewing flat results for a single dossier

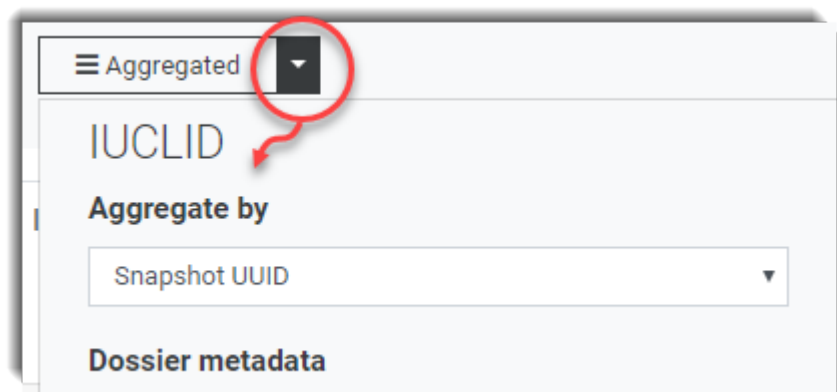


The metadata *Latest submission flag* can be ignored by industry users of Text analytics. It is always set to "true".

4.4.2. Search options for the aggregated view

Both the *Flat* and *Aggregated* views have a set of options that are accessible under the black arrowhead icon to the right of the toggle button. The options specific to the *Aggregated* view are shown below.

Figure 14: Aggregated view - options



4.4.2.1. Aggregate by

This is used to group the search results by the value of the property selected. *Snapshot UUID* groups search results by the dossier UUID. Industry users of *Text analytics* should not use the option *Latest submission flag*.

4.4.2.2. Dossier metadata

This is used to determine which properties are shown in the search results for the aggregated view.

4.5. Hyperlinks to dossiers in IUCLID

You may want to share with a colleague, a link from TA to a Dossier in IUCLID. However, depending on how IUCLID and Text Analytics are set up, the link may not work if it is used from a different machine or browser. If you come up against this limitation, first follow the link to the web interface of IUCLID, and then share the address that is in the address field of the browser. The general format of the address is:

```
https://<address of iuclid server>/iuclid6-web/browser/dossier/  
<UUID of Dossier>/SUBSTANCE/<UUID of Substance>
```

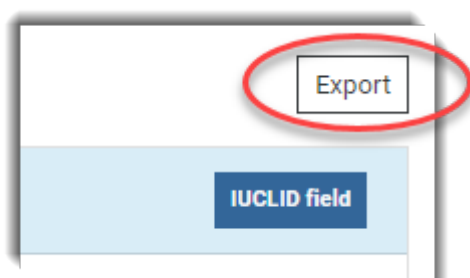
4.6. Saving and sharing search criteria

The search criteria used in a search can be saved as a bookmark in the browser used to view the interface. Simply run the search, and then create a bookmark. Thus, the bookmark functionality of the browser can be used to save and share searches.

4.7. Export

To export search results, clicking on the button labelled *Export* as shown below:

Figure 15: Export query results



The output can be controlled for either fields or attachments, depending on whether the search was set up to find them. The export of some data is mandatory, whereas other options can be turned on and off using tick boxes. In the example below, the box for *value* is ticked, which outputs the highlight that contains the matching text. The length of the highlight is set in the search settings before the search is carried out.

Figure 16: Controlling what is exported

Export Query Results

Fields

IUCLID Attachments

Select the data you would like to export

Field metadata

Select all

Unselect all

☒ Snapshot UUID

Maximum value length

☒ IUCLID path

☒ Parent entity type

☐ Label

☐ Detected language

☒ Parent entity
UUID/Content group

☐ Order in section

☐ All languages

☒ Value

☐ IUCLID type

The export format is MS Excel. The data is organised in worksheets. The first worksheet, *Legend*, describes the potential full structure of an export document. Dependent on what was exported, not all of the worksheets described may be present. The search query is in its own worksheet.

Caution

When exporting data, consider carefully whether it is confidential. Be aware that a browser, dependent on how it is set up, may automatically save exported files to the default downloads directory, as set in the browser. If that is not an appropriate place for confidential data, ensure that it is securely deleted.

5. Search using controls applied to the search terms – level 2

If the search requires more than simple keywords or a phrase, consider using the additional functionalities described below.

5.1. Requiring/Excluding terms

By default, all terms are optional, as long as one term matches. To make a term required place a plus sign "+" in front of it. To exclude a term place a minus sign "-" in front of it.

For example:

```
analyses report
```

This finds any field or attachment that contains one or more occurrence of "analyses" or "report". All other terms are optional.

For example:

```
Test Study +report -chemical
```

States that:

1. "report" must be present,
2. "chemical" must not be present,
3. "Test" and "Study" are optional. Their presence increases the relevance top those terms. For a brief description of relevance, see *section 4.1 Relevance*.

5.2. Boolean operators

The Boolean operators available are AND, OR and NOT. They are grouped using parentheses "(" and ")". Note that they must be in upper case. Where no Boolean operator is used, OR is implied between words and phrases in double quotes. Limits are applied to the complexity of search terms that include Booleans and parentheses. A search expression cannot begin with a parenthesis. The following example is possible:

```
report AND ((test OR report) AND (study OR report)) AND NOT chemical
```

Boolean operators can be entered manually into the search box, or selected from the automatic suggestion of functions box. NOT can be inserted either at the start of a term, or just before a quotation character, as shown below.

Figure 17: Inserting NOT from the suggestion of functions

5.3. Regular expression

A description of the syntax and functionality of regular expressions is out of the scope of this document. This section describes only how regular expressions behave within TA search queries. A regular expression is placed between two forward slash characters, like this: `/<regular expression>/`

Regular expressions need be used only where the other functionalities described for level 2 searches cannot create the required search query. An example is when searching for a specific pattern of text, such as exactly any three letters followed by exactly any four digits. The regular expression would be:

```
/[a-z]{3}[0-9]{4}/
```

This would match "ABC1974" but not "report_ABC1974.pdf".

The pattern must match exactly, so to find the same sequence anywhere inside strings would require wild cards at both ends:

```
/.*[a-z]{3}[0-9]{4}.*/
```

This would match both "ABC1974", and "report_ABC1974.pdf". The searches in this example are not case sensitive.

Any Unicode characters may be used in the pattern, but reserved characters must be escaped using the backslash character "\". The standard reserved characters are:

```
. ? + * | { } [ ] ( ) " \
```

6. Search using the query language of Text analytics – level 3

The query language of *Text analytics* is used where more precision is required than the search criteria described so far. Typical cases are:

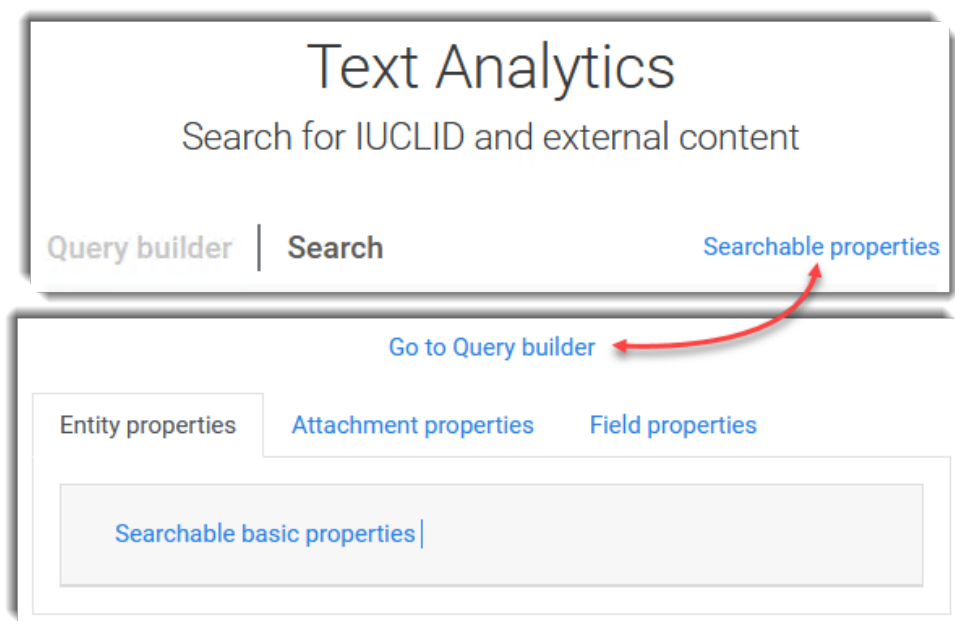
1. The values within only specific IUCLID fields are to be searched;
2. The parent document of the matches returned in the search should be only specific types of IUCLID entity, e.g. Reference substance;
3. The search results are to include IUCLID documents that are related to the parent document of the match, via the IUCLID hierarchy;
4. Searching in a language other than English is required.

The query language of *Text analytics* has its own syntax which is introduced in the next section. Phrases in the syntax can be entered into the search box under the tab *Search*, or created in a more automatic and deliberately limited way under the tab *Query builder*. The different ways that searches can be built under the tabs *Query builder* and *Search* are described in later sections.

6.1. Searchable properties of fields, entities, and attachments

The searchable properties within *Text Analytics* are documented under the link *Searchable properties* that is at the upper righthand side of the interface. To close *Searchable properties*, click on *Go to Simple Search*.

Figure 18: Opening and closing the list of Searchable properties



As can be seen above the listing of the properties is divided by type in to three tabs:

1. *Entity properties*: These refer to all of an entity or document that is searched;
2. *Attachment properties*: Properties of files that are attached to documents or entities,
3. *Field properties*: Properties of fields within documents or entities.

The meaning of an individual property and how it is used is indicated in the listing of properties. For brevity, this information is not duplicated in this manual. Some examples of *Field properties* are shown below.

Figure 19: Example of where to find a searchable property – inventory description

Entity properties Attachment properties Field properties		
Searchable Field properties		
iuclid_path	The unique identifier of the field or attachment as exists in the IUCLID definitions (see also F12 in IUCLID) e.g: field.iuclid_path:"FLEXIBLE_RECORD.MixtureComposition.Components.Components.TypicalConcentration" (KEYWORD)	Not analyzed
value_confidentiality	This is the confidentiality property of the confidentiality value. e.g: field.value_confidentiality:IP (KEYWORD)	Not analyzed
inventory_description	The description given for the inventory e.g: field.inventory_description:descr should return all the fields linked to an inventory with the specific description. (TEXT)	STANDARD
value	This is the property to search for the field value analyzed by a language analyzer. Note: the value is lower case stemmed and the punctuation is removed e.g: field.value:doses (TEXT)	LANGUAGE

The left-hand column contains the identifier of the property, shown in bold. In the centre there is a description of the property, an example of how it is used, and the data-type in parentheses. The data-type determines what type of value must be placed in the search criterion. The column on the right is the type of analysis applied, which can be *LANGUAGE*, *STANDARD*, or *Not analyzed*.

Note: Users of the production instance of TA at ECHA have extra types of searchable properties under entities and attachments. These are described in the documentation specific to ECHA.

A property is searched using a criterion of the general form shown below:

```
<identifier of property>:<search term>
```

The names of properties are case sensitive. The values of data-type KEYWORD are case sensitive.

Examples

Note that if the data type is KEYWORD, the search term must be a whole value that is defined either within TA or IUCLID. In the example below it must be one of the language codes in TA.

```
detected_language:"DE"
```

To find *Dossiers* that contain a *Substance* that has a particular UUID:


```
joinLevel:parent{entity.entity_type:"SUBSTANCE"} AND  
field.entity_uuid:"IUC5-18a7cdbe-73e0-4429-82d7-a499d4bb9892"
```

6.1.1. Searching properties with no language analysis applied

There are two exceptional searchable properties of type TEXT for which the language analyser, described in section 3.1.2 *Language analyser*, is not used. These are used in cases where language analysis prevents the required search from being carried out. For example, where the effects of stemming are to be avoided, or the search needs to be case sensitive. The properties are *value.std* for fields, and *content.std* for attachments. Note that these are equivalents of the properties *value* and *content* for which language analysis is performed. Some practical uses of not performing language analysis are given below.

6.1.1.1. Case sensitive search

Search in the values of fields for the uppercase word "REACH", ignoring the lowercase form "reach". This differentiates between the EU legislation REACH, and usage such as "did not reach the limit".

```
field.value.std:REACH
```

6.1.1.2. Search for a specific form of a word

Search in the values of fields for the word "assessment", ignoring other forms of the word such as "assess" and "assessed".

```
value.std:assessment
```

6.1.1.3. Wildcards

Using wildcards with language analysed text can produce unexpected results, and so it should be avoided. The idea of the language analyser is to make wildcards unnecessary.

The asterisk/star character * represents zero or more of any characters. It can be used to search for words that begin with a particular string but are not recognised by a language analyser,

e.g. To find all words that begin with "glutamyltrans" use:

```
field.value.std:glutamyltrans*
```

The question mark ? represents any single character.

e.g. to find "sulfur" but not "sulphur" use:

```
field.value.std:sul?ur
```

to find "sulphur" but not "sulfur" use:

```
field.value.std:sul??ur
```

6.2. joinLevel

joinLevel is used to specify the level in the IUCLID hierarchy at which certain search criteria are applied together. There are three types:

1. parent
2. relative
3. dossier

The syntax is:

```
joinLevel:<type>{ <criteria> }
```

6.2.1. joinLevel:parent

The search finds fields that are in a parent entity/document for which all the criteria in the curly brackets are true. In the example given below, the *joinLevel* term has been combined with a more general search criterion.

Example

```
field.value:nickel AND joinLevel:parent{field.label:/. *Discussion.*/ AND  
field.value:magnesium}
```

This query looks for fields for which:

1. The property *value* contains the word "nickel"
2. and the IUCLID parent document/entity contains a field that:
 - a. has a *value* that contains the word "magnesium"
 - b. and has a *label* that contains the word "Discussion"

6.2.1.1. joinLevel:parent{entity.entity_type}

Find a parent document by the type of entity. The syntax is:

```
joinLevel:parent{entity.entity_type:"<type of entity or document>"}
```

This type of criterion can be built up using the auto-suggest in the *Search* window, as described in section 6.5 *Search*.

Example

```
joinLevel:parent{entity.entity_type:"CONTACT"}
```

6.2.1.2. joinLevel:parent{entity.hierarchy}

Find a parent document by its location within a hierarchy The syntax is:

```
joinLevel:parent{entity.hierarchy:"<hierarchy>"}
```

The value of *hierarchy* can be:

```
DOSSIER.SUBSTANCE
```

```
DOSSIER.SUBSTANCE.ENDPOINT_STUDY_RECORD
```

```
DOSSIER.SUBSTANCE.TEMPLATE.ENDPOINT_STUDY_RECORD
```

```
DOSSIER.CATEGORY.SUBSTANCE
```

The value of hierarchy must be an exact match.

Examples

Find *Dossiers* of type UVCB that are in a *Category*.

```
field.label:/.*Type of substance.*/ AND field.value:"UVCB" AND
joinLevel:parent{entity.hierarchy:"DOSSIER.CATEGORY.SUBSTANCE"}
```

Find data that is in an endpoint study record that can be in a *Template* but is not in a *Category*.

```
field.label:/.*REACH Registration 10 - 100 tonnes.*/ AND
field.value:"chlorine" AND
joinLevel:parent{entity.hierarchy: ("DOSSIER.SUBSTANCE.ENDPOINT_STUDY_RECO
RD" OR "DOSSIER.SUBSTANCE.TEMPLATE.ENDPOINT_STUDY_RECORD") }
```

6.2.2. joinLevel:relative

This is useful for looking at relationships across multiple levels in the IUCLID hierarchy. There are two subtypes, one where its search criteria are applied higher in the IUCLID hierarchy (desc) and one where they are applied lower (asc). The number of levels searched can be controlled by passing an argument N which is an integer (2, 3, 4 ...). N = 2 searches the adjacent level, N=3 searches two levels, etc.

The OR operator cannot be used between *joinLevel:relative* clauses.

6.2.2.1. joinLevel:relative(desc:<N>)

This clause applies search criteria upwards in the IUCLID hierarchy.

In the example below, the first two criteria are applied to the values and labels of fields such that fields inside a *Reference substance* match. The clause *relative(desc:2)* is applied one level up the IUCLID hierarchy so the *Reference substances* must be in a composition in section 1.2 and the search hits are *Reference substances*.

Example

In the example below, the first criterion requires that the field label contains the value *Reference substance*. Therefore, the parent entities of the search hits will be *Reference substances*. The *Reference substances* must contain a field that contains both the values *isopentyl* and *ether*. The clause *relative(desc:2)* is applied upwards in the IUCLID hierarchy. The value of the fields must contain *1.2 Composition* so the *Reference substances* must be directly referred to by a *Composition* in section 1.2 of the *Dossiers*.

```
field.label:/.*Reference substance.*/
AND
field.value:*isopentyl* AND field.value:*ether*
```

AND

```
joinLevel:relative(desc:2){field.label:/.*1.2 Composition.*/}
```

6.2.2.2. *joinLevel:relative(asc:<N>)*

This clause applies search criteria downwards in the IUCLID hierarchy.

Example

In the example below, the first criterion requires that the field label contains the value 1.2 Composition. Therefore, the parent entities of the search hits will be *Compositions*. The clause `relative(asc:2)` is applied downwards in the IUCLID hierarchy. The value of the fields must contain *Reference substance* so the *Compositions* must contain *Reference substances* that also satisfy the other criteria; which are to have values that contain *isopentyl* and *ether*.

```
field.label:/.*1.2 Composition.*/
```

AND

```
joinLevel:relative(asc:2){field.label:/.*Reference substance.*/ AND  
field.value:*isopentyl* AND field.value:*ether*
```

6.2.3. *joinLevel:dossier*

The criteria must be true within the same dossier.

6.3. Searching with no stemming

By default, stemming is used, so `rat` finds both "rat" and "rats". To find only "rats", turn off stemming by using the property `value.std` for fields, and `content.std` for attachments.

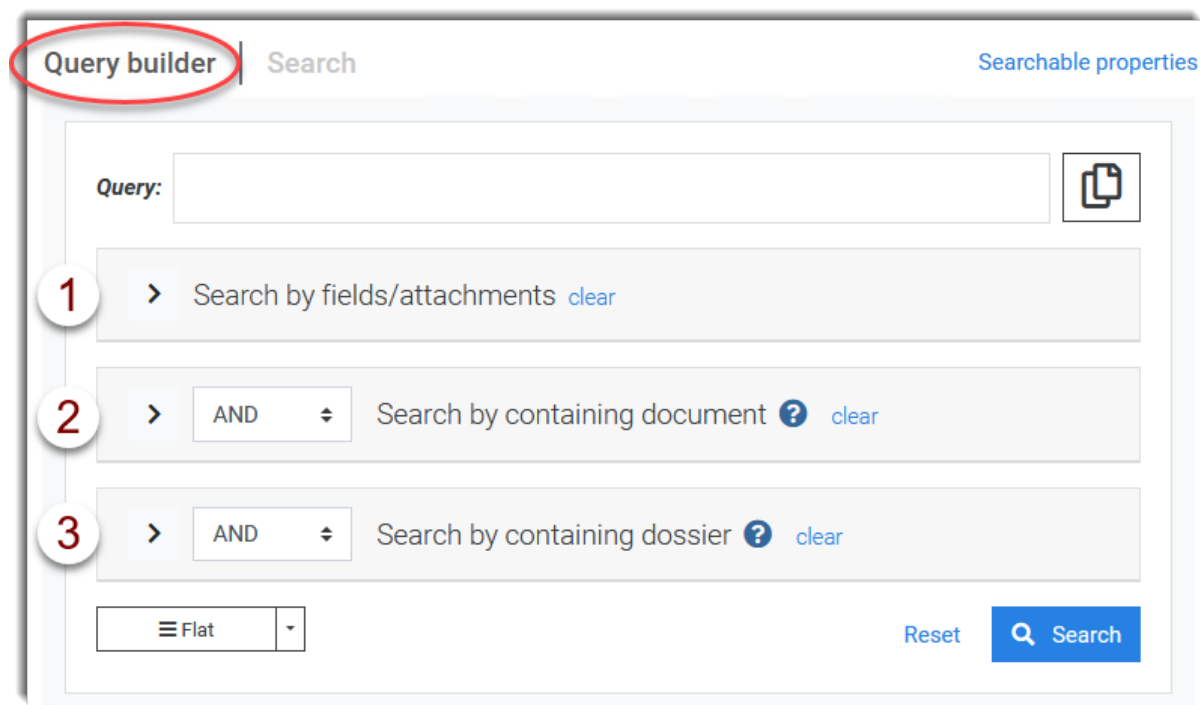
```
value.std:rats
```


```
content.std:rats
```

6.4. Query builder

Query builder is opened from its tab, as shown below.

Figure 20: Open Query builder



At the top of the page is the *Query* box which dynamically shows the overall query that is built from the individual search criteria defined in the sections below. The query cannot be edited in the box, but only from the structure beneath it. If the query needs to be refined beyond what is possible in *Query builder*, the query can be copied over to the *Search*, where it becomes editable. The copy feature will be described later in this document. It is accessed from the icon . A query cannot be transferred in the opposite direction.

The search criteria are divided in to three sections:

1. Fields/Attachments
2. Containing document
3. Containing dossier

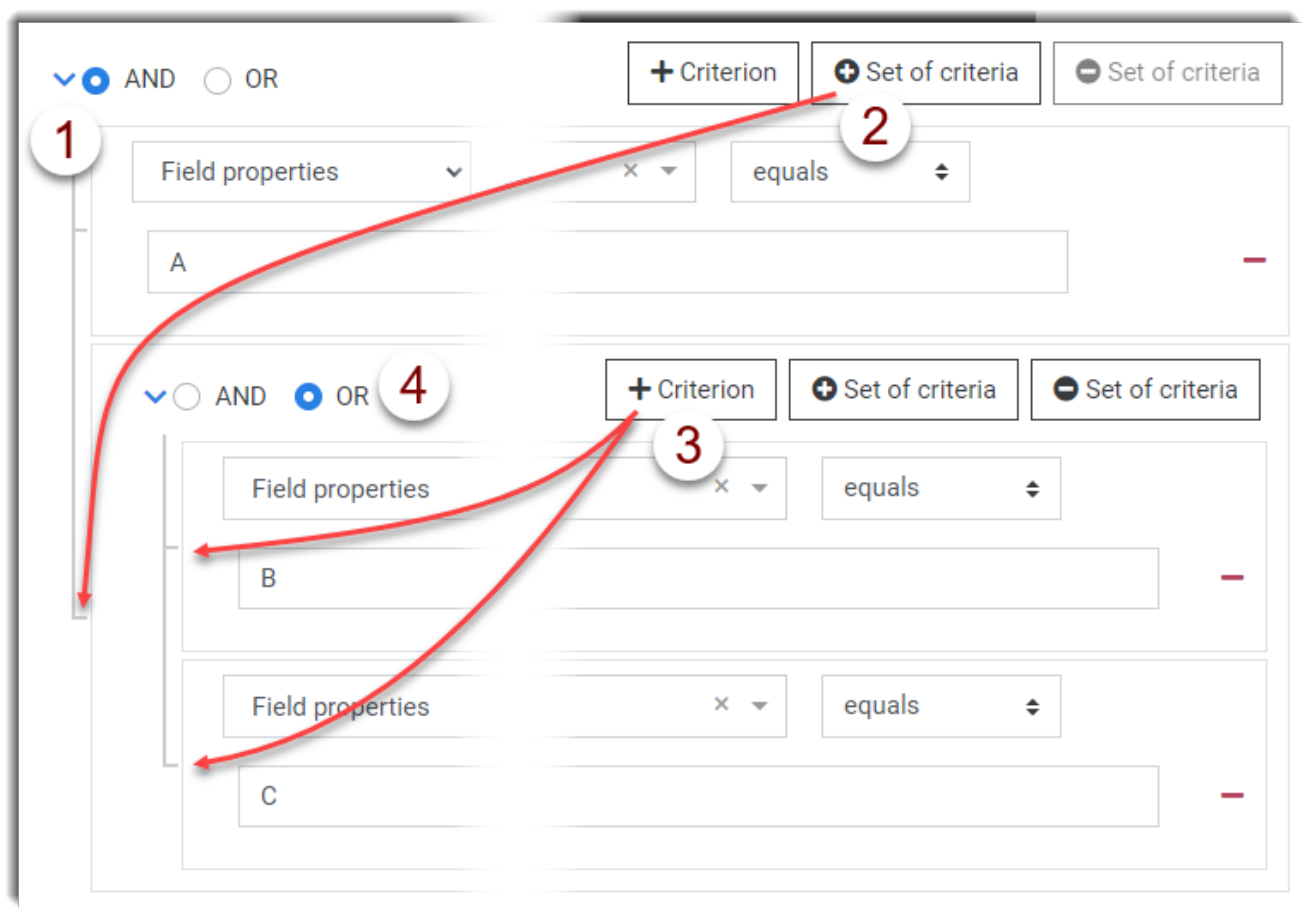
Features specific to each of the three sections are described in their own sub-sections later in this section, after more general features have been described. A section is opened or closed from the black arrow icon on its right. Criteria can be placed in sets which determine the locations of parentheses in the Boolean logic. A criterion is placed in a set on creation, by clicking on the button *+Criterion* at the top of that set. Sets are created using *+Set of criteria*. Sets can be placed within sets, but try to keep it simple. The features for building logic are introduced below, followed by an example.

Figure 21: Managing criteria in Query builder

The screenshot shows the 'Query builder' interface. At the top, there's a 'Query:' input field and a 'Searchable properties' link. Below this is a section titled 'Search by fields/attachments' with a 'clear' link. A black arrow (1) is at the top right of this section. Below the title, there are three buttons: '+ Criterion' (2), '+ Set of criteria' (3), and '- Set of criteria'. A Boolean operator selector (4) shows 'AND' selected with a blue arrow. Below this is a dropdown menu (5) labeled 'Field properties' with 'All languages' selected. A question mark icon (6) is next to it. To the right is a comparator dropdown (7) showing 'equals'. Below the dropdown is a text input field (8). A minus button (9) is at the bottom right of this section. At the bottom of the interface, there are two sections for 'Search by containing document' and 'Search by containing dossier', each with a Boolean operator dropdown (10) and a 'clear' link.

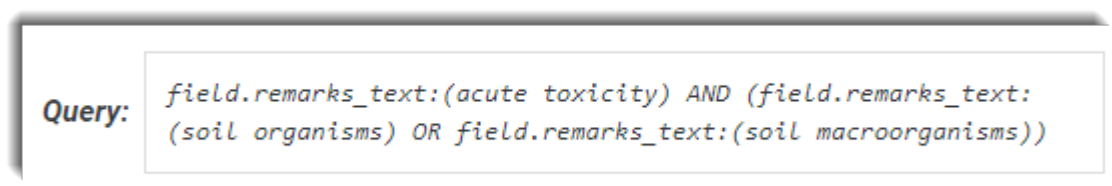
Legend for Figure 21

1. Open or close a section from the black arrow;
2. Add a criterion to the set which is immediately below the button;
3. Add or remove a set of criteria;
4. Define the Boolean operator between the criteria and/or sets of criteria. The blue arrow opens and closes the set of criteria;
5. Determine whether fields or attachments are searched;
6. Select from a drop-down menu, the property of a field or attachment that will be searched. A description of a property is available by hovering over its question mark icon . There is additional help under the link *Searchable properties* at the top right of the page;
7. Select a comparator. These depend on the data type of the property. The data types are given under *Searchable properties*;
8. Enter or select a value that the search must find;
9. Remove a criterion;
10. Define the Boolean operators between the three sections.

Figure 22: Example of creation of Boolean logic: A AND (B OR C)**Legend for Figure 22**

1. Start with a criterion for A. Leave the Boolean operator at the default, AND;
2. Add a set of criteria from A, which will contain B and C;
3. Add a criterion to the set for B;
4. Add a criterion to the set for C;
5. Change the Boolean operator in the set to OR.

Whilst building a query, check the logic from the *Query* window after each change is made. An example of a query with the structure A AND (B OR C) that searches for phrases contained in the field *Remarks* is given below:

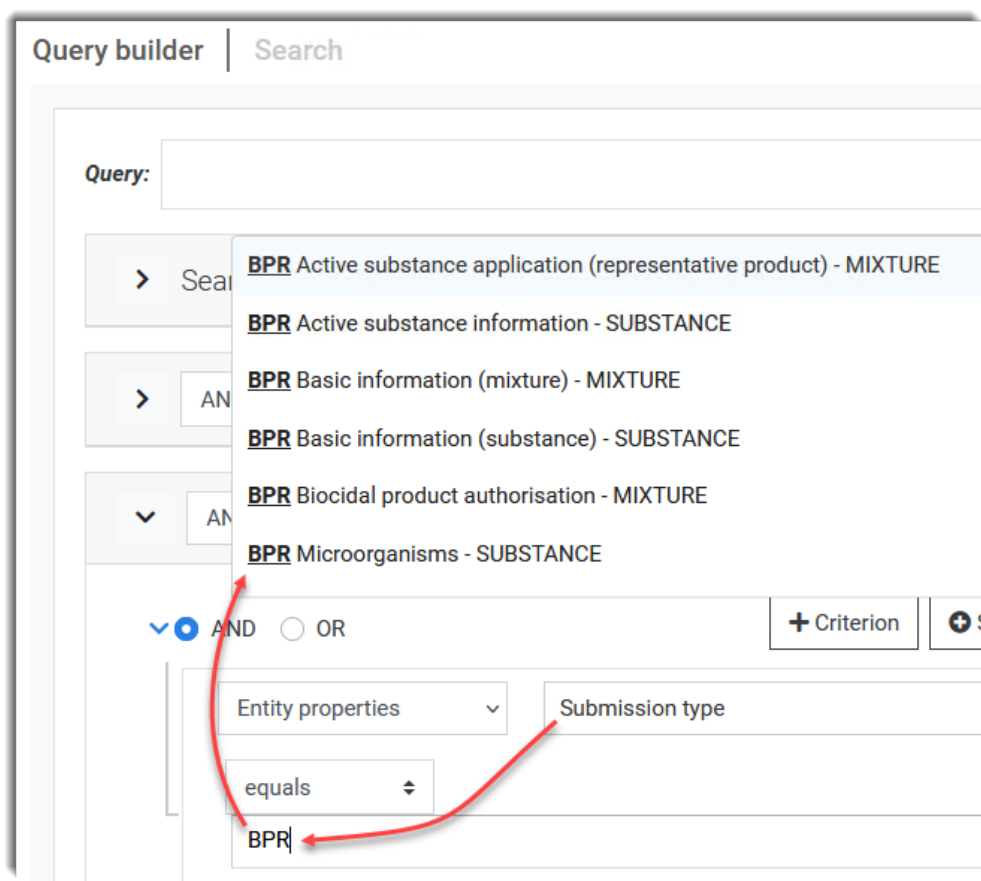
Figure 23: The Query window in Query builder

If building a complex query, it is recommended to build it up in stages, checking up at each stage that the search returns the expected hits.

The value to find is automatically surrounded by the appropriate separating characters for the data type of the property and the comparator. For example: double quotes, parentheses, and forward slashes where the comparator denotes a regular expression.

For some properties, the user is free to enter a text value, but for others a value is selected from a drop-down menu. The drop-down menu is used to present the user with only the values that exist in IUCLID. This means that the user can see what values are possible without having to remember them, and it helps to avoid error. For example, the types of entities are fixed in IUCLID, such as *Category*, *Dossier*, *Substance*. Entering a few characters searches for matches in the drop-down menu and presents only them. An example is shown for *working context* (Submission type) where “BPR” has been entered.

Figure 24: Search in a drop-down menu



6.4.1. The property Label and its autocomplete function


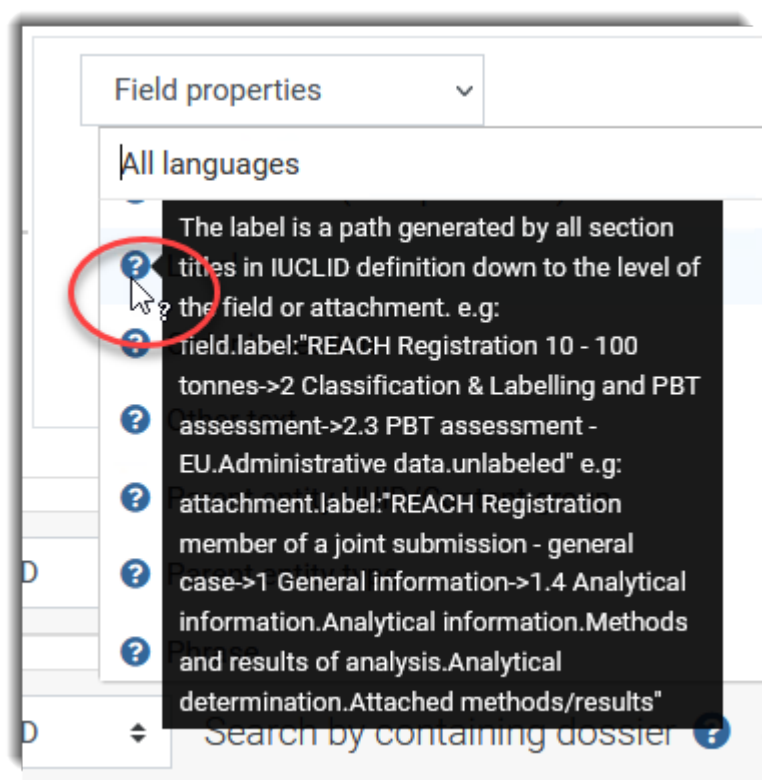
The property of a field named label is defined in the help as “a path generated by all section titles in IUCLID definition down to the level of the field or attachment.” This description can be viewed by hovering over the question mark icon , and via the link *Searchable properties* at the top right of the page.

Figure 25: The in-line help for properties

The field property *label* has an autocomplete function that presents the user with a drop-down menu containing the values that match the text entered into the box. At least four characters must be entered. The match is not sensitive to case. The wild card * can be used to represent zero or more of any character. Thus *Re*ach* matches both *treatments.attached* and **REACH**. Wild cards may be used to refine which labels are presented in the menu, but only one label can be selected. Therefore, wildcards cannot be used in the actual search itself, where the property is *label*. The maximum number of labels that can be offered at once is 21. If the label you want is not offered, refine the search criteria.

The currently selected option is high-lighted in blue. To select a label, click on it.

The label of a field in this context is a piece of text that starts with the dossier type. Note that dossier type is referred to as *Working Context* in the IUCLID interface. The labels contain the section numbering that is defined in the dossier type, and the names of the headings that lead down the IUCLID hierarchy. In *Text analytics* for the property *label* that correspond to fields end with a full stop followed by the value of *label* that is shown in the IUCLID interface.

Example of *label* in *Text analytics*:

```
field.label:"REACH Registration 10 - 100 tonnes->4 Physical and chemical
properties->4.2 Melting point / freezing point.Results and
discussion.Melting / freezing point.Melting / freezing pt."
```

Examples of *label* and *path* from the IUCLID format:

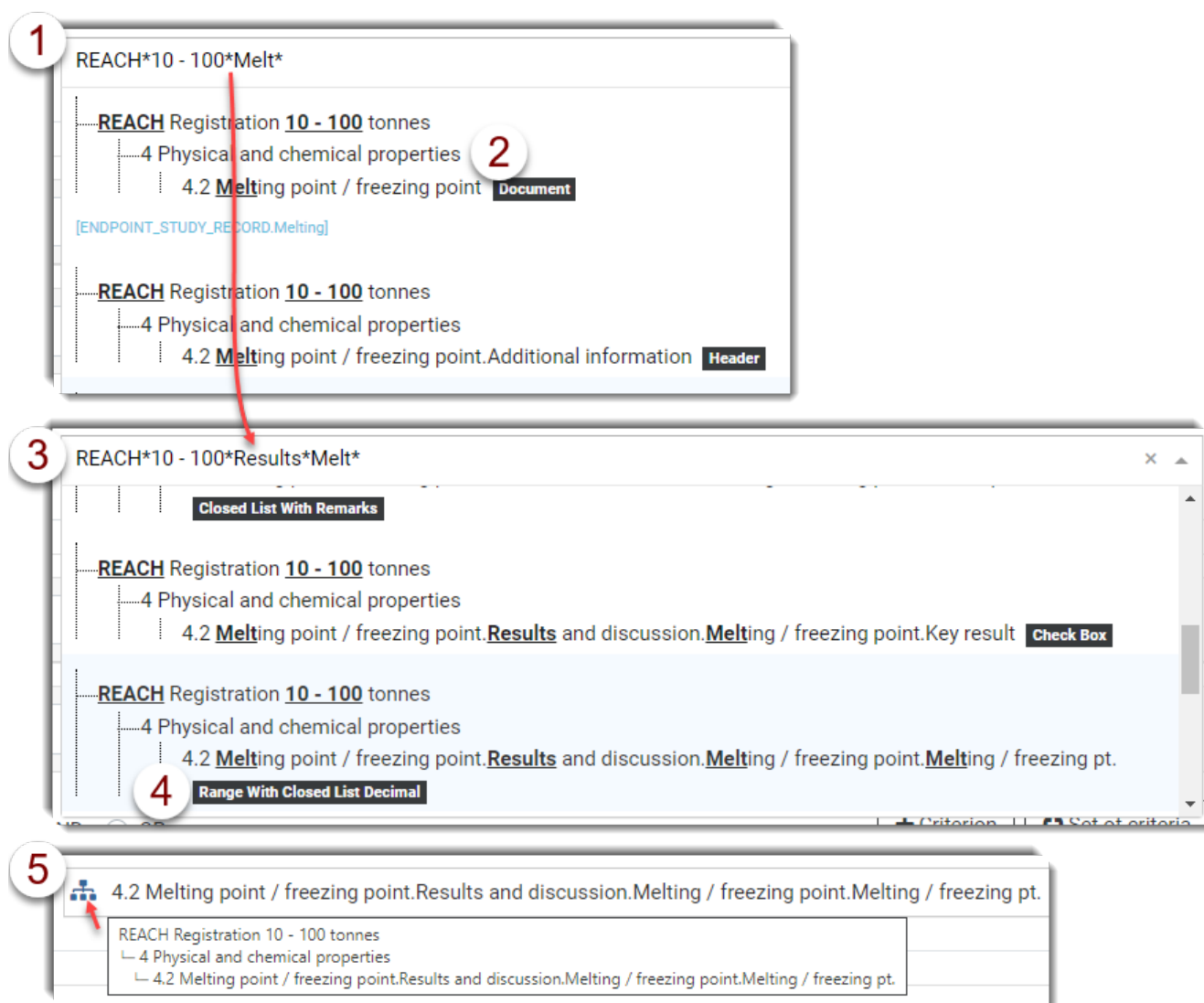
label = Melting / freezing pt.

path = ENDPOINT_STUDY_RECORD.Melting.ResultsAndDiscussion.MeltingPoint.MeltingPoint

An effect of the structure of the property `label`, and that a single value must be chosen, means that only one dossier type can be searched for, per criterion. To search for more, add criteria for them separated by OR, or copy the final query over to the Search where a wildcard can be applied to the dossier type.

Consider an example in which you know you want values of melting point in REACH 10 to 100 tonnes dossiers, but you do not know the exact value of the label to select. The process is described in the figure and steps below.

Figure 26: Selecting the label of a field



Legend for Figure 26

1. Start by entering `REACH*10 - 100*Melt*`;
2. This gives a list of labels, but none of them contain a numerical field, as indicated by the type of label given in a black box to the right of the label. The maximum number of labels that can

be shown at once is 21. The label we want must be outside this limit, so we need to refine the search;

3. The search is refined to `REACH*10 - 100*Results*Melt*`;
4. Now there is a label which looks correct judging from the section number and that it ends in the title of the field that is displayed in the IUCLID interface. Also it has the correct type, *Range With Closed List Decimal*, for the field in which melting points are recorded.
5. After having selected a label, its full value can be viewed by clicking on the hierarchy icon.

6.4.2. Fields / Attachments

The section *Fields / Attachments* is used to add criteria that search inside the values of IUCLID fields or the text within attachments. The general syntax is:

`field.<name of property>:<value to find>`

`attachment.<name of property>:<value to find>`

An example is shown below where the search hits are limited to documents that contain fields that have a particular label and the field contains a numerical value that is greater than 800.

Figure 27: An example of a criterion defined under Field / Attachment

☒ AND ☐ OR
 + Criterion + Set of criteria - Set of criteria

Field properties	Label	x	=
4.2 Melting point / freezing point.Results and discussion.Melting / freezing point.Melting / freezing pt. x			

Field properties	Range value (low bound)	x	>
800			

Query:

```
field.Label:"REACH Registration 10 - 100 tonnes->4 Physical and
chemical properties->4.2 Melting point / freezing point.Results and
discussion.Melting / freezing point.Melting / freezing pt." AND
field.value_range_Low:>800
```

6.4.3. Containing document

The section *Containing document* is used to apply limits to the parent document of the search hits. The general syntax is:

```
joinLevel:parent{field.<name of property>:<value to find>}
joinLevel:parent{attachment.<name of property>:<value to find>}
joinLevel:parent{entity.<name of property>:<value to find>}
```

An example is shown below where the search hits are limited to documents that were created after November 16th 2021.

Figure 28: An example of a criterion defined under Contained document

The screenshot shows a user interface for defining search criteria. It features a dropdown menu labeled 'Entity properties' with a downward arrow. To its right is a text input field containing 'Creation date' with a clear 'x' button and a dropdown arrow. Below these, there is a date range selector with a dropdown showing 'after', a date input field with '16-11-2021', and a calendar icon. A red minus sign is visible on the right side of the criteria bar. Below the criteria bar, a 'Query:' label is followed by a text box containing the generated query: `joinLevel:parent{entity.creation_date:>16-11-2021}`.

6.4.4. Containing dossier

The section *Containing dossier* is used to apply limits to the dossier that contains the search hits. The general syntax is:

```
joinLevel:dossier{field.<name of property>:<value to find>}
joinLevel:dossier{attachment.<name of property>:<value to find>}
joinLevel:dossier{entity.<name of property>:<value to find>}
```

An example is shown below where the search hits are limited to documents that are inside a dossier named "table_salt" of type *REACH Registration 10 – 100 tonnes* that is based on a *Substance*. Note that all the indexed fields that match will be presented as hits, including those in *Endpoint Study records* and *Legal entities* etc.

Figure 29: An example of a criterion defined under Containing dossier

Entity properties Entity name equals

table_salt

Entity properties Submission type equals

REACH Registration 10 - 100 tonnes - SUBSTANCE

Query: `joinLevel:dossier{entity.entity_name:"table_salt" AND entity.submission_type:"REACH Registration 10 - 100 tonnes - SUBSTANCE"}`

To filter for any dossier that has a submission type relating to the REACH regulation, select *like*, then *ALL REACH Registrations*. This inserts a regular expression into the query which matches working contexts that begin with “REACH Registration”.

Figure 30: Filter for all REACH registrations

Entity properties Submission type like

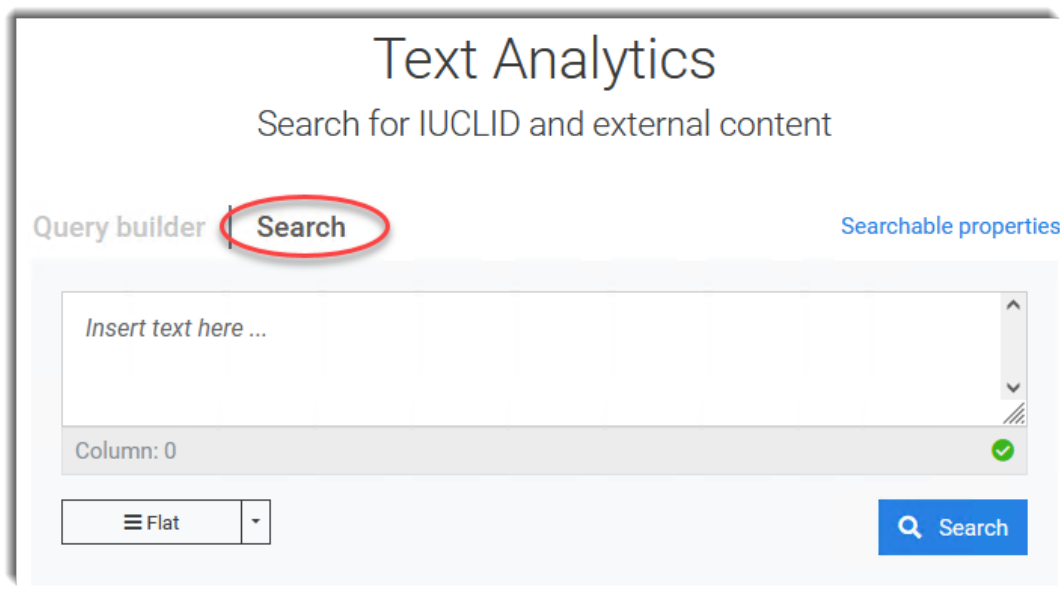
All REACH Registrations

Query: `joinLevel:dossier{entity.submission_type:/REACH Registration.*/}`

6.5. Search

Search is opened from its tab, as shown below.

Figure 31: Open the tab for Search

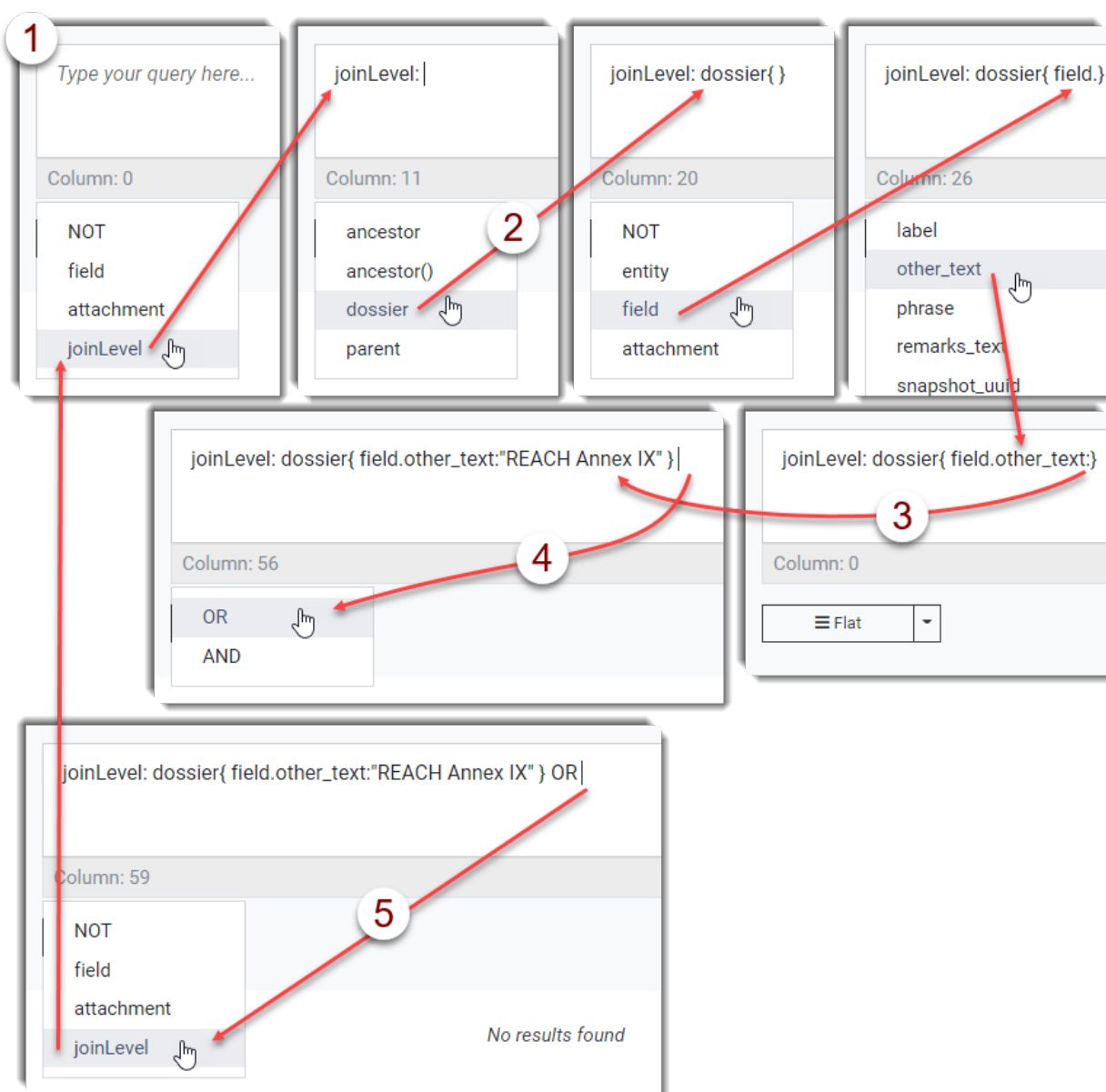


The *Search* is used to create queries that cannot be generated in the Query builder. These can range in complexity from simple keywords, to queries too complex to enter under Query builder. Queries can be copied over from the *Query builder* but not in the opposite direction. This copying can be used in a process whereby a query is started under *Query builder*, and then developed further under *Search*.

The query box for *Search* accepts free text, but there are features built in to help with the syntax of the query language. These are described in the following sections.

6.5.1. Search: Automatic syntax suggestions

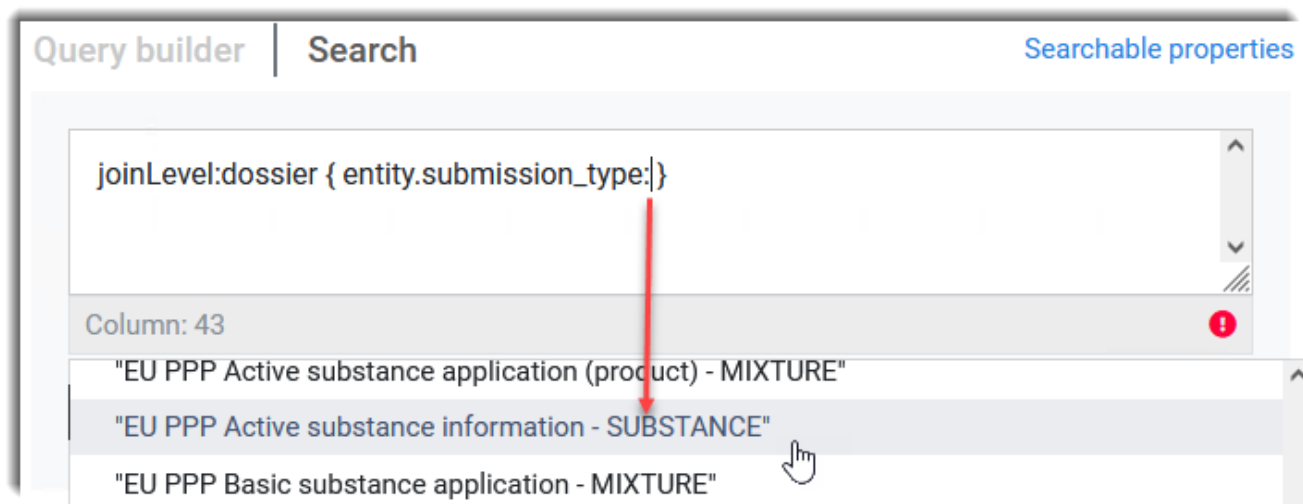
In the tab *Search*, placing the cursor where the next search term is to be entered causes a menu to pop up offering values that are syntactically correct in the current context. Starting from an empty window, an example is provided below to show the potentially cyclical nature of the process.

Figure 32: Building a search criterion from the automatic suggestions in the Search**Legend for Figure 32**

1. Start here by clicking in the search box and then selecting a suggested option from the drop-down menu;
2. Add levels to the criterion by selecting more suggested values;
3. If you get to a colon and there is no suggestion, manually enter a value to search for, wrapped in double quotes;
4. After a criterion is complete, Boolean operators are suggested to connect to the next criterion;
5. The initial menu is shown again at the start of the next criterion.

The suggested values can be of various types. For example following on from the sub-query: `joinLevel:dossier { entity. }`, there are *submission_type*, also known as *Working context*, and *Metadata*. The former allows search for an entity by the *Working context* of the containing dossier without having to remember all the options verbatim, or enter the required one manually. An example is provided below.

Figure 33: Search for an entity by the Working context (submission_type) of the containing dossier

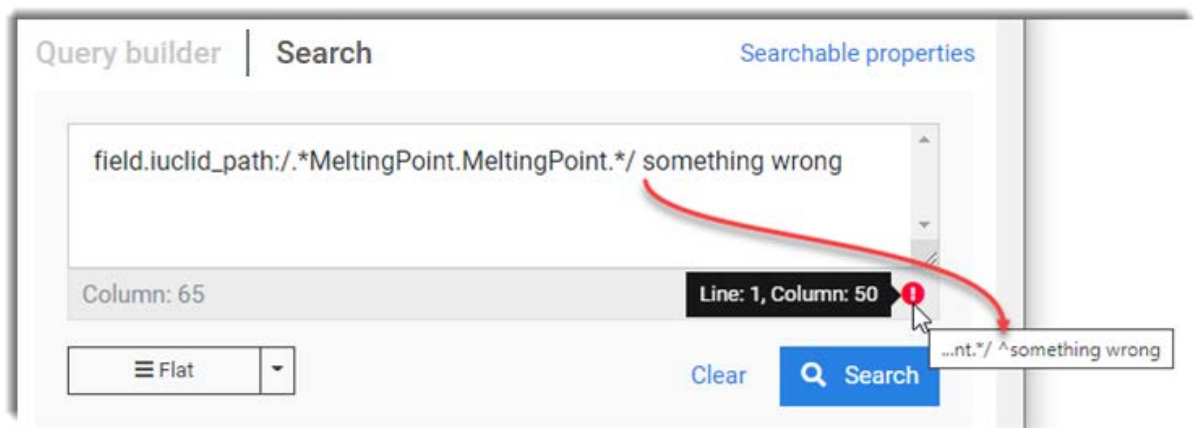


6.5.2. Search: Traffic light syntax indicator

Below the right-hand corner of the query window in the *Advanced* tab is an indicator that is green with a tick for correct syntax, and red with an exclamation mark for incorrect. Hovering over the red indicator shows where the problem begins. The line and column numbers are indicated in a black box. In a white box the point where the problem starts is indicated with a caret character (^) shown within the context of the query. If there is no text after the caret character, it means that something must be appended to the current query to give the correct syntax.

An example is shown below where the problem starts at on line 1 at column 50. The caret character indicates that the syntax breaks at the start of the word *something*.

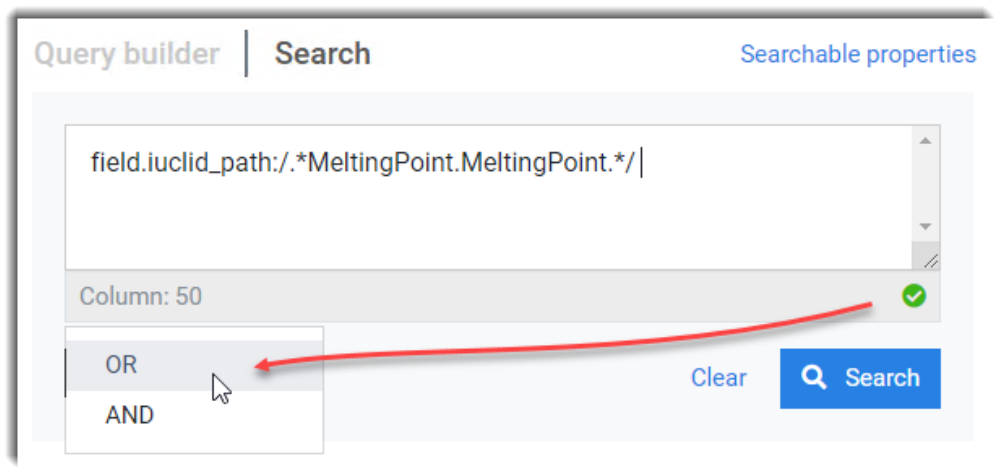
Figure 34: A red indicator of incorrect syntax, and where the break in the syntax is located



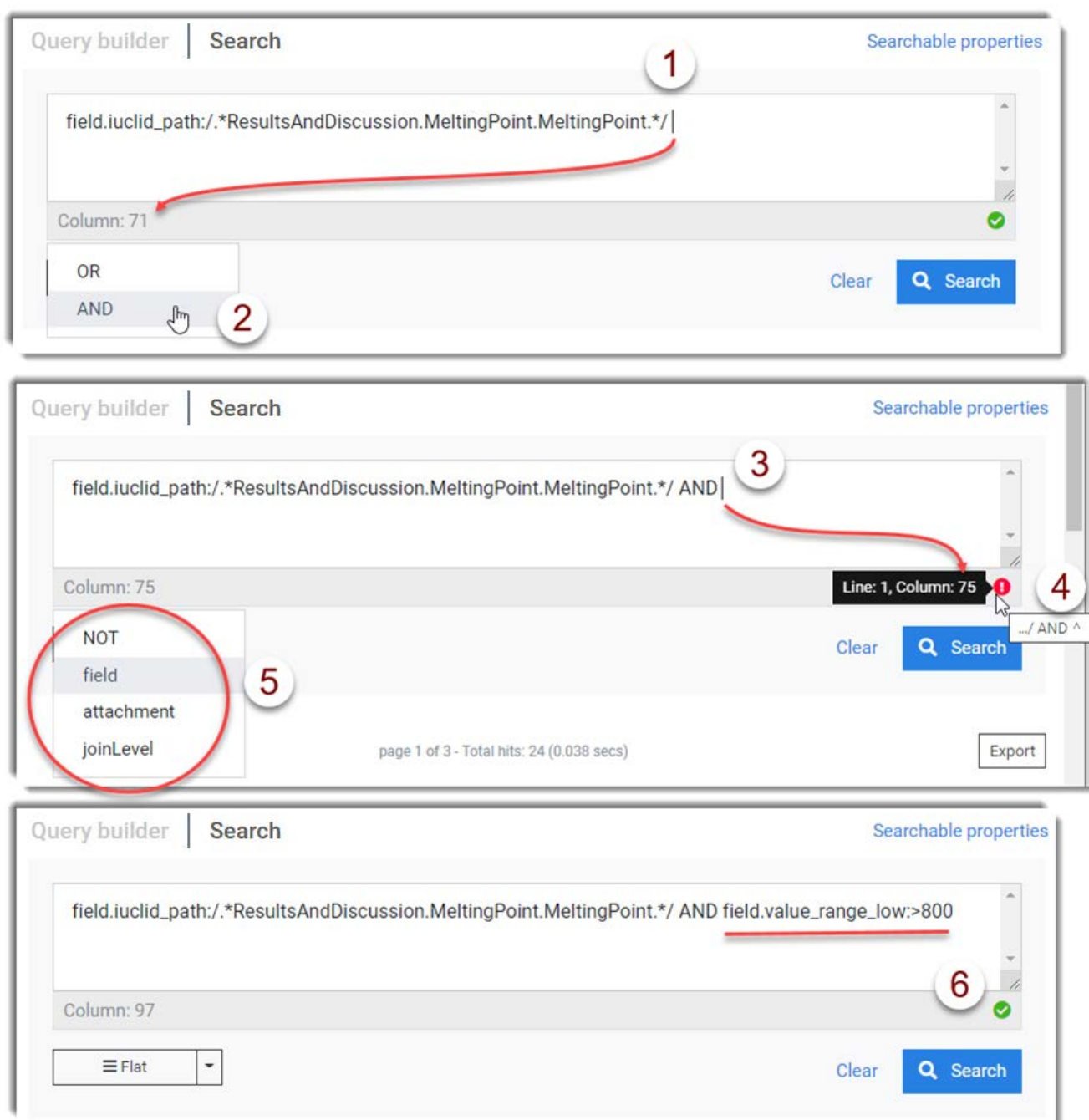
Deleting the text after the caret character returns the indicator to green.

In the example below, after deleting the text *something wrong*, the indicator has gone green, and the auto suggest feature offers a menu of Boolean operators which are correct in the context.

Figure 35: A green indicator of correct syntax



The traffic light indicator can be used in combination with the auto-suggest feature. An example is provided below.

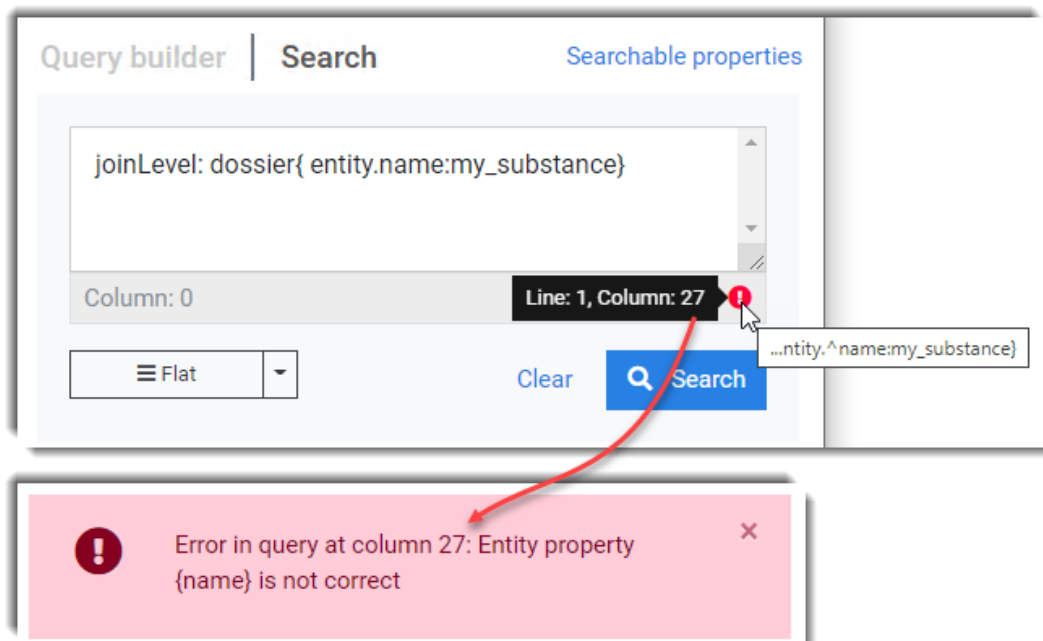
Figure 36: Traffic light indicator in combination with the auto-suggest feature**Legend for Figure 36**

1. The query has one term. The syntax is correct so the indicator is green;
2. The auto-suggest feature offers a choice of AND or OR to go after the term;
3. The operator AND is selected, and is therefore automatically appended to the query;
4. The indicator is now red and there is nothing after the caret character, meaning that something must be appended to the query;
5. The auto-suggest feature offers items to append to the query that give correct syntax;
6. Once a complete term has been added, the green icon indicates correct syntax.

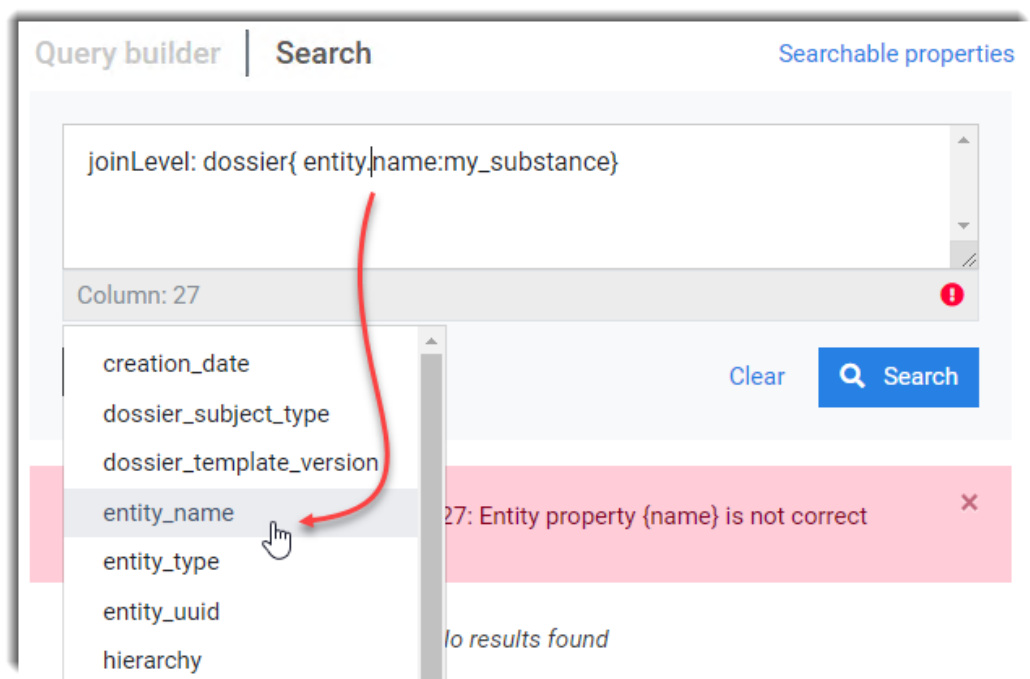
6.5.3. Search: Error messages

Clicking *Search* whilst there is a red indicator gives an error message, with information about why the syntax is incorrect. This can be used as a way of finding out why the syntax is wrong. In the example below, the problem starts at column 27, as stated in the indicator, and the error message. The error message also states that the entity property *name* is the problem.

Figure 37: Error message when searching through a red traffic light



One solution is to click into the column of the error, in this case 27. This causes the auto-suggest to provide a list of alternative values. In the example below it is evident that *name* must be replaced by *entity_name*.

Figure 38: Using autosuggest to correct queries

7. Worked examples of searches

Text to be entered into the search box is shown below in a monospace font like this.

7.1. Simple keyword search - level 1

For example, to find entities and documents that contain the whole words **rat** or **rats** or **liver** in any order use the following search terms.

```
rat liver
```

7.2. Intermediate searching - level 2

Make the order of the words matter by enclosing them in double quotes. For example, the following two are not the same.

```
"rat liver"
```

```
"liver rat"
```

Find a string anywhere in other strings, using the wild card star (*). For example to find the string **permethrin**, use:

```
*permethrin*
```

Look for multiple phrases that are not next to each other, using the Boolean operator AND. For example, to find all entities and documents that contain the word **permethrin** and the phrase **piperonyl butoxide**:

```
"permethrin" AND "piperonyl butoxide"
```

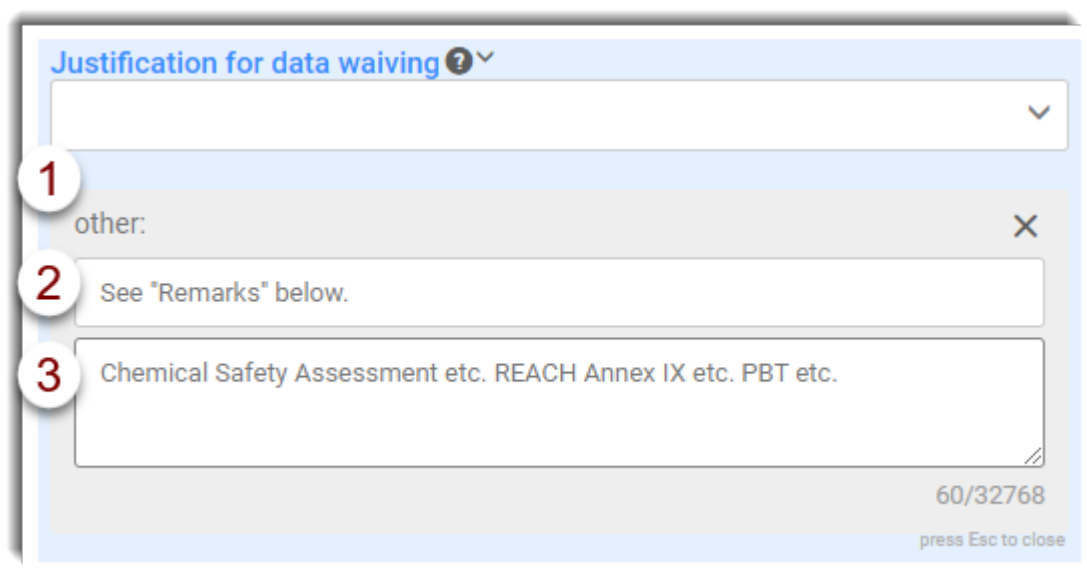
7.3. Search with the query language of text analytics - level 3

At this level, two examples are given. The first is created using only the *Query builder*. The second is a refinement of the first, created by transferring the query to the Search tab, and then editing it manually.

7.3.1. Example using only the Query builder

The search is to find documents in REACH dossiers under section 6.1.2 Long-term toxicity to fish that contain the text *REACH Annex IX* in the justification for data waiving where it has been set to have the value "other". The text can either be in the box that is provided for entering what the other justification is, or in the Remarks. The location in the interface and the corresponding properties are given below.

Figure 39: Location of the properties: *value*, *other_text*, and *remarks_text* in the web interface of IUCLID



Legend for Figure 39

The names of the properties in the numbered locations are:

1. *value*
2. *other_text*
3. *remarks_text*

In the first criterion select the label of the field using the auto complete function. For example, enter:

```
REACH Registration 10 - 100*6.1.2*Justification for data waiving
```

Create a set of criteria under the top criterion, and then add one criterion to it. Change the operator of the set to OR. In the lower criterion, select the property *Other Text* and then set the comparator to *contains*. Paste the text *REACH Annex IX* into the box for the search terms. Check that the query looks correct in the *Query* box. Click on the *Search* button. Note the number of hits. Add a

second criterion to the set, and then repeat the process for the property *Remarks Field*. Look at where the matches occurred to check that the result was as expected.

Figure 40: Example of Query builder

Search by fields/attachments [clear](#)

☒ AND ☐ OR

+ Criterion + Set of criteria - Set of criteria

Field properties Label equals 6.1.2 Long-term toxicity to fish.Administrative data.Justification for data waiving

☐ AND ☒ OR

+ Criterion + Set of criteria - Set of criteria

Field properties Other text contains REACH Annex IX

Field properties Remarks Field contains REACH Annex IX

Query:

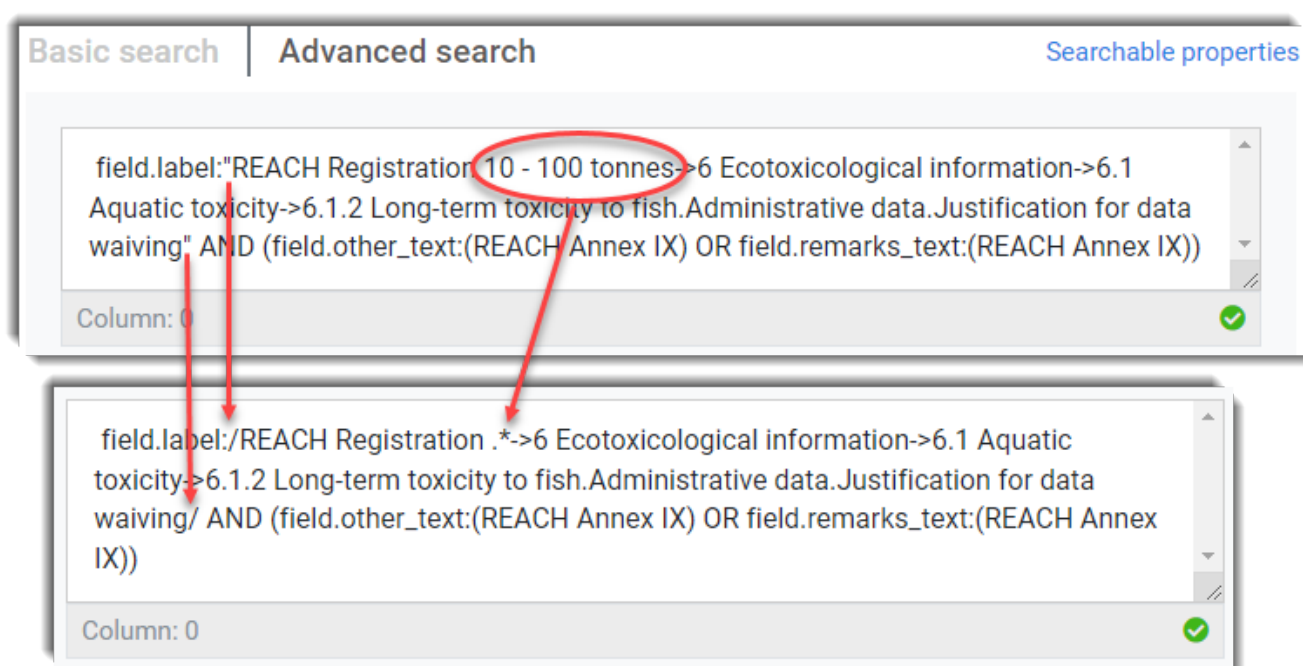
```
field.label:"REACH Registration 10 - 100 tonnes->6 Ecotoxicological information->6.1 Aquatic toxicity->6.1.2 Long-term toxicity to fish.Administrative data.Justification for data waiving" AND (field.other_text:(REACH Annex IX) OR field.remarks_text:(REACH Annex IX))
```

7.3.2. Example of manual refinement of a query generated in Query builder

This example is a refinement of the previous one. The query is changed to match all the dossier types for registration under the REACH regulation. First copy the query over to the advanced tab by clicking on the button to the right of the query in *Query builder*. Run the search and make a note of the number of hits. Change the criterion for the label into a regular expression by replacing each double quote character (") with a single forward slash character (/). Select the text in the search term that is common to all dossier types for registration under the REACH regulation, which is "10 - 100 tonnes". Replace it with a dot followed by a star (.*). Repeat the search, and then check whether the number of hits has increased.

The process is shown below.

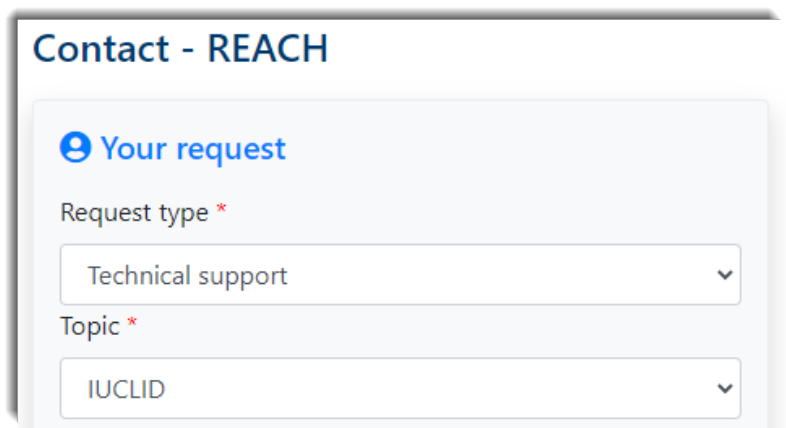
Figure 41: Example of refinement of a query using a regular expression in the label field



8. Getting help

To get help with specific queries in TA, you can submit a ticket to the IUCLID helpdesk at the address <https://echa.europa.eu/contact>. First select any of the legislations *REACH/CLP/Biocides/PIC*, then *Technical Support*, and then the topic *IUCLID*, as shown below.

Figure 42: How to direct a question to the IUCLID helpdesk



The screenshot shows a web form titled "Contact - REACH". Under the heading "Your request", there are two dropdown menus. The first, labeled "Request type *", has "Technical support" selected. The second, labeled "Topic *", has "IUCLID" selected. Both dropdowns have a downward arrow icon on the right side.

Appendix A. Quick training exercise on query syntax

Try out the following searches. Before trying to interpret the quantities of search results, check which options are set via the menu indicated below.



You can also try observing the effects of changing the settings. If necessary, adjust the example search terms to fit your data and needs.

Query	Comment
Exposure exposure "exposure"	These are equivalent, and so give the same number of hits.
Boolean logic	
exposure scenario exposure OR scenario	These are equivalent, and so give the same number of hits.
exposure AND scenario	Both words must be present but can be anywhere in the same document, entity, or attachment.
exposure AND NOT scenario	Exclude "scenario"
exposure -scenario	Exclude "scenario"
"exposure scenario"	Exact phrase.
Distance between terms	
"exposure scenario"~5	Finds those two words, with a maximum of 5 words in between
"exposure scenario"~5000	Finds those two words, with a maximum of 5000 words in between
IUCLID documents and entities versus Attachments	
field.value:exposure	IUCLID documents and entities. Check the search scope under the settings.
attachment.content:exposure	Text inside files attached to IUCLID documents and entities. Check the search scope under the settings.
attachment.mime:"application/pdf"	Filter by the type of attached files, using the mime type. For example PDF files have a mime type of <i>application/pdf</i> .
Attachment.filename:*IR*spect*	Filter by the filename of attached files. For example, look for infrared spectra. Note the use of the wild card *, which stands for zero or more of any character.

Stemming	
inventory inventories inventory inventories inventory OR inventories	Due to stemming, these should give the same results. The hits can be in either fields or attachments.
field.value:<any of the first three values above>, but for the fourth one, field.value:inventory OR field.value:inventories	Same as above but only in fields.
attachment.content:<any of the first three values above>, but for the fourth one, attachment.content:inventory OR attachment.content:inventories	Same as above but only in attachments.
field.value.std:inventory field.value.std:inventories	Stemming is not done so these differ from each other, and the rows above.
field.value.std:(inventory inventories) field.value.std:inventory OR field.value:inventories	Stemming is not done, but all the alternative forms are supplied so these give the same number of hits. Only in fields.
attachment.content.std:(inventory inventories) attachment.content.std:inventory OR attachment.content.std:inventories	Only in attachments.
Parsing at a full stop character	
field.value.std:"Repr.1B" field.value.std:"Repr. 1B"	The values are interpreted literally. If a space character is placed between the full stop and 1, the space must be present for a match to occur. Therefore the upper value on the left matches itself, but not the lower value.
field.value:Repr.1B field.value:Repr. OR field.value:1B	The value is parsed at the full stop and interpreted as separate words on either side of the full stop. Therefore the upper value on the left matches itself, and the lower value.
joinLevel	
joinLevel:dossier{entity.entity_type:"DOSSIER"}	Find all Dossiers, and all documents/entities in them.
joinLevel:parent{entity.entity_type:"DOSSIER"}	Find all Dossiers The parent must be a Dossier.
joinLevel:dossier{entity.snapshot_uid:<UUID>}	Find everything in a specific Dossier by Dossier UUID.
joinLevel:dossier{entity.metadata.ec_number:"919-284-0" OR entity.metadata.ec_number:"918-668-5"}	Find all hits in dossiers having these EC numbers as dossier metadata

<code>joinLevel:parent{entity.entity_type:"ENDPOINT_STUDY_RECORD" OR entity.entity_type:"TEST_MATERIAL_INFORMATION" OR entity.entity_type:"ENDPOINT_SUMMARY"}</code>	Find all hits from entities of type
<code>attachment.entity_uuid:"CLP" OR attachment.entity_uuid:"digitised_NONS"</code>	Find all hits in the External Content groups "CLP" or "digitised_NONS"
Filter by detected language	
<code>field.detected_language:de OR attachment.detected_language:de</code>	The German language has been detected in fields or attachments

Appendix B. What to do if the search results are unexpected

Check the following:

1. Start with a broad search and then narrow it down. *Text Analytics* is designed to respond quickly enough to allow the search criteria to be fine-tuned across multiple runs of the search.
2. Under *Search scope*, check the selection of *Fields* and *Attachments*.
3. Has what you are looking for been indexed yet? How new is what you expect to find?
4. What type of search is it: *Flat* or *Aggregated*?
5. Are the operators AND, OR, NOT, set correctly?
6. Are the double quotes used properly to search for an exact phrase?
7. Before using wildcards to search in fields and attachments, consider whether the language analyser is already doing what you need. Do not use wildcards in double quotes.
8. Are you searching the correct property? See the descriptions under *Searchable fields* for what data they contain.
9. Is language analysis affecting the results? If a specific property is being searched, check its data type under *Searchable fields*. The data types: TEXT and CLOB use the language analyser, but the others such as KEYWORD, do not.
10. When using the properties `content` and `value`, remember that the language analyser is used. If you do not want that, use `content.std` and `value.std` instead. Consider the effects on stemming, case sensitivity, and stop words.

Appendix C. Advanced topics

Some of the lesser used features of *Text analytics* have been omitted from this manual. A list of them is provided below.

1. Boosting: Affecting the order in which search results are presented.
2. Fuzziness
3. Proximity

4. Known limitations

If you need to know more about one of these, you can submit a request to the IUCLID helpdesk as described above in section *Getting help*.

Appendix D. Queries obsolete from v3.8.0 onwards

field.value:(gold AND nickel) to be written as: field.value:gold AND field.value:nickel
field.value:(gold OR nickel) to be written as: field.value:gold OR field.value:nickel
attachment.content:(gold AND nickel) to be written as: attachment.content:gold AND attachment.content:nickel
attachment.content:(gold OR nickel) to be written as: attachment.content:gold OR attachment.content:nickel
field.value:("Acute Tox. 2" OR "Acute Tox. 4") to be written as: field.value:"Acute Tox. 2" OR field.value:"Acute Tox. 4"
field.value:("recovered" OR "recycle" OR "reused" OR "waste" OR "dismissed") to be written as: field.value:"recovered" OR field.value:"recycle" OR field.value:"reused" OR field.value:"waste" OR field.value:"dismissed" field.value:(recovered OR recycle OR reused OR waste OR dismissed) to be written as: field.value:recovered OR field.value:recycle OR field.value:reused OR field.value:waste OR field.value:dismissed (equivalent to previous query)
attachment.content:("Acute Tox. 2" OR "Acute Tox. 4") to be written as: attachment.content:"Acute Tox. 2" OR attachment.content:"Acute Tox. 4"
attachment.content:("recovered" OR "recycle" OR "reused" OR "waste" OR "dismissed") to be written as: attachment.content:"recovered" OR attachment.content:"recycle" OR attachment.content:"reused" OR attachment.content:"waste" OR attachment.content:"dismissed"

<p>attachment.content:(recovered OR recycle OR reused OR waste OR dismissed)</p> <p>to be written as:</p> <p>attachment.content:recovered OR attachment.content:recycle OR attachment.content:reused OR attachment.content:waste OR attachment.content:dismissed (equivalent to previous query)</p>
<p>joinLevel:dossier{entity.submission_type:/REACH Registration.*}/ AND (field.value:("QSAR Toolbox" OR "OECD Toolbox") OR attachment.content:("QSAR Toolbox" OR "OECD Toolbox"))</p> <p>to be written as:</p> <p>joinLevel:dossier{entity.submission_type:/REACH Registration.*}/ AND (field.value:"QSAR Toolbox" OR field.value:"OECD Toolbox" OR attachment.content:"QSAR Toolbox" OR attachment.content:"OECD Toolbox")</p>
<p>field.value:(*nonyl* AND *phenol*)</p> <p>to be written as:</p> <p>field.value:*nonyl* AND field.value:*phenol*</p>
<p>field.value.std:("50-32-8" OR "63466-71-7" OR "601-032-00-3" OR "benzo[a]pyrene" OR "benzo[def]chrysene") AND joinLevel:parent{entity.entity_type:REFERENCE_SUBSTANCE} AND joinLevel:ancestor{field.label:/. *1.2 Composition.*}/</p> <p>to be written as:</p> <p>joinLevel:parent{entity.entity_type:REFERENCE_SUBSTANCE} AND joinLevel:relative(desc:2){field.label:/. *1.2 Composition.*}/ AND (field.value.std:"50-32-8" OR field.value.std:"63466-71-7" OR field.value.std:"601-032-00-3" OR field.value.std:"benzo[a]pyrene" OR field.value.std:"benzo[def]chrysene")</p>
<p>attachment.entity_uuid:("CLP" OR "digitised_NONS")</p> <p>to be written as:</p> <p>attachment.entity_uuid:"CLP" OR attachment.entity_uuid:"digitised_NONS"</p>
<p>joinLevel:dossier{entity.submission_type:/REACH Registration.*}/ AND joinLevel:dossier{field.label:/. *6 Ecotoxicological information.*}/ AND field.value:*} AND joinLevel:dossier{field.label:/. *7 Toxicological information.*}/ AND field.value:*} AND field.label:/. *3.5 Use and exposure information->3.5.3 Uses at industrial sites.*}/ AND field.value:("monomer" OR "polymer")</p> <p>to be written as:</p> <p>joinLevel:dossier{entity.submission_type:/REACH Registration.*}/ AND joinLevel:dossier{field.label:/. *6 Ecotoxicological information.*}/ AND field.value:*} AND joinLevel:dossier{field.label:/. *7 Toxicological information.*}/ AND field.value:*} AND field.label:/. *3.5 Use and exposure information->3.5.3 Uses at industrial sites.*}/ AND (field.value:"monomer" OR field.value:"polymer")</p>