

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

# REUSE OF IUCLID DATA

Edoardo CARNESECCHI, Dayana BUZLE

iDATA Unit

# 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)



### Generate report

Confidentiality Report (PPP) [PDF]

Confidentiality Report (PPP) [RTF]

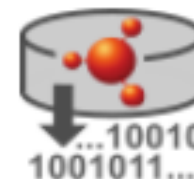
Documents D (GAP) [PDF]

Documents D (GAP) [RTF]

<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)

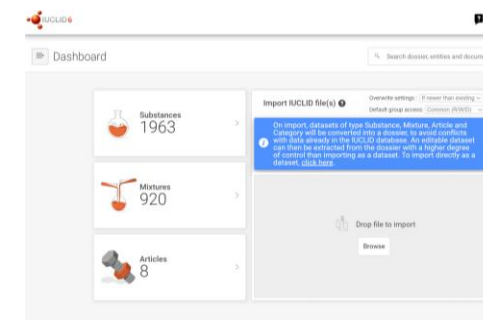
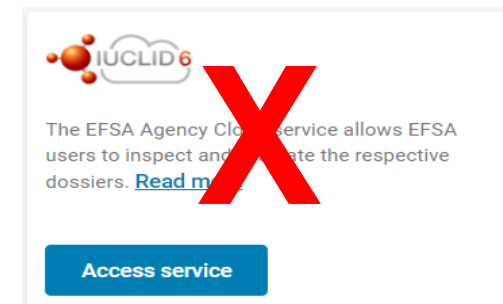


<https://iuclid6.echa.europa.eu/it/text-analytics>



# 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)?



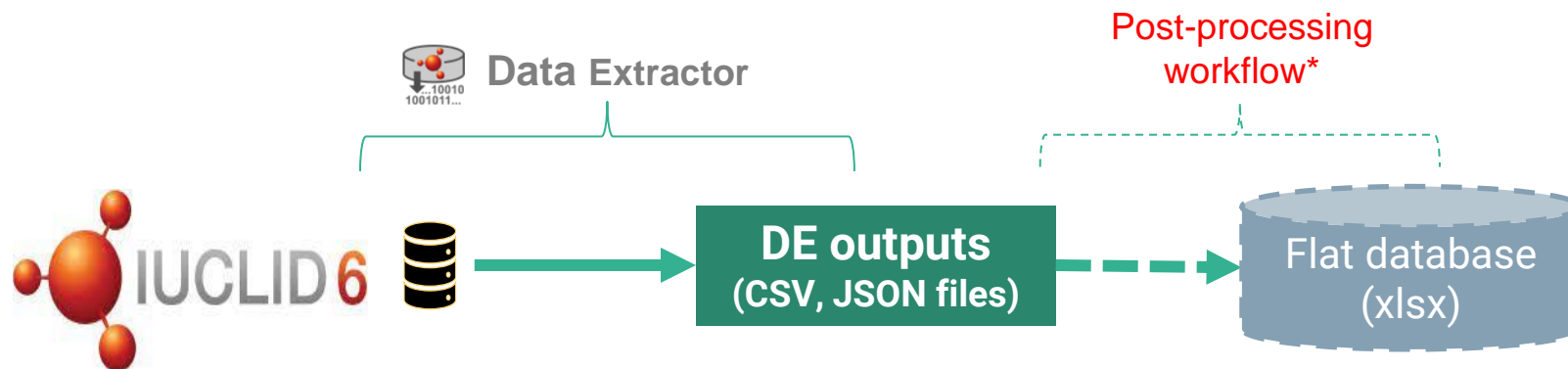
## What DE is:

- 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

Extraction #753: EU\_PPP Test

**Name \***  
EU\_PPP Test

**Format \***  
JSON and Normalise ✓

**Data settings**  
Replace new lines with: \n  
Column delimiter: ☒ <Tab> ☐ Other:   
Replace delimiter with: ☒ Other: \t  
☐ Remove HTML tags from rich text fields

**Description**  
Extraction Acute Toxicity Endpoints

**Targets \***  
Manual input From file

681de439-98c3-4e50-b771-00daa3f97c82.i6z / Document UUID 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_HSNO
<input type="checkbox"/> AU_IND_CHEM	<input type="checkbox"/> OECD
<input type="checkbox"/> DWD	<input type="checkbox"/> UK_REACH
<input type="checkbox"/> EFSA	

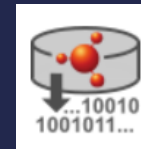
select ▼

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)

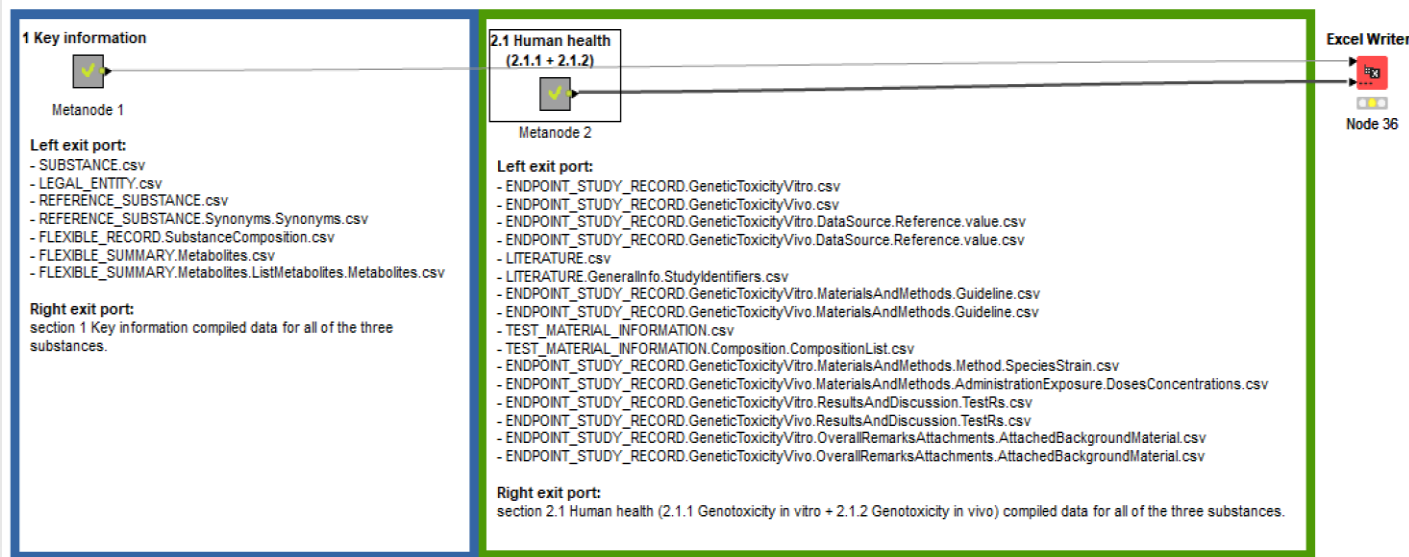
## ■ N = 7 CSV files - Substance info

- 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

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.

Query builder | Search | Searchable properties

Column: 0

Flat

Search scope

- ☒ IUCLID fields
- ☒ IUCLID attachments
- ☒ External content

IUCLID User groups

Common

Total highlights

3

Highlight length

50

☒ Only latest successful file

☒ Asset status active

☐ Collapse optional properties

OK Cancel Reset

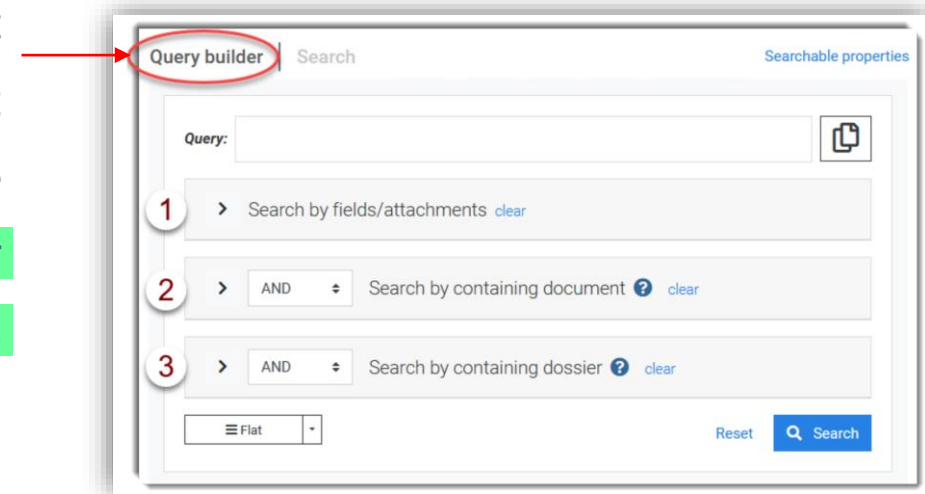
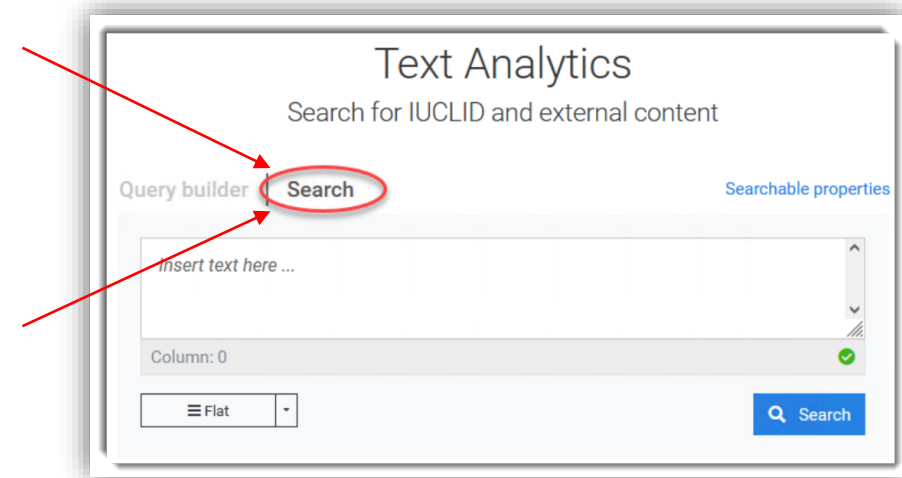
File type or application	File extension
PDF	pdf
Microsoft Word	doc, docx
Microsoft Excel	xls,xlsx
HTML (web page)	htm,html
RTF	rtf



# 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.

Text Analytics  
Search for IUCLID and external content

Query builder **Search** Searchable properties

aminopyralid

Column: 0

Flat

Clear Search

The UUID of the Dossier that contains the search result.

page 1 of 99 - Total hits: 988 (0.076 secs)

Export

parent entity type	matched content
FLEXIBLE_SUMMARY	Aminopyralid
parent entity UUID	label
7969a7ed-391e-4a3e-8d05-35054e4bb456	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

More options

IUCLID 6

Dashboard > Mixtures / Products > GF-1601

Filtered AminopyralidActiveSubstanceRenewal

4da31555-bf94-4920-bfba-75c4453df7fc

View Dossiers Validate

Type at least 3 characters

UUID: 4da31555-bf94-4920-bfba-75c4453df7fc

Hide empty fields

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](#).

Text Analytics

Query builder

Search

Searchable properties

Query: field.snapshot\_uuid: "4da31555-bf94-4920-bfba-75c4453df7fc" AND field.value: "dermal absorption"

Search by fields/attachments clear

AND OR

+ Criterion + Set of criteria - Set of criteria

Field properties

Snapshot UUID

x equals

4da31555-bf94-4920-bfba-75c4453df7fc

Field properties

Value

x equals

dermal absorption

page 1 of 1 - Total hits: 9 (0.39 secs)

Export

4da31555-bf94-4920-bfba-75c4453df7fc

IUCLID field

parent entity type

ENDPOINT\_SUMMARY

matched content

dermal absorption in vitro / ex vivo

parent entity UUID

bdf1df2-ddb8-4f52-803a-0b871c630c2b

label

-label not found-

More options

IUCLID type

Closed list with remarks

IUCLID path

ENDPOINT\_SUMMARY.DermalAbsorption.KeyValueCsa.Endpoint

All languages

en(1.0)

Optional properties

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

IUCLID6

Dashboard > Mixtures / Products > GF-1601

Filtered AminopyralidActiveSubstanceRenewal

4da31555-bf94-4920-bfba-75c4453df7fc

View Dossiers Validate

EU PPP Active substance application (product)

GF-1601

1 Identity of the plant protection product and applicant 8

2 Physical, chemical and technical properties of the plant protection product 38

3 Data on application 5

4 Further information on the plant protection product 2

5 Analytical methods 18

6 Efficacy data 2

7 Toxicological studies on the plant protection product 13

7.1 Acute toxicity 9

7.2 Data on exposure 2

7.3 Dermal absorption 2

Administrative data Link to relevant st... Description of key... Key value for che... Additional inform...

Key value for chemical safety assessment

Endpoint

dermal absorption in vitro / ex vivo

Type of information

experimental study

Justification

Species

human

Results

#	Concentration in g/L	Parameter	Absorption	Actions
1	0.02	percentage	2.1 %	
2	13	percentage	0.75 %	

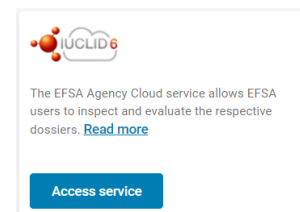
TOP

# 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



© 2004 Blackwell Publishing Ltd *Journal of Internal Medicine* 255: 103–110



- **DATA Team:**

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

# STAY CONNECTED

## SUBSCRIBE TO

[efsa.europa.eu/en/news/newsletters](https://efsa.europa.eu/en/news/newsletters)

[efsa.europa.eu/en/rss](https://efsa.europa.eu/en/rss)

[Careers.efsa.europa.eu](https://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://linkedin.com/company/efsa)



## CONTACT US

[efsa.europa.eu/en/contact/askefsa](https://efsa.europa.eu/en/contact/askefsa)

