Using OpenRefine for OSINT purposes

DataHarvest 2023 - June 1-4

Anuška Delić & Hervé Letoqueux

Advanced workshop

https://s.42l.fr/DH2023_2

Dataharvest – the European Investigative Journalism Conference

OpenRefine

Important TEXTRAZOR API.

We will need this!

- Take 5mn to create a free account on Textrazor : HTTPS://WWW.TEXTRAZOR.COM/
- Once done go to your profil and look for your API key
- This will allow you 500 requests per day, more than enough for today....

OpenRefine for OSINT.

A pimped version of OpenRefine with Docker...

- This version adds a plugin called vib-bits, some command line tools (cli) such as ddgr, and trafilatura to boost OpenRefine's potential.
- The container also embeds a TOR service that allows you to *torify* command line instruction (for anonymization).
- Easy to do with Linux, pain in the a** with Windows;
- That's why I crafted my own version of OpenRefine in a Docker container...
- We could consider OpenRefine as a hub for scraping, cleansing, enriching your data.

Important : Some of the things we'll see here will work on a regular version of Refine, but most won't. If you don't have Docker on your machine you can still use this version online : HTTP://35.210.183.250:80

What's special about this version?.

Plenty of command line tools...

- New tools & plugins : Whois, jq, trafilatura...
- You can add your own tools in the Dockerfile so it makes it extendable;
- You don't have to break any depencies on your own computer;
- There are 5 demos examples in the repository, for you to practice. We will only explore 2 or 3 of them.

Demos2 - import a json file and map it

Level Easy...

OpenRefine is really good at visually dealing with json files!

Let's use the ENI.JSON dataset and let me explain how I got this (OSINT baby!).....

- Create a new project :
  - **Browse** your file system, select the file, click **Next**.
  - Roll over your mouse on the preview window until the yellow area makes the point! Rename your project, add some tags.

At this point, your dataset is IN OpenRefine. Let's filter the data by country by selecting only FRA.

- Rename the column **latitudine** as **lat**, and **longitudine** as **lon**;
- Facet the **codice_nazione** column by **facet** and select FRA on the left
- Export your projet.
- Browse to OpenStreetMap, an opensource tool to create maps with your data. Create a new map. Import your fresh dataset.

Demos5 - Using Trafilatura and TextRazor

Level Medium...

Trafilatura is a powerful text scraper that works in command line. We will automatically download articles from a french disinformation website, cleanse this articles, and extract entities from it for text analysis.

- create a new project in OpenRefine by importing the csv file called MOUTONS.CSV.
- create a new column based on the url column with the trafilatura command **"trafilatura -u "+value** You can add a layer of anonymization, by using the *torify* command, which will wrap the command into TOR : "torify trafilatura -u "+value.
- create a new column based on this column by applying the provided jython script.

At this point, we scraped all the articles. Neat, Uh? Let's clean that!

- There are plenty of annoying characters in these texts such assmileys
- We cleanse them by applying a little python (jython script)
- We then create an other column , by curling the textrazor API (see demos5.pdf)

Others

Pleaser refer to DemosX.pdf files for other examples!

Questions

OpenRefine, je res super!
OpenRefine, c'est vraiment génial!