# Integrated Dietary Supplement Knowledge Base (iDISK)

Jake Vasilakes

December 12, 2018

# Contents

# 1. Introduction

## 1.1 Data Sources

| Resource | URL |
|---|---|
| NMCD | `https://naturalmedicines.therapeuticresearch.com` |
| DSLD | `https://dsld.nlm.nih.gov/dsld/` |
| NNHPD | `https://www.canada.ca/en/health-canada/services/drugs-health-products/natural-` |

# 2. Installation

# 3. Usage

# 4. Building iDISK from Source Data

As discussed previously, iDISK is built by integrating information from three dietary supplement resources: the Dietary Supplement Label Database (DSLD), the Natural Medicines Comprehensive Database (NMCD), and the Natural and Non-prescription Health Products Directorate (NNHPD). Ingredient and product information is extracted from each resource, merged, and formatted for iDISK. This chapter explains how to build a new iDISK version from these data sources.

## 4.1 iDISK Build Directory Structure

The overall structure of the `iDISK` build directory is the following:

```
iDISK/
  doc/
  lib/
  sources/
  versions/
```

- `doc/`: Contains all formal documentation of iDISK.
- `lib/`: Functions, scripts, and libraries that are common to all versions of iDISK.
- `sources/`: Contains a directory for each of the data resources, e.g. `NMCD/`.
- `versions/`: Contains a directory for each version of iDISK, e.g. `1.0.0` or `1.0.1`.

The directory structure for each resource follows. We use NMCD as an example.

```
sources/
  NMCD/
    08_01_2018/
      README
      download/
      import/
        preprocess/
          ingredients.jsonl
          products.jsonl
      scripts/
```

- NMCD/: Contains all data related to the initial download and processing of NMCD data.
- 08_01_2018/: The data in MM_DD_YYYY format when the source files were downloaded.
- README: Documentation for this download, including data version (if applicable), issues, etc.
- download/: Contains the downloaded data files in the original format.
- import/: Contains the data files in the standard iDISK JSON lines format.
- preprocess/: Contains files that hold and intermediate preprocessing in converting to the iDISK JSON lines format.
- ingredients.jsonl, products.jsonl: iDISK JSON lines files containing the converted ingredient and product information, respectively.
- scripts/: Scripts for processing and importing the data files.

The directory structure for each version follows. We use version 1.0.0 as an example.

```
versions/
  1.0.0/
    CHANGELOG.md
    scripts/
    tables/
    ingredients/
      matched/
        manual_review/
        nmcd_dsld.jsonl
      unmatched/
        nmcd.jsonl
```

- 1.0.0/: Contains all files related to the build of this version.
- CHANGELOG.md/: Changelog for this version.
- scripts: Scripts for building this version.
- tables: Contains the final iDISK tables. I.e. DSCONSO.RRF, DSSTY.RRF, DSSAT.RRF, DSREL.RRF

- `ingredients/`: Contains intermediate files related to the matching and processing of ingredient data.
- `matched/`: Contains data related to ingredients that were matched across one or more resources.
- `manual_review/`: Contains files necessary to conduct manual review of the automatically matched ingredients.
- `nmcd_dsld.jsonl`: iDISK JSON lines file containing NMCD ingredients that were matched to DSLD ingredients, after manual review.
- `unmatched/`: Contains iDISK JSON lines files for each resource, each containing the ingredients that were not found in any other resource.
- `nmcd.jsonl`: iDISK JSON lines file containing NMCD ingredients that were not found in any other resource.

## 4.2 Obtain the Source Data

The first step in the build process is to obtain the ingredient and product information contained in each of the resources. This is relatively straightforward for DSLD and NNHPD: DSLD provides means to download its ingredient and product data via a download button on the DSLD website and an API, respectively; NNHPD releases a full data extract containing both ingredient and product information. In contrast, NMCD does not provide a data release and, furthermore, requires a subscription to access the ingredient and product monographs. Below we detail how to obtain the ingredient and product data from each of these resources.

**DSLD:** DSLD provides two zip files containing ingredient and product information, respectively. These can be downloaded from `https://www.dsld.nlm.nih.gov/dsld/faq.jsp#9`. However, neither of these files contain sufficient information for inclusion in iDISK. The ingredients file does not contain the ingredient category, which will become an ingredient attribute in iDISK. The products file, besides missing a variety of attributes, does not list the ingredients for each product, which is necessary to create the `has_ingredient` relationships in iDISK. We therefore describe alternate means of obtaining the DSLD data:

1. **Ingredients:** For each letter in the alphabetical listing of ingredients on the DSLD website (`https://www.dsld.nlm.nih.gov/dsld/lstIngredients.jsp`) download the ingredient listings using "Export to Excel" button. Once all 27 listings have been downloaded, convert the Excel files to CSV and merge them. To automate the conversion and merging, we recommend installing `csvkit` (`https://csvkit.readthedocs.io/`), a collection of command line programs written in Python to perform various tasks on CSV files. Specifically, use `in2csv` to convert the Excel files and then concatenate them to merge.

2. **Products:** Use the DSLD API to download detailed product information. We provide a script, `sources/DSLD/scripts/dsld_api.py`, that queries the API given a list of product IDs. To get the product IDs, download and unzip the product listing (`all_lstProducts_csv.zip`) from the DSLD FAQ (URL given above) and extract the DSLD ID column. Run `dsld_api.py --help` for more information.

**NMCD:** To obtain the NMCD ingredient and product information it is first necessary to create an account on the NMCD website. With log in credentials, it is then possible to scrape the ingredient and product monographs to obtain structured data. We provide a set of web scrapers to do this, located in `sources/NMCD/download/nmcd_monographs`. The `README` in the `sources/NMCD/download/` directory contains detailed instructions on how to run the web scrapers.

**NNHPD:** Download the ingredient and product information directly from the Licensed Natural Health Product Database Data Extract (`https://www.canada.ca/en/health-canada/services/drugs-health-products/natural-non-prescription/applications-submissions/product-licensing/licensed-natural-health-product-database-data-extract.html`). Specifically, download the `NHP_Products` and `NHP_Medicinal_Ingredients` files.

## 4.3 Preprocess the Source Data

We then convert the source data from each resource into a common JSON lines data format. This format is detailed below. Scripts to do this conversion can be found in `RESOURCE_NAME/scripts/{extract_ingredients.py,extract_products.py`.

### 4.3.1 The iDISK JSON lines format

Each line of the JSON lines format corresponds to either an ingredient or a product and has the following format:

```
{
 "preferred_term": str,
 "src": str,
 "src_id": str,
 "term_type": str,
 "synonyms": [{"term": str,
               "src": str,
               "src_id": str,
               "term_type": str,
               "is_preferred": bool},
              {...}],
 "attributes": [{"atr_name": str,
                 "atr_val": str,
                 "src": str},
                {...}]
 "relationships": [{"rel_name": str,
                    "rel_val": str,
                    "src": str,
                    "attributes": [{...}]
                   }]
}
```

**N.B.** The `preferred_term` is not listed in the `synonyms`.
Descriptions of each field in the iDISK JSON format are below:

- `preferred_term`: The listed string for this concept in the resource.
- `src`: The name of the resource from which this concept was extracted, e.g. `NMCD`.
- `src_id`: The unique identifier assigned to this concept in the resource.
- `term_type`: The term type of the `preferred_term`. See the description of the term types in iDISK.
- `synonyms`: Synonyms for this concept as given in the resource. Descriptions of the subfields are below.
- `attributes`: Information extracted from this resource that will become attributes in iDISK. Each `key` corresponds to the attribute name, `value` corresponds to the attribute value. E.g. {`umls_semantic_type: plnt`}.
- `relationships`: Information extracted from this resource that will become relationships (and any attributes thereof) in iDISK. `rel_name` corresponds to the relation (e.g. `is_effective_for`) and `rel_val` corresponds to the object concept of the relation (e.g. `D0002123`). `attributes` follows the same format as above.

Below are the descriptions of the subfields of `synonyms`.

- `term`: The synonym string.
- `src`: The name of the resource from which this concept was extracted.
- `src_id`: The unique identifier assigned to this concept in the resource.
- `term_type`: The term type of the `preferred_term`. See the description of the term types in iDISK.
- `is_preferred`: Boolean value (`true` or `false`) specifying whether this `term` is preferred in this `src`, i.e. whether this `term` is a `preferred_term`.

In addition to converting the source data to this format, the `extract_ingredients.py` and `extract_products.py` scripts perform necessary preprocessing of the terms, including the removal of unwanted punctuation, stop words, and dosages (e.g. “10mg”). See the scripts themselves for more details.

## 4.4 Merge Ingredients from the Source Data

### 4.4.1 Manual review of the matched ingredients

## 4.5 Generate Attributes and Relationships

## 4.6 Create the iDISK Data Files