

Crawl Ingredient Database

I need to crawl a website to extra data for food ingredients.

The URL is:

<https://ndb.nal.usda.gov/ndb/foods?format=&count=&max=100&sort=&fgcd=&manu=&facet=&qlookup=&offset=0&order=desc>

If you change “max” to 9144 then you will retrieve all rows in the database so you don’t have to navigate pages.

NDB No.	Description	Food Group
01001	Butter, salted	Dairy and Egg Products
01002	Butter, whipped, with salt	Dairy and Egg Products
01003	Butter oil, anhydrous	Dairy and Egg Products
01004	Cheese, blue	Dairy and Egg Products

For each one of these rows, we need to extract the link:

```
<tr style="line-height:1.2em;" class="odd">
  <td style="padding:8px;width:10%;font-style:;><a href="/ndb/foods/show/3?
fg=&manu=&facet=&format=&count=&max=35&offset=&sort=&qlookup=" style="font-weight:normal;" title="Click to view reports for
this food">01003</a></td>
```

and use the number here: /ndb/foods/show/{num} and reformat request as:

<https://ndb.nal.usda.gov/ndb/foods/show/{num}?format=Full>

This will display the page:

Full Report (All Nutrients): 08113, Cereals, farina, enriched, cooked with water, without salt

[Return to Search Results](#) [Basic Report](#) [Statistics](#) [Report](#) [Download \(CSV\)](#) [Print \(PDF\)](#)

Food Group: Breakfast Cereals

4 **Carbohydrate Factor:** 4.12 **Fat Factor:** 0.84 **Protein Factor:** 4.05 **Nitrogen to Protein Conversion Factor:** 5.7

Nutrient values and weights are for edible portion.

☐ Apply Changes

Nutrient	Unit	1 Value per 100 g	# of Data Points	Std. Error	1 cup 240g	1 tbsp 15g
Proximates						
Water ¹	g	86.55	26	0.279	207.72	12.98
Energy	kcal	53	--	--	127	8
Energy	kJ	220	--	--	528	33
Protein ¹	g	1.82	26	0.048	4.37	0.27
Total lipid (fat) ¹	g	0.34	26	0.025	0.82	0.05
Ash ¹	g	0.38	26	0.023	0.91	0.06

Following are instructions how to parse this:

1. nbd_no = 5 digit code (e.g., 08113)
2. full_name = string after nbd_no (e.g., "Cereals, farina, enriched, cooked with water, without salt")
3. food_group = string (e.g., "Breakfast Cereals")
4. Zero or more factors (sometimes there won't be any, other times 1 or more). Store in factors array. (e.g., factors:[{factor:"Carbohydrate Factor", amt:4.12},{factor:"Fat Factor", amt:0.84},{factor:"Protein Factor", amt:4.05},{factor:"Nitrogen to Protein Conversion Factor", amt:5.7}])
5. Nutrients. We should parse all rows in this table (including the detail rows under "Lipids" and "Amino Acids"). We only need three columns per nutrient:
nutrients:[
 {nutrient:"Water", unit:"g", per100g:86.55},
 {nutrient:"Energy", unit:"kcal", per100g:53},
 {nutrient:"Energy", unit:"kJ", per100g:220 },
 {nutrient:"Protein", unit:"g", per100g:1.82 },
 etc. (there are around 100-120 nutrients in total)
]

That's it. Data should be exported into a valid json file which should contain 9144 ingredient objects.