## Опис зібраних даних з Wikipedia API і їх обробки

**Зібрання даних відбувалося в наступній послідовності**:

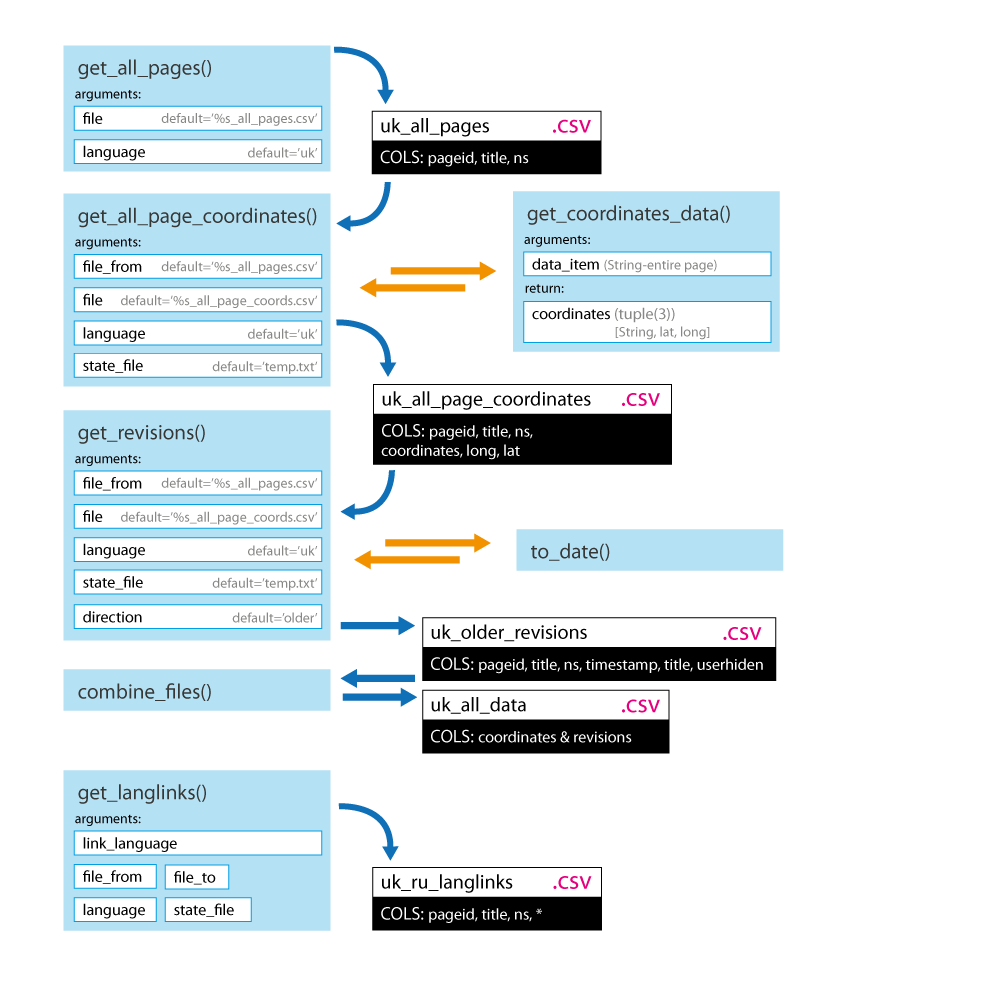
1. Обирання мови до якої буде відбуватись зібрання інформації.
2. Отримання всіх сторінок з Вікіпедії даної мови. (Результатом є назва (title) сторінки і pageid, які використовуються для ідентифікації сторінки в Вікіпедії і відповідно для збору більше інформації про дану сторінку).
3. Отримання повного вмісту сторінки

Весь текст сторінки є в вигляді спеціального формату, використовуваного Вікіпедією, для визначення окремих її частин (як назва, коротка інформація, заголовки, фотографії, посилання на інші сторінки Вікіпедії, джерела даної сторінки, короткий зміст сторінки, географічні координати)

1. З отриманого вмісту сторінки, необхідно виділити координати, розкодувати їх і зберегти.

Отримання координат здійснювалось водночас з безпосереднім отриманням вмісту сторінки, адже збережений повний вміст сторінок займав би занадто багато місця на комп’ютері і був би складним для обробки.

1. Отримання Даних про першу редакцію сторінки, зокрема про автора редакції та час редакції.
2. Отримання зв’язків між частинами Вікіпедії різних мов (langlinks).   
   Виконується, якщо вже завантажено інформацію (попередні пункти) про 2 мови.

**Візуалізація процесу зібрання і обробки інформації**:

**Опис функцій обробки інформації:**

Основна частина обробки інформації була здійснена за допомогою функції **get\_coordinates\_data**().

Функція приймає весь вміст однієї сторінки і повертає знайдені і декодовані координати. Знаходить координати, які знаходяться, як в назві сторінки (основні координати і мають display=title атрибут заданий), так і просто координати, які знаходяться на сторінці і несуть додаткову інформацію.

Координати на сторінці Вікіпедії позначаються наступним чином: {{coordinates …}} або {{coord … }}. Всередині вони можуть містити значення ‘missing’ або ‘unknown’. Такі координати не обробляються.  
Далі функція розглядає внутрішній вміст координат. Параметри розділені ‘|’ і можуть задавати координати різними способами:

* Рядком цифр (0-360 | -90-90), які вказують координати число з плаваючою точкою безпосередньо в градусах.
* Набором параметрів, які визначають градуси, мінути і секунди, а також напрямок (‘n’ / ‘s’ / ‘w’ / ‘e’). Кількість параметрів може бути від 4 до 8.
* Координати також можуть містити додаткову інформацію, як назву місця, спосіб відображення координат, джерело інформації. Ці параметри функція знаходить і упускає.

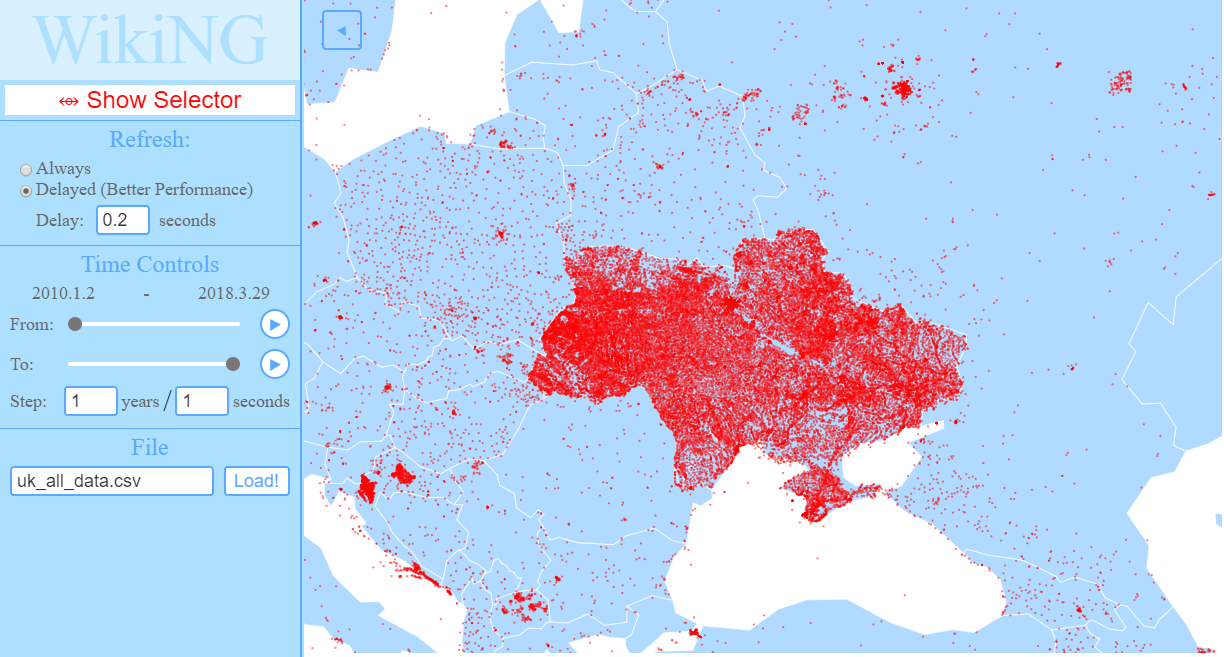
Функція **to\_date**() перетворює стрічку часу в стандартному для Вікіпедії форматі в об’єкт Date в python, з яким потім можна здійснювати операції за допомогою pandas.

**В результаті було отримано наступні кількості інформації**:

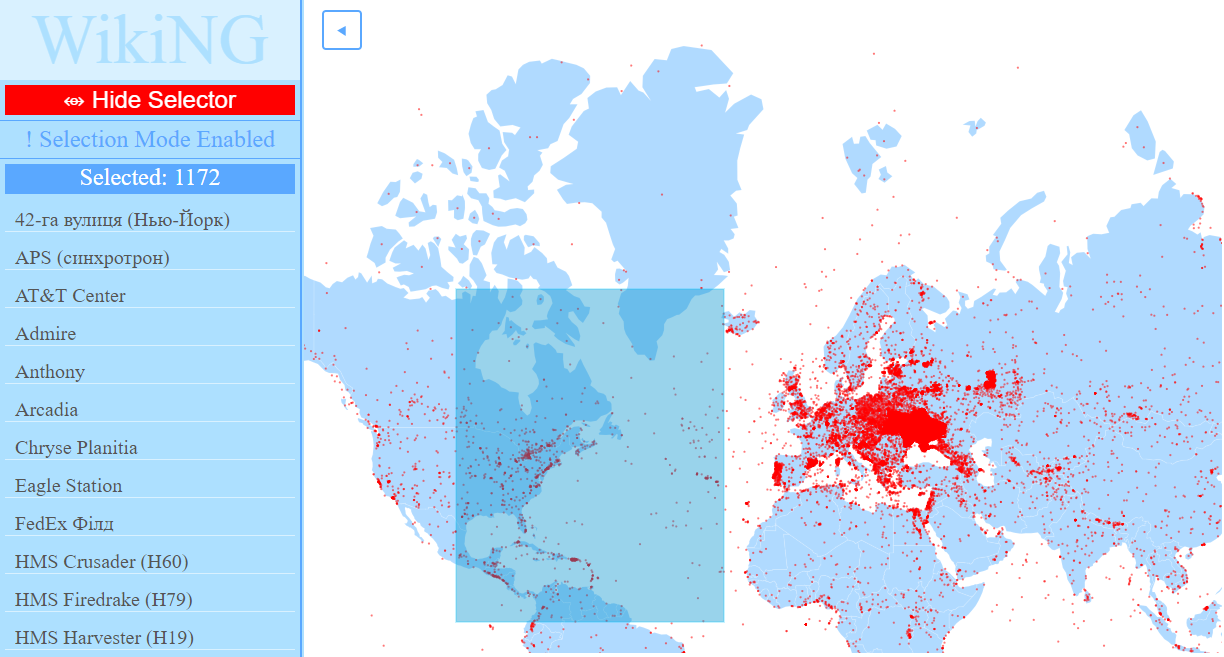
|  |  |
| --- | --- |
| Назва файлу | Кількість рядків таблиці |
| uk\_all\_pages.csv | 781 260 |
| uk\_all\_page\_coords\_clear.csv | 62 977 |
| uk\_older\_revisions.csv | 62 977 |
| uk\_all\_data.csv | 62 977 |
| uk\_ru\_langlinks.csv | 234 |

**Візуалізації результату**

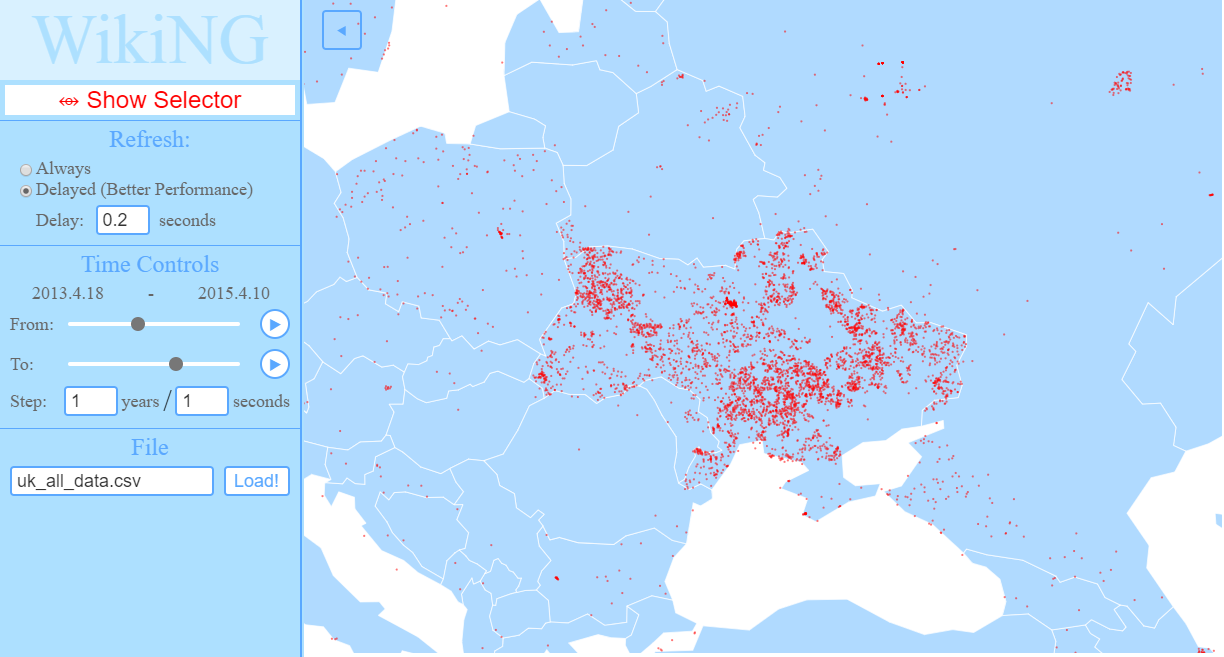
Подальша обробка інформації здійснюється при візуалізації і за бажанням користувача.  
Візуалізація карти Вікіпедії здійснюється в вигляді веб-додатку, який завантажує інформацію з сервера Flask і виводить її на екран використовуючи можливості бібліотеки d3.js та canvas-елементу html і взаємодію з ним і іншими елементами DOM через js.

**Відображення всієї інформації** uk\_all\_data.csv **на карті** за допомогою програми:

Додаток має наступні додаткові можливості (завдяки до того зібраній інформації):

1. **Переміщатись** по карті і **приближувати** певні її території – карта автоматично заново виводить інформацію, більш детально для даного регіону.  
   В залежності від кількості виведеної інформації можна оптимізувати переміщення і загрузку карти для більшої швидкості роботи або красивіших переходів.
2. **Виділяти певні території** карти. Карта підрахує кількість сторінок в виділеній області, представить список всіх назв сторінок і дасть можливість перейти на відповідну сторінку Вікіпедії, представлену на карті.
3. Контролювати вивід інформації в залежності від **часу створення сторінки** на Вікіпедії.

2 слайдери вводу інформації дозволяють вказати межі після і до яких дат виводити інформацію (картинка знизу показує сторінки української Вікіпедії від квітня 2013 до квітня 2015, як вказано в розділі Time Controls зліва вікна.

На початку виводиться також інформація про час створення першої сторінки з геолокацією. Цю сторінку можна переглянути відповідно перетягнувши слайдери. Це буде [Площа Рамуса ді Азаведу](https://uk.wikipedia.org/wiki/%D0%9F%D0%BB%D0%BE%D1%89%D0%B0_%D0%A0%D0%B0%D0%BC%D1%83%D1%81%D0%B0_%D0%B4%D1%96_%D0%90%D0%B7%D0%B5%D0%B2%D0%B5%D0%B4%D1%83_(%D0%BE%D0%BA%D1%80%D1%83%D0%B3_%D0%91%D1%83%D1%82%D0%B0%D0%BD%D1%82%D0%B0%D0%BD)) (Бразилія).  
Кнопки біля слайдерів дозволяють візуалізувати зміни, автоматично переміщуючи слайдери зліва направо.

1. Додатково зібрана інформація дозволяє фільтрування сторінок в залежності від автора редакції – зареєстрований користувач, анонім, бот і інші можливості.