

This document contains additional details about the development of the food/allergen resource described in the abstract submitted to the American Medical Informatics Association (AMIA).

Wiki data were found at:

https://en.wikipedia.org/wiki/Wikipedia:Database_download

Where do I get the dumps?

English-language Wikipedia

- Dumps from any Wikimedia Foundation project: dumps.wikimedia.org and the [Internet Archive](#)
- English Wikipedia dumps in SQL and XML: dumps.wikimedia.org/enwiki/ and the [Internet Archive](#)
 - [Download](#) the data dump using a BitTorrent client (torrenting has many benefits and reduces server load, saving bandwidth costs).
 - pages-articles-multistream.xml.bz2 – Current revisions only, no talk or user pages; this is probably what you want, and is over 19 GB compressed (expands to over 86 GB when decompressed).
 - pages-meta-current.xml.bz2 – Current revisions only, all pages (including talk)
 - abstract.xml.gz – page abstracts
 - all-titles-in-ns0.gz – Article titles only (with redirects)
 - SQL files for the pages and links are also available
 - All revisions, all pages: **These files expand to multiple terabytes of text. Please only download these if you know you can cope with this quantity of data.** Go to [Latest Dumps](#) and look out for all the files that have 'pages-meta-history' in their name.
 - To download a subset of the database in XML format, such as a specific category or a list of articles see: [Special:Export](#), usage of which is described at [Help:Export](#).
 - Wiki front-end software: [MediaWiki \[1\]](#).
 - Database backend software: [MySQL](#).
 - Image dumps: See below.









Specifically, we visited this page:

https://meta.wikimedia.org/wiki/Data_dump_torrents#English_Wikipedia

and downloaded the [enwiki-20241201-pages-articles-multistream.xml.bz2](#) (22.64 GiB) file.

English Wikipedia [\[edit \]](#)

For more info on multistream vs non multistream, please see [en:Wikipedia:Database download#Should I get multistream?](#)

- **2024-12-01**
 - [enwiki-20241201-pages-articles-multistream.xml.bz2](#)  (22.64 GiB) on academictorrents.com ([magnet](#))
- 2024-11-01
 - [enwiki-20241101-pages-articles-multistream.xml.bz2](#)  (22.54 GiB) on academictorrents.com ([magnet](#))
- 2024-10-01
 - [enwiki-20241001-pages-articles-multistream.xml.bz2](#)  (22.43 GiB) on academictorrents.com ([magnet](#))
- 2024-09-01
 - [enwiki-20240901-pages-articles-multistream.xml.bz2](#)  (22.31 GiB) on academictorrents.com ([magnet](#))
- 2024-08-01
 - [enwiki-20240801-pages-articles-multistream.xml.bz2](#)  (22.21 GiB) on academictorrents.com ([magnet](#))
- 2024-07-01
 - [enwiki-20240701-pages-articles-multistream.xml.bz2](#)  (22.10 GB) on torrage.info ([magnet](#))
 - [enwiki-20240701-pages-articles-multistream.xml.bz2](#)  (22.10 GiB) on academictorrents.com ([magnet](#))
- 2023-12-20
 - [enwiki-20231220-pages-articles-multistream](#)  (21.15 GB) ([magnet](#))
- 2023-08-20
 - [enwiki-20230820-pages-articles-multistream](#)  (20.8 GB) ([magnet](#))

About this overall data file:

Number of rows: 1,523,494,532

We identified articles that had **main_ingredient** as part of the food info box, which is where the key information of interest was listed.

Wikipedia has author guidelines on articles, including for foods:

https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Food_and_drink/Tools/guidelines

The key area we looked into were Infoboxes:

https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Infoboxes

And there are guidelines for food infoboxes:

https://en.wikipedia.org/wiki/Template:Infobox_food

which has a category for **main_ingredient** and a separate one for **minor_ingredient**

We only focused on the **main_ingredient** for this work.

Number of rows in the dataset that contain **main_ingredient** : 6764

However, some of them are essentially blank, and others have 'garbage' text

We noted that **name** is a metadata element and **title** is the html web page title. For the articles we extracted that had a **main_ingredient**:

- 16 had a name but not a title (so the name was used instead in these cases)

- 19 had no name or title
- 279 had a title but no name
- 5713 rows had title and name that matched exactly

We used title when possible, but for those with a name but not a title, we used the name.

Some light clean-up was done of the names like removing some descriptive text in parentheses and words like "Draft:" that was in some of them.

Additional processing was done using Chat GPT. Details are below:

We used Chat GPT 3.5 turbo (gpt-35-turbo), API version 2024-10-21.

Max tokens: 1024

Temperature: 0

Frequency penalty: 0

Top-p: 0.95

Step 1.

We first used it to obtain a list of ingredients for each of the food names derived from Wikipedia.

System prompt: “You are an expert in cooking and recipes”

User prompt: “List all of the main ingredients in the food known as ” + *foodName* + “. List each ingredient separated by a pipe (|). Do not list minor ingredients. Do not provide any introductory text. Do not provide any other definitions or context.”

(In the code, the *foodName* was substituted for each food name).

This was run on January 23, 2025.

Step 2.

We then used the Chat GPT API to obtain a list of allergens for each food, by providing it with the name of the food as well as the ingredients.

System prompt: “Act as an expert nutritionist and allergist”

User prompt: “Below is a list of common ingredients for a food known as “ + *currentFoodName* + “. Please look through each ingredient carefully and make an overall assessment about which common allergens are likely to be present in the overall food. The list of possible allergens are: milk, eggs, fish, shellfish, tree nuts, peanuts, wheat, soy. The main ingredients in this food are as follows, bar (|) delimited: “ + *currentIngredientList* + “. Output your results in a format like the following example, where YES means the allergen is likely to be present and NO means the allergen is unlikely to be present. Provide the result to match the example output. Do not add any additional commentary, details in parentheses, etc. Example output: food name: XXXXX | milk: no | eggs: yes | fish: no | shellfish: yes | tree nut: no | peanuts: yes | wheat: no | soy: no”

(In the code, the *currentFoodName*, would be substituted for the name of the food. The *currentIngredientList* would be substituted for the ingredient list derived from Step 1.)