8th meeting of the PSN IUCLID sub-group
22 November 2023

REUSE OF IUCLID DATA

Edoardo CARNESECCHI, Dayana BUZLE

iDATA Unit

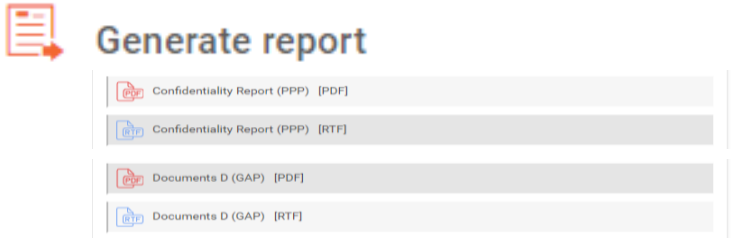
EUROPEAN FOOD SAFETY AUTHORITY

OUTLINE

- Overview of IUCLID tools
- Data Extractor (DE)
- Text Analytics (TA)
- EU Survey for Member States Competent Authorities

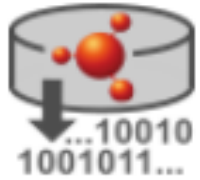
IUCLID TOOLS FOR DATA REUSING

- **Report Generator:**
  - Extracting data as predefined report (PDF, RTF, CSV)



https://iuclid6.echa.europa.eu/it/reports

- **Data Extractor (DE)**:
  - web-based user interface
  - **extracts data/datasets** according to a set of user-defined rules



https://iuclid6.echa.europa.eu/it/data-extractor

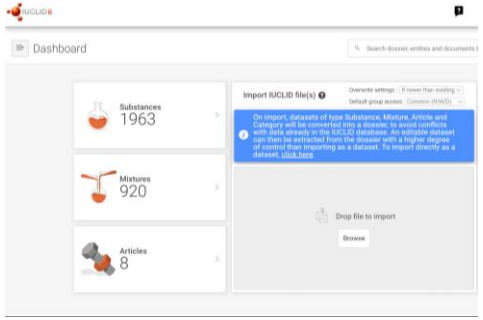
- **Text Analytics (TA):**
  - web-based user interface
  - **Search engine** (e.g., text, attachment)

https://iuclid6.echa.europa.eu/it/text-analytics

IUCLID TOOLS FOR DATA REUSING – STATE OF THE ART

- Currently, **DE** and **TA** are NOT supported/implemented in **EFSA Agency IUCLID** (secure instance).

- At EFSA, **DE** and **TA** can be run by EFSA staff only in the **IUCLID (ECHA) Test instance** for testing purposes only.



- **DE** and **TA** can be downloaded from ECHA-IUCLID website as industry account and run on a local instance.

DATA EXTRACTOR (DE)?

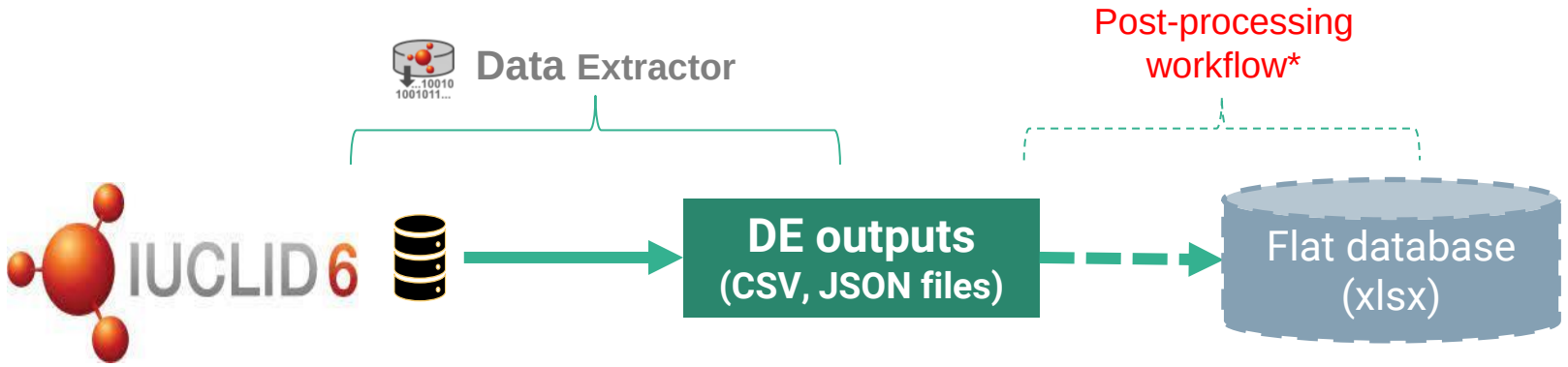
What DE is:

- A powerful, flexible web-based tool for data experts to **retrieve (raw) data** from any IUCLID database

What DE does:

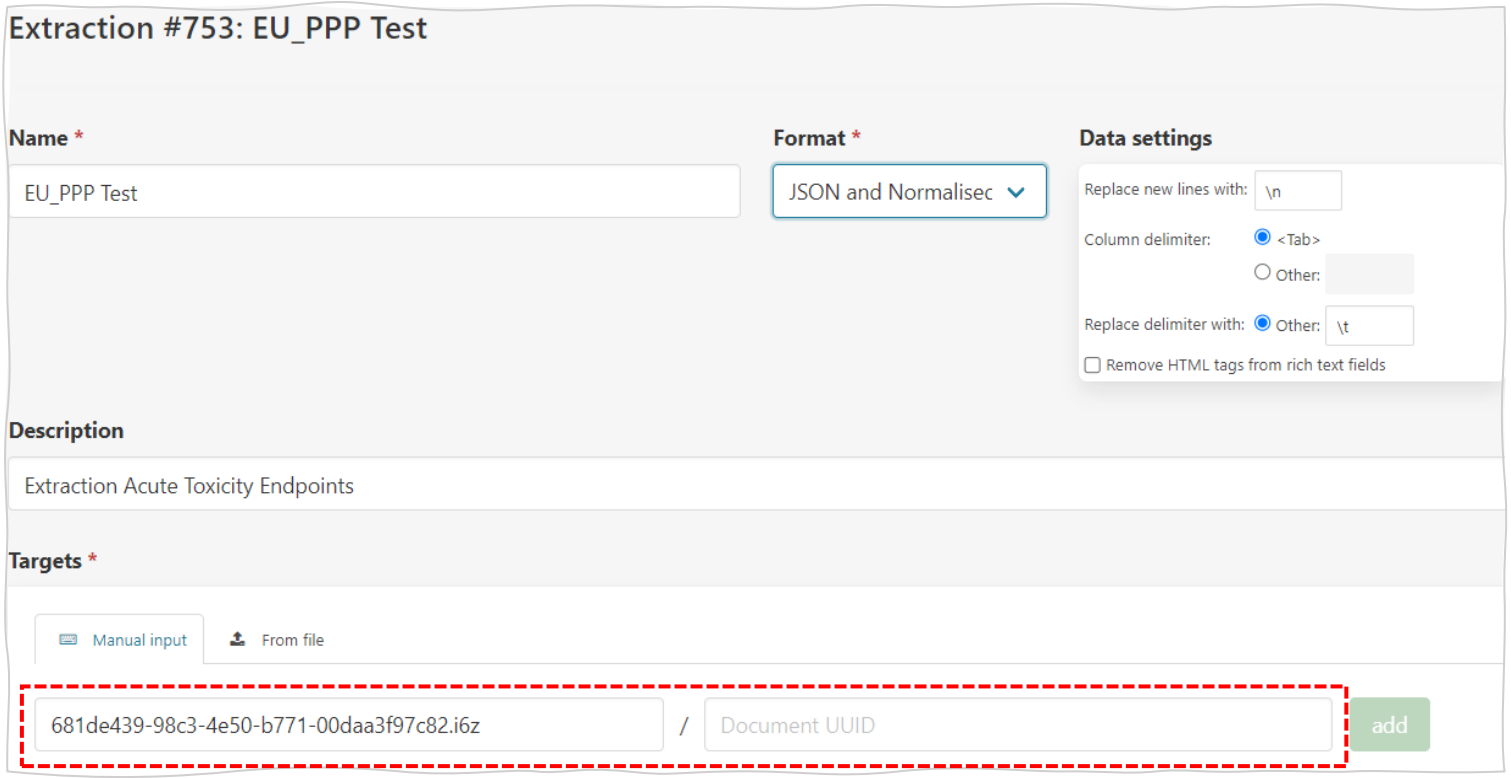
- extracts all **field data** and **field/document attachments** of defined IUCLID sections
- extracts from a **high number** of dossiers/datasets
- provides extracted data in **flat format*** i.e., JSON and normalised CSV, which is versatile as input for ad hoc data analysis

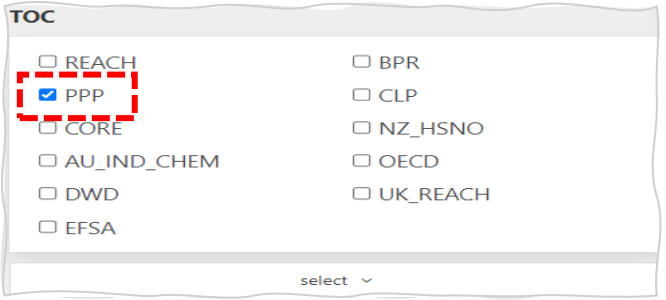
***** Users need to use **programming tools** (e.g., KNIME, Python) to develop a post-processing workflow for any ad-hoc analysis on the flat files

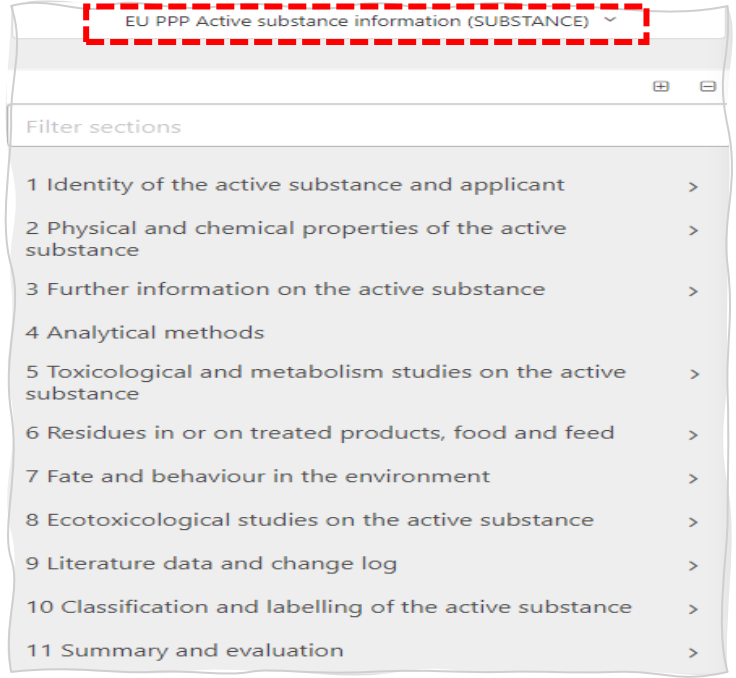


HOW DOES DE WORK? EXAMPLE OF GENOTOXICITY DATA EXTRACTION

- **INPUT**:
  - Target dossier/document **UUIDs**;
  - Table of Content (TOC) & IUCLID docs/fields





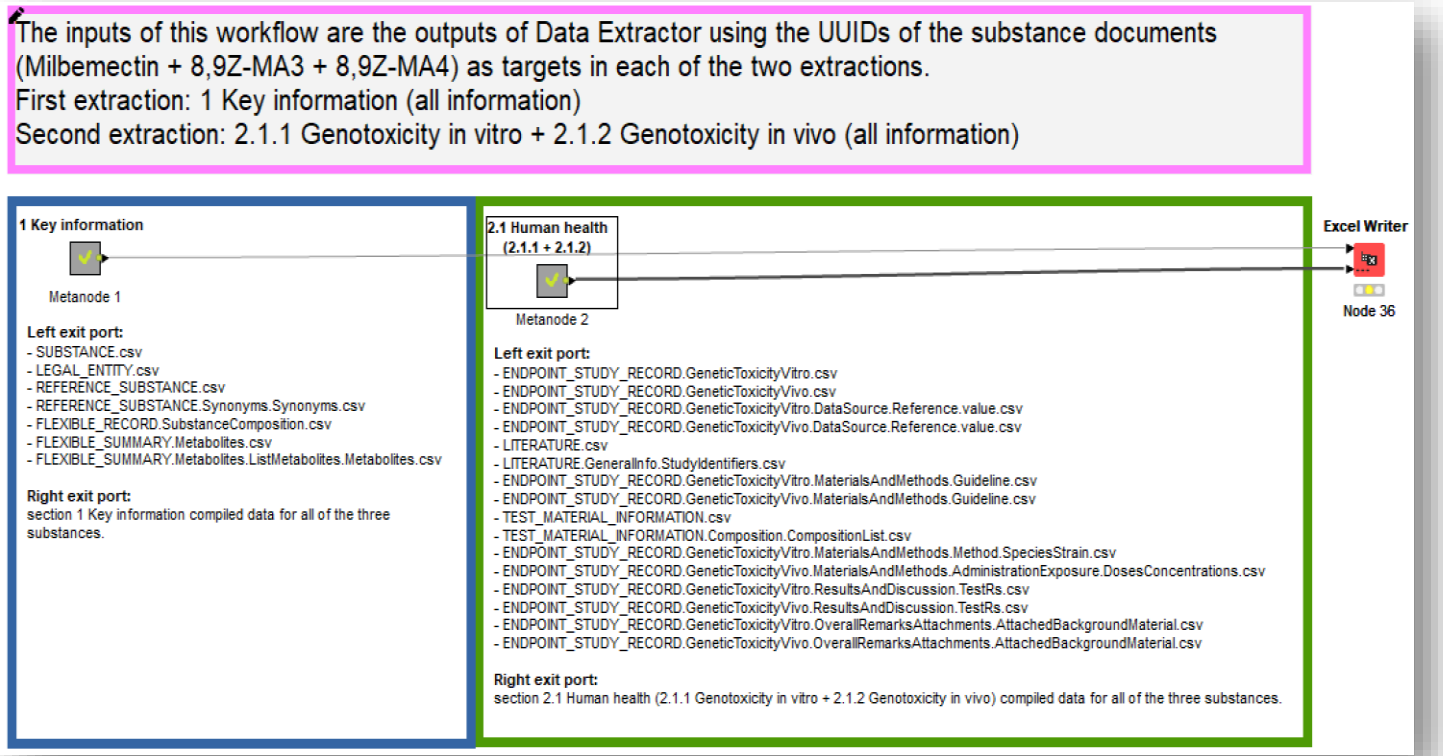


POST PROCESSING WORKFLOW TO AGGREGATE DE OUTPUT FILES (CSV) AS A FLAT FILE

➢ **OUTPUT: multiple CSV** files (based on the IUCLID fields selected as input)

- **N = 7 CSV** files - Substance info
  - FLEXIBLE_RECORD.SubstanceComposition
  - FLEXIBLE_SUMMARY.Metabolites
  - FLEXIBLE_SUMMARY.Metabolites.ListMetabolites.Metabolites
  - LEGAL_ENTITY
  - REFERENCE_SUBSTANCE
  - REFERENCE_SUBSTANCE.Synonyms.Synonyms
  - SUBSTANCE

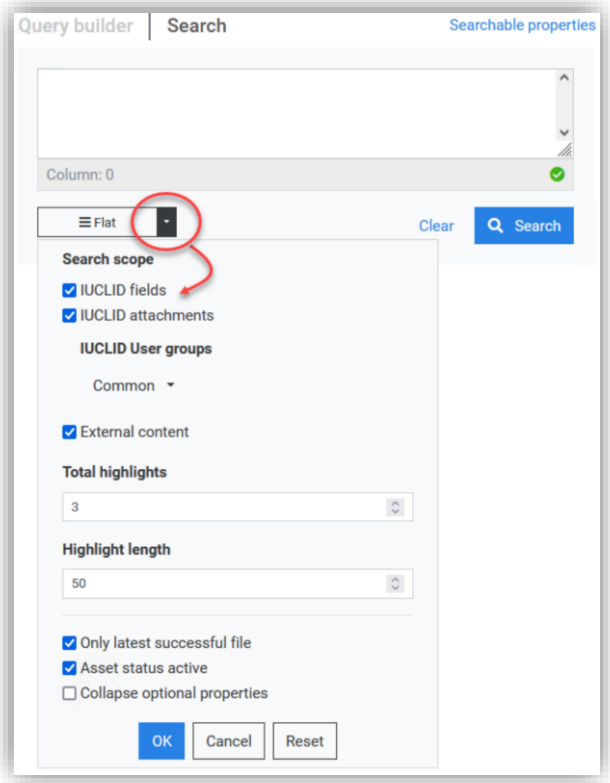
- **N = 17 CSV** files - Human health (Genotox)
  - ENDPOINT_STUDY_RECORD.GeneticToxicityVitro
  - ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.DataSource.Reference.value
  - ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline
  - ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Method.SpeciesStrain
  - ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.OverallRemarksAttachments.AttachedBackgroundMaterial
  - ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs
  - ENDPOINT_STUDY_RECORD.GeneticToxicityVivo
  - ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.DataSource.Reference.value
  - ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.MaterialsAndMethods.AdministrationExposure.DosesConcentrations
  - ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.MaterialsAndMethods.Guideline
  - ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.OverallRemarksAttachments.AttachedBackgroundMaterial
  - ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.ResultsAndDiscussion.TestRs
  - LITERATURE
  - LITERATURE.GeneralInfo.StudyIdentifiers
  - SUBSTANCE
  - TEST_MATERIAL_INFORMATION.Composition.CompositionList
  - TEST_MATERIAL_INFORMATION



Docs **UUIDs** as **keys** to aggregate/merge CSV files

WHAT IS TEXT ANALYTICS (TA)?

- **Search engine** enabling execution of queries on **IUCLID Dossier(s)**, including:
  - ✓ **IUCLID fields** (text fields, pick lists, numbers, checkboxes, etc.)
  - ✓ **IUCLID attachments** (including scanned documents and images)
  - ✓ **Documents uploaded from external sources** (other than IUCLID)
- TA provides **multiple languages query support**; user can perform queries against content that may be written in various languages.
- TA is **fast** and **efficient**: results are returned in a few seconds.
- TA supports **Optical character recognition** (OCR): TA extracts the text from the scanned file and the original attached PDF can be downloaded.



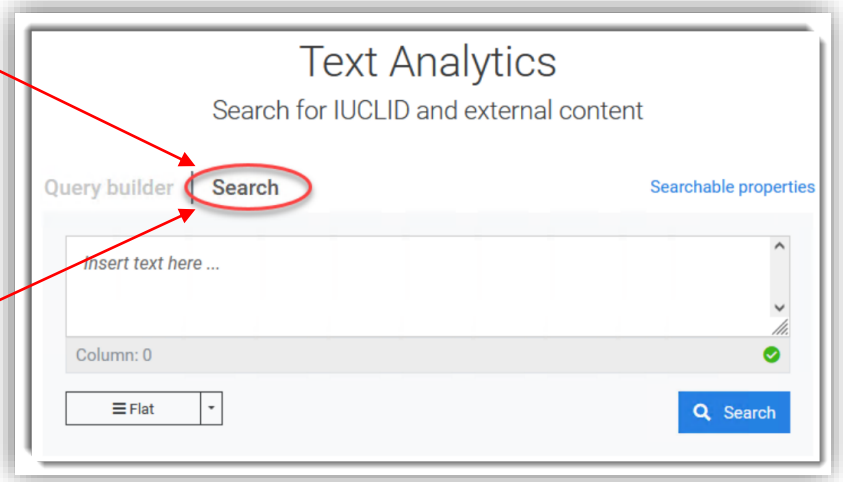
File type or application	File extension
PDF	pdf
Microsoft Word	doc, docx
Microsoft Excel	xls, xlsx
HTML (web page)	htm, html
RTF	rtf

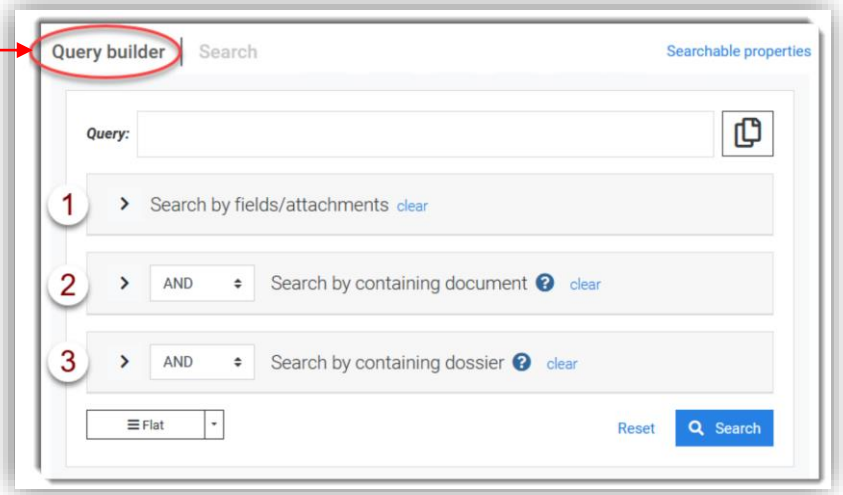
THREE WAYS TO SEARCH IN TEXT ANALYTICS

1. **Search for a phrase – level 1** (lowest complexity): Enter words related to the intended search result (similar to using a web-browser). If user needs to do a simple search, or is not familiar with the terminology/data structure of IUCLID.

2. **Search using controls applied to the search terms – level 2** (intermediate complexity): Entering a search term in text AND controls in the search criteria (e.g., wildcards, Boolean logic, and/or groupings). If user is looking for specific information, e.g., excluding certain search results.

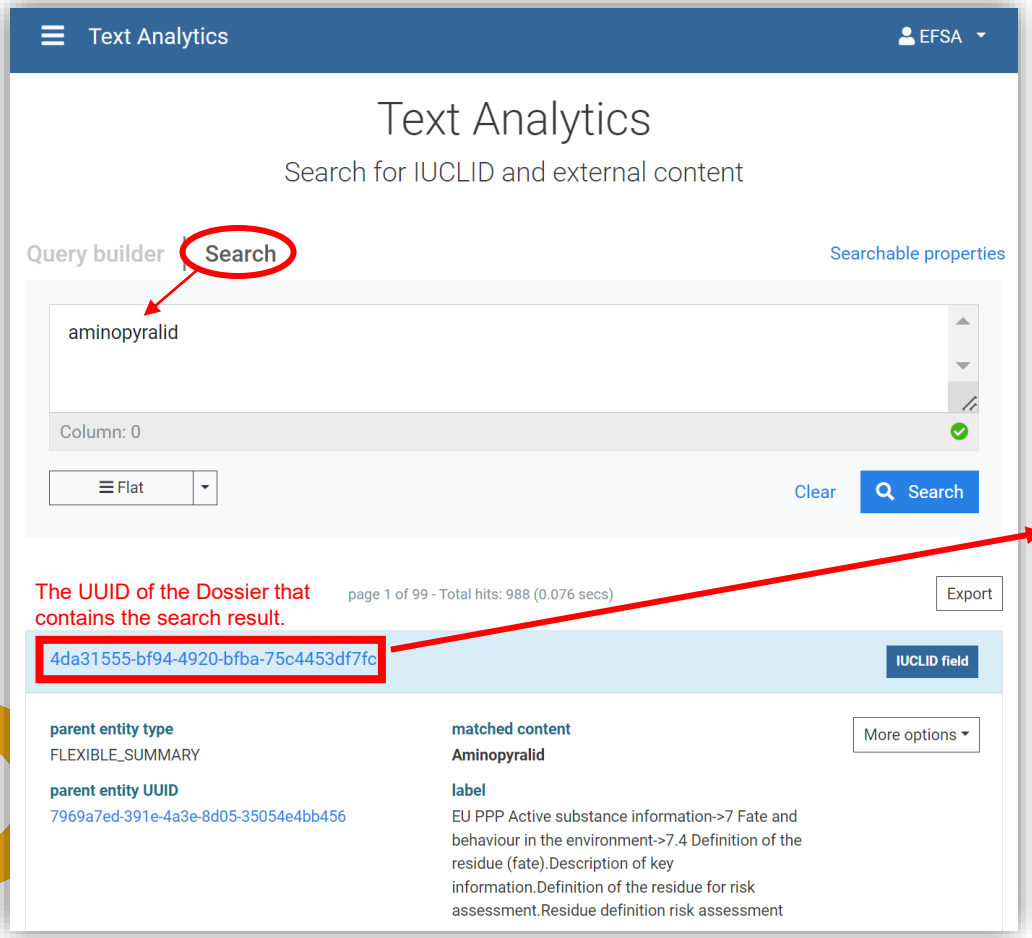
3. **Search using the query language of *Text Analytics* – level 3** (highest complexity): building a specific query using "*Query builder*" tab. A refinement of the query created using *Query builder* can be done by transferring the query to the *Search* tab, and then editing it manually. If user is looking for something highly specific, or wants to control exactly what type of data is found.

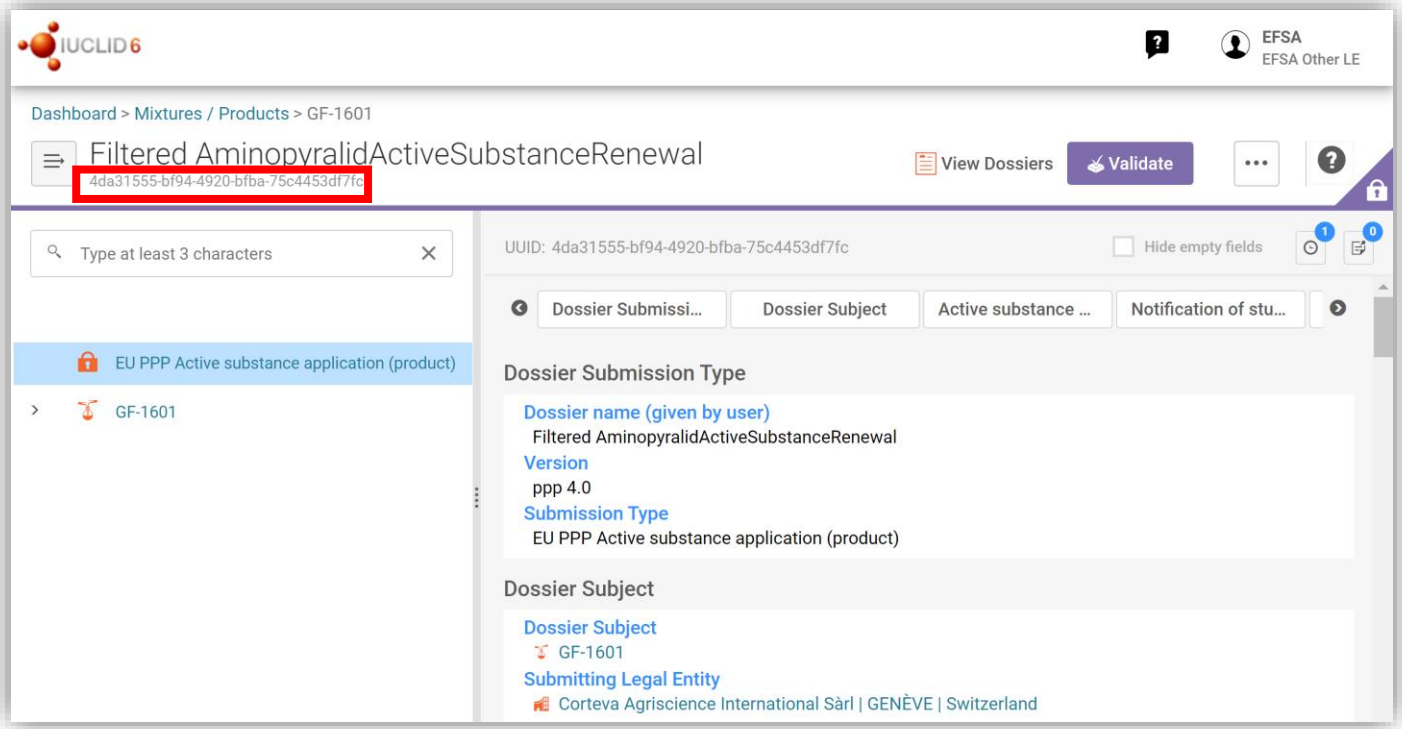




EXAMPLE: SEARCHING FOR A DOSSIER UNDER EU PPP WORKING CONTEXT (*SEARCH* TAB – LEVEL 1)

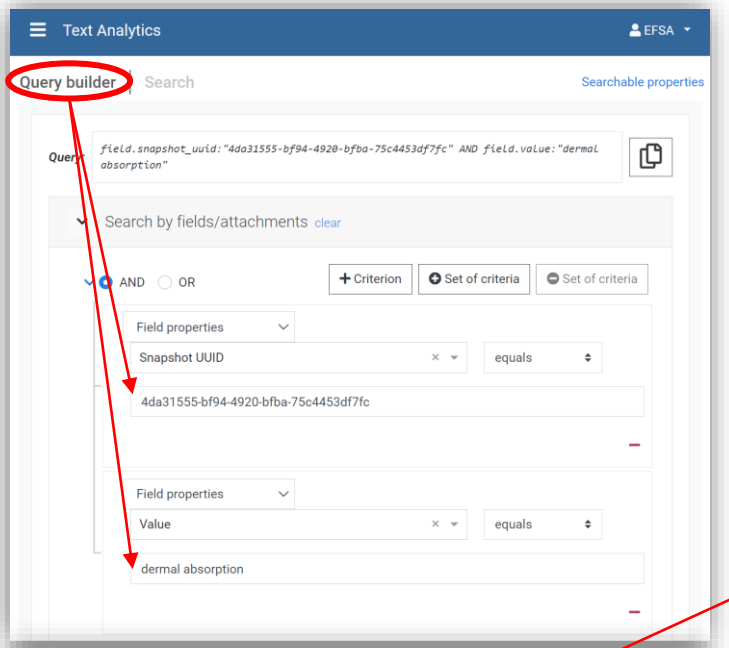
- In the “Search” tab, type ***aminopyralid*** to find the dossier shown in the web interface of IUCLID.
- A new window in IUCLID will be opened when clicking on the UUID of the dossier that contains the search result.

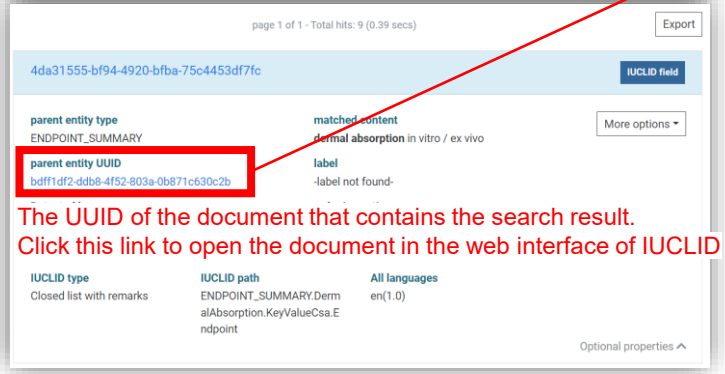




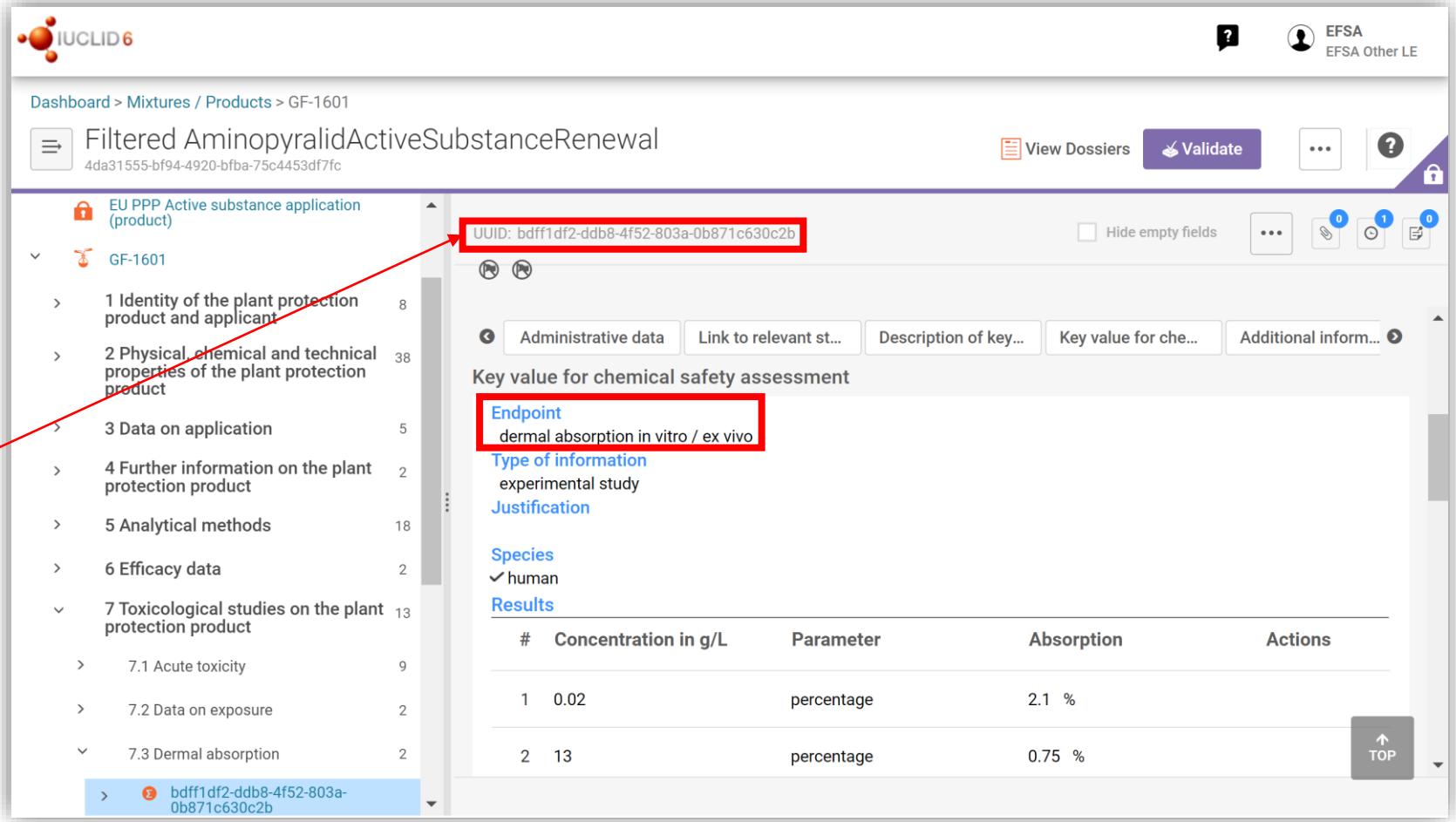
EXAMPLE: SEARCHING FOR A DOSSIER UNDER EU PPP WORKING CONTEXT (*QUERY BUILDER* TAB – LEVEL 3)

- Go to the "***Query builder***" tab and type the UUID of the dossier (Aminopyralid) AND dermal absorption into the dedicated search boxes, to find the ENDPOINT SUMMARY document containing the values for "dermal absorption" in IUCLID. A new window will be opened by clicking on the UUID of the document.





The UUID of the document that contains the search result.
Click this link to open the document in the web interface of IUCLID

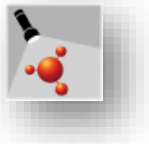


EU SURVEY FOR RISK ASSESSORS – HAVE YOUR SAY!

Gathering MSs/Risk Assessors feedback on:

- EFSA Agency IUCLID **data reuse** needs

- **Searching** and **analysis** needs in EFSA Agency IUCLID to support the RA process
- Current and future use of **IUCLID tools** (DE, TA, Report Generator)



Survey for risk assessors (Member States and EFSA staff) to gather use cases/needs on IUCLID data reuse

Not published

b20c74c1-57e9-a5fb-f946-856b70f32049

Survey will be shared via email/PSN Teams channel

ACKNOWLEDGMENTS

- **DATA Team**:

  T. ALASUVANTO, X. COMPAS, A. FRONTINI, P. KARAMERTZANIS, U. PIRNAR

STAY CONNECTED

SUBSCRIBE TO
efsa.europa.eu/en/news/newsletters
efsa.europa.eu/en/rss
Careers.efsa.europa.eu – job alerts

LISTEN TO OUR PODCAST
Science on the Menu –Spotify, Apple Podcast and YouTube

FOLLOW US ON TWITTER
@efsa_eu @methods_efsa
@plants_efsa @animals_efsa

FOLLOW US ON LINKEDIN
Linkedin.com/company/efsa

FOLLOW US ON INSTAGRAM
@one_healthenv_eu

CONTACT US
efsa.europa.eu/en/contact/askefsa

efsa
EUROPEAN FOOD SAFETY AUTHORITY