Exploring GitHub Data for Geospatial Government Research: Understanding the Limitations of Inadequate Quality Control

JAYDEEP MISTRY
Department of Geography
University of Waterloo
jaydeep.mistry@uwaterloo.ca

ABSTRACT

GitHub is an online platform that allows for open collaboration between government members and public contributors. Over the past few years, there was an increase in the number of governments who were adopting the use of GitHub to host their own software projects publically because of the nature of GitHub being friendly to open source projects. This brings a need to research how governments are using the platform, for what projects, and how their use differs spatially. To perform geospatial government research of their use of GitHub, there is need for the data to be complete and not missing information. It is found that the data that is automatically generated from the GitHub platform tends to be complete and accurate, while the voluntarily provided data by the governments is often missing some information that is geospatial or contextual.

1. Introduction

Over the past decade, there have been significant advancements in the realm of Open Data (Janssen et al., 2012), spatial analysis (Anselin, 2012), and collaboration (Palomino et al., 2017) for solving issues in various fields such as policy planning (Taeihagh, 2017), ecology (Steiniger & Geoffrey, 2009), web mapping (Neset et al., 2016), CyberGIS (Wang, et al., 2013), and more. Research has been conducted on the use case of existing tools, or the examination of emerging tools in the industry that can support multi-scale, multi-temporal, and multi-dimensional geospatial data management or analysis (Palomino et al., 2017). Although working together on a software project is not immediately taken into consideration for being a geospatial and temporal problem, any team developing a software solution needs to face these issues if they want to work together on a project without being bound by any team member's location or time of day. Open collaboration has brought new platforms that can allow collaboration from members inside and outside an organization to develop software that can be shared from the web (Mergel, 2015). However, due to barriers in individual expertise of software use, or organization-level adoption of the platform from IT constraints, only select platforms are adopted by governments (Longo & Kelly, 2016).

A platform that was adopted by governments in the recent years is called GitHub. It is a web-based and version control software project repository hosting service that allows users within and outside an organization to work together on projects, review changes, comment on issues, and more. Although it is possible to make GitHub accounts and projects be private and only visible to approved users, it has been mostly used to host open-source

projects where all of the data is copyrighted under a public license, but anyone can contribute their changes to the project or use the code themselves. Due to it being very friendly to open-source projects, it has become a very useful and powerful tool for governments to use because it allows them to work together on projects while having it be accessible to the public, and still control who gets to make changes (Longo & Kelly, 2016).

With the rise in the adoption of GitHub by governments for government related uses, there is a need to research how those governments are using it, for what projects, and if their use of the platform is differs between governments of different regions; i.e. its use in North America versus Europe. Although there have been previous studies that have tried to analyze GitHub use by governments, there have only been some that have tried to analyze the quantitative data available from GitHub. Thus, the research goal of this paper is to use the GitHub data to explore how many governments are using the platform, and also analyze its completeness of spatial and contextual information for use in future Geospatial Research. It will do so by answering the following research objectives:

1. How many governments are using GitHub accounts?
2. How many government GitHub accounts are geo-locatable?
3. What is the completeness of the contextual government GitHub data?

2. Methods

All of the data was gathered using Python libraries in a Jupyter Notebook. Figure 1 illustrates the process of gathering the data. The first step involved using GitHub's Rest API to make hundreds of web request for data on specific GitHub accounts which are listed as official government accounts on GitHub's own website (government.github.com/community). As the GitHub API would answer the request with the data on the GitHub accounts, they were stored as a table using a Python library called Pandas. For each account, there was a field which listed the geographic location of where the account was in the world. Since the location data on these accounts was just in plain text, it had to be geocoded, meaning that it has to be converted to a latitude/longitude pair which could be placed accurately on a world map. Using Google Maps API to geocode each account, the geospatial dataset was ready and stored into a Microsoft Excel for further data visualization.

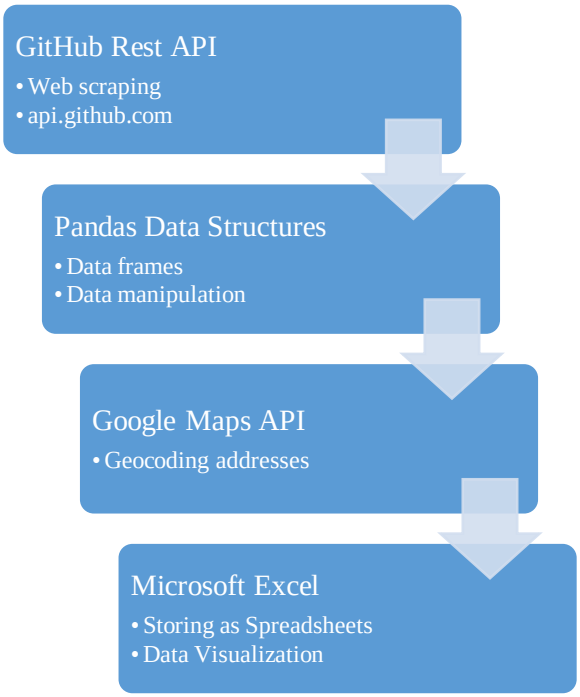

Figure 1: Spatial Data Gathering Workflow

3. Results

3.1 Research Objective 1

The first objective is to see how many governments are using GitHub accounts. After web scraping all 770 GitHub accounts

listed as official government accounts on GitHub's own webpage, they were plotted based on their date of creation. Figure 2 shows the plot of those accounts from the early days of using GitHub in 2009, and up to the end of 2018. Since the creation of GitHub, its adoption in governments was increasing year over year until 2014 where that increase plateaued. It is also important to note that a single government organization could own multiple GitHub accounts.

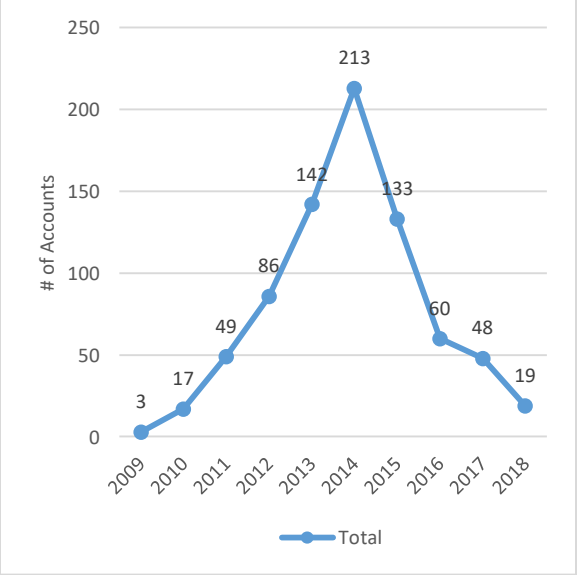

Figure 2: Number of GitHub Accounts by Year of Creation

However, ownership of the GitHub accounts to their real world government organization is often not listed and rather implied by the name of the account. For example the account @thecityofcalgary is owned by the City of Calgary in Canada, but the @web-boew account is owned by the federal Government of Canada.

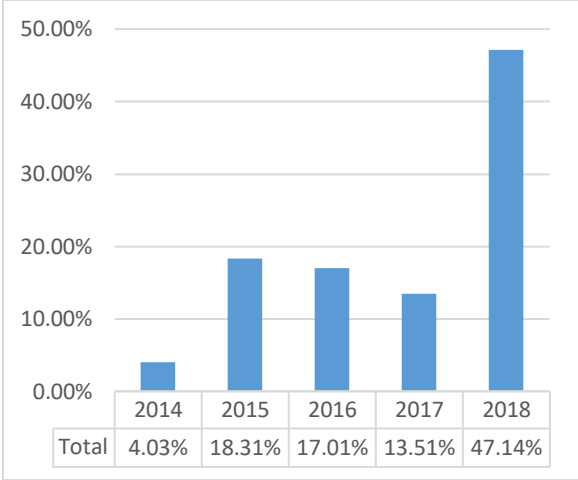

Figure 3: Percentage of Government Accounts by Year of Update

Figure 3 shows that at least 60% of the accounts have been updated since the beginning of 2017, whereas a sum of 40% of the accounts have not been updated since 2016. Although an account could have been created a while ago, it is possible that it could have genuinely not needed to be updated in any way, thus the date of creation and update are not the best indicators of account activity. It is possible to look at the performance of individual repositories of each government account, but that would require analyzing over 27,000 repositories which is beyond the scope of this paper.

3.2 Research Objective 2

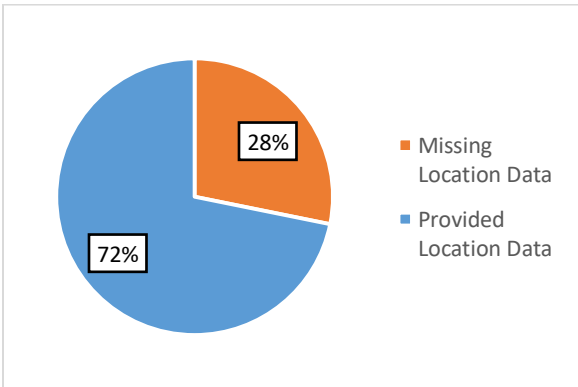

Figure 4: Proportion of Accounts with Location Data

The second objective was to determine how many government accounts are geo-

locatable. Figure 4 illustrates that 28% (217) of the government GitHub accounts were completely missing location data. The only way to tell what country those accounts belong to would be from further web-scraping the GitHub webpage which lists these official government accounts and recording what country the account was listed under.

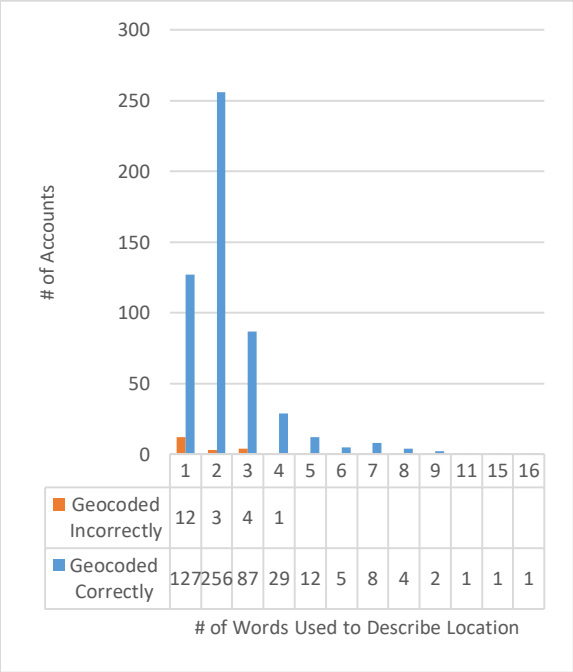

Figure 5: Number of Accounts Geocoded to Correct Country

One of the biggest issues with the location data, other than being empty, is that some are geo-located to the incorrect country because their GitHub data only listed a few word which were generic enough to be places in other countries. Figure 5 shows that there were 12 accounts who were geocoded incorrectly because there was only one word used to describe their location. An example is the account for the Canterbury Regional Council which only used the word 'Canterbury' in the location field. Since there are various places called Canterbury across the world, the Google Maps API geocoded the account to a place in the United States instead of New Zealand, which is where the account is actually from.

3.3 Research Objective 3

The third objective was to determine the completeness of the contextual information of the government GitHub data. For this paper, the contextual information being assessed was limited to four aspects of the government GitHub data available about the organization: name, description, email, and location. Figure 6 illustrates that only about 32% of the accounts have given all 4 of the contextual information items. Almost 40% of the accounts are completely missing at least one item, and over 28% are missing more than one item.

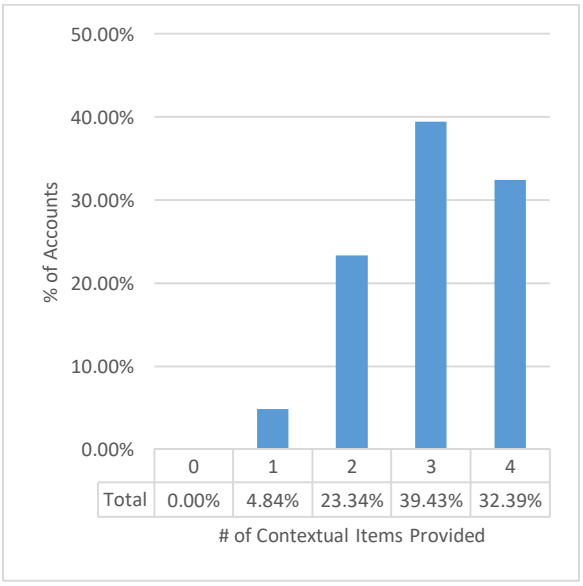

Figure 6: Percentage of Accounts by the Amounts of Contextual Information Provided

4. Conclusion

In conclusion, only after having adequate information is it possible to use the government GitHub account data for geospatial research that might investigate who, when, and where these accounts are coming from. Data such as the date of account creation and last update are accurate to perform analysis because they are automatically recorded by the platform as the changes happened. However, data

such as the contextual information that the governments can voluntarily add to their accounts is often missing some information, for example the geographic location of the organization.

There is a need for better quality control of the voluntarily provided government data because it is often incomplete or lacking some parts which should not be the case for a public facing government resource that citizens of their community may want to view or interact with. Having complete geospatial and contextual GitHub data on these government accounts can not only benefit the public, but also allow for future geospatial government research into various fields of GIScience, open collaboration, open source software, and much more.

Acknowledgements

I would like to acknowledge my graduate supervisor Dr. Peter A. Johnson for supporting me in my own Masters studies. I would also like to thank him for funding me through scholarships awarded by the Social Sciences and Humanities Research Council of Canada (SSHRC).

References

Anselin, L. (2012). Anselin, L. (2012). From SpaceStat to CyberGIS: Twenty years of spatial data analysis software. International Regional Science Review, 131-157.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information systems management, 258-268.

Longo, J., & Kelly, T. M. (2016). GitHub use in public administration in Canada: Early experience with a new collaboration tool. Canadian Public Administration, 598-623.

Mergel, I. (2015). Open collaboration in the public sector: The case of social coding on GitHub. Government Information Quarterly, 32(4), 464-472.

Neset, T. S., Opach, T., Lion, P., Lilja, A., & Johansson, J. (2016). Map-based web tools supporting climate change adaptation. The Professional Geographer, 103-114.

Palomino, J., Muellerklein, O. C., & Kelly, M. (2017). A review of the emergent ecosystem of collaborative geospatial tools for addressing environmental challenges. Computers, Environment and Urban Systems, 79-92.

Steiniger, S., & Geoffrey, H. J. (2009). Free and open source geographic information tools for landscape ecology. Ecological Informatics, 183-195.

Taeihagh, A. (2017). Taeihagh, A. (2017). Crowdsourcing: a new tool for policy-making? Policy Sciences, 629-647.

Wang, S., Anselin, L., Bhaduri, B., Crosby, C., Goodchild, M., Liu, Y., & Nyerges, T. L. (2013). CyberGIS software: a synthetic review and integration roadmap. International Journal of Geographical Information Science, 2122-2145.