Work Package 5 EU-CEG data and enhanced laboratory capacity for regulatory purposes

EU-CEG Reference Tables

Author: ANSES
September 2024

Doc. Ref. N°: D5.9
Type: Document (R)
Dissemination: Public (P)

JATC 2 JOINT ACTION ON TOBACCO CONTROL

Co-funded by the European Union's Health Programme under Grant Agreement n°: 101035968 - JA-01-2020 - HP-JA-2020 / HP-JA-2020-2

Version	Date	Authors	Comments
1	01 Sept. 2024	ANSES	Submitted version

Table of contents

Introduction

Manufacturers and importers of tobacco and related products must submit key information about the products they wish to market to the competent authorities of EU Member States.

Among other things, they are required to provide the complete composition of their products, including a list of their ingredients and their quantities. This information is entered by the declarants as free text in the notification file. As a result, there is significant variability in the naming of chemical substances between declarants and even between different declarations.

Although most chemical substances can be identified using authoritative reference databases (such as CAS Registry Numbers), these do not always enable correct identification of a substance, and errors in CAS numbers are possible. Similarly, there is considerable diversity in the possible naming of chemical substances.

Thus, for the competent authorities of Member States, it is challenging to extract from all declarations the list of products that contain a given substance or family of substances, for purposes such as regulatory enforcement. Additionally, descriptive analysis of a set of declarations requires that the identification of ingredients be standardized.

This report proposes a method and a first mapping table between the declared chemical substances (primarily ingredients) in EU-CEG. It provides a list of so-called 'reference substances' that can be used for subsequent analyses.

Initially, separate tables were created depending on whether they concerned tobacco or vaping products, ingredients, or emissions. Ultimately, the various tables were consolidated into a single table through a unified process.

1 Method

The source information to be aligned was gathered from several datasets: French EU-CEG data (Nov. 2016 to Jul. 2024), EU-CEG data shared as part of the first Joint Action JATC (2017-2020), EU-CEG data shared as part of the current Joint Action JATC-2 (Jul. 2023), data from chemical analysis campaigns of products on the market, and data from the literature.

Each ingredient (chemical substance) as declared is characterized by a CAS number and a name. To facilitate file transfers during the curation process and prevent information loss due to character encoding changes between different systems, a unique identifier in the form of a unique hash key (SHA1) is calculated from the pair (CAS number, name), and a table containing all the triples (identifier, CAS number, name) is created. This is the starting point for the alignment table, which must now be curated, primarily through manual effort.

The curation process includes several steps. Identifiers known from previous curation are directly assigned to a reference substance (with CAS number and substance name). For known CAS numbers, whether they come from previous curation or are found in easily accessible databases such as PubChem[1], a reference substance is suggested. These automatic assignments are subject to human verification. Similarly, ingredients that cannot be automatically aligned undergo a manual review to assign a reference substance.

When a reported CAS number does not match the substance name provided by the submitter, it is decided not to assign a reference substance (NA). In some cases, the CAS number is missing or contains a typographical error (e.g., digit inversion), and a CAS number and substance name are then assigned. When no reference CAS number is available, this field is left blank, but a substance name

1 https://pubchem.ncbi.nlm.nih.gov/

is always assigned.

This results in a mapping table between the reported ingredients (identifier, CAS, name) and the reference substances after curation. A unique identifier is also calculated for the latter, following the same principle as for the reported ingredients.

Upon reviewing this initial step, it became clear that further refinement was necessary. Indeed, some substances turned out to be identical (due to the existence of different CAS numbers for the same chemical entity) or stereo-isomers that should be grouped together to subsequently provide more concise descriptive analyses.

Additionally, this step allows the separation of molecular compounds from more complex substances that cannot be defined by a single chemical structure, such as plant extracts, plants, mixtures, polymers, and substances defined by their functional properties in the product (cooling agent, pH modifier)...

This second step enables the creation of an alignment table between reference substances and generic substances. The latter are characterized by their name and type, as previously mentioned (molecular compound, mixture, polymer, plant extract, flavor, etc.).

For molecular compounds, the PubChem identifier is provided, along with the empirical formula, molecular weight, SMILES[2] notation of the isomer or canonical form, and the InChIKey[3]. This allows reference to additional information about the substance, for example, from the PubChem database[4]. Plants and plant extracts are characterized by information on their genus and species. Finally, substances designated as flavors were aligned with the flavor wheel established for e-liquids[5].

This additional information completes the mapping table.

2 Results

The mapping table is provided in ANNEX 1, in the form of an SQL script to be executed to create a SQLite database table (named MapIngSubst) in EC.db or TP.db described in deliverable D5.2 of this JATC-2 project.

It contains 57,200 entries corresponding to 4,036 reference substances, of which 1,948 are ingredients of vaping products recorded in EU-CEG and 904 are ingredients of tobacco and plant-based smoking products. Seventy-five of these substances have a CMR classification. The list of these reference substances is provided in ANNEX 2.

2 https://en.wikipedia.org/wiki/Simplified_Molecular_Input_Line_Entry_System
3 a hashed version of the full standard InChI (https://en.wikipedia.org/wiki/International_Chemical_Identifier)
4 For instance, PuchemID:89594 refers to 'Nicotine': https://pubchem.ncbi.nlm.nih.gov/compound/89594
5 Krüsemann, Erna J. Z., Sanne Boesveldt, Kees de Graaf, et Reinskje Talhout. 2019. « An E-Liquid Flavor Wheel: A Shared Vocabulary Based on Systematically Reviewing E-Liquid Flavor Classifications in Literature ». Nicotine & Tobacco Research: Official Journal of the Society for Research on Nicotine and Tobacco 21 (10): 1310-19. https://doi.org/10.1093/ntr/nty101.

3 Conclusion

This alignment table of reported ingredients with reference substances is a necessary step to deepen the analysis of the declared composition in EU-CEG for tobacco products and related products. Whether for research or regulatory enforcement, it opens the possibility to perform general descriptive statistics or to identify products that may potentially be non-compliant with regulations or pose a risk.

To do this, it is essential to use the table as a starting point to connect with other reference frameworks, notably the list of substances with harmonized classification under the CLP Regulation.

This work should also encourage declarants to strive for more consistent reporting of the substances they use in their product formulations.

For the services of the European Commission, it can serve as a starting point to integrate this reference system into an update of the EU-CEG reporting system, with the goal of improving the reliability of the data entered by manufacturers.

Finally, for the competent authorities of EU Member States, a process must still be defined to update and maintain this controlled vocabulary for chemical substances contained in products over time.

4 ANNEX 1

-- MapIngSubst.sql version 2024-08-24b

THIS FILE IS ATTACHED TO THIS REPORT

```
::
:: EC-update-MapIngSubst.bat version 2024-08-24
::
:: Copyright (c) euceg@anses.fr 2022-2024 (JATC2-WP5)
:: Except specific files which bear a different mention, this programme is licensed under the EUPL-
1.2 or later
:: You may obtain a copy of the license at https://joinup.ec.europa.eu/collection/eupl/eupl-text-
eupl-12
::
:: This activity has received funding from the European Union's Health Program (2014-2020) under
grant agreement N°101035968 (JA-01-2020 - HP-JA-2020 - HP-JA-2020-2).
:: The content of this document represents the views of the author only and is his/her sole
responsibility; it cannot be considered to reflect the views of the European Commission and/or the
European Health and Digital Executive
:: Agency (HaDEA) or any other body of the European Union. The European Commission and the Agency do
not accept any responsibility for use that may be made of the information it contains.
::
:: USE: run this script to updtate EC.db database with new reference ingredients table (MapIngSubst)
::
@echo off
:: <eucegfolder> containing subfolders csv and db
set eucegfolder=c:\euceg
::
setlocal enabledelayedexpansion
if exist %eucegfolder%\db\sqlite set PATH=%eucegfolder%\db\sqlite;%PATH%
cd /d %eucegfolder%\db
::
echo ------ EC cleanup ------
sqlite3 EC.db ".echo on" ".read sql/MapIngSubst.sql" ".echo off" ".exit"
echo ------ END ------
pause
```

```
::
:: TP-update-MapIngSubst.bat version 2024-08-24
::
:: Copyright (c) euceg@anses.fr 2022-2024 (JATC2-WP5)
:: Except spTPific files which bear a different mention, this programme is licensed under the EUPL-
1.2 or later
:: You may obtain a copy of the license at https://joinup.TP.europa.eu/collTPtion/eupl/eupl-text-
eupl-12
::
:: This activity has rTPeived funding from the European Union's Health Program (2014-2020) under
grant agreement N°101035968 (JA-01-2020 - HP-JA-2020 - HP-JA-2020-2).
:: The content of this document represents the views of the author only and is his/her sole
responsibility; it cannot be considered to reflTPt the views of the European Commission and/or the
European Health and Digital ExTPutive
:: Agency (HaDEA) or any other body of the European Union. The European Commission and the Agency do
not accept any responsibility for use that may be made of the information it contains.
::
:: USE: run this script to updtate TP.db database with new reference ingredients table (MapIngSubst)
::
@echo off
:: <eucegfolder> containing subfolders csv and db
set eucegfolder=c:\euceg
::
setlocal enabledelayedexpansion
if exist %eucegfolder%\db\sqlite set PATH=%eucegfolder%\db\sqlite;%PATH%
cd /d %eucegfolder%\db
::
echo ------ TP cleanup ------
sqlite3 TP.db ".echo on" ".read sql/MapIngSubst.sql" ".echo off" ".exit"
echo ------ END ------
pause
```

5 ANNEX 2