

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 [\[https://iuclid6.echa.europa.eu/fi/data-extractor\]](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
 - [\[https://iuclid6.echa.europa.eu/fi/reach-study-results\]](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

The screenshot displays the IUCLID 6 website interface. At the top, the IUCLID 6 logo is visible alongside a search bar and a 'Sign in' link. A navigation menu includes 'Home', 'IUCLID product', 'Download', 'Support', and 'News'. The breadcrumb trail indicates the current location: 'IUCLID > Support > Get IUCLID data > REACH Study Results'. On the left, a list of data sources is provided, with 'REACH Study Results' highlighted. The main content area is titled 'REACH Study Results' and contains a paragraph explaining that the data is a collection of non-confidential substance data submitted to ECHA under the REACH regulation. Below this, a 'Downloads' section features a light blue box with information about data refreshes and a list of download links: 'REACH Study Results as IUCLID 6.6 dossiers (11 GB)' and 'List of substances (11 MB)'. The 'Content of the data' section describes the types of studies included. At the bottom, a preview of the 'REACH Complete' dataset structure is shown, with a list of sections: '1 General information', '2 Classification & Labelling and PBT assessment', '3 Manufacture, use and exposure', '4 Physical and chemical properties', '5 Environmental fate and pathways', '6 Ecotoxicological information', and '7 Toxicological information'. The ECHA logo is in the bottom right corner.

Obtaining Data Extractor

- Download Data Extractor, the installation instructions 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



The screenshot shows the IUCLID 6 website interface. At the top, there is a navigation bar with links for Home, IUCLID product, Download, Support, and News. A search bar is located in the top right corner. The main content area is titled "Data Extractor" and features a list of links on the left: IUCLID format, Planned releases, Template manager (ITEM), Data validation, Report generator, Data filtering, Public REST API, Data Uploader, Data Extractor (highlighted), Text Analytics, and QSAR Toolbox integration. The main text describes the IUCLID Data Extractor as an advanced tool for extracting data from IUCLID. It mentions that the tool is installed separately but connected to an IUCLID Server. Below this, it states that installers are provided for MS Windows and Linux. A "Documentation" section lists links for Installation Instructions, User Manual, and Release Notes. A "IUCLID Data Extractor installer (7th July 2023 - v1.11.3)" section provides details about the installation package and its components. At the bottom, there is a note about the lack of an updater tool and a button to sign in to download files.

IUCLID 6

Search the IUCLID 6 website

Home IUCLID product Download Support News

IUCLID > IUCLID product > Data Extractor

- IUCLID format
- Planned releases
- Template manager (ITEM)
- Data validation
- Report generator
- Data filtering
- Public REST API
- Data Uploader
- **Data Extractor**
- Text Analytics
- QSAR Toolbox integration

Data Extractor

 **IUCLID Data Extractor** is an advanced tool that extracts data from IUCLID in accordance with a set of user-defined rules. It is installed separately from, but connected to, an instance of IUCLID Server. IUCLID Data Extractor has its own web-based user interface, separate from that of IUCLID, but modelled on the IUCLID data structure.

Installers are provided for MS Windows and Linux in the same downloadable package. IUCLID Data Extractor requires its own database, which can be either the embedded H2 database supplied with it, or an Oracle database which must be obtained from the vendor. It is possible to install IUCLID Data Extractor and IUCLID on either the same machine, or separate ones connected over a network.

For more information see the documentation below.

Documentation

- [Installation Instructions](#) (PDF , < 2 MB)
- [User Manual](#) (PDF , < 2 MB)
- [Release Notes](#) (PDF , < 1 MB)

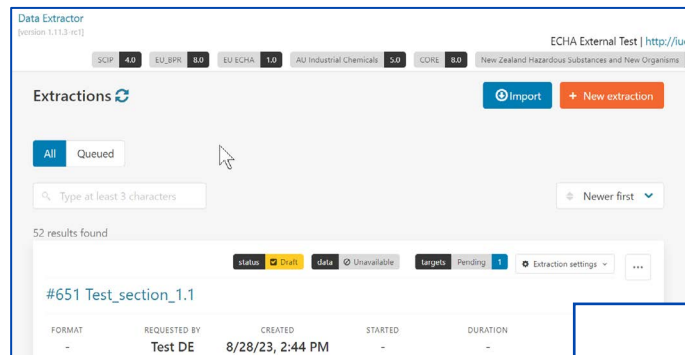
IUCLID Data Extractor installer (7th July 2023 - v1.11.3)

The installation package contains the software components needed to deploy IUCLID Data Extractor on either a Windows or Linux server, and to connect it to an installation of IUCLID Server 6.7. Data Extractor can use either its own embedded H2 database, or a separate Oracle database which must be obtained separately, from its vendor. Documentation is provided on this page from the links above.

There is no updater tool for IUCLID Data Extractor. To use the current version, make a fresh installation.

Please sign in to download files

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

The screenshot shows the configuration page for 'Extraction #651: Test_section_1.1'. The page has a header with 'Extractions' and 'Extraction #651: Test_section_1.1'. There are buttons for 'Reset', 'Save', and 'Create'. The 'Name' field is 'Test_section_1.1' and the 'Format' is 'Normalised'. The 'Data settings' section includes options for 'Replace new lines with:', 'Column delimiter:' (set to '<Tab>'), 'Replace delimiter with:', and a checkbox for 'Remove HTML tags from rich text fields'. The 'Description' field contains the text 'Extract section 1.1, some sections relevant for structure based identification'. The 'Targets' section has a 'Manual input' tab and a 'From file' tab. A 'Choose a file...' button is present, with a callout stating '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'. Below this, it says 'At least one target must be provided' and '0 target(s)'. The 'Fields' section has a callout stating 'When you start selecting IUCLID fields to be extracted, the fields are listed here'. At the bottom, it says 'Attachments' and '0 attachment(s)'.

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

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

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

The screenshot displays the configuration interface for an extraction job, organized into several panels:

- TOC (Table of Contents):** A list of sections with checkboxes for selection. ☒ REACH, ☐ BPR, ☐ PPP, ☐ CLP, ☐ CORE, ☐ NZ_HSN0, ☐ AU_IND_CHEM, ☐ OECD, ☐ DWD, ☐ UK_REACH, ☐ EFSA. Below this is a dropdown for "REACH group 2".
- Filter sections:** A section for filtering the data.
- # Dossier Header:** A list of sections with checkboxes. ☒ 1.1 Identification, ☐ 1.2 Composition, ☐ 1.3 Identifiers, ☐ 1.4 Analytical Information, ☐ 1.5 Joint submission, ☐ 1.6 Sponsors, ☐ 1.7 Suppliers, ☐ 1.8 Recipients, ☐ 1.10 Assessment approach (assessment entities), ☐ 2 Classification & Labelling and PBT assessment, ☐ 3 Manufacture, use and exposure, ☐ 4 Physical and chemical properties, ☐ 5 Environmental fate and pathways, ☐ 6 Ecotoxicological information, ☐ 7 Toxicological information, ☐ 8 Analytical methods, ☐ 11 Guidance on safe use, ☐ 12 Literature search, ☐ 13 Assessment reports, ☐ 14 Information requirements.
- Substance:** A dropdown menu for selecting the substance.
- Identification -> Reference substance:** A section for identifying the reference substance. It includes checkboxes for "Select All", "DataProtection", "Reference substance name", "IUPAC name", and "Description".
- Inventory:** A section for inventory management. It includes checkboxes for "Inventory number", "No inventory information available - Justification", "CAS number", and "CAS name".
- Synonyms:** A section for synonyms. It includes checkboxes for "DataProtection", "Identifier", "Identity", and "Remarks".
- Molecular and structural information:** A section for molecular and structural information. It includes checkboxes for "DataProtection", "Molecular formula", "Molecular weight", "SMILES notation", "InChI", "Structural formula", "Include attachment", "Remarks", "Chemical structure files", "Structure file", "Include attachment", and "Remarks on structure file".
- Related substances:** A section for related substances.

Download the Extracted Data

The image shows two side-by-side screenshots of the 'Extractions' interface, with a large blue arrow pointing from the left screenshot to the right one, indicating a transition in the extraction job's status.

Left Screenshot (In Queue):

- Buttons: Import, + New extraction
- Tabs: All, Queued
- Search: Type at least 3 characters
- Results: 1 results found
- Filters: status In Queue, data Unavailable, targets Pending 1, Extraction settings
- Job #651 Test_section_1.1
- Table:

FORMAT	REQUESTED BY	CREATED	STARTED	DURATION	EXPIRES	PRIORITY
Normalized	Test DE	8/28/23, 2:44 PM	-	-	-	1

Right Screenshot (Finished):

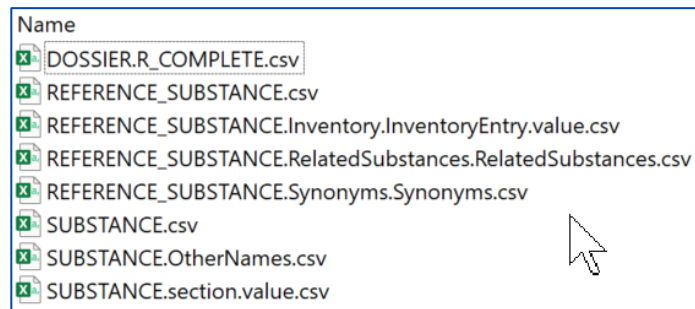
- Buttons: Import, + New extraction
- Tabs: All, Queued
- Search: Type at least 3 characters
- Results: 2 results found
- Filters: status Finished, data Available, targets Processed 1, Extraction settings
- Job #651 Test_section_1.1
- Table:

FORMAT	REQUESTED BY	CREATED	STARTED	DURATION	EXPIRES
Normalized	Test DE	8/28/23, 2:44 PM	8/29/23, 3:25 PM	00 00 01 days hours min	9/18/23, 3:26 PM

- 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"

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

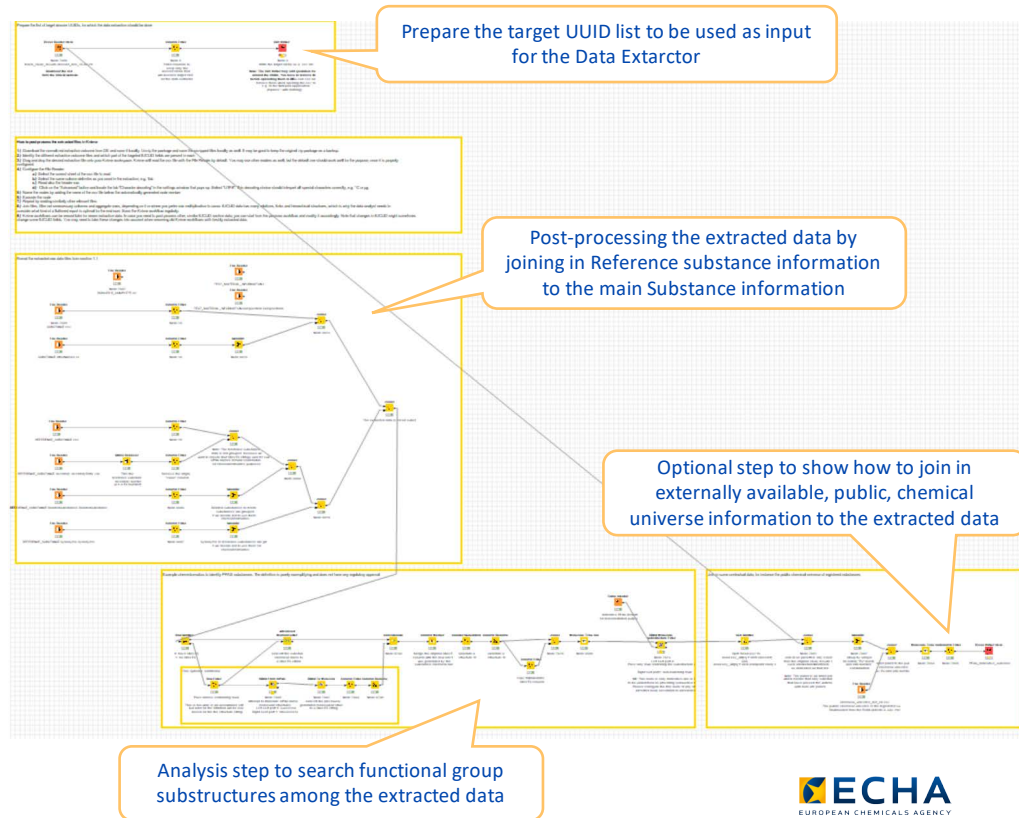
- IUCLID data is nested, the extraction essentially produces a different table for every level in the hierarchy
- The aim of the post-processing is to join the files and build a single report
- This single report will inevitably result in an increased number of rows; the data-analyst can determine the 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: [\[https://www.knime.com/\]](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

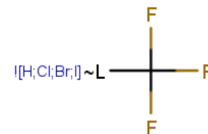
- The use case material contains an KNIME workflow shown to the right
- View the example workflow from the top left to the bottom right and read the added clarifications
 - The input files refer to the relevant extracted flat data files
 - The input files are joined by the identified foreign key columns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



Cheminformatics

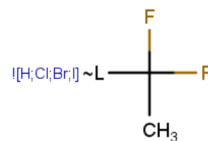
- We identify substances by applying a sub-structure search to the structural information available in the dossier
- For dossiers that do not have a structure given as a SMILES string, we attempt to convert IUPAC names to SMILES strings
- We use cheminformatics nodes from the open source RDkit, but others 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](https://www.instagram.com/one_healthenv_eu)



[@EU_ECHA](https://twitter.com/EU_ECHA)



[@EUECHA](https://www.facebook.com/EUECHA)



[EUchemicals](https://www.youtube.com/EUchemicals)