Cochrane Crawler Technical Overview

Design Reasoning

The crawler makes Cochrane Library review data accessible outside the website. It's designed to be lightweight, easy to run, and requires no extra setup beyond basic PHP.

The design focuses on simplicity and reliability. It is easy to use, captures all results, and saves files safely in the project's root directory.

High-Level Flow of Application Logic

Step 1: Startup

The user runs:

```
php crawler.php
```

The program starts by launching the crawler, which gets everything ready to connect to the Cochrane Library website.

Step 2: Topic Discovery

The crawler requests the Cochrane Library's topic list page and parses it for available topics. These are displayed in a numbered menu for the user:

```
0. Allergy & intolerance
1. Blood disorders
2. Cancer
...
36. Wounds
```

Step 3: Topic Selection

The user selects the number of their chosen topic from the list.

Step 4: Review Fetching

The crawler goes through all the review pages for the chosen topic, collects the important details like title, authors, and date, and shows progress in real time on the screen.

Console output shows progress in real time:

```
Fetching results for 'Allergy & intolerance'...
  -> Processing page 1 of 3
  -> Processing page 2 of 3
```

```
  -> Processing page 3 of 3
63 reviews have been found.
```

Step 5: File Export

The user is asked to choose a filename, and the reviews are then saved to a file in the project folder.

Example:

```
Enter a filename to save the results (e.g. cochrane_reviews.json):
allergy_reviews.json
Reviews saved to: /project/allergy_reviews.json
```

Each review is written on a single line in the following pipe-delimited format.

```
http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD010112.pub2/full|Allergy &
intolerance|Polyunsaturated fatty acid supplementation in infancy for the
prevention of allergy|Tim Schindler, John KH Sinn, David A Osborn|2016-10-28
```

3. Potential Extensions

The crawler could be redesigned to work with multiple research sites, not just Cochrane. In that approach, each site would have its own rules for collecting data, while the overall process for the user, such as choosing a topic, running the crawl, and saving results, would remain the same. Future updates could also add filtering options so users can narrow results by date, review type, status, language, or content type. This would allow researchers to quickly focus on the most relevant information, such as reviews from the past year or those with updated conclusions, without needing to sort through unnecessary data.

4. Conclusion

The crawler provides a simple and reliable way to collect review data from the Cochrane Library. It allows users to obtain organized information in just a few steps from the command line. The design also makes it easy to expand in the future, for example by adding new fields, changing how the output is saved, or connecting it to a larger system.