Files description

collatedsources.csv

It is a table of book metadata with following columns:

[ID] : list index from 1 to 146

[in_corpus] : boolean (Y/N) if digitized copy of the book is available in *collatedbooks_v1.zip* file. Non-available book copies were removed from the list.

[language] : identifies a language of the digitized book copy. All copies other than French (fr) were removed.

[book_code] : the value is a combination of [book_category_code] transformed into 2-digit value (1 → 01, 2 → 02, etc.) and of [book_number] also transformed into 2-digit value. Hence, the book_number=2 in book_category=3 becomes the value 03_02 in the list. The [book_code] exactly matches the file name of the digitized book copy. Hence, each record in the table matches exactly one file (digitized book copy) in the *collatedbooks_v1.zip*. Filename structure follows exactly the pattern: *[book_code].txt*. Example: The record with book_code=03_02 refers to the file *03_02.txt*.

[book_category_code] : is a numeric value attributed to each categorical value of [book_category].

[book_category_name] : is a given categorical value

[book_number] : is a given numeric number of the book in a given category.

[author] : name of the author of the book, available as digitized copy in *collatedbooks_v1.zip*

[book_title] : original title of the book

[place] : place of the original book print

[year_edited] : year of print, original or closest approximation

collatedbooks_v1.zip

The ZIP file contains 148 TXT files with raw text in French. Each file metadata is represented by 1 row in *collatedsources.csv* table. Name of the each file matches the pattern: [book_code].txt