SCIENTIFIC DATA

OPEN

Data Descriptor: The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products

Received: 16 October 2017
Accepted: 30 April 2018
Published: 10 July 2018

Kathie L. Dionisio[1], Katherine Phillips[1], Paul S. Price[1], Christopher M. Grulke[2], Antony Williams[2], Derya Biryol[1,3], Tao Hong[4] & Kristin K. Isaacs[1]

Quantitative data on product chemical composition is a necessary parameter for characterizing near-field exposure. This data set comprises reported and predicted information on more than 75,000 chemicals and more than 15,000 consumer products. The data's primary intended use is for exposure, risk, and safety assessments. The data set includes specific products with quantitative or qualitative ingredient information, which has been publicly disclosed through material safety data sheets (MSDS) and ingredient lists. A single product category from a refined and harmonized set of categories has been assigned to each product. The data set also contains information on the functional role of chemicals in products, which can inform predictions of the concentrations in which they occur. These data will be useful to exposure and risk assessors evaluating chemical and product safety.

Design Type(s)	data integration objective
Measurement Type(s)	physicochemical characterization
Technology Type(s)	digital curation
Factor Type(s)	chemical product
Sample Characteristic(s)	

[1]U.S. Environmental Protection Agency, National Exposure Research Laboratory, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, USA. [2]U.S. Environmental Protection Agency, National Center for Computational Toxicology, 109 T.W. Alexander Drive, Research Triangle Park, NC 27709, USA. [3]Oak Ridge Institute for Science and Education, Oak Ridge, TN 37830, USA. [4]ICF International, 2635 Meridian Pkwy #200, Durham, NC 27713, USA. Correspondence and requests for materials should be addressed to K.I. (email: Isaacs.kristin@epa.gov).

Background & Summary

Evaluating chemical safety and sustainability over the life cycle of products requires drawing upon the various data streams and impact assessment tools from the life cycle assessment (LCA) field, along with improved exposure models that rapidly and reliably characterize human health risks of chemicals from direct and indirect exposure pathways. Near-field exposure to chemicals in consumer products has been identified as a significant source of exposure for many chemicals[1,2]. Quantitative data on product chemical composition is a necessary parameter for characterizing near-field exposure.

Currently there are limited data on the composition of products available in a format which allows for large-scale modeling of near-field exposures to chemicals in consumer products. Previous efforts at the U.S. Environmental Protection Agency (U.S. EPA) have aggregated data on consumer product composition (the Chemical/Product Categories Database (CPCat)[3], and the Consumer Product Chemical Profiles database (CPCPdb[4])) and functional use of chemicals (i.e., the role a chemical plays in a product, e.g. solvent vs. fragrance; the Functional Use database (FUse)[5,6]). Here we describe the Chemicals and Products Database (CPDat), a data set of consumer product composition and functional use, including information on >75,000 chemicals and >15,000 consumer products. CPDat incorporates CPCat, CPCPdb, and FUse in full, and streamlines the three data sets with a consistent scheme for categorizing products and chemicals. In addition, CPDat includes seven distinct newly acquired data sets on product composition (both reported values and quantitative predictions based on ingredient list labels) and functional use. The newly acquired data sets have expanded the scope and number of records in the database which the user has to draw from, dependent on their data needs. Further, harmonization of the existing databases allows for a richer data set which describes the full extent of the chemical-product data relevant for exposure modeling of consumer products.

Each product-related record (i.e., a unique piece of data corresponding to a specific consumer product, e.g., an MSDS sheet for a particular product) in CPDat is linked to a unique product category. Product categories exist as a hierarchical, harmonized nomenclature for categorization of consumer products. Product records in CPDat contain chemical information, e.g., a product record may include the list of chemicals included in the formulation of that product. Each of these chemicals has been mapped to a unique, curated chemical record in the EPA's Distributed Structure-Searchable Toxicity (DSSTox) Database, which underlies the publicly available CompTox Chemistry Dashboard (https://comptox.epa.gov/dashboard). The Dashboard is a resource which includes an abundance of information on the chemical's structure, properties, and toxicology. All CPDat data are available via the 'Exposure' tab in the CompTox Dashboard, providing an easily accessible, central repository for product composition and functional use data.

As described in Fig. 1, CPDat currently includes the following types of data:

- Quantitative chemical composition
  - Reported data on composition of a large number of consumer products, including weight fractions for each component, taken from publicly available Material Safety Data Sheets (MSDS)
  - Predicted weight fractions of chemicals in consumer products, derived from publicly available ordered product ingredient lists and ingredient disclosures, which are documents that provide a list of the ingredients in a product, and the functional use that some or all of the ingredients serve in the product[7]
- Consistent product and chemical categorization
  - General categorization of chemical usage (e.g. industrial uses, broad categorizations of use)[3]
  - Consumer product category specific categorization
- Chemical functional use
  - Reported chemical functional use data from publicly available government, manufacturer, and industry sources
  - Reported data on functional uses of chemicals within consumer products from manufacturers
  - Predicted data on functional use of chemicals, based on quantitative structure-use relationship (QSUR) modeling methods[6]

The data included in CPDat will inform exposure assessments of consumer products for commercial chemicals. Information about exposure is a critical component of risk evaluations and provides real-world context to hazard (i.e., toxicity data). These data can be reused in aggregate (i.e., single chemical, multi-product) and cumulative (i.e., multi-chemical) assessments by federal and state governments, academia, and industry.

Methods

Data included in the CPDat data set represent an aggregation of publicly available data on chemical-use categorization, consumer product composition (including qualitative ingredient lists and quantitative

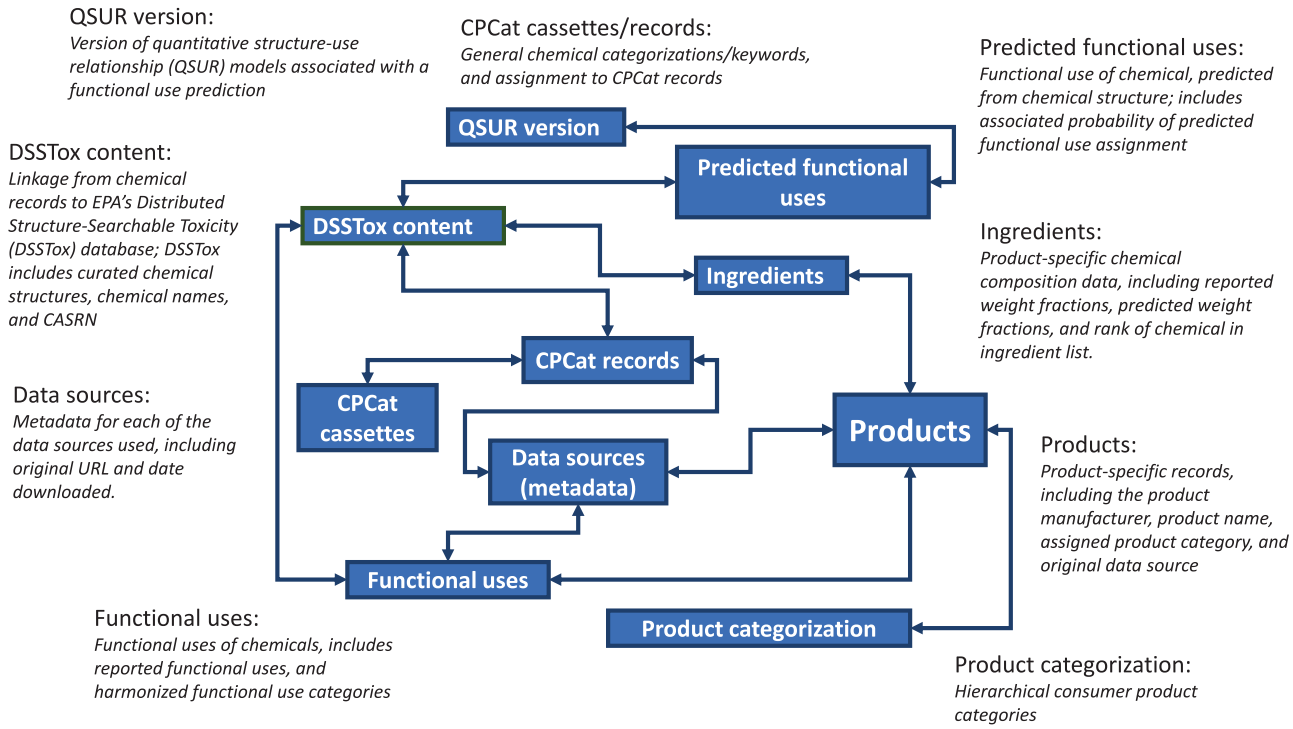

Figure 1. **CPDat data structure.**

composition data, i.e., weight fraction), and functional use of chemicals. All data included in CPDat were acquired or generated from publicly available sources as detailed below. CPDat represents the aggregation and harmonization of three existing data sets with newly available related data. Methods and included data sources presented here reflect the data set as of August 8, 2017. We anticipate that additional data will be added to the data set over time. The addition of new data will be documented in the public repository on both the 'News' (https://comptox.epa.gov/dashboard/news_info) and 'Downloads' (https://comptox.epa.gov/dashboard/downloads) pages of the Dashboard, as appropriate. The three existing data sets are described briefly below, with full details included in the associated publications. Additional details and methods are provided for the new data and sources included in CPDat.

Curation of chemicals

Each data record in CPDat is associated with a chemical name and/or Chemical Abstracts Service Registry Numbers (CASRN). To enable integration of CPDat data with other DSSTox data and into the Dashboard, it was necessary to map CPDat chemicals to the appropriate DSSTox chemical identifier[8]. CPDat names and CASRN were searched against the complete list of potential chemical identifiers in DSSTox. Potential DSSTox generic substances for each CPDat chemical record were scored based on the quality of the automatically generated mapping, taking into account the trust associated with the various identifiers in the DSSTox database (e.g., a match against a validated synonym is scored higher than a match against an ambiguous synonym). A DSSTox generic substance with the top score was mapped to each CPDat chemical record.

Existing datasets incorporated into CPDat

Chemical/Product Categories (CPCat) Database. The CPCat relational database was constructed from a variety of publicly available data sources on chemicals and associated categorical groupings. CPCat integrates information from 11 major national and international sources. Original data sources and methods are described in detail in Dionisio *et al.*[3]. A harmonized set of CPCat terms (general use keywords, sometimes but not always describing specific consumer products) and cassettes (unique keyword combinations) were manually linked to each chemical included in the data set, based on information on chemical usage available from the original source. Cassettes are comprised of one or more CPCat terms, all terms within a cassette must be interpreted together to reflect the categorical information provided by the original data source.

Consumer Product Chemical Profiles database (CPCPdb). CPCPdb was the first effort to create a publicly available data set of consumer product composition and weight fractions to be used in high-throughput exposure modeling. The data set focused on MSDS available on the internet as Adobe PDF documents for a single retailer. Full details on methods used to extract data from the MSDS are available

Data source	Description	URL	Download date	Data included	Collected files	CPDat product records[a]
Unilever	Manufacturer including multiple leading brands of personal care and household products	https://pioti.unilever.com/PIOTI/EN/p2.asp	7/17/2015	Product specific, weight fraction-ordered ingredient lists and functional use disclosures	948	846
Unilever MSDS USA	Manufacturer including multiple leading brands of personal care and household products	http://www.unilevermsdsusa.com/	10/8/2015	Quantitative product composition	577	282
Proctor & Gamble	Manufacturer including multiple leading brands of personal care and household products	http://www.pgproductsafety.com/productsafety/sds/SDS_2015/	9/19/2015	Quantitative product composition	2403	1799
Proctor & Gamble Product Safety	Manufacturer including multiple leading brands of personal care and household products	http://www.pg.com/productsafety/search_results.php?searchtext=*&submit=Search&submit=Search	7/15/2015	Product-specific ingredient and functional use disclosures	250	226
Drugstore.com	Online retail site for personal care and household products. Website no longer functional.	http://www.drugstore.com/	9/1/2015	Product-specific ingredient lists	4635	4635
Church & Dwight	Manufacturer including multiple leading brands of personal care and household products	http://www.churchdwight.com/brands-and-products/msds-ingredient-search.aspx	7/16/2015	Product-specific ingredient and functional use disclosures	35	34
Palmolive	Single brand website for liquid dish soaps	http://www.palmolive.com/ingredients#soft-touch-aloe	7/17/2015	Product-specific ingredient and functional use disclosures	16	16

Table 1. Additional data sources included in CPDat (beyond the CPCPdb, CPCat, and FUse data sets). [a]Database record count varies from collected files due to removal of products with no listed ingredients or no manufacturer-provided CASRN, duplicated products, and multi-component products.

in Goldsmith *et al.*[4]. Briefly, MSDS were downloaded as PDF files, and programmatic tools including optical character recognition software were used to extract the product ingredient information. Data were stored in a MySQL database custom-designed for this purpose. Each entry was manually curated using a custom, web-enabled interface. Products were assigned to a product category and subcategory based on the categorization used on the retailer's online shopping website, with the product categorization automatically retrieved from the retailer's website using a freely available data extraction program.

Functional Use database (FUse). FUse is a collection of chemicals in consumer products linked by the function (or role) that the chemical is known to have in products (e.g., fragrance, solvent, etc.). The database was created via automated collection of publicly available functional use data provided by manufacturers of consumer products, in ingredient repositories for product formulators, or from government regulators. The reported functional uses were harmonized to a set of consistent 'root' terms (e.g., 'whitening agent' and 'whitener' were mapped to the same term). For modelling purposes (functional use prediction), the functions were further harmonized using a hierarchical cluster analysis. A full description of the methods and compilation of FUse can be found in Isaacs, *et al.* and Phillips *et al.*[5,6].

Additional data

CPDat also includes data from the sources listed in Table 1. These data sources include product composition data (quantitative weight fraction) not included in previous databases curated by the authors, product-specific ingredient list data (qualitative), and functional use data. Data sources listed in Table 1 were all obtained from public internet sites of reputable manufacturers or retailers.

Proctor & Gamble. For the Proctor & Gamble data source, MSDS were automatically downloaded as PDF files and converted to TXT files using R scripts (http://www.R-project.org/). Composition data associated with each product were extracted from the TXT files automatically using custom Python (https://www.python.org/) scripts. This was possible because the MSDS followed a common format. If a chemical composition table spanned two pages, the words 'Page xx' were removed manually. Manual extraction of data was also required for Duracell products, which used a different MSDS template. Extracted data records were stored in CSV files.

Unilever MSDS USA. The PDF MSDS in the Unilever MSDS USA repository were downloaded manually, with data then extracted automatically using custom R and Python scripts as above, with extracted data stored in a CSV. For the Unilever MSDS USA data source, product type was available on the website, but not on the MSDS, thus this field was manually added to the data record, for use in assigning the product to a product category.

Proctor & Gamble Product Safety. The ingredient and functional use disclosures provided by Proctor & Gamble Product Safety were automatically downloaded as PDF files and converted to text files using the RCurl (http://CRAN.R-project.org/package = RCurl) and XML (http://CRAN.R-project.org/package = XML) packages for the R programming language. Pertinent information from the converted text files were parsed and stored in a CSV file via custom Python scripts. For some products, the name "fragrance" was listed as an ingredient with its functional use listed as a hyperlink to another PDF containing a list of chemicals used by the manufacturer as fragrances in products. In these cases, the hyperlink was removed and the functional use was replaced with "fragrance".

Unilever. The ingredient disclosures provided by Unilever were obtained by downloading the HTML file of the disclosure for each product of each brand provided by the manufacturer. The ingredients and functional uses provided on the disclosures were listed in separate columns of an HTML table, which were parsed and stored in a CSV file. The files were automatically downloaded and parsed using the RCurl and XML packages for the R programming language. For some products, the ingredient disclosure was duplicated on the website (e.g., the HTML table contained records for twice the number of chemicals in a product, when in reality each chemical was listed with the same functional use twice in the table). Therefore, duplicated records of an ingredient-functional use pair were dropped from the parsed HTML table in such a way that ingredient order in the table was preserved.

Church & Dwight. Ingredient disclosures were automatically downloaded from the Church & Dwight website as PDF files using the RCurl and XML packages for the R programming language. PDFs were converted to text files which allowed for parsing of relevant information (i.e., product name, ingredient name, and ingredient functional use in product) into a CSV file via custom Python scripts.

Palmolive. All products and their ingredient disclosures were available in a single table from the Palmolive website. This information was manually copied and pasted into a text file which was then parsed into a CSV file via a custom Python script.

Drugstore.com. Reported ingredient data for products were obtained from Drugstore.com. Lists of links to HTML files for available products were obtained using custom R scripts employing the R packages RCurl and XML. The HTML file for each product was subsequently obtained and parsed using custom R text parsing scripts. The data were manually curated, with 2-component products identified. Each ingredient name and rank in the reported ingredient list were obtained for use in future weight fraction predictions. The product name and Drugstore.com product category was retained for use in harmonized product categorization.

Generated data fields. Data fields generated after data collection (assignment of product category, predicted weight fraction from ingredient lists, and harmonized and predicted functional use) are detailed below; all other data fields are replicated exactly as they were in the original data source.

For each product-specific record, the product was assigned to a product category. The product categories used are a refined set of categories based on those previously developed for use in linking products to exposure scenarios[9], but further refined for product form (e.g., liquid versus spray cleaner) and population of use (e.g., children's sunscreen). The product categories are reproduced in Table 2 (available online only). The assignment of products to categories was performed using the following process. Product names were first passed through a custom R script which performed a preliminary categorization of products that included a pre-identified list of key words (e.g. if product name included 'shampoo', assign product to 'shampoo' category). All product category assignments were then reviewed manually for accuracy. Assignment of product categories were completed manually for any product where a preliminary categorization was not possible.

Product-specific records linked to quantitative composition data were obtained in one of two ways. If the manufacturer provided quantitative composition data for a specific product, data were replicated in CPDat exactly as provided by the manufacturer. However, some sources included qualitative ingredient list data, providing a list of ingredients (by chemical name and/or CASRN) in the product without information on the weight fraction of each ingredient. When it was known that these ingredient lists were arranged in a specific order (e.g., personal care products, which are mandated by law to list ingredients in decreasing weight fraction order), a model detailed in ref. 7 was applied to the ingredient list to obtain the 5th and 95th percentile bounds on the weight fraction for each ingredient in the list.

Reported functional uses contained in FUse were harmonized across the various data sources. The harmonized functional use reported in CPDat is intended 1) to remove redundancy in functional uses reported by different sources (e.g.., "foaming aid" and "foam boosting agent" would be collapsed into "foam boosting agent") and 2) to provide a single descriptive functional use for each chemical in FUse.

For example, many chemicals are reported to have functional uses of both "cleanser" and "surfactant", as cleansers are typically surfactants, these chemicals would have a single harmonized functional use of "surfactant". The details of the harmonization process via hierarchical clustering have been previously described[5,6]. These harmonized uses were then used to train and validate a suite of quantitative structure-use relationship (QSUR) models that can predict a chemical's likely functional use(s) for chemicals with no known functional use[5,6]. The full details of the development and validation of these models has been published previously[5,6]. Briefly, the models were developed from publicly-available structural descriptors using Random Forest classification and validated using 5-fold cross-validation and y-randomization[5,6]. For any chemical structure, the valid Random Forest model for a function returns the probability (0-1) of the chemical performing the given function. In CPDat, we report results from 41 valid function QSUR models for chemicals modelled in Phillips *et al.*[6]. The performance of these 41 valid models is provided in Table 3 (available online only).

Code availability

Scripts to perform data downloads are available at: https://github.com/HumanExposure/CPDatManScripts.

Data Records

The data repository where the CPDat consumer product and functional use data are stored, the U.S. EPA's CompTox Chemistry Dashboard (hereafter referred to as the 'Dashboard'), is available publicly at https://comptox.epa.gov/dashboard. The Dashboard is a freely accessible web-based application and data hub integrating data for ~760,000 chemical substances (as of August 2017), most of these with their related chemical structures. Various types of data are associated with the substances including both experimental and predicted physicochemical and environmental fate and transport properties, toxicity data, bioassay data (i.e. the ToxCast data[8]), exposure data and external links to relevant public data sources. When viewing a chemical's landing page through the Dashboard, the CPDat data are available under the 'Exposure' tab, through the 'Product & Use Categories,' 'Chemical Weight Fraction,' and 'Chemical Functional Use' sub-tabs. Note that the CPDat data covers only a subset of the chemicals included in the Dashboard, but includes data on more than 75,000 chemicals (unique CAS-chemical name pairs) and more than 15,000 products (unique product names).

The CPDat data presented in the Dashboard is stored in a MySQL relational database maintained by the CPDat team. A copy of the MySQL database as of August 2017 is archived in FigShare (Data Citation 1). Users may install MySQL and download the data set (Data Citation 1) for manipulation and data extraction; however, users are encouraged to use the Dashboard as an access point, for ease of data manipulation and accessibility. Addition of new data, and updates to the underlying CPDat data set are ongoing, with data updates pushed to the public facing Dashboard periodically. Notification of updates will be posted on both the 'News' (https://comptox.epa.gov/dashboard/news_info) and 'Downloads' (https://comptox.epa.gov/dashboard/downloads) pages of the Dashboard, as appropriate. Details presented below represent data records archived in the MySQL database (Data Citation 1), and available on the Dashboard as of August 2017.

Within the MySQL CPDat database, two types of data are stored related to each data record: metadata, and record-specific data. Metadata refers to descriptions and information which relate to all, or to large subsets, of data records, as compared to record-specific data which is specific to a single data record (e.g., weight fraction of a chemical in a particular product). The mapping of the CASRN and/or chemical name provided by the data source to the DSSTox unique identifier underlies the presentation of data in the Dashboard itself. The 'official' CASRN, chemical name, and additional information about the chemical can be found on each chemical's landing page within the Dashboard.

Metadata

To ensure data provenance and tracking, metadata about each source the data records were obtained from has been linked to each data record. Metadata includes a brief description of the data source, the date the dataset was downloaded, and the web URL the data was downloaded from. Additionally, descriptions defining the product categorizations are reproduced in Table 2 (available online only). The specific metadata or definition associated with each data record is available in a pop-up window when hovering over or clicking on the data source name or the product categorization in associated tables in the Dashboard.

Record-specific data

Record-specific data can include different pieces of information depending on the data source. All information provided by the original data source is included. A summary of the pieces of data included by data source is provided in Table 4. Record specific data available in the Dashboard may include:

Product name. Specific name of the product, as provided by the original data source (may or may not include brand name, or name of manufacturer).

	Product use categorization		Record-specific data		
Source	CPCat cass.	Prod. Cat.	Weight fraction	Ingredient list	Functional use
CPCat	X				
CPCPdb		X	X		
FUse	NA	NA			X
Unilever		X		X	X
Unilever MSDS USA		X	X		
Proctor & Gamble		X	X		
Proctor & Gamble Product Safety		X		X	X
Drugstore.com		X		X	
Church & Dwight		X		X	X
Palmolive		X		X	X

Table 4. Data records included in CPDat as of August 8, 2017.

Product use category. Product category assigned to the product, indicating the general category and product type assigned to each data record based on information provided in the original data source. Also see 'Categorization type' below.

Categorization type. Indicates the type of categorization assigned to the data record (CPCat Cassette or product category). Additional detail regarding CPCat Cassette categorizations can be found in Dionisio *et al.*[3], detail regarding product categories can be found in Methods and Table 2 (available online only).

Minimum/Maximum weight fraction. For 'MSDS' data, the minimum and maximum weight fraction for each listed ingredient, as detailed on the associated MSDS. A minimum and maximum value is required because many MSDS report weight fraction as a range. Where an MSDS reported a single value, the minimum and maximum weight fraction is the same. For 'Ingredient List' data, the 5th and 95th percentiles of the predicted weight fraction (See Isaacs *et al.*[7] for detail on how weight fractions were predicted based on ordered ingredient lists).

Data type. Form of product-specific data provided by original source (MSDS or Ingredient List).

Reported functional use. The functional use of the chemical in the product, as defined by the original data source.

Harmonized functional root. A simple abbreviation of a functional use allowing harmonization of different reporting names (e.g., 'flavorant' versus 'flavouring').

Harmonized functional use. The harmonized functional use, as determined by cluster analysis of the reported functions for each chemical. See ref. 5 for detailed methods.

Probability of harmonized functional use. Probability ranging from 0 to 1 that a chemical is likely to have one of the 41 harmonized functional uses with a valid QSUR model[6].

Source. The original data source from which each specific data record was obtained.

The record-specific data available under each sub-tab of the 'Exposure' section on a chemical's landing page in the Dashboard is summarized in Table 5.

Technical Validation

The main quality objective for CPDat was ensuring that data included in the data set accurately reflected the data as provided in the raw data source. Efforts to curate the data set have not focused on checking for errors that may have been made by the original provider of the data (e.g., an error the manufacturer made in listing a weight fraction on a MSDS). Rather, quality assurance focused on ensuring that data records were transcribed accurately from the original data source and represented appropriately in the repository. A manual review of data (5% for Unilever USA MSDS, 1% for Proctor & Gamble MSDS) automatically extracted from PDF files was completed to ensure accuracy, with no errors identified. Note that for the Unilever USA MSDS data source, occasionally product names listed on the MSDS did not match the product name on the corresponding website link and in these cases, it was assumed that the MSDS contained the correct product name. In some cases, errors were identified in the process of migrating data from CSV files into the MySQL database (e.g., duplicated chemicals on a single product's ingredient list). In these cases, custom Python scripts were written to check all data and correct for these errors.

Product & Use Category sub-tab	Chemical Weight Fraction sub-tab	Chemical Functional Use sub-tab
Product or Use categorization	Product name	Harmonized Functional Use
Categorization type	Product category	Reported Functional Use
Number of unique products	Minimum weight fraction	Probability of Associated Functional Use
	Maximum weight fraction	
	Data type	
	Source	

Table 5. Record specific data fields available under each sub-tab of the 'Exposure' section of the Dashboard.

The mappings of chemical records from CPDat to DSSTox unique chemical identifiers currently in use were obtained from a semi-automated mapping process. Within the semi-automated process, a baseline level of trust in the mapping must be met for the mapping to be maintained. Refinement of algorithms used in the semi-automated mapping process, and additional manual verification of mappings by a trained DSSTox curation team are always ongoing.

Usage Notes

Though data included in CPDat can be used in many ways in future analyses, users should be aware of limitations of the data set, and appropriate usage of the data. Though a wide range of consumer products are included in CPDat, the products included in CPDat are necessarily limited to those for which data were publicly available. Thus, for example, personal care products are more heavily represented due to Food and Drug Administration reporting requirements for cosmetics. Additionally, each product category will contain multiple products, each with their own formulation. Though the range of different formulations within each product category allows the user to gain insight into similarities and differences (e.g., a particular chemical may be present in all formulations for a particular category), the formulations should not be assumed to be representative of all formulations, nor should they be assumed to accurately represent market share within a product category.

There are additional factors a user must consider, specifically when interpreting data obtained from MSDS. Formulations obtained from MSDS are associated with a date, representing the date the manufacturer issues the MSDS detailing composition of that particular formulation. Users should take care to note these associated dates and filter data as appropriate for their intended use since formulations of a product do change over time. Further, manufacturers are not required to report all ingredients in a product (only those which are potentially hazardous), although manufacturers may choose to report all ingredients. Quantitative composition information may also be withheld if the ingredient is considered a trade secret. Additionally, some chemicals and products are exempt from the hazard communication standards, e.g., solid articles or products covered under other legislation, such as foods, pharmaceuticals, and tobacco products[10]. Lastly, with the exception of a few data sources contained within CPCat, ingredients associated with product-specific records in CPDat are drawn from ingredient lists or MSDS, and by nature represent 'intended' ingredients in products. Therefore, CPDat does not currently capture the range of unknown contaminants which may be present in consumer products, for example, chemicals that may migrate from the plastic bottle of hand lotion into the hand lotion itself.

CPDat includes predictions of functional use based on machine-learning classification models developed from chemical structures[6]. As with any QSAR method, these predictions are by their nature limited by the depth and breadth of the chemical space of the training set on which the classification models rely. Any use of these predictions should consider these limitations in evaluating their fitness-for-purpose. The training set used to develop the models is composed of the reported function data (also included in CPDat); the modeling methodology and model performance has been previously discussed in detail[6].

References

1. Wallace, L. A. Comparison of risks from outdoor and indoor exposure to toxic chemicals. *Environ Health Perspect* **95,** 7–13 (1991).
2. Wambaugh, J. F. *et al.* High Throughput Heuristics for Prioritizing Human Exposure to Environmental Chemicals. *Environ Sci Technol* **48,** 12760–12767 (2014).
3. Dionisio, K. L. *et al.* Exploring consumer exposure pathways and patterns of use for chemicals in the environment. *Toxicol Reports* **2,** 228–237 (2015).
4. Goldsmith, M.-R. *et al.* Development of a consumer product ingredient database for chemical exposure screening and prioritization. *Food and Chemical Toxicology* **65,** 269–279 (2014).
5. Isaacs, K. K. *et al.* Characterization and prediction of chemical functions and weight fractions in consumer products. *Toxicol Reports* **3,** 723–732 (2016).
6. Phillips, K. A., Wambaugh, J. F., Grulke, C. M., Dionisio, K. L. & Isaacs, K. K. High-throughput screening of chemicals as functional substitutes using structure-based classification models. *Green Chemistry* **19,** 1063–1074 (2017).

7. Isaacs, K. K., Phillips, K. A., Biryol, D., Dionisio, K. L. & Price, P. S. Consumer Product Chemical Weight Fractions from Ingredient Lists. J Expo Sci Environ Epidemiol **28**, 216–222 (2018).
8. Richard, A. M. *et al.* The ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem Res Toxicol* **29,** 1225–1251 (2016).
9. Isaacs, K. K. *et al.* SHEDS-HT: An Integrated Probabilistic Exposure Model for Prioritizing Exposures to Chemicals with Near-Field and Dietary Sources. *Environ Sci Technol* **48,** 12750–12759 (2014).
10. Electronic Code of Federal Regulations, §1910.1200 Hazard communication. https://www.ecfr.gov/cgi-bin/text-idx?SID = 54 ea58abd39079c15cfd14ae9f72e0cc&mc = true&node = se29.6.1910_11200&rgn = div8 (2018).

Data Citation

1. Williams, A. *Figshare* http://dx.doi.org/10.23645/epacomptox.5352997 (2017).

Acknowledgements

The authors would like to acknowledge Ann Richard, Inthirani Thillainadarajah, and David McKee of the U.S. EPA for their efforts in the DSSTox project and chemical curation work which allowed CPDat data to be incorporated with the CompTox Chemistry Dashboard. We also thank the technical staff at ICF International for their work verifying the accuracy of data collected from public sources.

This research was supported in part by the Research Participation Program at the Office of Research and Development, US Environmental Protection Agency, administered by the Oak Ridge Institute for Science and Education through an interagency agreement between US Department of Energy and US Environmental Protection Agency. The information in this document has been funded wholly or in part by the US Environmental Protection Agency. It does not signify that the contents necessarily reflect the views of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use. The paper has been subjected to the Agency's review process and approved for publication.

Author Contributions

K.L.D. drafted the manuscript, led integration of the CPDat data with the CompTox Dashboard, and curated chemicals. K.P. downloaded and extracted raw data, built and maintained underlying MySQL database, curated chemicals, and contributed to text. P.S.P. conceptualized the integration of the existing data sets into CPDat, and advised on project concept and planning. C.G. advised on design of the MySQL database, and led efforts to map CPDat chemical identifiers to DSSTox chemical identifiers. A.W. leads development of the CompTox Dashboard. D.B. downloaded and extracted Drugstore.com product data. T.H. wrote scripts to download and extract Unilever and Proctor & Gamble product data. K.K.I. initiated and led the collection and curation of MSDS, ingredient list, and functional use data; downloaded, extracted, and processed raw data; advised on database design; generated weight fraction predictions from ingredient lists; developed product categories; mapped products to categories; and contributed to text.

Additional information

Tables 2 and 3 are available only in the online version of this paper.

Competing interests: The authors declare no competing interests.

How to cite this article: Dionisio, K. L. *et al.* The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products. *Sci. Data* 5:180125 doi: 10.1038/sdata.2018.125 (2018).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files made available in this article.

© The Author(s) 2018