



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

REUSE OF IUCLID DATA



Edoardo CARNESECCHI, Dayana BUZZLE

iDATA Unit

OUTLINE



Overview of IUCLID tools



Data Extractor (DE)



Text Analytics (TA)



EU Survey for Member States Competent Authorities



IUCOLID TOOLS FOR DATA REUSING

➤ Report Generator:

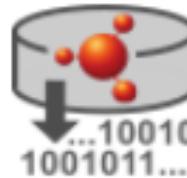
- Extracting data as predefined report (PDF, RTF, CSV)



<https://iucclid6.echa.europa.eu/it/reports>

➤ Data Extractor (DE):

- web-based user interface
- **extracts data/datasets** according to a set of user-defined rules



<https://iucclid6.echa.europa.eu/it/data-extractor>

➤ Text Analytics (TA):

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

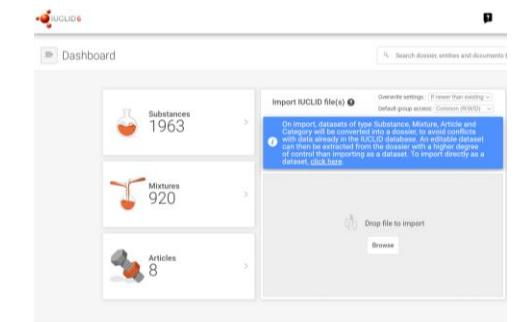


<https://iucclid6.echa.europa.eu/it/text-analytics>

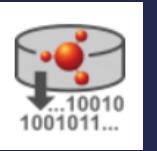


IUCLID TOOLS FOR DATA REUSING – STATE OF THE ART

- 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)?



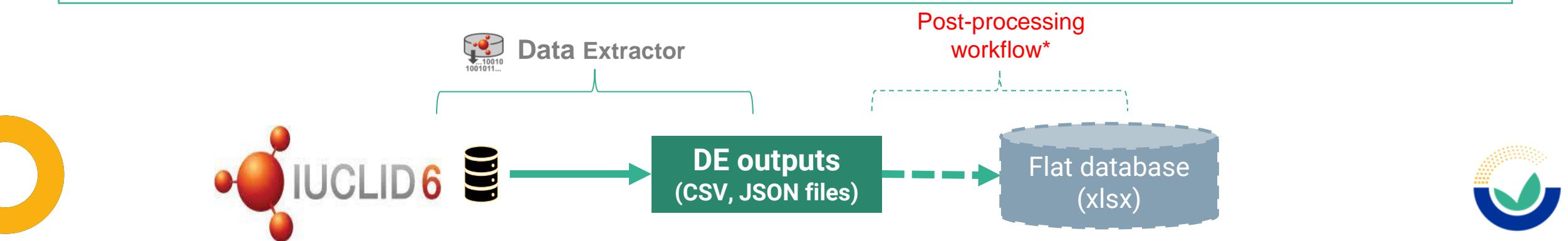
What DE is:

- 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

Extraction #753: EU_PPP Test

Name *	Format *	Data settings
EU_PPP Test	JSON and Normalisec	Replace new lines with: <code>\n</code> Column delimiter: <input checked="" type="radio"/> <Tab> <input type="radio"/> Other: <code>\t</code> Replace delimiter with: <input checked="" type="radio"/> Other: <code>\t</code> <input type="checkbox"/> Remove HTML tags from rich text fields
Description	Extraction Acute Toxicity Endpoints	
Targets *	<input type="checkbox"/> Manual input <input type="checkbox"/> From file 681de439-98c3-4e50-b771-00daa3f97c82.i6z / Document UUID <input type="button" value="add"/>	

TOC

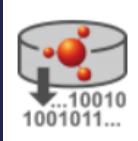
<input type="checkbox"/> REACH	<input type="checkbox"/> BPR
<input checked="" type="checkbox"/> PPP	<input type="checkbox"/> CLP
<input type="checkbox"/> CORE	<input type="checkbox"/> NZ_HSN
<input type="checkbox"/> AU_IND_CHEM	<input type="checkbox"/> OECD
<input type="checkbox"/> DWD	<input type="checkbox"/> UK_REACH
<input type="checkbox"/> EFSA	

select ▾

EU PPP Active substance information (SUBSTANCE) ▾

Filter sections
1 Identity of the active substance and applicant >
2 Physical and chemical properties of the active substance >
3 Further information on the active substance >
4 Analytical methods >
5 Toxicological and metabolism studies on the active substance >
6 Residues in or on treated products, food and feed >
7 Fate and behaviour in the environment >
8 Ecotoxicological studies on the active substance >
9 Literature data and change log >
10 Classification and labelling of the active substance >
11 Summary and evaluation >

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)

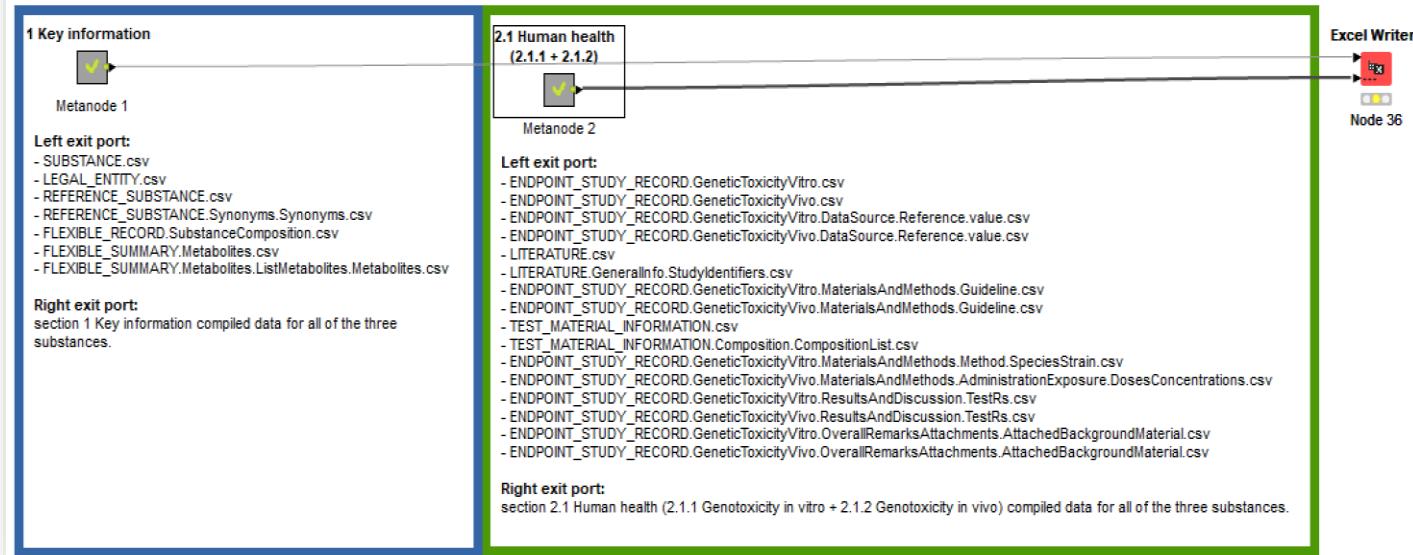
▪ N = 7 CSV files - Substance info

- CSV FLEXIBLE_RECORD.SubstanceComposition
- CSV FLEXIBLE_SUMMARY.Metabolites
- CSV FLEXIBLE_SUMMARY.Metabolites.ListMetabolites.Metabolites
- CSV LEGAL_ENTITY
- CSV REFERENCE_SUBSTANCE
- CSV REFERENCE_SUBSTANCE.Synonyms.Synonyms
- CSV SUBSTANCE

▪ N = 17 CSV files - Human health (Genotox)

- CSV ENDPOINT_STUDY_RECORD.GeneticToxicityVitro
- CSV ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.DataSource.Reference.value
- CSV ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Guideline
- CSV ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.MaterialsAndMethods.Method.SpeciesStrain
- CSV ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.OverallRemarksAttachments.AttachedBackgroundMaterial
- CSV ENDPOINT_STUDY_RECORD.GeneticToxicityVitro.ResultsAndDiscussion.TestRs
- CSV ENDPOINT_STUDY_RECORD.GeneticToxicityVivo
- CSV ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.DataSource.Reference.value
- CSV ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.MaterialsAndMethods.AdministrationExposure.DosesConcentrations
- CSV ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.MaterialsAndMethods.Guideline
- CSV ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.OverallRemarksAttachments.AttachedBackgroundMaterial
- CSV ENDPOINT_STUDY_RECORD.GeneticToxicityVivo.ResultsAndDiscussion.TestRs
- LITERATURE
- LITERATURE.GeneralInfo.StudyIdentifiers
- SUBSTANCE
- TEST_MATERIAL_INFORMATION.Composition.CompositionList
- TEST_MATERIAL_INFORMATION

The inputs of this workflow are the outputs of Data Extractor using the UUIDs of the substance documents (Milbemectin + 8,9Z-MA3 + 8,9Z-MA4) as targets in each of the two extractions.
First extraction: 1 Key information (all information)
Second extraction: 2.1.1 Genotoxicity in vitro + 2.1.2 Genotoxicity in vivo (all information)



Docs UUIDs as keys to aggregate/merge CSV files



WHAT IS TEXT ANALYTICS (TA)?



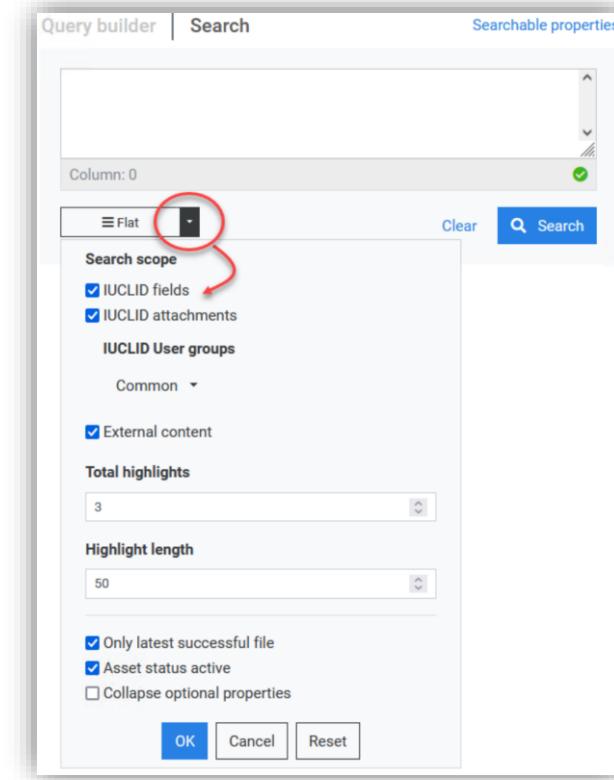
- **Search engine** enabling execution of queries on **IUCLID Dossier(s)**, including:

- ✓ **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.**

Text Analytics
Search for IUCLID and external content
Searchable properties

Query builder | **Search**

Insert text here ...
Column: 0
Flat

Search

Text Analytics
Search for IUCLID and external content
Searchable properties

Query builder | **Search**

Query:

1 > Search by fields/attachments

2 > AND Search by containing document

3 > AND Search by containing dossier

Flat

Reset **Search**

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.

The screenshot illustrates the search process for a dossier under the EU PPP working context. On the left, the 'Text Analytics' interface shows a search query 'aminopyralid' in the 'Query builder' search bar. The 'Search' button is highlighted with a red circle. Below the search bar, the text 'The UUID of the Dossier that contains the search result.' is displayed in red, followed by the UUID '4da31555-bf94-4920-bfba-75c4453df7fc' in a red box. On the right, the 'IUCLID 6' dossier detail page is shown for dossier 'GF-1601'. The UUID '4da31555-bf94-4920-bfba-75c4453df7fc' is highlighted in a red box in the search bar. The dossier summary includes the title 'EU PPP Active substance application (product)' and the identifier 'GF-1601'. The 'Dossier Submission Type' section lists the dossier name as 'Filtered AminopyralidActiveSubstanceRenewal', version 'ppp 4.0', and submission type 'EU PPP Active substance application (product)'. The 'Dossier Subject' section lists the subject as 'GF-1601' and the submitting legal entity as 'Corteva Agriscience International Sàrl | GENÈVE | Switzerland'.

Text Analytics

Search for IUCLID and external content

Query builder

Search

aminopyralid

Column: 0

Flat

Clear

Search

Export

IUCLID field

The UUID of the Dossier that contains the search result.

4da31555-bf94-4920-bfba-75c4453df7fc

parent entity type: FLEXIBLE_SUMMARY
parent entity UUID: 7969a7ed-391e-4a3e-8d05-35054e4bb456

matched content: Aminopyralid

label: EU PPP Active substance information->7 Fate and behaviour in the environment->7.4 Definition of the residue (fate). Description of key information. Definition of the residue for risk assessment. Residue definition risk assessment

Searchable properties

Dashboard > Mixtures / Products > GF-1601

Filtered AminopyralidActiveSubstanceRenewal

4da31555-bf94-4920-bfba-75c4453df7fc

Type at least 3 characters

EU PPP Active substance application (product)

GF-1601

Dossier Submission Type

Dossier name (given by user): Filtered AminopyralidActiveSubstanceRenewal
Version: ppp 4.0
Submission Type: EU PPP Active substance application (product)

Dossier Subject

Dossier Subject: GF-1601
Submitting Legal Entity: Corteva Agriscience International Sàrl | GENÈVE | Switzerland

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 screenshot illustrates the search process for a dossier under EU PPP Working Context. On the left, the 'Text Analytics' interface shows the 'Query builder' tab selected. A red box highlights the 'Query' field containing the search criteria: 'field.snapshot_uuid:"4da31555-bf94-4920-bfba-75c4453df7fc" AND field.value:"dermal absorption"'. Below this, the 'Search by fields/attachments' section is expanded, showing two search criteria: 'Snapshot UUID' equals '4da31555-bf94-4920-bfba-75c4453df7fc' and 'Value' equals 'dermal absorption'. The search results page shows a single hit with the UUID '4da31555-bf94-4920-bfba-75c4453df7fc'. A red box highlights the 'parent entity UUID' field, which also contains 'bdff1df2-ddb8-4f52-803a-0b871c630c2b'. A red arrow points from this UUID to the corresponding document in the IUCLID 6 interface on the right.

The right side shows the IUCLID 6 document detail view for dossier 'GF-1601'. The document title is 'Filtered AminopyralidActiveSubstanceRenewal' with UUID '4da31555-bf94-4920-bfba-75c4453df7fc'. The document structure includes sections for EU PPP Active substance application (product), GF-1601, and various toxicity studies. A red box highlights the 'UUID: bdff1df2-ddb8-4f52-803a-0b871c630c2b' in the document header. Another red box highlights the 'Endpoint' section, which lists 'dermal absorption in vitro / ex vivo' as the type of information for an experimental study. The 'Results' table shows two entries for dermal absorption, with values 0.02 and 13, and absorption percentages of 2.1% and 0.75% respectively.

EU SURVEY FOR RISK ASSESSORS – HAVE YOUR SAY!

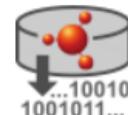


➤ Gathering MSs/Risk Assessors feedback on:

- 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)



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



FOLLOW US ON TWITTER

[@efsa_eu](https://twitter.com/efsa_eu) [@methods_efsa](https://twitter.com/methods_efsa)
[@plants_efsa](https://twitter.com/plants_efsa) [@animals_efsa](https://twitter.com/animals_efsa)



FOLLOW US ON INSTAGRAM
[@one_healthenv_eu](https://www.instagram.com/one_healthenv_eu)



LISTEN TO OUR PODCAST

Science on the Menu – Spotify, Apple Podcast and YouTube



FOLLOW US ON LINKEDIN

[Linkedin.com/company/efsa](https://www.linkedin.com/company/efsa)



CONTACT US

efsa.europa.eu/en/contact/askefsa

