


# Design of MapDB database

## 1. Used Package/Library

Language	Kotlin 
Database Library	mapdb:3.0.8
HTML parsing Library	jsoup:1.19.1
Logging Framework	kotlin-logging-jvm:7.0.3, slf4j-api:2.0.3, logback-classic:1.4.14
Test Libraries	mockito-core:2.1.0, junit-jupiter-api:5.12.1, mockito-junit-jupiter:5.16.1, mockito-kotlin:4.1.0

## 2. Data Structure

### Page Class

```
data class Page(  @Wayd
    val id: String,
    val url: String,
    val title: String? = null,
    val content: String = "",
    val lastModified: String? = null,
    val size: Int = 0,
    val links: List<String> = emptyList()
)
```

id	the UUID of the page
url	the url to the page

title	the title of the page
content	the content(body) of the page
lastModified	the last modified time of the page
size	the size in Bytes of the page
links	the links appeared on the page

## Post Class

```
data class Post (  👤 Wayd +1
    val pageID: String,
    val frequency: Int,
    val position: List<Int>,
) : Serializable
```

-> Worked as the value of the inverted indexes

pageID	the UUID of the page
frequency	the frequency of the keyword
position	position of the keyword (Not implemented in phase 1)

## Keyword Class

```
data class Keyword(  👤 Wayd
    val wordID: String,
    val frequency: Int
) : Serializable
```

-> Worked as the value of the forward indexes

wordID	the UUID of the word
frequency	the frequency of the word

### 3. Supporting Database

#### **CrawlerDB (crawler.db):**

- **UrlToPageId Map:**
  - Maps URLs to unique page IDs.
  - Example:  
https://www.cse.ust.hk/~kwtleung/COMP4321/Movie/265.html maps to dfb710fd-e3a6-493d-9bec-7f9f749b68ed.
- **PageProperties Map:**
  - Stores metadata for each page, including title, last modified date, size, and URL.
  - Example: Page 8b30cbc9-d72d-46a9-bc22-d9ecd9bc7e1d has title book1, last modified 2025-03-22T16:37:09.238804Z, size 1620, and URL https://www.cse.ust.hk/~kwtleung/COMP4321/books/book1.htm.
- **Parent-Child Link Map:**
  - Tracks hyperlinks between pages (parent page to child pages).
  - Example: Page 0eae24f3-3760-43c8-8b5c-d7c4a02208ce links to pages like 97eadfac-6321-4da7-9931-482857fca454 and c3b8e144-fa98-4b25-a2a9-21bfc8a68313.

#### **IndexerDB (indexer.db):**

stopwords and symbols in a page are removed before performing stemming using Porter algorithm. The results are stored into the following maps

- **InvertedTitle Map:**
  - Maps word IDs (UUIDs) to posts (pages) where the word appears in the title.
  - Example: Word ID 0419cf5e-1aa6-4850-8dfa-1531112e71c6 appears in pages 0eae24f3-3760-43c8-8b5c-d7c4a02208ce and

c3b8e144-fa98-4b25-a2a9-21bfc8a68313, each with a frequency of 1.

- **InvertedBody Map:**

- Maps word IDs to posts where the word appears in the body of the page.
- Example: Word ID  
009aad20-a162-449c-89a5-c9ed0ac71767 appears in page f36a3654-4b76-486d-9547-13fcd730ab0d (frequency 1) and c3b8e144-fa98-4b25-a2a9-21bfc8a68313 (frequency 2).

- **WordToWordID Map:**

- Maps actual words to their corresponding word IDs.
- Example: Word simply maps to  
15abf012-8b56-45df-9523-b5f8d5b5285a.

- **WordIDToWord Map:**

- Reverse mapping of word IDs to words.
- Example: Word ID  
7fa9e4a4-37f3-4955-ba47-2d182ea5ab02 maps to shriek.

- **Forward Index Map:**

- Maps page IDs to a list of keywords (word IDs) with their frequencies on that page.
- Example: Page 8b30cbc9-d72d-46a9-bc22-d9ecd9bc7e1d has keywords like  
cebd09e0-57a8-4fcc-909d-a12a89fb0a37 (frequency 5).