ECHA
EUROPEAN CHEMICALS AGENCY

REACH Study Results & Data Extractor

Use Case: Searching for a particular functional group

December 2023

European Chemicals Agency

Use case: Searching for a particular functional group from the REACH Study Results

→ The purpose is to

- Show in practice how the Data Extractor can be used
- Provide an illustrative example of how the extracted results can be post-processed and aggregated in KNIME
- Show how the chemical identifiers in the REACH registrations can be analysed with freely available cheminformatics tools
- The REACH Study Results can be analysed further, e.g. to extract and summarise the toxicity information that together with the generated molecular structures can be used for developing models
    - such use cases will be available in the future

Practical use case demonstration

→ Please read through the general support material before proceeding to this use case

- Installation
- Use manuals
- See [https://iuclid6.echa.europa.eu/fi/data-extractor]

→ The use case has been prepared using

- KNIME v4.2.2
- Data Extractor v1.11
- Current version of REACH Study Results and IUCLID in October 2023
- The applications and the Reach Study Results should apply with later than October 2023 versions. The use case will not be updated to later versions, unless if significant changes are identified

Please note the following

- This use case gives only an indicative idea of how substances that contain a given functional group can be identified from public data; there is no guarantee that the results are complete
- This use case has no regulatory or legal relevance
- The use of freely available tools does not mean that ECHA endorses or recommends them; the user can choose different tools for the processing of the results extracted with Data Extractor

Obtaining the REACH Study Results

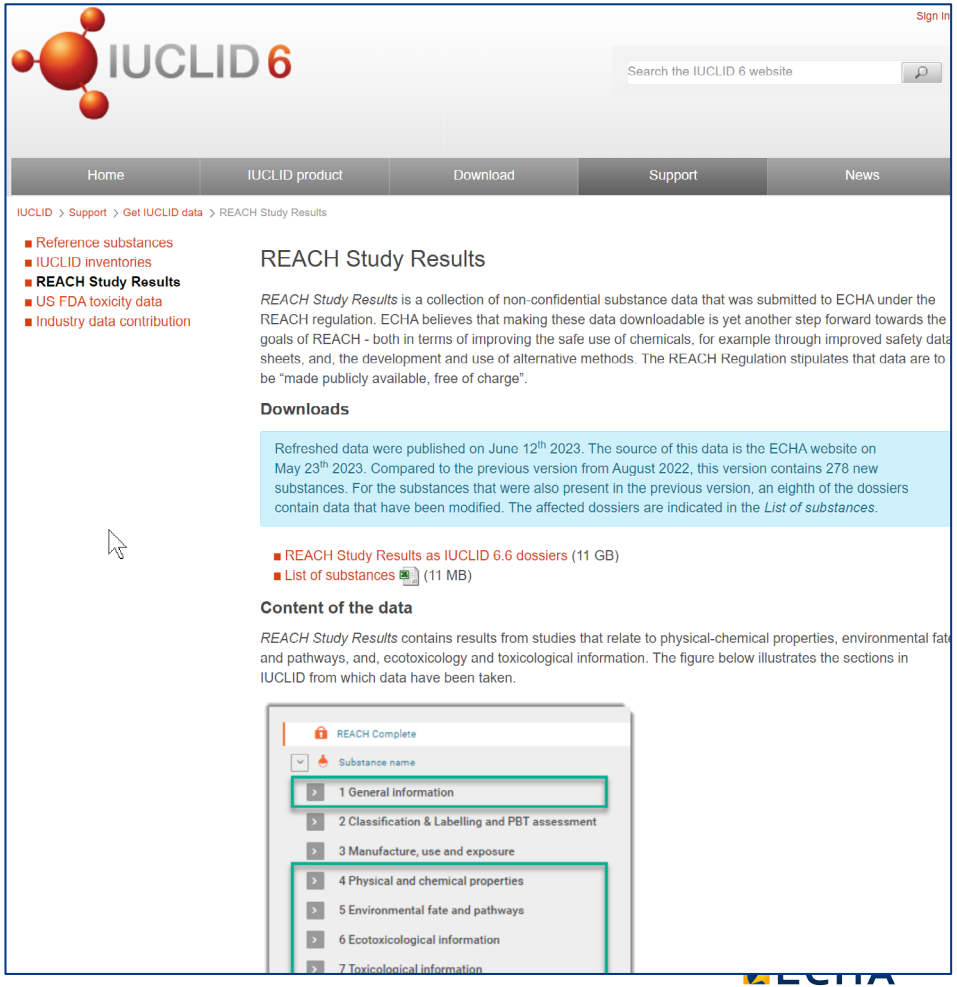
→ Create a user account or sign in on the IUCLID website

→ Download the REACH Study results dataset and the list of substances in this dataset

- [https://iuclid6.echa.europa.eu/fi/reach-study-results]

→ Load the dataset into a IUCLID instance using the support material available

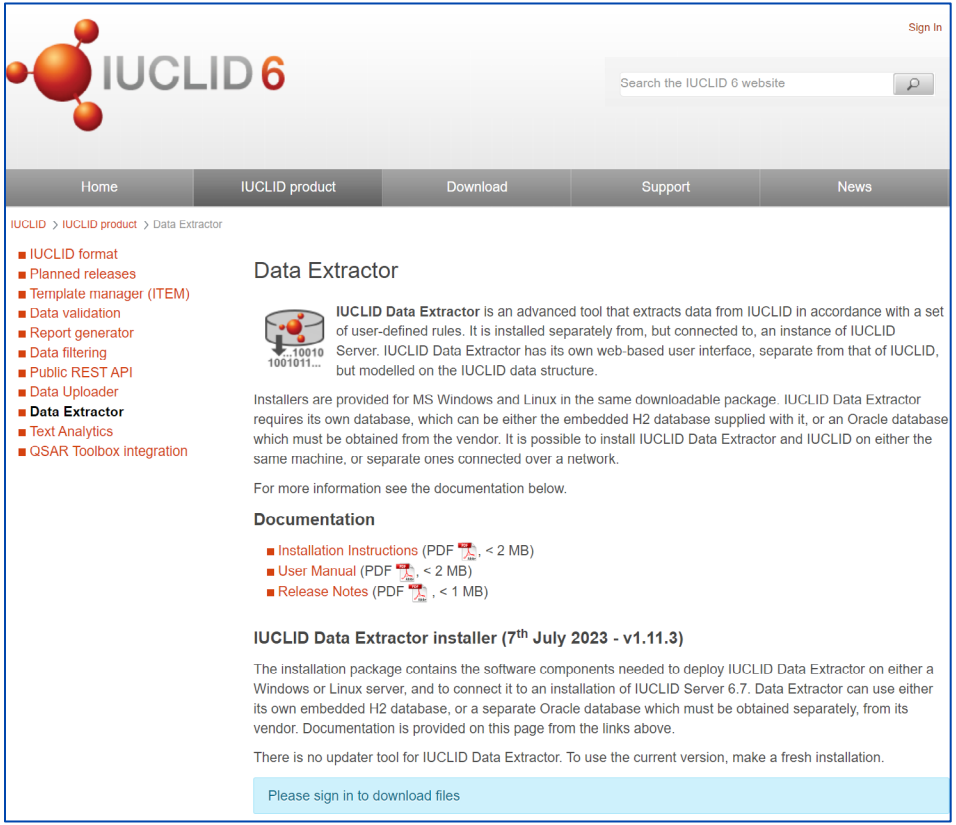
→ The figure on the right shows which IUCLID sections are included in the REACH Study Results



Obtaining Data Extractor

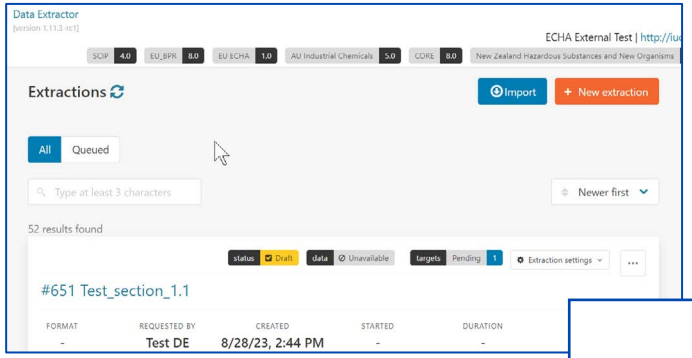
→ Download Data Extractor, the installation instructors and the use manual

→ Please refer to the instructions on how to connect the Data Extractor with the IUCLID instance, in this case, the REACH Study Results



ECHA
EUROPEAN CHEMICALS AGENCY

Configure the Extraction job



- Open a new extraction job on the main Data Extractor page
- Configure the new extraction job as described in the picture to the right
  - The list with dossier (target) UUIDs can be downloaded together with the REACH Study Results.
  - [https://iuclid6.echa.europa.eu/fi/reach-study-results]
  - NB! Prepare a .csv input file consisting only of the target UUID column. Import the UUIDs from that csv file.

Extractions

Extraction #651: Test_section_1.1

Reset | Save | + Create

Name *: Test_section_1.1

Format *: Normalised

Data settings: Replace new lines with: | Column delimiter: <Tab> / Other: | Replace delimiter with: Other: | Remove HTML tags from rich text fields

Description: Extract section 1.1, some sections relevant for structure based identification

Targets *: Manual input | From file | Choose a file... | At least one target must be provided | 0 target(s)

Fields *: At least one field must be provided | 0 field(s)

Attachments: 0 attachment(s)

The normalised extraction format may be optimal for manual post-processing of the data

You can use space here, if no particular visible character is desired

The default TAB column separation, should work well

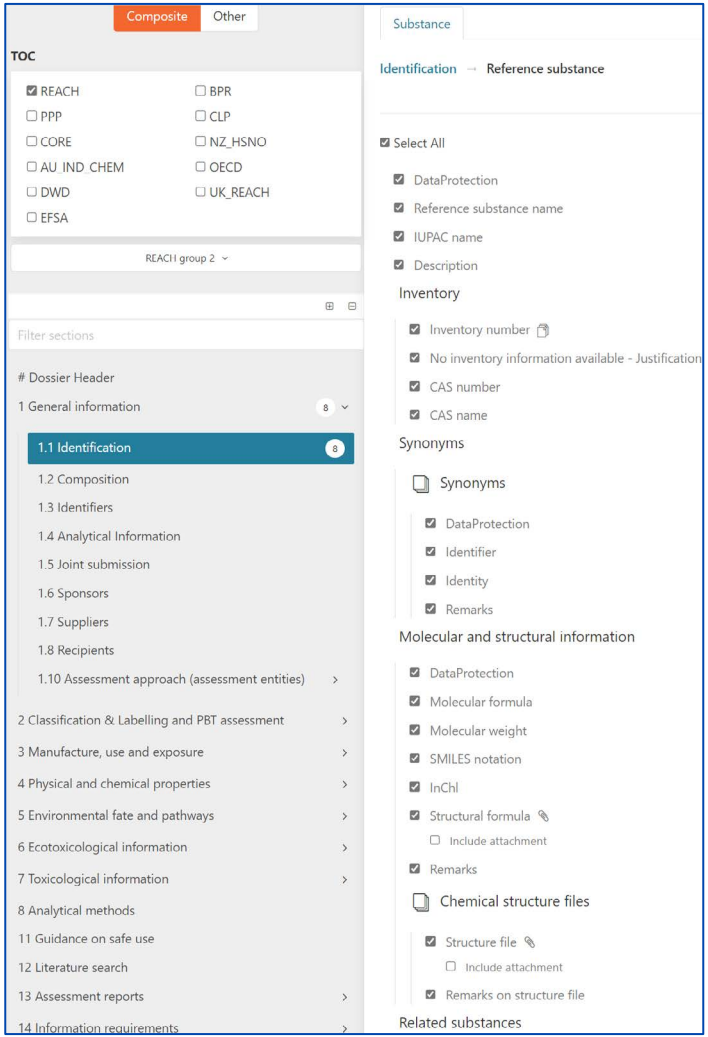
You can use space here, if no particular visible character is desired

Import your list of target UUIDs here. In this case the list of dossier UUIDs for the REACH Study Results. You may download the csv list from the IUCLID website. Do not upload other columns but the target UUIDs

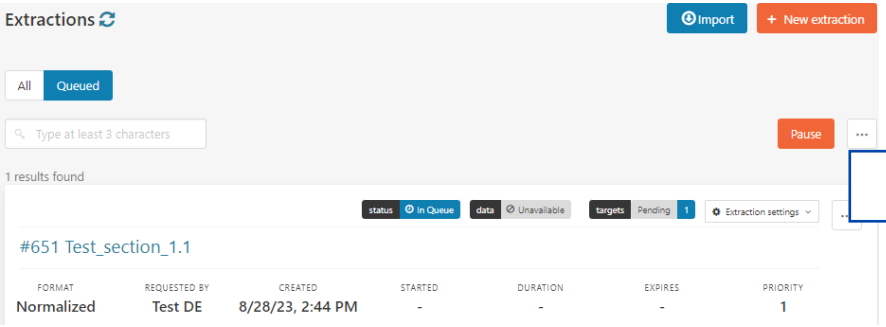
When you start selecting IUCLID fields to be extracted, the fields are listed here

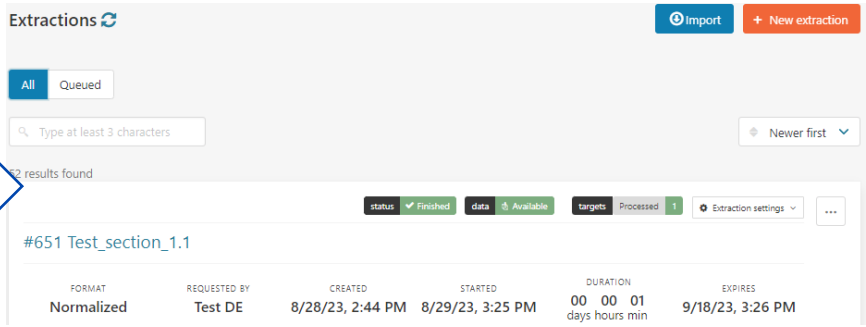
Configure the Extraction job

- Select the relevant Table of contents, that is, the regulation and working context
- For the REACH Study Results please choose "REACH">"Substance"> "REACH group 2"
- Select the fields for which the data need to be extracted
- In this example the relevant data is in IUCLID section 1.1; we will extract data from the reference substance that is linked to the dossier substance (use the "Select All" button)
- Data from other sections can be extracted in the same job. However, to make the extraction faster and to make the post-processing easier, we recommend extracting data from limited number of documents in the same extraction job; in case you need to execute multiple extraction jobs please use the queue functionality
- Push the "+ Create" button on top of the page to initiate the extraction



Download the Extracted Data



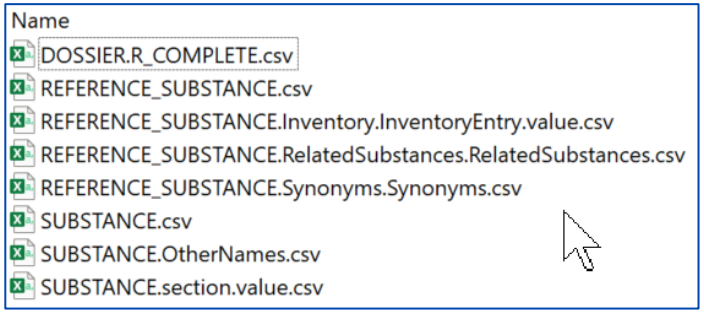


- Once the status of the extraction job has turned from “In queue”>”In progress ..%”>”Finalised”, the data can be downloaded
- Download the extracted data by clicking the data “Available” button
- In case any targets have not been successfully extracted, you will see statuses other than “processed”

ECHA
EUROPEAN CHEMICALS AGENCY

Download the Extracted Data

- The extracted data can be downloaded as a zip file; the name of the zip file starts with the number of the extraction job
- Please save the zip file locally, unzip and save the included flat data in the location of your preference
- In this use case the flat data files that will be used in the post-processing are displayed in the picture
- The more sections and fields an extraction job contains, the more flat files it will produce



Post-process the Extracted Data

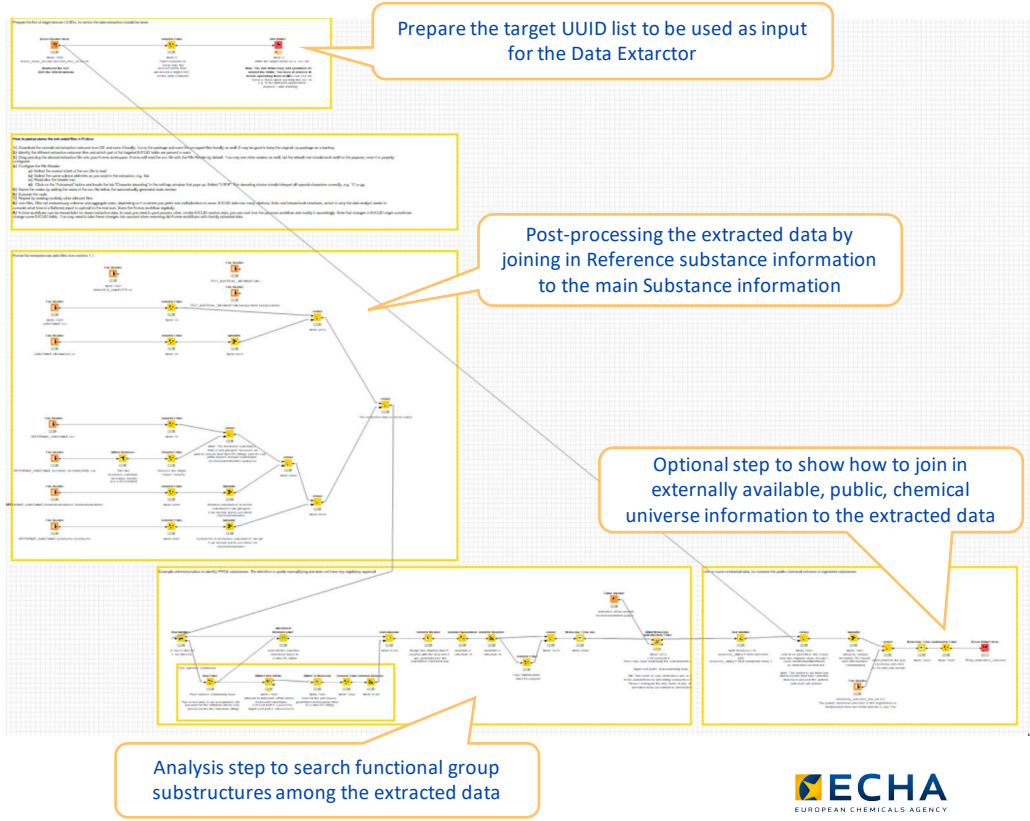
- IUCLID data is nested, the extraction essentially produces a different table for every level in the hierarchy
- The aim of the post-processing is to join the files and build a single report
- This single report will inevitably result in an increased number of rows; the data-analyst can determine the aggregation level and the final multiplicity of rows, depending on the needs
- It may be helpful to keep the original IUCLID instance open when post-processing the data. This is because usually the data analyst can recognise the location of the extracted data in IUCLID as the datasets are joined
- It is important to identify which flat data file contains which part of the extracted data and which column can be used as the joiner key.
- Identifying the joining strategy is facilitated from the schema map file, which is created for each extraction job. The schema map file is named “schema.csv”. More information about the schema map will be provided in future.

KNIME as a data post-processing tool

- One among many available third-party data processing tools is the KNIME Analytics Platform
  - For more information about KNIME, go to: [https://www.knime.com/]
- KNIME is downloadable free of charge, and you can use the standard set of included data processing nodes, or buy commercially maintained nodes
- KNIME offers extensive functionality for data wrangling, such as filtering and joining operations
- KNIME supports cheminformatics and visualisation
- For this use case we assume that the user is using a locally deployed KNIME instance
- The KNIME Analytics Platform receives timely updates and sometimes the user will need to perform some updates on existing workflows if attempting to run them with the latest KNIME version.
- There are plenty of KNIME tutorials available, e.g. at www.knime.com

KNIME as post-processing tool

- The use case material contains an KNIME workflow shown to the right
- View the example workflow from the top left to the bottom right and read the added clarifications
  - The input files refer to the relevant extracted flat data files
  - The input files are joined by the identified foreign key columns
  - The functional group filtration is made based on SMILES/SMARTS structure strings

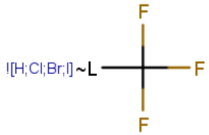


Cheminformatics

→ We identify substances by applying a sub-structure search to the structural information available in the dossier

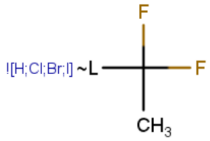
- For dossiers that do not have a structure given as a SMILES string, we attempt to convert IUPAC names to SMILES strings
- We use cheminformatics nodes from the open source RDkit, but others are available

→ Our indicative example of criteria for sub-structure based SMARTS filtering



SMARTS: FC(F)(F)[!#1!#17!#35!#53]

Fluorinated methyl group that cannot be connected to halogens or H.
The three C-F bonds enable the fourth bond to be a single bond



SMARTS: [#6]C(F)(F)-[!#1!#17!#35!#53]

Fluorinated methylene group that cannot be connected to halogens or H.
It must have a C-C single bond in addition to the two C-F single bonds

→ You may consult freely available material, e.g. online, about working with SMILES and SMARTS or other structure formats

Aspects of the workflow

- The output of the workflow is a list of substances that match the functional group sub-structure filter
- The output file format can be selected by using the appropriate writer node in KNIME
- Additional row aggregations can be made in the workflow before the report is exported
- The whole KNIME workflow can be exported for archiving purposes with or without the included data

Summary of the use case

- The use case provides an example of how to extract data from IUCLID dossiers in bulk. This may be applied to REACH Study Results, which is available on the IUCLID website
- The output of the extraction has been compiled in KNIME for post-processing purposes
- These data have been further processed by running a set of substructure searches using the KNIME cheminformatics functionality

Thank you

echa.europa.eu/subscribe

Connect with us

echa.europa.eu/podcasts

European Chemicals Agency

@one_healthenv_eu

@EU_ECHA

@EUECHA

EUchemicals