Complex food data extraction from [USDA FoodData Central](https://nam10.safelinks.protection.outlook.com/?url=https%3A%2F%2Ffdc.nal.usda.gov%2F&data=05%7C02%7Cyuzeguo%40ufl.edu%7C326d37ef36034d09e18808dd715b4d5e%7C0d4da0f84a314d76ace60a62331e1b84%7C0%7C0%7C638791359126563613%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=5eh7IkMDUtO89WVGTXG8aU0lSPRsz1lm6k8A5fPuVsk%3D&reserved=0)

Goal: extract nutritional value and ingredients (with weight) to serve for further healthy eating index.

We also need an FDC ID,  a unique ID that we can easily refer back to for further information.

There are five databases under the USDA food data search engine.

            Foundation Foods ( for simple food)

            SR Legacy Foods (Stopped updating in 2019)

            Survey foods (FSDDN) (have ingredient composition, most recommended)

            Experimental food (have supplement, not usually used in the BEACH study)

            Branded Food (measure method unknown)

To extract the data:

1. Search the complex food name in the food box. Try Require All Words first, if none of the database have any similar food. Then untick the box and click search again.

Sometimes you may find that the name is not directly matched but the ingredient is totally match when you click the Ingredients tab under the FSDDN database. Therefore, do not give up if you do not see a direct match of the complex food name.

If you can use the FSDDN database, always use it in terms of timeliness and comprehensiveness.

1. Copy and paste targeted information into the spreadsheet. Unmatched is labeled as 0, similar is 1, and direct match is 2.

To make the spreadsheet neat, it is better to integrate the ingredients with weight in one cell. You can extract the ingredient table to an Excel, and use R to integrate the information to the text for one cell.

**# Function to format ingredients into a single-cell recipe string**

format\_ingredients\_for\_spreadsheet <- function(df,

                                                col\_names = c("Ingredient Description", "Ingredient Code", "Ingredient Weight (g)"),

                                                item\_separator = " | ") {

**# Apply to each row: build a formatted string with name, code, and weight**

  ingredient\_lines <- apply(df[col\_names], 1, function(row) {

    paste0(

      row[1],                                  # Ingredient name

      " (Code: ", row[2],                      # Ingredient code

      ", Weight: ", row[3], "g)"               # Weight in grams

    )

  })

**# Combine all ingredient lines into one long string, separated by " | "**

  result <- paste(ingredient\_lines, collapse = item\_separator)

  return(result)

}

 Also, **wrap** the text in a cell to avoid the text occupying the next cell.

These information are used for HEI calculation, which is the proximity to the Dietary Guideline of Americans, with a maximum of 100. Since we know the total calories, the ingredient, weight, the the nutritional value of the complex food, we are able to calculate the index of these food that a mother consume during pregnancy and it may facilitate further study.