Expansion and Refinement of Chemical Use Data for Characterizing Exposure Pathways

Sakshi Handa[1], Kristin Isaacs[1], David Lyons[1], Jonathan Wall[1], Allison Larger[1,2], Beth Horton[1,2] and Kathie Dionisio[3]

[1]Center for Computational Toxicology and Exposure, Office of Research and Development, U.S. Environmental Protection Agency

[2]Oak Ridge Associated Universities

[3]Chemical Safety for Sustainability Research Program, Office of Research and Development, U.S. Environmental Protection Agency

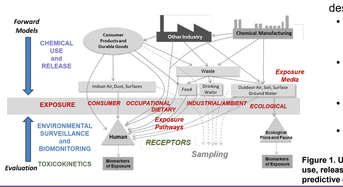
Sakshi Handa | handa.sakshi@epa.gov | 919-541-0636

Abstract

The US EPA is responsible for evaluating thousands of chemicals for the potential risks they may pose to humans and ecosystems, which necessitates information on hazard and exposure potential for each chemical. To support chemical decision-making, EPA's Office of Research and Development (ORD) must identify and characterize relevant exposure pathways - the path of a chemical from source to a receptor. How a chemical is used (e.g. in a consumer, occupational, or industrial context) is critical to determining exposure pathways. ORD has developed a data management and curation application, called Factotum, which facilitates the rapid collection and distribution of high-quality chemical and exposure related data from public documents via curation, quality assurance, visualization and data delivery tools. Within Factotum, there has been a significant focus on chemical composition of consumer products, functional role of chemicals within products and processes, and presence of chemicals on reported specific or general use lists. Factotum has facilitated the expansion of these use databases to include new information related to occupational use of chemicals, literature measurement of chemicals in key media, and population use patterns for consumer products. Factotum also includes new category schema for the classification of products used in industrial and occupational settings. Ongoing efforts are made to broaden the scope of the data while ensuring data quality, through the addition of new data sources, data curation and cleaning, manual Quality Assurance (QA) workflows, and chemical curation to harmonized chemical identifiers. To date, Factotum has been used to collect and curate data from 511,898 documents, representing over 3.9 million individual chemical records and 29,391 unique chemical substances. These methods have rapidly expanded the scope and quantity of data in EPA's Chemicals and Products Database (CPDat). These expanded use data are being integrated with other exposure-relevant data streams, including chemical monitoring and release data, to rapidly inform EPA and State agencies workflows for assessing potential exposures via different pathways.

Introduction

- EPA is charged with evaluating risks associated with chemicals in commerce, including consumer products.
  - As of February 2021, there are **41,864** active chemicals on the EPA's Toxic Substances Control Act (TSCA) Inventory.
- Evaluating chemicals for risk to humans or the environment requires information on hazard and **exposure potential.**
- To support chemical decision-making, EPA's Office of Research and Development (ORD) must develop robust, well-documented, and accessible datasets to inform exposure assessments.



- ORD needs sustainable methods for cataloging and documenting available data describing:
  - ***Chemical use and release.*** Provides critical information for identifying chemical sources, exposure pathways, and relevant predictive models for a given chemical.
  - ***Media occurrence, environmental surveillance, and biomonitoring***. Provides data for directly supporting decision-making and evaluating predictive models.
  - ***Exposure estimates***. Predictions of chemical intake in mg/kg/day can be compared with hazard information to inform risk.
  - ***Toxicokinetics***. Provides real-world exposure context to *in vitro* high-throughput screening data and biological receptor monitoring information.

Figure 1. Understanding and characterizing exposure pathways for commercial chemicals requires chemical use, release, monitoring, and toxicokinetic data. These data inform the development and evaluation of predictive exposure estimates.

Approach

- **EPA-ORD is building an integrated system for collecting, curating, storing, and annotating exposure-relevant information.**
- **ChemExpoDB** refers to a data base that holds raw and curated data on use of chemicals.
- Documents and datasets can be curated into ChemExpoDB using ORD's web-based **Factotum** application.
- Curated data from ChemExpoDB is releases as ORD's Chemicals and Products database (CPDat)[1], and available via the CompTox Chemicals Dashboard.
- Factotum includes web services that can supply ChemExpoDB data to internal EPA clients.

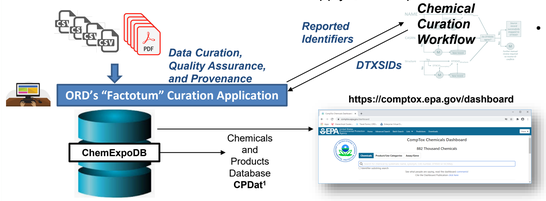


Figure 2. The ChemExpoDB / Factotum System

- Factotum and ChemExpoDB make use of the extensive chemical curation workflows built by ORD to support the CompTox Chemicals Dashboard.[2] These methods maps hundreds of thousands of chemical identifiers to unique substance identifiers (DTXSIDs), where a substance can be any single chemical, mixture, polymer, etc.

Methods

The Factotum Application

Factotum improves ORD's ability to rapidly compile and distribute useful, high-quality, chemical and exposure related data via curation and cleaning, quality assurance and tracking, visualization, and data delivery tools.

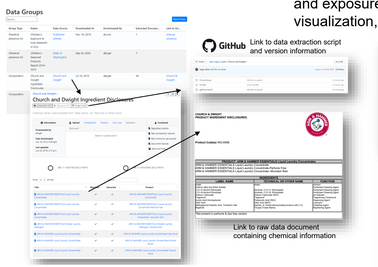


Figure 3. Factotum provides tools for loading, extracting, and managing chemical data documents and associated data extraction scripts.

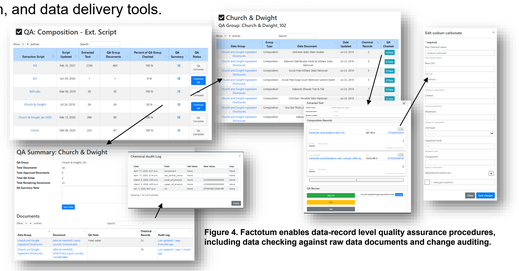


Figure 4. Factotum enables data-record level quality assurance procedures, including data checking against raw data documents and change auditing.

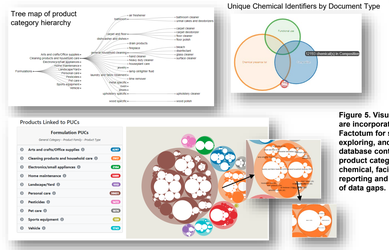


Figure 5. Visualization tools are incorporated in Factotum for searching, exploring, and visualizing database content by product category and chemical, facilitating reporting and identification of data gaps.

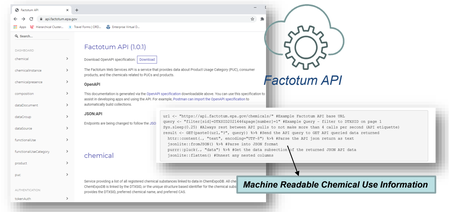


Figure 6. Factotum contains an Application Programming Interface that can deliver data in batch/bulk form to EPA stakeholders.

Expansion of CPDat Scope and Data

- Factotum tools are being used to expand the scope of data in CPDat:
  - CPDat contains extensive data on consumer formulations, functional use, and general chemical list presence.
  - New data related to occupational exposure, chemical occurrence in environmental media, and consumer product use patterns have been curated.
  - New automated natural language processing (NLP) classifier models for curating data documents to consumer product categories used in exposure assessments have been developed, allowing for greater curation capability.
    - **(See SOT Abstract Number/Poster Board 2650/ P116)**
  - New categorization systems for consumer articles and products used in occupational settings have been developed.

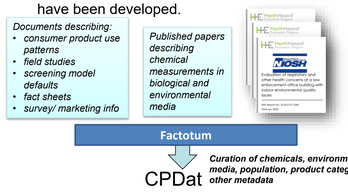


Figure 7. Factotum is allowing for the rapid curation of literature monitoring studies, NIOSH Health Hazard Evaluation reports, and documents related to consumer habits and practices into CPDat.

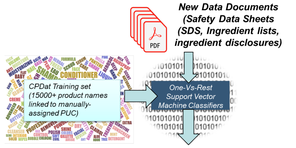


Figure 8. Machine learning classifiers were built using the existing manually curated CPDat data as a training set, allowing for automated curation of new documents to Product Use Categories (PUCs) developed specifically for exposure modeling.[3]

Results

- The Factotum application has facilitated the rapid expansion of the volume and scope of data in the CPDat database (Figure 9).
  - To date, Factotum has been used to upload more than **500,000** data documents (Table 1) including those containing information on product chemical composition (e.g., SDS, ingredient lists, voluntary ingredient disclosures); documents disclosing chemical functional use; documents describing public chemical lists; Health Hazard Evaluations; published papers describing measured chemicals in media; and reports/documents describing consumer product habits and practices (i.e. consumer product use patterns).
  - **3.9 million** chemical-specific records have been extracted from these documents; to date **1.9 million** of these records have been curated to harmonized substance identifiers (DTXSIDs); chemical curation is ongoing.
  - The curated data records cover **29,390** unique DTXSIDs
  - Over **608,000** individual products with chemical ingredients have been linked to product categories.

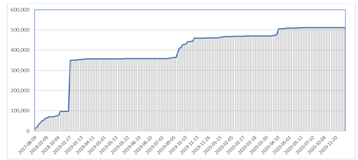


Figure 9. Data documents uploaded within Factotum by date.

Group Type	Documents	Raw Chemical Records	Curated Chemical Records
Composition	473,146	3,735,296	1,880,037
Functional use	33,775	34,680	12,766
Chemical presence lists (CPCat Categories)	2,178	134,129	75,355
HHE Report	1,304	5,421	1,078
Literature monitoring	1,175	2,340	In process
Habits and practices	202	NA	NA

Table 1. Data curated into CPDat via Factotum as of February 2021 by document type.

Impact and Next Steps

- The Factotum/ChemExpoDB system will improve the volume, timeliness, quality, and accessibility of exposure data available for use in decision-making by EPA stakeholders and ultimately the public.
- The development of webservices will support timely and sustainable reporting of new data to systems that surface exposure information to the public; the feasibility of public webservices is currently being investigated.
- The QA tools in Factotum are allowing QA procedures to be formalized and documented, which is critical when data may be used in regulatory decision-making.
- The automated NLP models for assigning chemical data documents to relevant consumer product categories will increase the rate at which these data are available for use in exposure assessments.
- Existing ORD databases of toxicokinetic data are planned for Factotum integration.
- Data models for storing EPA exposure model results are under development. In the future, standards for documenting and reporting exposure model results should be developed and incorporated into Factotum. These standards should include methods for capturing model code versions and input data sets and parameters, allowing for reproducibility of any results.
- Additional workflows are being developed for the harmonization of reported chemical functional uses to internationally recognized functional use categories.

References

1. Dionisio KL, Phillips K, Price PS, et al. The Chemical and Products Database, a resource for exposure-relevant data on chemicals in consumer products. *Sci Data*. 2018;5:180125. Published 2018 Jul 10. doi:10.1038/sdata.2018.125.
2. Williams AJ, Grulke CM, Edwards J, et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. J Cheminform. 2017;9(1):61. Published 2017 Nov 28. doi:10.1186/s13321-017-0247-6.
3. 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;30(1):171-183. doi:10.1038/s41370-019-0187-5.