ASSOCIATION OF CANADIAN MAP LIBRARIES AND ARCHIVES

BULLETIN

Geospatial Data and Software Reviews

Meg Miller
University of Manitoba

2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository: *Johns Hopkins University Center for Systems Science and Engineering*

Introduction

The following will examine the ***2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository*** compiled by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). This data is openly available to the public for educational and academic research purposes from JHU CSSE's GitHub (https://github.com/CSSEGISandData/COVID-19) and is the basis for the popular data dashboard tracking global cases of the Novel Coronavirus seen below in Figure 1.

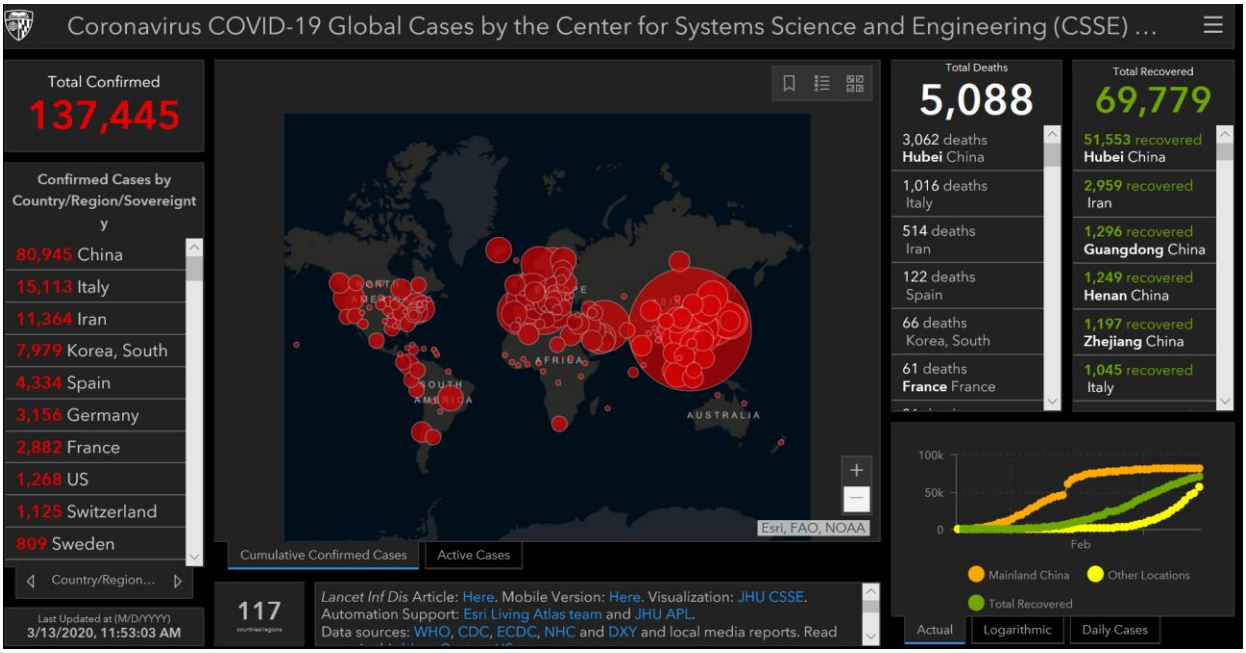

Figure 1: Johns Hopkins Coronavirus dashboard

Background

Dong, Du and Gardner summarized the history of this repository in their correspondence with The Lancet published online in February 2020. What started out as a manual data collection effort using

Google Sheets, became unsustainable as the pandemic unfolded and the team switched to a semi-automated method of data collection pulling from an aggregation platform (DXY) created by Chinese medical professionals, and manually supplementing and cross-checking with other open government resources from various countries. The repo is now accessible through GitHub (https://github.com/CSSEGISandData/COVID-19).

Data details

The repository is composed of three main folders of .csv files and supporting documentation:

- *archived_data*- contains older datasets which have inconsistencies in certain attributes such as time zone and update frequency
- *csse_covid_19_data*- contains two aggregate datasets created by Johns Hopkins with new files uploaded each day at 23:59 UTC. Attributes included are: Province/State/City/Other (Other delineates events such as the passengers coming from the Diamond Princess cruise ship), Country/Region, Last Update, Confirmed, Deaths, Recovered, Latitude, Longitude.
    - *csse_covid_19_daily_reports*- contains a report for each day with confirmed cases, deaths, and recovered numbers. Chinese and Canadian data is broken down at a provincial level (see figure 2 below), American data at a state level, all other data is at captured at a country level.

	A	B	C	D	E	F	G
1	UID	Province/State	Country/Region	Lat	Long	NumberConfirmed	
2	1	Anhui	Mainland China	31.82571	117.2264	830	
3	2	Beijing	Mainland China	40.18238	116.4142	337	
4	3	Chongqing	Mainland China	30.05718	107.874	468	
5	4	Fujian	Mainland China	26.07783	117.9895	261	
6	5	Gansu	Mainland China	36.0611	103.8343	83	
7	6	Guangdong	Mainland China	23.33841	113.422	1151	
8	7	Guangxi	Mainland China	23.82908	108.7881	210	
9	8	Guizhou	Mainland China	26.81536	106.8748	109	
10	9	Hainan	Mainland China	19.19673	109.7455	136	
11	10	Hebei	Mainland China	38.0428	114.5149	218	
12	11	Heilongjiang	Mainland China	47.862	127.7622	331	
13	12	Henan	Mainland China	33.88202	113.614	1073	
14	13	Hubei	Mainland China	30.97564	112.2707	29631	

Figure 2 Example of data from daily report dataset.

    - *csse_covid_19_time_series*- contains three reports (confirmed, deaths, recovery) with a new entry being added to each of these reports each day.

- *who_covid_19_situation_reports*- contains situational pdf reports and data created by the World Health Organization. Containing documentation discusses why the CSSE data is more accurate and current than that of the WHO.

Robust documentation is provided to users in plain language explaining field types, definitions and any ambiguity surrounding them (i.e.: what does confirmed mean from a reporting standpoint), data modification records and links to the visual dashboard as well as references to all of the primary data sources.

Users

At the University of Manitoba, the programs who have displayed the most interest in proactively seeking out this dataset are users from Health Science and Environmental Science. Both graduate and undergraduates have been interested in using it in their final projects where they can pick their own topics or as a dataset to teach themselves a visualization tool during the lab drop-in hours we run. The faculty who were looking to access the data were attempting this when the pandemic was in its early days before the data was posted to GitHub. They were looking for support in finding and cleaning it so that it could easily integrate into their teaching materials.

Library usage

Outside of discovery and access requests three themes of instruction have been offered using this dataset. All sessions were well attended by a cross-section of programs, skill levels and professional levels (students, faculty and staff). Duration ranged from 50-120mins.

- Data Integration
- Online Data Visualization Considerations/ Web Mapping
- Introduction to Software (QGIS, ArcGIS Online)

Data Integration is centered on the idea that in the real world a researcher would be combining a variety of data from different sources to support their own research data. In the session a report from the *csse_covid_19_daily_reports* dataset is selected to act as a stand-in for 'researcher data'. Using the GitHub pages we explore the documentation and discuss things such as field types, naming conventions, data modification, versioning and metadata. OpenRefine is used to do some basic data cleanup tasks. Changes that would need to be made to the file to join it to a country boundaries file in GIS software are also discussed. Moving forward this session could be team taught with the RDM Librarian or Digital Archivist.

The Web Mapping session acts as an introduction to data visualization considerations for the web, and has also been done as a brown bag at the Health Science Campus. Discussion is held around best practices in data visualization (colour, cognitive load, accessibility concerns etc.). In the second half of the session the *csse_covid_19_daily_reports* data is loaded into ArcGIS Online and used to create a map depicting confirmed cases for a colourblind user, participants chose what visualization methods they wish to use (proportional symbology, choropleth etc). A future session has been requested to do the same but using Leaflet.

The final way this dataset has been used in University of Manitoba Libraries is as the dataset in Introduction to Software 'X' sessions. So far it has been run using ArcGIS Online, and QGIS with a request for a version done in R. The differences in raster, vector and tabular datasets are explored

using Natural Earth Data and the *csse_covid_19_time_series* data for confirmed cases. The Coronavirus data is added as a table, and joined to the Natural Earth country boundary file. Selecting by attribute allows us to highlight all the cases that are associated with the Diamond Princess Cruise ship. The output is a choropleth map where we look at how different types of classification can impact the reader's perception of a situation as seen in Figure 3. Discussion is also had about importance of taking population into consideration when visualizing epidemiological data; and how easy it is to make accidental population distribution maps.

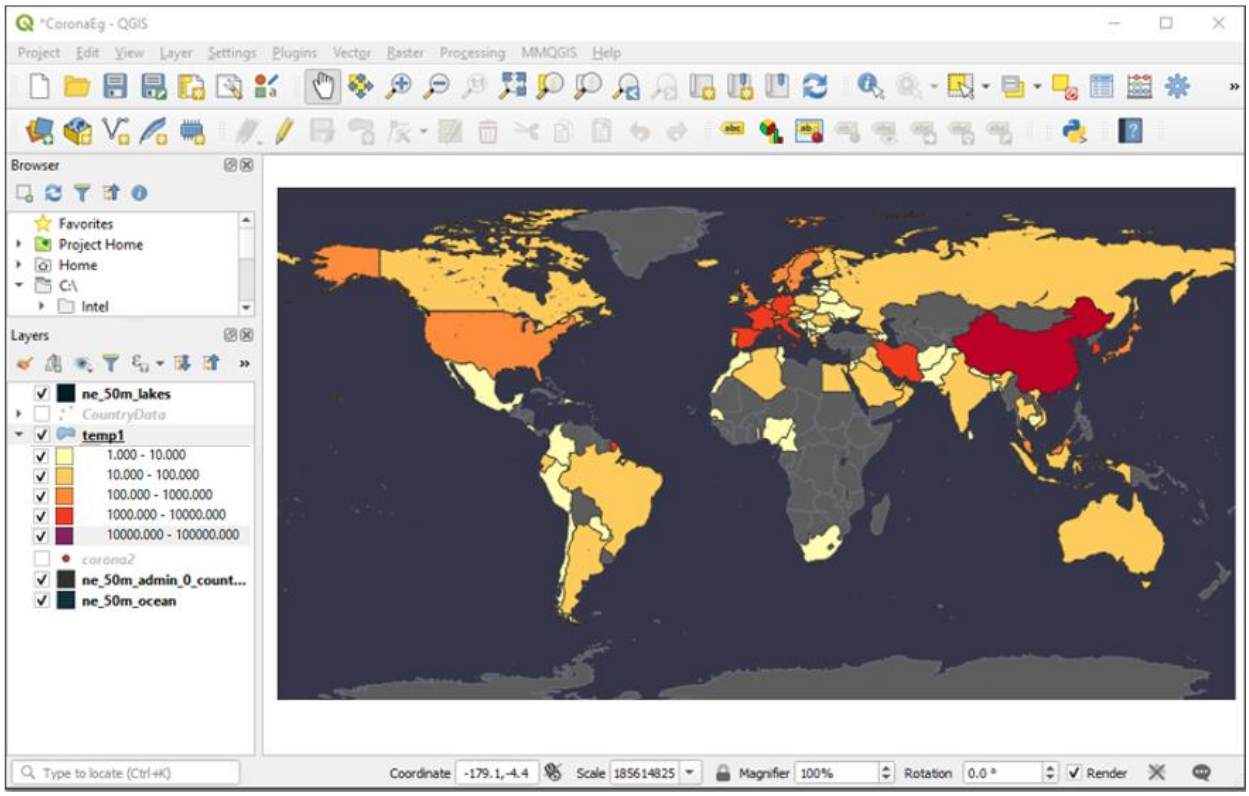

Figure 3: Mapping confirmed cases of coronavirus using QGIS

These sessions are the ones that get the most feedback about the usefulness of using a current dataset that users can engage with.

For the two software related workshops it was necessary to do some initial data cleanup to the files in terms of field names, removing unnecessary columns, and country name disambiguation. As time has gone on the creators have been working to get their data into a more easily usable state, so there is much more consistency (and all changes are documented).

Conclusion

The *2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository* compiled by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) is of very high quality when it comes to aggregated open data sets. From its beginning as a Google/ Kaggle Sheet to the repo's current form in GitHub, the creators have evolved along with the public health issue

making use of a stable platform, clearly documenting their choices and changes in plain language, and working to make the dataset as clean and accurate as possible. All of these things make it a very useful tool for teaching in both theory and practical applications.

Note: For some context I am a single person offering support for Data Visualization services, my position is new so I had no learning materials to start with. A current event is something many people from different areas of campus can engage with. Hopefully by discussing the ways which

this dataset was made use of allows for an easier entry point in adopting a new resource.

Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*; published online Feb 19. https://doi.org/10.1016/S1473-3099(20)30120-1

Meg Miller *is the GIS & Data Visualization Librarian at the University of Manitoba. In her role she assists the campus community in communicating their research through mapping and other data visualization methods. Her research interests involve exploring how non-traditional users of GIS make use of and learn the technology as well as navigate the interdisciplinary relationships and skillsets associated with their projects.*