**December 2020 Release of the Chemical and Products Database**

**Created:** 2020-12-16

Jonathan Taylor Wall, Kathie Dionisio, and Kristin Isaacs, Center for Computational Toxicology and Exposure, Office of Research and Development, US EPA. Contact: Isaacs.Kristin@epa.gov

Table of Contents

## Overview

The data included in this version of the Chemicals and Products Database (CPDat) updates the information included in the 2018 release (Dionisio et al., 2018). Since 2018, the CPDat team has worked to implement a new data curation and management system, called Factotum, that will increase the volume and reliability of data related to chemical use. This download is the first version of CPDat curated within Factotum. CPDat data curated within Factotum will also be regularly released via the CompTox Chemicals Dashboard (https://comptox.epa.gov/dashboard). This data release contains: 1) Additional machine-readable data (e.g., document metadata) not available via the Dashboard, and 2) Data associated with new curation activities since the most recent Dashboard release. Specifically, this release reflects curation of new data sources (including new sources related to chemical functional use and occupational use of chemicals) and additional curation of chemicals and products.

The bulk CPDat download is made up of data and dictionary files which represent the four different overall document types curated into CPDat via Factotum: 1) Consumer Product Composition, 2) Chemical List Presence, 3) Functional Use, and 4) Health Hazard Evaluation (HHE) Reports. The data files contain the data records associated with each document type, while the dictionary files contain additional metadata on documents, chemicals, functions, and product use categories (PUCs), and include curation to harmonized data where appropriate and available. The dictionaries can be combined with the data files via chemical, PUC, document, and/or functional use IDs for further analysis. This document provides descriptions of each document type, the contents of the dictionary and data files, and some data notes.

## Document Type Descriptions

The following document types defined below are curated within CPDat. The curation of these documents results in composition, list presence, functional use, and occupational use data files. Often, each data file is associated with a single document type (e.g., composition data is curated from composition documents), but functional use data may be derived from multiple document types (as described below).

### *Composition Documents*

Documents from product manufacturers concerning ingredient and chemical composition of consumer products, typically in the form of SDS, MSDS, and product ingredient disclosure documents. May also include functional use information. Curators extract chemical names and other identifiers, composition information, and any functional use data.

### *List Presence Documents*

Documents from many different sources describing the presence of chemicals on lists within public documents related to chemical use, exposure scenario, chemical occurrence, or regulatory purview. These document types were formerly associated with EPA's CPCat database (Dionisio et al. 2015, 2018). The documents are tagged with keyword sets describing the list (see Data Notes below). May also include functional use information.

### *Functional Use Documents*

Documents from product manufacturers or other organizations, that describe the functional use or purpose of chemicals in products.

### *Health Hazard Evaluation (HHE) Reports*

Health Hazard Evaluation Reports from the National Institute for Occupational Safety and Health (NIOSH). No specific measurement data are included in CPDat for these documents, but the presence in this dataset can be interpreted as indicating use of the associated chemicals in an occupational setting.

## Data Notes

### *Curation of Chemicals*

Each unique chemical record associated with a data document has been assigned a chemical record ID, associated with a reported chemical name and Chemical Abstract Service Registration Number (CAS), if available. Each of these individual chemical records are eventually officially curated using automated workflows within EPA's Distributed Structure-Searchable Toxicity (DSSTox) database (Williams et al., 2017), where they are assigned a preferred name, CAS, and a DSSTox Chemical Substance ID (DTXSID). Here, chemical records not yet officially curated into DSSTox (i.e., awaiting formal curation) are assigned provisional preferred identifiers and DTXSID based on the official curation of other CPDat records with identical reported identifiers.

### *Consumer Product Composition Data*

The consumer products represented in the composition data have primarily been manually curated to product use categories (PUC; Isaacs et al. 2020) developed explicitly for exposure assessment and modeling. Some products have been provisionally assigned to PUC via analysis of product name, using automated text recognition methods that are currently under development. These products could be filtered out if the user wishes; they may provide additional data but also may contain occasional errors. In this version of the data, multiple products associated with the same document and formulation (for example, different Universal Product Codes, or UPCs, associated with the same chemical list) have been removed.

The formulation PUC codes used here are the 3-level hierarchical codes described in Isaacs et al. (2020), with some small additions/refinements that have been made as additional products have been curated. The PUC codes from Isaacs et al. are included in the PUC dictionary. PUCs from Isaacs et al. that had the same text identifier for both the PUC_level2 (product family, prod_fam) and PUC_level3 (product type, prod_type) now report solely the PUC_level2 identifier and contain an empty PUC_level3 (simply for convenience and to save memory; essentially the redundant PUC_level3 has been dropped). In addition, some very general article and industrial/occupational PUCs have been added as some data sources for these types of products have now introduced into CPDat. The article PUCs are based on the OECD harmonized article categories (OECD, 2017) and provisional industrial/occupational product PUCs were developed from examination of the occupational products present. It is anticipated that the industrial/occupational product PUCs will be refined in the future.

### *List Presence Data*

The list presence data reported here include specific keywords that define and describe the presence of chemicals on lists contained within public documents. These keywords can be considered an update/refinement to the terms previously developed for EPA's Chemical and Product Categories (CPCat) database (Dionisio et al. 2015, 2018). These refinements better align the assignment of keywords with other CPDat data streams, namely functional use and consumer product composition data. CPCat included specific functional uses as terms, whereas we now explicitly recognize reported functional uses associated with list documents as part of our functional use data. In addition, CPCat contained large number of keywords associated with different consumer product types or categories. We now have harmonized and updated these terms to the product use categories (PUCs) that we have developed for use with consumer product composition data (Isaacs et al. 2020). For example, a chemical list from a public document, denoting chemicals used in a specific type of personal care product, would be assigned a keyword identical to the PUC that would be assigned to similar products referenced in composition documents. A large number of pharmaceutical-related keywords in CPCat were collapsed to a single "pharmaceutical" keyword here. Finally, we have now also assigned individual keywords to an overarching higher-level category (keyword "kind") as defined here, to aid in organization or summarization of the data:

**General use**: related to general chemical use

**PUC – article**: keyword is an product use category (PUC) of kind "article"

**PUC – formulation**: keyword is an product use category (PUC) of kind "formulation"

**Location:** related to location/origin of the document/list

**Manufacturing**: related to the manufacturing process

**Foods & Agriculture**: related to food and agriculture

**Specialty list**: keyword refers to a recognized, specialty list of chemicals, e.g., a regulatory list

**Subpopulation**: keyword denotes specific population associated or affected by the document/list

**Media**: associated with a specific environmental media, e.g., as in a measurement study

**PUC – industrial**: keyword is product use category (PUC) of kind industrial/occupational

**Modifiers**: keyword modifies other keywords assigned to the same list/chemical combination

The keywords associated with any chemical list are to be interpreted as a group. The "modifier" keyword kind therefore modifies the meaning of other keywords assigned to the chemical list in question. For example, the modifier "detected", in the presence of the words "drinking_water" and "pesticides", denotes that the list includes chemicals that act as pesticides and were specifically detected in drinking water.

There are currently approximately 140 keywords assigned to public lists in CPCat. It is anticipated that this list will grow as additional public chemical documents and lists are curated.

### *Functional Use Data*

The functional use data reported here expands on the function data included in EPA's Functional Use Database (FUse, Phillips et al. 2017) and the previous version of CPDat (Dionisio et al. 2018) Many of the reported functional uses have been curated to the harmonized technical function categories developed by the Organisation for Economic Co-Operation (OECD, 2019). Curation of functional uses to OCED technical functions is ongoing.

### *Quantitative Structure Use Relationship Model Predictions*

In addition to the newly curated data available in this CPDat release, we have also included an updated set of predictions from the functional use quantitative structure use relationship (QSUR) models described in Phillips et al. (2017). The models themselves have not changed, but we now provide predictions for many additional chemical structures now available from DSSTox. For any individual chemical (DTXSID), predictions are provided for all QSUR models for which the chemical structure fell within the model domain of applicability. Since the models have not been updated, they are based on the harmonized functional uses described in Phillips et al., as opposed to OECD technical functions reported in this version of CPDat. Future refined versions of the QSUR models may be based on the newly curated functional use data reported here. Please see Phillips et al (2017) for more information.

## Data Summary

| **Data Type or Subset** | **Count** |
|---|---|
| | |
| ***Composition Data*** | |
| Total number of records | 451972 |
| Number of records curated to DTXSID | 260599 |
| Number of records with quantitative information | 216758 |
| Number of unique products | 61131 |
| Number of unique chemicals (DTXSIDs) | 4896 |
| | |
| ***Functional Use Data*** | |
| Total number of records | 243894 |
| Number of records curated to DTXSID | 149740 |
| Number of records curated to OECD function categories | 146421 |
| Number of unique chemicals (DTXSIDs) | 12316 |
| | |
| ***List Presence Data*** | |
| Total number of records | 326250 |
| Number of records curated to DTXSID | 92410 |
| Total number unique documents | 2179 |
| Number of unique chemicals (DTXSIDs) | 20815 |
| | |

| | |
|---|---|
| ***HHE Data*** | |
| Total number of records | 6122 |
| Number of records curated to DTXSID | 1304 |
| Total number unique documents | 3814 |
| Number of unique chemicals (DTXSIDs) | 493 |
| | |
| ***QSUR Models*** | |
| Total number of records | 1183705 |
| Number of QSUR models (harmonized functions) | 37 |
| Number of unique chemicals (DTXSIDs) | 414044 |

## Data Field Descriptions

The following are tables which list and describe the data fields found in every dictionary and data file. The data files can be joined with their dictionary files using their unique, corresponding ID columns.

### *Data Dictionaries*

**Chemical Dictionary**

| Field | Description |
|---|---|
| chemical_id | Unique chemical *record* ID unique ID for a chemical record (from any data document type). |
| raw_chem_name | Chemical name, as provided exactly on original data document |
| raw_casrn | CAS, as provided exactly on original data document (so, for example, may be listed as multiple CAS for one chemical on original data document) |
| preferred_name | Preferred chemical name, per DSSTox |
| preferred_casrn | Preferred CASRN for a chemical, per DSSTox |
| DTXSID | DSSTox substance identifier (unique substance identifier) |
| curation_level | C= Individual chemical record is officially curated within EPA's DSSTox<br>PR=Provisional curation based on official curation of other records with reported name and CAS pair. |

**Document Dictionary**

| Field | Description |
|---|---|

| | |
|---|---|
| document_id | Unique document ID |
| title | The title of the data document. e.g., if the data document is a journal article or report, this might be the title of the article or report; if the data document is an MSDS sheet, this may be the name of the product represented in the MSDS sheet. If no title is provided when the data document is registered by the curator, the title defaults to be the same as the file name of the data document. |
| subtitle | Document's subtitle |
| doc_date | Date on the data document |

**Functional Use Dictionary**

| Field | Description |
|---|---|
| chemical_id | Unique chemical *record* ID unique ID for the chemical record (from any data document type) associated with a reported functional use. |
| functional_use_id | Unique functional use ID associated with the chemical record |
| report_funcuse | Reported functional use |
| oecd_function | OECD function category (OECD, 2017) |

**List Presence Dictionary**

| Field | Description |
|---|---|
| list_presence_id | Unique list presence ID |
| name | The list presence tag name |
| definition | Detailed description of the list presence keyword |
| kind | The list presence keyword kind (higher-level category for the keyword); see Data Notes |

**PUC Dictionary**

| Field | Description |
|---|---|
| puc_id | Unique Product Use Category (PUC) ID |
| gen_cat | The general category assignment (PUC level 1) of a product (see Isaacs et al. 2020) |
| prod_fam | The product family assignment (PUC level 2) of a product (see Isaacs et al. 2020) |
| prod_type | The product type assignment (PUC level 3) of a product (see Isaacs et al. 2020) |
| puc_code | Corresponding code from Isaacs et al. 2020. PUCs without a code have been added since |

| | |
|---|---|
| | the PUCs were published, or correspond specifically to higher-tier PUCs (to which products can be specifically assigned in CPDat if they don't fall into a more refined category). |
| description | Detailed description of the PUC |
| kind | PUC kind:<br>F: formulation<br>A: article<br>O: industrial/occupational |

### *Data Files*

**Product Composition Data**

| Field | Description |
|---|---|
| document_id | Corresponding document dictionary ID (for merging with document dictionary) |
| chemical_id | Corresponding chemical dictionary ID (for merging with chemical dictionary) |
| functional_use_id | Corresponding functional use dictionary ID (for merging with functional use dictionary) |
| product_id | Unique product ID |
| brand_name | Brand name of product |
| prod_title | Product Name |
| puc_id | Corresponding PUC ID (for merging with PUC dictionary) |
| classification_method | PUC Classification method (MA = Manual curation to product category, MB = Batch assigned manually in Factotum; PR= Provisional assignment to PUC based on automated analysis of product name; method under development); see Data Notes |
| raw_min_comp | Raw reported weight fraction/composition data: minimum. May be reported as percent, decimal fraction, or other form. |
| raw_central_comp | Raw reported weight fraction/composition data: central or point value. May be reported as percent, decimal fraction, or other form. |
| raw_max_comp | Raw reported weight fraction/composition data: maximum. May be reported as percent, decimal fraction, or other form. |
| clean_min_wf | Cleaned weight fraction data: minimum. Cleaning of data to a harmonized form (decimal fraction between 0-1) is ongoing. |

| | |
|---|---|
| clean_central_wf | Cleaned weight fraction data: central or point value. Cleaning of data to a harmonized form (decimal fraction between 0-1) is ongoing. |
| clean_max_wf | Cleaned weight fraction data: maximum. Cleaning of data to a harmonized form (decimal fraction between 0-1) is ongoing. |

**Functional Use Data**

| Field | Description |
|---|---|
| document_id | Corresponding document dictionary ID (for merging with document dictionary) |
| chemical_id | Corresponding chemical dictionary ID (for merging with chemical dictionary) |
| functional_use_id | Corresponding functional use dictionary ID (for merging with functional use dictionary) |

**Health Hazard Evaluation (HHE) Data**

| Field | Description |
|---|---|
| document_id | Corresponding document dictionary ID (for merging with document dictionary) |
| chemical_id | Corresponding chemical dictionary ID (for merging with chemical dictionary) |

**List Presence Data**

| Field | Description |
|---|---|
| document_id | Corresponding document dictionary ID (for merging with document dictionary) |
| chemical_id | Corresponding chemical dictionary ID (for merging with chemical dictionary) |
| list_presence_id | Corresponding list presence dictionary ID (for merging with list presence dictionary) |

**Quantitative Structure Use Relationship (QSUR) Data (independent from other datasets; see Data Notes)**

| Field | Description |
|---|---|
| DTXSID | DSSTox substance identifier (unique substance identifier) |
| preferred_name | Preferred chemical name, per DSSTox |
| preferred_casrn | Preferred CASRN for a chemical, per DSSTox |
| harmonized_function | Harmonized function category (as defined in Phillips et al., 2017) |

| | |
|---|---|
| probability | Probability returned by the QSUR model for the harmonized function (see Phillips et al. 2020) for details |

## References


Dionisio KL, Frame AM, Goldsmith MR, Wambaugh JF, Liddell A, Cathey T, Smith D, Vail J, Ernstoff AS, Fantke P, Jolliet O, Judson RS. Exploring consumer exposure pathways and patterns of use for chemicals in the environment. Toxicol Rep. 2015 Jan 2;2:228-237. doi: 10.1016/j.toxrep.2014.12.009. PMID: 28962356; PMCID: PMC5598258.

Dionisio KL, Phillips K, Price PS, Grulke CM, Williams A, Biryol D, Hong T, Isaacs KK. The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products. Sci Data. 2018 Jul 10;5:180125. doi: 10.1038/sdata.2018.125. PMID: 29989593; PMCID: PMC6038847.

Isaacs KK, Dionisio K, Phillips K, Bevington C, Egeghy P, Price PS. Establishing a system of consumer product use categories to support rapid modeling of human exposure. J Expo Sci Environ Epidemiol. 2020 Jan;30(1):171-183. doi: 10.1038/s41370-019-0187-5. Epub 2019 Nov 11. PMID: 31712628.

Organisation for Economic Co-operation and Development. Internationally Harmonised Functional Product and Article Use Categories ENV/JM/MONO(2017)14. 2017.

Phillips KA, Wambaugh JF, Grulke CM, Dionisio KL, Isaacs KK. High-throughput screening of chemicals as functional substitutes using structure-based classification models. Green Chem. 2017;19(4):1063-1074. doi: 10.1039/C6GC02744J. PMID: 30505234; PMCID: PMC6260937.

Williams AJ, Grulke CM, Edwards J, McEachran AD, Mansouri K, Baker NC, Patlewicz G, Shah I, Wambaugh JF, Judson RS, Richard AM. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J Cheminform. 2017 Nov 28;9(1):61. doi: 10.1186/s13321-017-0247-6. PMID: 29185060; PMCID: PMC5705535.