The Activity Crawler API

This tool is designed to empower the MOPO content team to rapidly build a large library of Activities from open-data and publicly available information.

This python-based API will take as input a search string consisting of the name of an activity and quite often location information such as the city and state of the activity (as part of the same string). You can build this API out in Django independently (or even just regular python) and we'll work on integration with our main Django API later.

It will then fetch information from a series of publicly available data sources to construct an initial dossier about an Activity. This dossier, the output of this API, will be presented as a JSON that combines significant information about the Activity consisting of multiple fields.

The JSON will be consumed by a user interface that allows the content team member to make any fine-tuning adjustments to the Activity metadata, including selecting pictures that the API returns, and then finalizing the Activity into a finished Activity ready for display on the MOPO mobile app.

For v1 of the crawler here are the data sets we have researched and recommend for first use which we need to get the fields for the the JSON. For a list of fields please see the MOPO Gold Standard doc. These are activities so we don't have "event start date" or "event end datetime" but will have opening hours and phone number. You may not be able to get those fields reliably which is OK… you can generate a google maps link for the venue and we can use AI to parse out operating hours and phone number so those two fields are not as critical.

The sample query string in this example is "nevada state railroad museum carson city"

https://nominatim.openstreetmap.org/ui/search.html?q=nevada+state+railroad+museum+carson+city
This web page also produces a clean JSON with this url structure:
https://nominatim.openstreetmap.org/search.php?q=nevada+state+railroad+museum+carson+city&format=jsonv2

If you follow the json you'll see an osmid and can supply that as the next query param
https://nominatim.openstreetmap.org/ui/details.html?osmtype=W&osmid=407063554&class=tourism
With JSON alternative:
https://nominatim.openstreetmap.org/details.php?osmtype=W&osmid=407063554&class=tourism&addressdetails=1&hierarchy=0&group_hierarchy=1&format=json

Note you can already construct the following fields: Activity name, street number, street address, city, state and zip and lat and long.

Please crawl from the links on those pages the following:
https://www.wikidata.org/wiki/Q14705239
Links to to commons category and images:
https://commons.wikimedia.org/wiki/Category:Nevada_State_Railroad_Museum
See pic detail page (link from image urls) EXAMPLE (and please grab the creative commons licensing string, and please offer to the content team member the LIST of images on the wikimedia page along with their CAPTIONS and AUTHORS - the person who created the image)
https://commons.wikimedia.org/wiki/File:4-4-0_Inyo.jpg
(links to: https://scholia.toolforge.org/topic/Q14705239)
And Links to Wikipedia:
https://en.wikipedia.org/wiki/Nevada_State_Railroad_Museum
and scholia (grab wikipedia summary?)
https://scholia.toolforge.org/topic/Q14705239

YOUR OUTPUT should be a JSON object about this Activity (the Nevada State Railroad Museum in Carson City) that you are able to populate by crawling the links above, with the fields in this JSON object matching as many of the fields in this “MOPO Gold Standard” document as possible.
https://docs.google.com/spreadsheets/d/1aygK4wf51_6M-umAWLLp-XFI2UAlwTXmc725C9_hOh4/edit?usp=sharing

Please note the spreadsheet is just to give you an idea of what the fields are that we need. The output should be a json, not a csv or spreadsheet.

Please note that for the Experience image field, you should supply an array (because wikimedia supplies a set of images, and we need a human to choose one of more of them) for the images, and each image should be represented by a 1) a url, 2) the license type (usually some kind of creative commons license) and 3) the caption (the short description of each image) as supplied in the wikimedia page.

For the Experience description field, please capture the first paragraph from Wikipedia (if first paragraph is less than 50 words then include second paragraph as well)

So the main idea is to crawl these pages, get as much information about as an Activity as possible, and put those fields into a JSON.

Please also share the python file you create to make the JSON