CC-GPX: Extracting High-Quality Annotated Geospatial Data from Common Crawl

Ilya Ilyankou*
ilya.ilyankou.23@ucl.ac.uk
UCL SpaceTimeLab
London, UK

Meihui Wang
meihui.wang.20@ucl.ac.uk
UCL SpaceTimeLab
London, UK

Stefano Cavazzi
stefano.cavazzi@os.uk
Ordnance Survey
Southampton, UK

James Haworth
j.haworth@ucl.ac.uk
UCL SpaceTimeLab
London, UK

ABSTRACT

The Common Crawl (CC) corpus is the largest open web crawl dataset containing 9.5+ petabytes of data captured since 2008. The dataset is instrumental in training large language models, and as such it has been studied for (un)desirable content, and distilled for smaller, domain-specific datasets. However, to our knowledge, no research has been dedicated to using CC as a source of annotated geospatial data. In this paper, we introduce an efficient pipeline to extract annotated user-generated tracks from GPX files found in CC, and the resulting multimodal dataset with 1,416 pairings of human-written descriptions and `MultiLineString` vector data from the 6 most recent CC releases. The dataset can be used to study people's outdoor activity patterns, the way people talk about their outdoor experiences, as well as for developing trajectory generation or track annotation models, or for various other problems in place of synthetically generated routes. Our reproducible code is available on GitHub: https://github.com/ilyankou/cc-gpx.

CCS CONCEPTS

• **Information systems** → **Geographic information systems**; **Web mining**; • **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → Geographic visualization.

KEYWORDS

Common Crawl, GPS, GPX, GIS, hiking, user-generated routes

ACM Reference Format:
Ilya Ilyankou, Meihui Wang, Stefano Cavazzi, and James Haworth. 2024. CC-GPX: Extracting High-Quality Annotated Geospatial Data from Common Crawl. In *The 32nd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '24), October 29-November 1, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3678717.3691215

1 INTRODUCTION

The Common Crawl (CC) corpus[1] is the largest publicly available web crawl dataset containing 9.5+ petabytes of data dating back from 2008 [2]. Due to its size, the dataset is widely used to train large language models, including GPT-3 [8].

In the past, researchers have studied the contents of Common Crawl for undesirable content [6], extracted documents to create the largest Spanish-language web crawl corpus [3], extracted parallel texts for machine translation [7], PDF documents [9] and Word documents with layout annotations [10], and built the 'Colossal Clean Crawled Corpus' (C4) dataset[2] among many other things. To our knowledge, we were first to quantify geospatial in CC [4].

While websites and apps such as Strava[3], AllTrails[4], and Outdooractive[5] contain large collections of curated and user-generated routes that are often accompanied by textual descriptions, these typically come with substantial licensing restrictions. CC, on the other hand, provides free access to all its data collected in accordance with website policies (subject to fair use restrictions), and exposes researchers to niche and varied websites that are harder to come across any other way.

In this paper, we introduce an efficient pipeline to identify, download, and clean GPX files that contain user-generated tracks, such as those produced by recording hiking, running, or cycling activities, along with human-written textual descriptions of those tracks. We run our process on the six most recent CC releases (spanning just over a year) to generate a multi-lingual, multi-activity dataset containing over 1,400 pairs of annotated `MultiLineString` features. When extended to all CC data back to 2008, the dataset is likely to contain over 10,000 samples.

The resulting dataset can be used to study the way people talk about outdoor activities, to facilitate the generation of human-like descriptions of activity tracks, to study outdoor activity trajectories, or in place of synthetically generated routes in any domain, including leisurely navigation.

*Corresponding author.

SIGSPATIAL '24, October 29-November 1, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1107-7/24/10
https://doi.org/10.1145/3678717.3691215

[1] https://commoncrawl.org/
[2] https://paperswithcode.com/dataset/c4
[3] https://www.strava.com/
[4] https://www.alltrails.com/
[5] https://www.outdooractive.com/en/

2 DATA COLLECTION AND PROCESSING

Hikers, cyclists, and in particular runners often track their activities using GPS-enabled mobile devices, such as Garmin smart watches [5], and sometimes add detailed annotations describing their experience. Alternatively, people may plan active journeys using websites such as `cycle.travel`, and add notes before or after the journey. In this paper, we decided to focus on the popular XML-based `.gpx`[6] (GPX Exchange Format) files that are most commonly used to share the activities, and ignore less popular formats such as `.fit` or `.tcx` which are very rare in Common Crawl [4].

2.1 Downloading GPX files from CC

We use index tables from the 6 most recent releases, `CC-MAIN-2023-*` and `CC-MAIN-2024-10`, to identify GPX files. Each index table contains a list of URLs and MIME-types of all web pages and files scraped by the CCBot in that release, and is broken into 300 parts. We use `duckdb`[7] to query each part to locate files whose detected MIME type is `ilike '%gpx%'` or whose filename extension is '.gpx' (typically both conditions are true at once).

For each GPX file, we record its Web Archive (WARC) file location, together with the offset and length in bytes. The offset and length indicate the location of the GPX file within a WARC file. This allows us to use Python's `requests`[8] library with the *Range* HTTP header to download individual GPX files from CC without the need to download large WARC files first:

```
requests.get(
    f'https://data.commoncrawl.org/{warc_file}'
    headers={
        'Range': f'bytes={offset}-{offset+length-1}'
    })
```

This massively speeds up the data acquisition process; in fact, it takes approximately 40 minutes to download all GPX files from a single CC release, which typically contains around 90 terabytes of compressed data spread across 3 billion files. We identified 112,953 GPX files across the six most recent CC releases. We were able to successfully download 111,102 files (98.4% of those identified). Among downloaded files, 102,103 (or 91.9%) came from unique URLs, indicating that few GPX files appear in more than one CC release. After removing duplicate GPX files that come from different URLs, we are left with 94,170 GPX files.

2.2 Identifying high-quality tracks

We use `gpxpy`[9] Python library to parse response strings into GPX objects that can be analysed. A typical GPX file consists of tracks; tracks consist of segments; segments consist of points. GPX files representing simple recorded point-to-point activities, such as hiking or cycling trips, would normally contain a single track that consists of one or more segments (for example, a runner may choose to record each lap as a separate segment). Multi-track GPX files that we inspected were typically used to record multi-day activities such as races around Europe. We decided to keep only single-track GPX files that represent activities no longer than 100 km (62.1 mi) because we believe longer activities cannot be adequately described in a few paragraphs, and are unlikely to be relevant to most people. We also remove activities shorter than 0.5 km (0.31 mi).

We keep tracks that have at least 1 GPS point per 100 m (328 ft) on average (~87% of all tracks satisfy this constraint). The threshold was determined empirically by analysing individual GPX files.

2.3 Identifying activity descriptions

We extract descriptions from each track's `<desc>` tag. We use `BeautifulSoup`[10] Python library to remove occasional HTML tags present in text, replace characters such as newlines and tabs with single spaces, and use regular expressions to remove text inside square and curly brackets (tags likely added by some apps).

We only keep the tracks that have an associated description between 50 and 2,000 characters long after the manipulations above. The upper character limit represents ~99th percentile of the description length distribution.

To identify high-quality track and activity descriptions, we employ Llama-3 8B Instruct, one of the most capable smaller open-source language models as of April 2024 [1]. We use the following prompt: *Does the text in triple quotes represent a high-quality and insightful route or track description, or an activity description such as hiking, cycling, or racing? Respond with 'True' or 'False'. If you are unsure, say 'False'. Text: “‘*`{text}`*”’*, and remove tracks for which *False* is returned.

Positive examples (tracks included in the dataset):

- 'A lovely 4 hour walk from Ockley railway station to Dorking, via the beautiful view from Leith Hill Tower.'
- 'Quieter roads and backstreets, quirky interest but still direct.'
- 'This trail leading to the borough of Kukleny takes in significant industrial buildings evoking the glory of the iron and leather industries, as well as family houses, a funeral hall, and a Cubist vocational school.'

Negative examples (tracks excluded from the dataset):

- 'Cheese-making is a centuries-old tradition in Gruyère and firmly rooted in the heritage. Immerse yourself in a walk with a historic feel.'
- 'File with points/tracks from Locus Map Classic/3.65.2'
- 'Note check opening times of Manor Park Cremitorium'

2.4 Removing personal information

While the GPX files available in CC would typically be uploaded and openly shared by the users in the open web, we chose to remove some personal information from the descriptions.

We use regular expressions to mask occasional emails and URLs with tokens *<EMAIL>* and *<URL>*. We identify phone numbers using Google's `libphonenumber`[11] Python library, and mask those with *<TELEPHONE>*. We also remove timestamps from points due to both privacy and data quality reasons (many tracks lack timestamps, and, unlike missing elevation data, timestamps are hard to recreate).

We then use Llama-3 8B Instruct with the prompt *Does the text in triple quotes contain any personally identifiable information, such as someone's address or name? Respond with 'True' or 'False'. If you are*

[6] http://www.topografix.com/GPX/1/1/
[7] https://duckdb.org/
[8] https://github.com/psf/requests
[9] https://github.com/tkrajina/gpxpy
[10] https://www.crummy.com/software/BeautifulSoup/
[11] https://github.com/google/libphonenumber

unsure, say ‘True’. Text: “‘`{text}`’” to further identify and remove tracks whose descriptions may contain personal information.

2.5 Translating descriptions to English

The majority of GPX files we identified are from European websites, and their descriptions are not in English. We use `pycld2`[12] Python library to identify the original language. We remove tracks whose language is identified as *Unknown* as those often contain majority numbers or incoherent text, as well as all tracks whose descriptions are in rare languages.

Initially, we experimented with several popular open-source LLMs, including Mistral and the Llamas (2 & 3), to translate descriptions into English, a task LLMs are shown to be good at [11]. Unfortunately, we discovered that LLMs occasionally modify formatting and add unnecessary context (e.g., *‘The text is in Ukrainian. Here’s the English translation: ...’*). Extensive prompt-engineering was not able to remedy these issues. We ultimately chose to use `argos-translate`[13], a popular offline Python translation library based on `OpenNMT`[14], an open source neural machine translation system, to translate all non-English descriptions into English.

2.6 Adding missing elevation

Just over half of all relevant GPX files contained device-recorded point elevation data. To produce a consistent dataset with all tracks represented by 3D points (lat, lon, elevation), we use the Shuttle Radar Topography Mission[15] digital elevation model, acquired via the `SRTM.py`[16] Python library, to approximate point elevations in tracks where elevation data is not originally available. We note the source of the elevation data, *GPS* (device-recorded) or *DEM* (SRTM), in the *elev_source* field (see Table 1).

3 DATASET & POTENTIAL APPLICATIONS

The final dataset consists of 1,416 tracks (or around 1.5% of the initially deduplicated GPX files) and accompanying descriptions in both the original language and English, along with other properties described in Table 1. The examples of two tracks with descriptions are shown in Figures 1 and 2.

The GPX files come from 135 unique domain names. Track descriptions come in 11 languages, with the majority being French (802), German (330), and English (110). The original descriptions are between 50 and 1999 characters long (for translated, between 34 and 1866), with a median of 303 characters (for translated, 285.5). The tracks originate in 25 countries, with most popular being France (710), Austria (164), Germany (133), Switzerland (116), and Italy (81). The shortest track is 607 m (1991 ft) while the longest is 99.3 km (61.7 mi). The average track length is 20.5 km (12.7 mi), while the median is 12.1 km (7.5 mi). Histograms of route length (in km) and description length (in characters) are shown in Figure 3.

The tracks in the resulting dataset represent real-life outdoor activities that were either recorded (i.e., completed) or planned using GIS software. Thus, these routes can be used in place of synthetically generated trajectory datasets, as well as for training trajectory generation ML models.

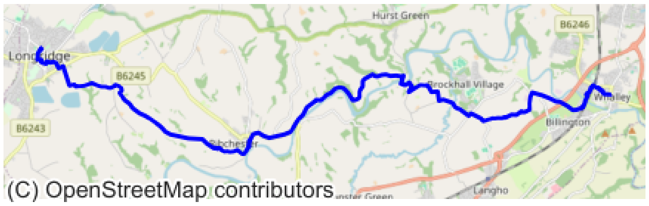

Figure 1: A 18.2 km (11.3 mi) route in the UK. The description reads: *‘Longbridge to Whalley Slowway following part of the Ribble Way. Difficult to find a good crossing of the A59. The crossing chosen crosses the road from footpath to footpath in a place with good visibility. The road junctions/bridges were actually worse as would need to walk along a fast road with no pavement rather than just cross once at right angles. This crossing sets up good sections without roads. Good spacing of waypoints at Old Langho and Ribchester.’*

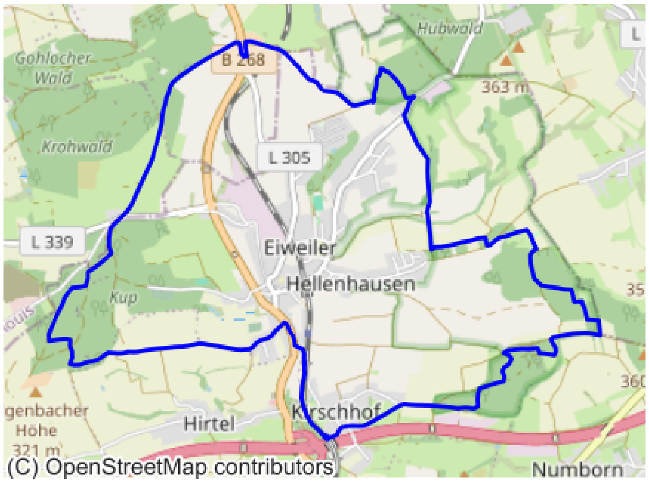

Figure 2: A 13.8 km (8.6 mi) circular route in Germany. The description translated from German reads: *‘The path is very well marked with a black deer beetle (Hootzemann) on white ground. My starting and finishing point was the Schützenhaus Eiweiler near the Großwald brewery.’*

The description-`MultiLineString` pairs can be used to study how people describe their outdoor activities, such as what kinds of landmarks, landscapes, and experiences they choose to mention or ignore, and get an insight into turn-by-turn decisions of users. These pairs can also be used to fine-tune language models to generate human-like descriptions of routes, or to compute location embeddings. Our dataset can become especially powerful when combined with Point of Interest (POI) and/or road segment datasets, such as those from OpenStreetMap[17] or Overture[18].

[12] https://github.com/aboSamoor/pycld2
[13] https://github.com/argosopentech/argos-translate
[14] https://opennmt.net/
[15] https://www.earthdata.nasa.gov/sensors/srtm
[16] https://github.com/tkrajina/srtm.py
[17] https://welcome.openstreetmap.org/
[18] https://overturemaps.org/

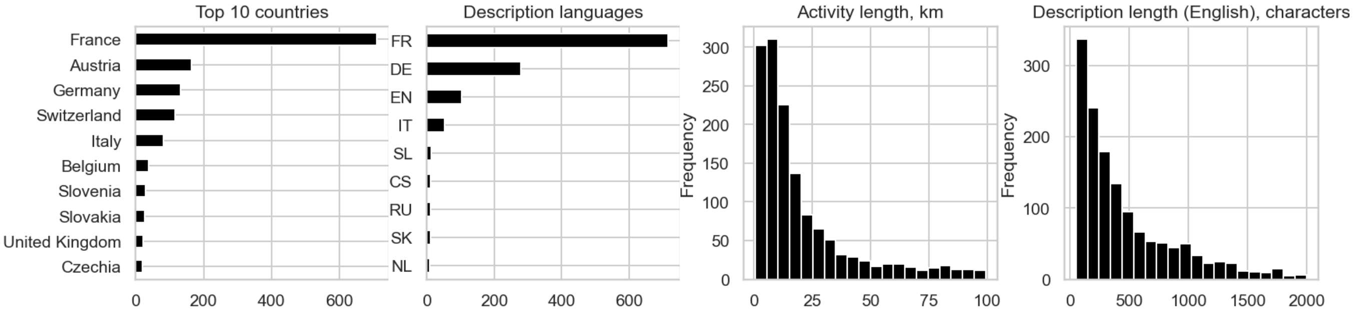

Figure 3: Select dataset properties.

Table 1: Dataset property descriptions

#	Property	Description
1	url	URL of the GPX file
2	warc_file	CC WARC file with GPX file
3	warc_offset	GPX file position in WARC
4	warc_len	GPX file byte length
5	country	Country name as determined by the first point in the track intersecting boundaries from https://www.geoboundaries.org/
6	desc	Original track description
7	desc_lang	Track description language code, as determined by `pycld2`
8	desc_en	Track description translated into English
9	elev_source	*GPS* if elevation is recorded by device; *DEM* if determined later from Shuttle Radar Topography Mission
10	elev_highest	Track's highest point, m
11	elev_lowest	Track's lowest point, m
12	uphill	Cumulative elevation gain, m
13	downhill	Cumulative elevation loss, m
14	length_2d	Track length disregarding elevation, m
15	length_3d	Track length accounting for elevation, m
16	is_circular	*True* if start and end points are within 350 m from each other, *False* otherwise
17	geometry	MultiLineString Z geometry in GPS coordinates: *(lat, lon, elevation)*

4 CONCLUSION

In this paper, we demonstrated that Common Crawl can be used as a novel source of annotated geospatial data. As a case study, we built an efficient pipeline to extract annotated outdoor activity tracks from GPX files that can be executed on a medium-powered laptop in a matter of hours. The resulting multimodal dataset contains textual descriptions of outdoor activities paired with `MultiLineString` vector geospatial data and can be used to study people's outdoor activity habits, and the relationship between what people experience and what they *say* they experience. The dataset can be useful as a fine-tuning set for various GeoAI models, from trajectory to human-like description generators. We recognise that our case study captures just a fraction of all annotated geospatial data available in CC, and we intend to extend the scope in future work.

ACKNOWLEDGMENTS

This work was supported by Ordnance Survey & UKRI Engineering and Physical Sciences Research Council [grant no. EP/Y528651/1].

REFERENCES

[1] Meta AI. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/

[2] Stefan Baack. 2024. Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI. (Feb. 2024).

[3] Asier Gutiérrez-Fandiño, David Pérez-Fernández, Jordi Armengol-Estapé, David Griol, and Zoraida Callejas. 2022. esCorpius: A Massive Spanish Crawling Corpus. http://arxiv.org/abs/2206.15147 arXiv:2206.15147 [cs].

[4] Ilya Ilyankou, Meihui Wang, Stefano Cavazzi, and James Haworth. 2024. Quantifying Geospatial in the Common Crawl Corpus. https://doi.org/10.48550/arXiv.2406.04952 arXiv:2406.04952 [cs].

[5] Armağan Karahanoğlu, Rúben Gouveia, Jasper Reenalda, and Geke Ludden. 2021. How Are Sports-Trackers Used by Runners? Running-Related Data, Personal Goals, and Self-Tracking in Running. *Sensors* 21, 11 (Jan. 2021), 3687. https://doi.org/10.3390/s21113687 Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[6] Alexandra Sasha Luccioni and Joseph D. Viviano. 2021. What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus. http://arxiv.org/abs/2105.02732 arXiv:2105.02732 [cs].

[7] Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the Common Crawl. In *Smith, Jason R; Saint-Amand, Herve; Plamada, Magdalena; Koehn, Philipp; Callison-Burch, Chris; Lopez, Adam (2013). Dirt cheap web-scale parallel text from the Common Crawl. In: 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, August 2013. Association for Computational Linguistics, 1374-1383.* Association for Computational Linguistics, Sofia, Bulgaria, 1374–1383. https://doi.org/10.5167/uzh-80038

[8] Alan D Thompson. 2022. What's in my AI? A Comprehensive Analysis of Datasets Used to Train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher. (2022).

[9] Michał Turski, Tomasz Stanisławek, Karol Kaczmarek, Paweł Dyda, and Filip Graliński. 2023. CCpdf: Building a High Quality Corpus for Visually Rich Documents from Web Crawl Data. In *Document Analysis and Recognition - ICDAR 2023*, Gernot A. Fink, Rajiv Jain, Koichi Kise, and Richard Zanibbi (Eds.). Springer Nature Switzerland, Cham, 348–365. https://doi.org/10.1007/978-3-031-41682-8_22

[10] Maurice Weber, Carlo Siebenschuh, Rory M Butler, Anton Alexandrov, Valdemar R Thanner, Georgios Tsolakis, Haris Jabbar, Ian Foster, Bo Li, and Rick Stevens. 2023. WordScape: a Pipeline to extract multilingual, visually rich Documents with Layout Annotations from Web Crawl Data. (2023).

[11] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. http://arxiv.org/abs/2304.04675 arXiv:2304.04675 [cs].

Received 29 May 2024; accepted 23 August 2024