Geographic Information Retrieval of Early COVID19-Reports

Evaluating a compact GIR process on WHO Situation Reports (20. Jan – 9. Feb 2020)

Maximilian Lengenfelder
GIUZ
University of Zurich, Switzerland
maximilian.lengenfelder@uzh.ch

ABSTRACT

This project implemented a compact Geographic Information Retrieval (GIR) process to extract and map place names from the 20 earliest WHO COVID19 Situation Reports (20 Jan–9 Feb 2020). PDFs were parsed with pdfplumber, a prompt-guided large language model (Google Gemini) was used to identify locations explicitly linked to reported cases and capture simple numeric mentions, and extracted toponyms were geocoded with GeoNames (with an in-run cache to reduce duplicate queries). Outputs comprised a master CSV, a static overview map, a cumulative animation showing mentions over time, and a per-report diagnostic map of a single report used for qualitative validation.

The main finding is that this rather simple approach reliably reconstructs the broad, coarse-scale spatial information reported by WHO. Especially the Wuhan-cluster at the beginning of the pandemic and the progressive spread over international locations have been well identified, making it useful for obtaining rapid situational overviews. Recurring limitations are format-driven errors, ambiguous gazetteer matches, and occasional LLM misinterpretation. These motivate lightweight manual inspection and correction, documentation of prompts and model versions for reproducibility, and further quantitative validation.

KEYWORDS

Geographic Information Retrieval GIR, Geoparsing (toponym recognition & resolution), WHO Situation Reports, COVID19, GeoNames geocoding, Large Language Model LLM extraction

1 Introduction and Motivation

Nowadays, there is as much information to be found online as probably never before (Bruggmann 2017). It comes from different sources and exists in even more forms. A pretty obvious part of information that is freely available on the internet consists of all sorts of social media posts. Then, similar to that, there are lots of blogs, articles and newspapers. But there is also more information consisting of actual knowledge and valuable information present. Wikipedia articles are one well known representative. But also, the amount of scientific literature that can be found and used online has increased over the last decades. This opens many possibilities, as the availability of such a vast amount of information can act as a fundamental foundation for analysing and targeting many questions. While the content and the quality of said information may strongly vary, also influencing what can be done with it, there is a lot of data, that has a geographic component. This can be implicit or explicit. Meaning the geographic component may be hidden, so it is not part of the actual information (post, text, content in general), but can be found and retrieved over detours, like when it is hidden in the metadata/description, a geotag, some sort of gps/coordinate-information that is passively saved through the usage of the internet and the stations for example. Other than that, geographic information can also be present explicitly. This means it is directly present in the content (e.g. Text, paper, report, news article, etc.). in either way, this adds a very strong component that can be used for analysis and answering various questions. But to make use of such information, it is necessary to retrieve it in a first instance. That is, where GIR (Geographic Information Retrieval) comes into play. Basically, it concerns with the (as the name already suggests) retrieval of geographic information, mainly from structured and unstructured text (Jones and Purves 2008, Purves et al. 2018).

In that context, this project looks at the GIR process and how well simple and commonly used methods are able to perform. As foundation, WHO situation reports of the early COVID19 pandemic are used. From the beginning, the WHO daily released a short (a few pages) report on communicating the situation (World Health Organization 2020). Those reports are in English, well-structured and they include the necessary geographic information in the form of places, where cases have been recorded. This means, they contain rich and valuable geographic information, offering a big potential for GIR. The main motivation of this project is, that such information is very valuable to assess the situation and make the right decisions. But as the situation in such crisis are constantly and rapidly evolving, there might not be the time to manually handle all of the available geographic information, let alone in a consistent way. Therefore, GIR could be a big help for extracting the necessary information, so that it can then be directly used to map and analyse the situation from a geographical standpoint: this could include things beginning from simply displaying the current distribution, up to complex analysis where spatial patterns and trend are identified. Even more, if working correctly, those processes could also be automated. But

with the later aspect being a bit more in the future, the first part would be to assess whether, and to what extent GIR can be used for extracting the foundational information.

2 Literature review and research question

Geographic Information Retrieval (GIR) geoparsing provide a practical tool set for extraction spatial information from structured and unstructured texts. Simplified, it happens usually in two stages. First, text elements that mention and consist of places and contain spatial data have to be found and extracted (toponym recognition). Second, those extracted entries have to be geocoded, meaning they have to matched with real world places, so that their coordinates can be added (toponym resolution) (Purves et al. 2018; Jones and Purves 2008). A good review and on overview, summarizing the field and discussing this two-step process, as well as commonly used methods and challenges is provided by Purves et al. (2018), as well as an older article of Jones and Purves (2008).

Toponym recognition is commonly performed with Named Entity Recognition (NER). Ready to use NER systems (for example spaCy's pretrained models) are fast and easy to use, which is why many applied GIR pipelines include them. Whereas transformer-based NER models (BERT-style models) can give better recall on noisy or domain-specific text, but they are heavier and require more resources to run (Purves et al. 2018; Raza and Schwartz 2023). For reproducible and not overly heavy implementations, the trade-off often favours a small, reliable NER with light normalization heuristics rather than a resource-intensive fine-tuned model (Bruggmann 2017, Purves et al. 2018).

The harder part usually is toponym resolution, meaning the mapping of a place name such as "Springfield" or "Hubei" to a unique place identifier and coordinates. Practical systems typically combine a gazetteer lookup (for example GeoNames or OpenStreetMap) with simple ranking heuristics (prefer populated places, prefer matches in the same country context, prefer high-population candidates when ambiguous, etc.). More advanced approaches use learned ranking and contextual embeddings to score candidates, resulting in those methods to improve the accuracy. However, they require more training data and infrastructure. Because of these trade-offs, many applied GIR pipelines use gazetteer + caching strategies, so that ambiguous matches can be reviewed manually if needed (Pérez and Aybar 2024).

Evaluation work in geoparsing is now reasonably well established. Standard metrics often used are precision, recall and F1 for the recognition stage, while resolution is usually evaluated by geodesic distance between predicted and ground-truth coordinates, or by the exact identifier matching (Gritta et al. 2019, Purves et al. 2018). Practical evaluation guides highlight, that evaluation should be task oriented. For mapping tasks, it is often more relevant to inspect, whether the extracted places reproduce the same spatial patterns at the scales of interest (countries, provinces, major cities) instead of looking, if every low-level placename was perfectly resolved. This motivates the recommendation of using a mixed evaluation protocol. Automated counts and geocoding-success rates, plus a focused manual check of at least one representative report (Gritta et al. 2019, Purves et al. 2018).

There are several practical restrictions, that the literature highlights and that directly affect this project. First, place-name ambiguity is omnipresent and should include disambiguation logic in any pipeline (Purves et al. 2018). Second, geoparsing performance is spatially heterogeneous. Systems generally perform better for large, well-known cities and for countries that are well represented in gazetteers, and worse for small settlements or under-represented regions (Liu et al. 2022). Third, the textual format matters. Tabular, list-like, or tightly packed table text (frequent in institutional situation reports) can degrade both NER and resolution performance (Gritta et al. 2019). Fourth, mapping place mentions from reports to epidemiological conclusions requires caution, because mentions in situational reports can describe case locations, travel histories, administrative units, or policy contexts and may not be a direct measure of incidence (Smith and Mennis 2020, Rastogi et al. 2021). Finally, geocoding tools differ in coverage and behaviour, wherefore results can vary depending on the choice of the geocoder (Pérez and Aybar 2024).

Applied work during the COVID-19 pandemic shows how GIR and GIS techniques were used to produce fast situational overviews and dashboards. These efforts underline GIR's practical value for early response and crisis management, while also highlighting the need for cautious validation and explicit statements about what the maps represent (reported locations vs. confirmed infection points) (Smith and Mennis 2020, Rastogi et al. 2021). Related NLP work on clinical texts demonstrates, that entity extraction can be effective when combined with careful post-processing and domain rules (Raza & Schwartz, 2023). (Smith and Mennis 2020, Rastogi et al. 2021, Raza and Schwartz 2023)

Altogether, the literature supports that GIR can provide useful spatial insights for the early COVID19 spread. Expectations set by the literature are logical. Good performance is likely at the country level and for major cities, while smaller towns and table-formatted text will be more error-prone (Purves et al. 2018, Liu et al. 2022). That motivates focusing on the most informative spatial scales, using a cache for geocoding decisions, and validating a representative report by hand (Report 7 in this project).

Based on that, and the initial idea and motivation, the research question is: "To what extent is GIR able to perform for extracting and mapping the early COVID-19-cases locations mentioned in WHO situation reports?"

3 Methods

The basic workflow of this project consists the following: (1) gathering the 20 earliest WHO situation reports, (2) extracting the text from the pdfs, (3) using a LLM like Gemini to find and extract the spatial information and some additional information, (4) using GeoNames to geocode those extracted places, (5) visualising the results, (6) assessing the results. The foundation is

the WHO Situation Reports 1-20, starting from 20. January 2020 and reaching to 9. February 2020 (World Health Organization 2020). Each of those reports contains a few pages with geographic information and placenames. Additionally, they include a temporal element in form of their release date, allowing to showcase a trend over time. Before diving deeper into the methods, it should be mentioned that the initial plan to use SpacyNER for the extraction of the places (toponym recognition) has been aborted. While playing around with AI in order to improve the result and remove irrelevant and wrong places, it showed that the AI itself was able to do this distinction. Moreover, also additional information like the linked COVID19 cases of a location could be identified. Therefore, instead of using the traditional method, AI was automatically implemented in the process. More details on that will be mentioned in the exact step later on.

3.1 Gathering the 20 earliest WHO situation reports

This step was done manually. All the necessary reports were downloaded and saved locally. However, it would also be possible to automate that process, accessing them over URLs for example. But since they don't follow the same concept, it has not worked as easily as it should for gaining an advantage. And manually making a list of all the correct URLs wouldn't a big improvement.

3.2 Extracting the text from the PDFs

Each PDF is opened and the text is extracted with pdfplumber. The text is then stored in memory together with the parsed report date (from the filename, but it would also be possible to get this out of the report) and the file name. Files that cannot be parsed are skipped and a warning is returned.

3.3 Using a LLM like Gemini to find and extract the spatial information and some additional information

Using Google Gemini and a predefined, specific prompt, the desired information gets extracted and returned. The prompt requires several important things. A filter, to only return locations explicitly associated with cases, outbreaks or active transmission. An exclusion, to get rid of locations mentioned for administrative or organizational reasons (meetings, office locations). The context, to include parent context (country or state) for each location to improve geocoding. And include counts of the cases, if present, otherwise use 1 as a placeholder. The response text is then cleaned (remove code fences) and parsed into rows with fields such as name, context, cases, mentions, status. Each extracted row is augmented with the report date and a report index.

Prompt:
You are an expert epidemiological data analyst. Your task is to extract spatial data from the following WHO Situation Report.
CRITICAL INSTRUCTIONS:

1. FILTER:
Only extract locations (cities, provinces, regions, or countries) that are explicitly linked to virus cases, outbreaks, or active transmissions.
2. EXCLUDE:
Do not extract locations that are only mentioned for administrative reasons (e.g., "A meeting was held in Geneva," "The regional office in Cairo announced...").
3. MAPPING DATA:
For every location, provide a "context" (the parent country or state) to ensure 100% accuracy during geocoding.
4. NUMERIC DATA:
If a specific case count is mentioned for that location, extract it as an integer. If no specific number is given, use 1 as a placeholder for the mention.

OUTPUT FORMAT:
Return strictly a JSON list of objects. No prose, no markdown code blocks.

Example:
[{{"name": "Wuhan", "context": "Hubei, China", "cases": 444, "mentions": 3, "status": "new"}}, {{"name": "Tokyo", "context": "Japan", "cases": 1, "mentions": 1, "status": "imported"}}]

LLM extraction is used because it can interpret table-like text and nearby context easier and better, than a simple version of spacyNER, as it has shown while playing around to get a better result (Raza and Schwartz 2023). However, LLMs may also make mistakes, making it important to inspect the result (as it would be necessary with the other methods too).

Because Gemini is a proprietary LLM, these extraction steps should be documented carefully to support reproducibility and validation (Raza and Schwartz 2023, Purves et al. 2018).

3.5 Using GeoNames to geocode extracted places

Each extracted location is geocoded with GeoNames . A simple in-run cache avoids duplicated queries, and the code pauses for 0.3s between requests, to be polite to the API. Avoiding duplicates is important since there is a limit for the service of GeoNames and it would also take unnecessarily longer. Rows without successfully added coordinates are dropped from mapping outputs but remain available for auditing in the master CSV. GeoNames is a widely used open gazetteer and provides consistent country/context matching, especially when context is included (Pérez and Aybar 2024).

3.6 Visualising the results

Use the final CSV file with the extracted and geocoded information from all reports to visualize the results.

Static overview map:
Creating a map, showing all locations with recorded cases. The size and colour of the dots represent the frequency, the places are mentioned.

Cumulative animation:
Create an animation, showing how the places add up over time. If a place is mentioned several times, the dot gets darker with every new mention.

Single-report diagnostics (Report 7):
Creating a map, showing only places from report 7. This will be used for qualitative validation by comparing it to the content of the actual report.

3.7 Assessing the results

Use a qualitative comparison of the actual reports with the result of the GIR process, to assess what worked correctly and where errors occurred.

4 Results

As a result, we get a CSV file, including all the important information, like name, context, cases, date, report index, mentions, status, latitude, longitude (fig. 1). Based on that, we also get several maps. One is a static overview map, showing all the places, where COVID19 cases have been recorded according to the 20 earliest WHO situation reports (fig. 2). They are represented through dots, where the colour and size refer to the frequency a place is mentioned in the reports. In addition, we get an animation, showing the temporal distribution of those locations (fig. 3). It is cumulative, and the more a place was mentioned, the darker the colour of the dot is. Lastly, there is a static map, only including entries from report number 7 (fig. 4). It shows the places with the according placename. This map is meant for validation purposes to compare the result manually with the places found in the text it the respective report.

China	China	2744	2020-01-27	7	3.0	active outbreak	35	105
Wuhan	Hubei, China	1	2020-01-27	7	3.0	epicenter	30.58333	114.26667
Hong Kong SAR	China	8	2020-01-27	7	2.0	confirmed	22.25	114.16667
Taipei	Taiwan, China	4	2020-01-27	7	1.0	confirmed	24	121
Japan	Japan	4	2020-01-27	7	2.0	confirmed	35.68536	139.75309
Republic of Korea	Republic of Korea	4	2020-01-27	7	2.0	confirmed	36.5	127.75
Viet Nam	Viet Nam	2	2020-01-27	7	2.0	local transmission	16.16667	107.83333
Singapore	Singapore	4	2020-01-27	7	2.0	confirmed	1.28967	103.85007
Australia	Australia	4	2020-01-27	7	2.0	confirmed	-25	135
Malaysia	Malaysia	4	2020-01-27	7	1.0	confirmed	2.5	112.5
Thailand	Thailand	5	2020-01-27	7	2.0	confirmed	15.5	101
Nepal	Nepal	1	2020-01-27	7	1.0	confirmed	28	84
United States of America	United States of America	5	2020-01-27	7	3.0	imported	33.45876	-90.15081
Canada	Canada	1	2020-01-27	7	1.0	confirmed	60.10867	-113.64258
France	France	3	2020-01-27	7	1.0	confirmed	46	2
New South Wales	Australia	1	2020-01-27	7	1.0	confirmed	-33.86785	151.20732

Figure 1. Extract of the resulting table including all the extracted and geocodes places and information. (name, context, cases, date, report index, mentions, status, lat, lon).

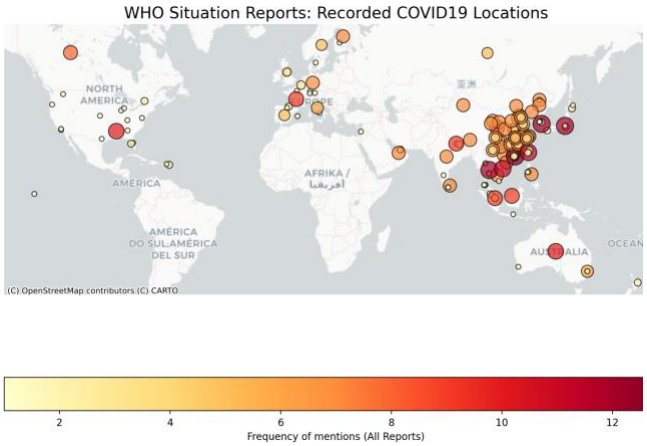

Figure 2. Overview map, showing the places of recorded COVID19 cases mentioned in the 20 earliest WHO situation reports. Colour and size represent the frequency of mentions in the reports.

Figure 3. Animation of the spread of the locations with recorded cases.

Figure 4. Locations of recorded cases in report 7.

5 Discussion

The resulting table (fig. 1) and the maps (fig. 2, 3, 4) provide a clear and summarizing view of the geographic situation described in the WHO Situation Reports 1-20. The static map and the cumulative animation both show the expected early concentration of cases in China (with Wuhan clearly visible as the main cluster) and the stepwise spreading of locations outside China (fig. 2, 3). First in neighbouring Asian countries, then Europe, the Middle East and North America. These large-scale patterns are also represented in the WHO reports. The static overview map nicely shows the hotspots in China, and then later in the USA and central Europe (fig. 2). The animation shows, how the spread increased first in China and Asia, and later, after the first cases have been recorded in the rest of the world, that same trend on increasing spread is visible (fig. 3). This is a good example that such a process can help to obtain fast and valuable information needed for communication, information and decision-making purposes.

However, a more detailed comparison reveals some errors, and with that shows the main, recurring limitations. This is especially noticeable when taking the report 7 as an example and comparing the static map with the information found in the actual report (fig. 4). While most of the locations match, there are some errors. A good example is the inclusion of New South Wales in the static map. When looking at the report, it gets clear that this place has nothing to do with the COVID19 cases. It comes from the resources part in that report (fig. 6). While there have been records in that region, only Australia is mentioned in the text (fig. 5), whereas New South Wales is the source of some cases recorded in that region (fig. 6). This is a very good illustration of how the bigger picture might be correct, but often, some small, but mentionable mistakes are included. Therefore, manual inspection and validation should never be neglected. Overall, the qualitative comparison between the reports and the generated maps support, that a basic GIR process can reconstruct the reported spatial footprint at a coarse scale (countries and major cities). The Report 7 map proved particularly useful as a practical validation instrument, highlighting what worked and what resulted in mistakes. The maps and the process therefore function well as reported-location summaries for situational awareness and obtaining a first impression of the current situation.

WHO Regional Office	Country/Territory/Area	Confirmed Cases
Western Pacific	**China***	**2761**
	Japan	4
	Republic of Korea	4
	Viet Nam	2
	Singapore	4
	Australia	4
	Malaysia	4
South-East Asia	Thailand	5
	Nepal	1
Region of the Americas	United States of America	5
	Canada	1
European Region	France	3
Total Confirmed cases	Total	2,798

*Confirmed cases in China include cases confirmed in Hong Kong SAR (8 confirmed cases), Macau SAR (5 confirmed cases) and Taipei (4 confirmed cases).

Figure 5. Table from actual report 7. Countries, territories or areas with reported, confirmed cases of COVID19, 27 January 2020.

Figure 6. Screenshot showing a part from the resources of report 7, causing a wrong entry, not directly mentioned in the report.

6 Conclusion

This project shows that a compact, reproducible GIR process, combining context-aware extraction with an LLM, GeoNames geocoding and straightforward visualization, can effectively show the main spatial patterns and the dispersion, reported in early WHO Situation Reports. Direct comparison with the original reports demonstrates good agreement at coarse spatial scales. The Wuhan/China cluster and the later international mentions are consistently reflected in both, the maps as well as the reports. Thus, in answer to the research question, GIR is able on a useful, coarse-scale extent, to extract and map the early COVID19 locations reported by WHO, and the derived visualisations are valuable for decision and communicative purposes.

At the same time, the comparison of the results to the actual reports (especially report 7) highlights, that these maps must be used and interpreted with caution. They can include a smaller number of mistakes, mainly resulting from irrelevant information. This shows, that while the extraction with LLM has worked surprisingly well, it still generates some false entries. To increase robustness and practical value, you would have to tackle those limitations. Manual inspection could be used in a last step, and besides that, a manual check should in anyway be included before trusting such results. In addition, you could also combine several Methods to increase the accuracy and assess, in which context which method works best. Also, a quantitative validation should be included in addition to a qualitative one. This would be especially important, if don't just want to show coarse tends.

Concluding, it can be said that a rather simple GIR process was able to extract and map the spatial information in the COVID19 WHO situation reports. But while the bigger image looked good, a more detailed inspection revealed the errors and the limitations of

the used methods. However, since dealing with such information could be very helpful in the future and in the management during other crisis, further studies should focus on refining the methods and not only assessing and comparing different approaches on an academic level but also providing specific guidelines and maybe applications that can be used in the future.

REFERENCES

[1] Bruggmann, A. 2017. Visualization and interactive exploration of spatio-temporal and thematic information in digital text archives. Ph.D. Dissertation. University of Zurich. Zurich Open Repository and Archive. https://doi.org/10.5167/uzh-148756

[2] Gritta, M., Pilehvar, M. T., and Collier, N. 2019. A Pragmatic Guide to Geoparsing Evaluation. Language Resources and Evaluation. Preprint on arXiv.

[3] Jones, C. B., and Purves, R. S. 2008. Geographical information retrieval. International Journal of Geographical Information Science 22, 3 (2008), 219–228. https://doi.org/10.1080/13658810701626343

[4] Liu, Z., Janowicz, K., Cai, L., Zhu, R., Mai, G., and Shi, M. 2022. Geoparsing: Solved or Biased? An Evaluation of Geographic Biases in Geoparsing. AGILE GIScience Series 3 (2022), Article 9. https://doi.org/10.5194/agile-giss-3-9-2022

[5] Liu, Z., Janowicz, K., Cai, L., Zhu, R., Mai, G., and Shi, M. 2022. Geoparsing: Solved or Biased? An Evaluation of Geographic Biases in Geoparsing. AGILE GIScience Series 3 (2022), Article 9. https://doi.org/10.5194/agile-giss-3-9-2022

[6] Pérez, V., and Aybar, C. 2024. Challenges in geocoding: An analysis of R packages and web scraping approaches. ISPRS International Journal of Geo-Information 13, 6 (2024), Article 170. https://doi.org/10.3390/ijgi13060170

[7] Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., and Murdock, V. 2018. Geographic information retrieval: Progress and challenges in spatial search of text. Foundations and Trends® in Information Retrieval 12, 2–3 (2018), 164–318. https://doi.org/10.1561/1500000034

[8] Rastogi, A., Padhi, A., Syed, S., Keshan, P., and Gupta, E. 2021. Mapping the footprints of COVID-19 pandemic. Journal of Family Medicine and Primary Care 10, 7 (2021), 2467–2476. https://doi.org/10.4103/jfmpc.jfmpc_2105_20

[9] Raza, S., and Schwartz, B. 2023. Entity and relation extraction from clinical case reports of COVID-19: A natural language processing approach. BMC Medical Informatics and Decision Making 23, 1 (2023), Article 20. https://doi.org/10.1186/s12911-023-02117-3

[10] Smith, C. D., and Mennis, J. 2020. Incorporating geographic information science and technology in response to the COVID-19 pandemic. Preventing Chronic Disease 17 (2020), E58. https://doi.org/10.5888/pcd17.200246

[11] World Health Organization. 2020. Coronavirus disease (COVID-19) Situation Reports. World Health Organization. Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports