IUCLID-Decoder: Python code to extract the chemical property data for the substances registered under REACH

Paulina Körner[1] and Juliane Glüge[1]
[1]Institute of Biogeochemistry and Pollutant Dynamics, ETH Zurich, 8092 Zurich, Switzerland

Abstract

The IUCLID-Decoder is a package in python that enables users to extract chemical property data from registration dossiers that have been submitted to the EU Regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH). The data is extracted from files that the European Chemicals Agency (ECHA) offers for download. These files, the so called 'REACH Study Results' contain more than 4 million IUCLID 6 format (i6d) files with study results. The 'iuclid_parser_utils.py' function in the IUCLID Decoder package extracts and decodes the data and standardizes the units. The output is an SQLite database and (if set to 'True') an excel file with the database.

Statement of need

The European Chemicals Agency (ECHA) publishes on its website (ECHA 2024b) the non-confidential substance data that have been submitted by the registrants under REACH. However, ECHA reserves the right to block systematic automated data collection activities including scraping, data mining, and extraction and re-utilization of the whole or a substantial part of the website and the ECHA databases, where justified and subject to applicable law. To be still able to access the data, ECHA offers to download the 'REACH Study Results' via the IUCLID website (ECHA 2024a). The REACH Study Results contain results from studies that relate to physicochemical properties, environmental fate and pathways, and ecotoxicology and toxicological information. Data from free text fields are not provided, but most of the other data. As the data of the currently more than 20 000 substances come in over 4 million i6d-files, with a structure similar to xml-files, a script is needed to import them into a database. The IUCLID-Decoder package that is available on our GitHub repository (https://github.com/pkoerner6/IUCLID-Decoder) enables the user to extract the information from the i6d-files and to compile them into a database or optionally an excel file. In addition, the information that is available in the REACH Study Results as numerical code including information on units, methods, reliability or study types is converted into text and the units of the study results are standardized as far as possible.

Description

The REACH Study Results come as zipped i6z-file. When using Linux or Mac, the path to the unzipped i6z-file can be given to the iuclid_parser_utils function and the function can be run without considering where to place it. Under Windows, all i6z-files should be unzipped before running the function und the iuclid_parser_utils should be place in the same folder as the REACH Study Results. More information on the expected document structure is provided in the GitHub repository.

The REACH Study Results contain one main folder per registered substance. This folder then contains the so-called 'manifest' as well as individual study results and the files that are needed to decode the results. The iuclid_parser_utils.py function extracts the information property by property. To do this, it first opens the manifest of each substance and looks up the file ID for the property of interest. This property file is then opened and the data for the property of interest are read out. Additionally, data

on the substance identity such as the European Community (EC) number, the Chemical Abstract Service Registry Number® (CAS RN®) and the substance name are extracted. SMILES are not available and would need to be added later on separately. Before saving the data into a dataframe, the data are decoded, and the units are standardized.

It is recommended to check the IUCLID website regularly, as updated REACH Study Results are uploaded once or twice a year. This data can be easily transferred to a new database using the IUCLID decoder package.

The iuclid_parser_utils.py function has already been used in our previous work where we analyzed the physicochemical property data (Glüge and Scheringer 2023) and the bioconcentration data (Glüge et al. 2023a) in the ECHA database. We have also used it in Körner et al. (2024) to check certain REACH study results on ready-biodegradation. Information on SMILES and curated SMILES for most of the organic mono-constituent substances registered under REACH are available in Glüge et al. (2023b)

Acknowledgements

We thank Martin Scheringer for the initial idea of the project and Stefan Glüge for support in between. PK and JG acknowledge funding from the Swiss Federal Office for the Environment.

References

ECHA. 2024a. IUCLID6 - REACH Study Results. Available: https://iuclid6.echa.europa.eu/reach-study-results.

ECHA. 2024b. REACH Dissemination Platform - Registered Substances Factsheets. Available: https://echa.europa.eu/de/information-on-chemicals/registered-substances.

Glüge J, Escher BI, Scheringer M. 2023a. How error-prone bioaccumulation experiments affect the risk assessment of hydrophobic chemicals and what could be improved. Integr Environ Assess Manag 19:792–803; doi:10.1002/ieam.4714.

Glüge J, McNeill K, Scheringer M. 2023b. Getting the SMILES right: identifying inconsistent chemical identities in the ECHA database, PubChem and the CompTox Chemicals Dashboard. Environ Sci Adv 2:612–621; doi:10.1039/D2VA00225F.

Glüge J, Scheringer M. 2023. Evaluation of Physicochemical Property Data in the ECHA Database. J Phys Chem Ref Data 52; doi:10.1063/5.0153030.

Körner P, Glüge J, Glüge S, Scheringer M. 2024. Critical insights into data curation and label noise for accurate prediction of aerobic biodegradability of organic chemicals. Environ Sci Process Impacts; doi:10.1039/D4EM00431K.