The Journal of Open Source Software

Drug Ontology Parsing Engine (DOPE): An R Package for Querying of a Comprehensive Ontology of Substances of Abuse

Raymond R. Balise[1], Layla Bouzoubaa[1], Gabriel Odom[3], Aneesh Chandramouli[1], Sean X. Luo[2], and Daniel J. Feaster[1]

1 University of Miami, Miller School of Medicine, Department of Public Health Sciences, Biostatistics Division **2** Division Substance Use Disorder, Department of Psychiatry, Columbia University **3** FIU, Robert Stempel College of Public Health and Social Work, Biostatistics

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

One important barrier in both substance use disorder treatment delivery and research is the myriad of different ways substances of abuse are categorized and described by clinicians and patients, as well as related stakeholders such as law enforcement officials and regulators. Substances of abuse can have a "street name", which refers to the jargon for a particular compound that may be circumscribed to geographical region or cultural contexts. They also are broadly categorized by their pharmacodynamic properties in the context of law enforcement, such as "depressants" and "hallucinogens". While such categories do not have a strict correspondence in pharmacology, they are clinically relevant and are widely used by addiction treatment professionals. Finally, there are more classical classifications of substances of abuse in their chemical structure ("opioid", "benzodiazepine"). These classification schemes have corresponding vocabulary in standard biomedical informatics databases, such as SNOMED Clinical Terminology. While all of these schemes are useful for specific tasks, what is needed is a structured vocabulary that captures the organization of abused drugs used by law enforcement and clinical investigators. That is, a drug name ontology is needed. The purpose of our work is to provide software tools, in particular a library of functions for the R programming language, which allow a "crosswalk" between generic, brand and street drug names, with their drug classes, which occur in many clinical documents such as physician notes and patient self-report transcripts. This will allow both clinicians and researchers to better appreciate the patterns of substance use in individuals, especially for substances that are not standard items of query in other surveying methods, such as in a Urine Drug Screen (UDS). To address this need, we introduce a new R software package, the Drug Ontology Parsing Engine (`DOPE`).

`DOPE`, provides clinical investigators, substance use disorder researchers and treatment providers with access to a large and comprehensive database of street names of substances of abuse tracked by the United States Drug Enforcement Agency. Unfortunately, these data, including the categories, classes, brands and street names of controlled substances, are scattered across many dozens of web pages and PDF files. The `DOPE` R package integrates these data with a second publicly available lexicon (http://www.noslang.com/drugs/dictionary/) to provide these data in a computable format. The website for the package (URL will be here) is designed to be a teaching tool. It includes vignettes which describe the code, with the details to support someone who wants to teach the data creation process, including tutorials on web scraping with the `rvest` package and data cleaning using the `tidyverse` R package suite.

To date, there has been only one other ontology for substance abuse. Cameron et al.

(2013) built the Drug Abuse Ontology (DAO) to facilitate processing of free text notes using a web-based platform (PREDOSE). It defined drug classes and object properties, and used lexicons from across the web, many of which no longer exist. Unfortunately, neither the DAO nor the PREDOSE web platform is currently available. This R package has the advantage of a portable database that can be consistently updated with new DEA vocabularies as they are collected during enforcement, as well as standardization and consistency in dissemination through the CRAN R Package repository.

Ontology Details

An inspection of the DEA website (https://www.dea.gov/factsheets and https://www.dea.gov/documents/2018/07/01/2018-slang-terms-and-code-words) reveals that the drugs it tracks fall into 11 categories (i.e., stimulants, depressants, designer drugs, etc.), 41 classes (e.g., amphetamines, cocaine, etc.), and hundreds of drug synonyms that can either be brand or street names (e.g., Adderall, Ritalin, 777, Zip, etc.). Many of these synonyms can be grouped into generic or chemical drug names. Figure 1 shows the general structure of the data in `DOPE`. Discrepancies within the files at the DEA create a few oddities in the structure. Most notable are the location of two benzodiazepines, alprazolam (i.e., Xanax) and clonazepam (i.e., Klonopin), which are placed in the hierarchy directly under the category of depressants. That is, they are represented in the hierarchy at the same level as benzodiazepines, even though they are benzodiazepines. This confusing discrepancy can be seen in Figure 2. This unexpected twist in the hierarchy is handled by the built-in drug search functions described below. To see details of the tables, please see the package website (URL GOES HERE).

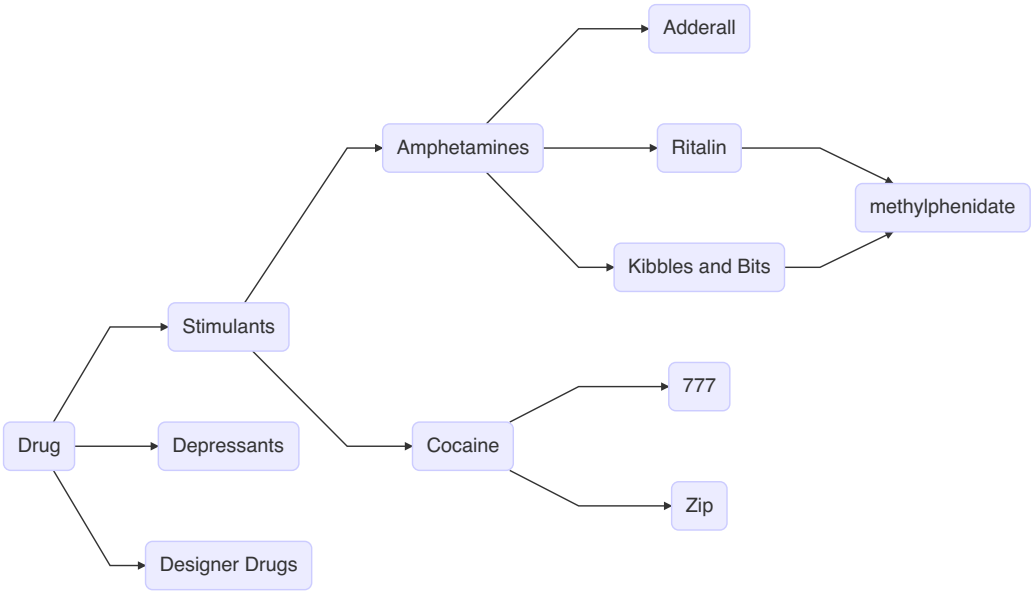

Figure 1: Structure of Drug Categories, Classes, Synonmyms and Names

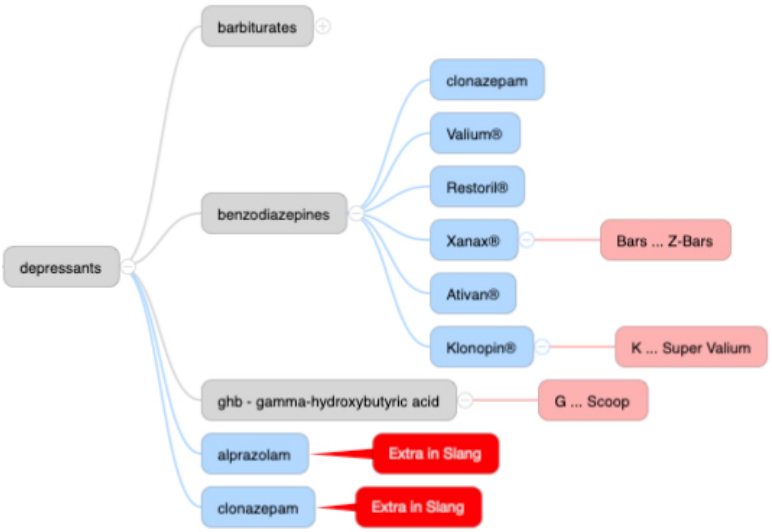

Figure 2: DEA Organization of Benzodiazepines

Data Sets in DOPE

Figure 3 displays the *entity relationship* diagram with the tables, variables and the relationships between the 7 tables in the package: dea_brands, dea_controlled, dea_factsheets, dea_street_names, lookup_df, noslang_raw, noslang_street_names.

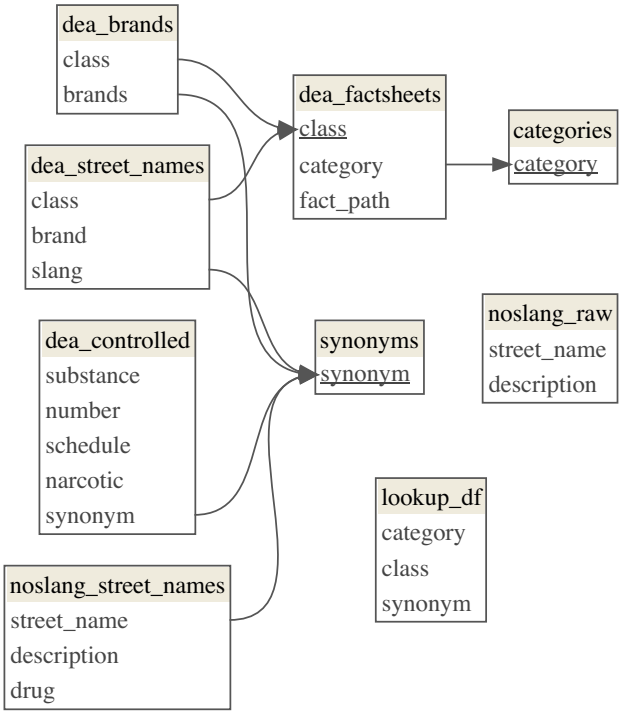

Figure 3: Entity Relationship Diagram for Tables in DOPE

The Journal of Open Source Software

Functionality

The DOPE package includes two functions, `lookup()` and `compress_lookup()`. `lookup()`, when given a drug name (or a set of drug names), will return the drug category, class and generic/chemical drug name. `lookup()` is a vectorized function designed to work using pipes within the "tidyverse" framework. With this function, casual users can discover the identities of unknown drugs. Data scientists can use this function as part of a data processing workflow to identify patterns of drug use in a corpus such as free text medical notes, social media posts or drug history "timeline follow-back" files. DOPE can be used to translate common slang into drug names. For example DOPE unambiguously translates "horse" into "heroin". While context is critical to avoid false positives, DOPE can be used look for drug references hidden in text. For example, when a user reports that yesterday they had "a", "cheese", "pizza", "with", "a", "soda", DOPE returns a table indicating that the person actually ingested cocaine, lsd and either heroin, marijuana or methamphetamine. `compress_lookup()` can be used to drop columns and drop duplicates. That is, a user who wishes to only know the categories of a set of drugs processed by `lookup()` can call `compress_lookup(lookup("cheese", "pizza", soda"), compressClass = TRUE, compressSynonym = TRUE)`. As shown below, `compress_lookup()` is also designed to work with pipes.

```
library(DOPE)
lookup("a", "cheese", "pizza", "with", "a", "soda")
      category           class synonym
1   stimulants         cocaine    soda
2       heroin          heroin  cheese
3 hallucinogen             lsd       a
4 hallucinogen             lsd   pizza
5     cannabis       marijuana  cheese
6   stimulants methamphetamine  cheese
```

```
lookup("cheese", "pizza", "with", "a", "soda") %>%
  compress_lookup(compressCategory = FALSE,
                  compressClass = TRUE,
                  compressSynonym = TRUE)
      category
1   stimulants
2       heroin
3 hallucinogen
4     cannabis
```

Future Directions

As mentioned above, there are inconsistencies in the structure/organization of the drugs tracked by the DEA. A future release of DOPE will attempt to further harmonize the data by adding an additional level to the hierarchy corresponding to the generic drug names or the chemical names of substances that have never been patented. That may help resolve the discrepancies. However, not all discrepancies can be resolved. In particular, the slang terms may not have unique mappings to other levels of the hierarchy of terms. However, knowing the multiple mappings is essential to finding the user's meaning of the term. As additional lexicons become available, we plan to add in more sources of slang and allow users to select different, ideally geographically specific, lexicons.

The DOPE package provides a convenient way to lookup drug information, but it can be used for much more. Two members of the DOPE development team (Balise & Bouzoubaa) have prototyped a web-enabled application using the Shiny (Chang, Cheng, Allaire, Xie,

& McPherson, 2020) R package to automate spell checking of self-reported drug use information. The application reads in a text file or Excel workbook, reports the frequency of known drug words and then spell checks the remaining words with the (Ooms, 2018) R package using a custom dictionary based on the National Library of Medicine. The authors plan to integrate the `DOPE` package as the lexicon for the next iteration of the spell checker application in which the dictionary will be updated with more comprehensive and programmatically attainable data. As an additional functionality of the spell checker, the `DOPE` package will be a fundamental component to categorizing user input drug information for researcher analyses. Further, `DOPE` can be used as part of a pipeline to identify and categorize drugs mentioned in free text on social media platforms or in clinical notes. The DOPE engine could also be embedded into a real-time interviewing tool, which could prompt users to resolve multiple mappings of slang terms. The ability to automate the processing of such text and spoken communication is critical for researchers seeking to identify patterns of substance abuse and how these patterns impact response to addiction treatment. `DOPE`, used in conjunction with the myriad currently available R modeling packages will enable clinical investigators to address these needs.

Acknowledgements

Work for this project was supported by the National Institute on Drug Abuse grant (NIDA UG1 DA013720 and UG1DA013035) HEAL Initiative Supplement CTN-0094: Individual Level Predictive Modeling of Opioid Use Disorder Treatment Outcome. Dr. Luo is funded in part by an additional grant from NIDA (K23DA042136). The authors thank Ryan from noslang.com for giving us permission to scrape their drug dictionary.

References

Cameron, D., Smith, G. A., Daniulaityte, R., Sheth, A. P., Dave, D., Chen, L., Anand, G., et al. (2013). PREDOSE: A semantic web platform for drug abuse epidemiology using social media. *J Biomed Inform*, *46*(6), 985–97. Journal Article. doi:10.1016/j.jbi.2013.07.007

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2020). *Shiny: Web application framework for r*. Retrieved from https://CRAN.R-project.org/package=shiny

Ooms, J. (2018). *Hunspell: High-performance stemmer, tokenizer, and spell checker*. Retrieved from https://CRAN.R-project.org/package=hunspell