Is VGI Big Data?

Peter Mooney and Adam C. Winstanley
Department of Computer Science,
Maynooth University,
Co. Kildare, Ireland.

Summary (100 words)

Volunteered Geographic Information (VGI) has become a popular source of geographic data for GIS practitioners in recent years. VGI datasets are characterised as being: large in volume, subject to dynamic changes and updates, collected through crowdsourcing architectures using a variety of devices and technologies and contain a mixture of structured and unstructured information. Can we call VGI a form of Big Data? Are VGI datasets developing characteristics that make processing them using traditional data processing applications and techniques difficult and unsatisfactory? We explore this question with reference to a number of sources of VGI.

KEYWORDS: (5) VGI, Big Data, GI Data Processing, Crowdsourcing

INTRODUCTION

Volunteered Geographic Information (VGI) continues to gain research attention (Mooney and Corcoran, 2014). VGI is a special case of user-generated content (UGC), usually having an explicit or implicit embedded spatial component. The large quantities and diverse information generated as VGI presents a number of challenges for developing methodologies to use it in research, applications and for understanding its societal implications (Elwood et al, 2012). In this abstract we explore the question "Is VGI Big Data?" by presenting an overview of three popular sources of VGI. Does carrying out research with VGI require GIScientists and practitioners to equip themselves with a new set of tools, skills and methodologies capable of extracting knowledge from these very large dynamic datasets as defined as Big Data? Clear definitions of what exactly Big Data is are very difficult to find (Goodchild, 2013). Kitchin and Laurialt (2014) state that new forms of Big Data are produced predominantly through new information and communication technologies (ICTs). Prior to 2008 data were rarely considered in terms of being 'small' or 'big' (Kitchin and Laurialt, 2014). All data were, in effect, what is now sometimes referred to as 'small data' regardless of their volume. Big Data is associated with data from sensors and software that digitize and store a broad spectrum of social, economic, political, and environmental patterns and processes. Miller and Goodchild (2014) write that geographical Big Data is produced by in-situ sensors carried by individuals in phones, attached to vehicles, embedded in sensing infrastructure and georeferenced social media. Boyd and Crawford (2012:663) argue that "there is little doubt that the quantities of data now available are often quite large, but that is not the defining characteristic of this new data ecosystem".

Using VGI for GIS research and application development has been growing in popularity over the past number of years. Cinnamon and Schuurman (2013) state three principal methods by which VGI is collected and generated: (1) by using geo-aware mobile devices, (2) annotating geographic features using geoweb mapping interfaces, and (3) by extracting or inferring location information from ambient geospatial data in social media (photos, videos, blog posts, tweets, etc) (Stefanidis et al, 2013). Ambient geospatial data, as opposed to VGI generated by (1) or (2) are often messy, consisting of data that are unstructured, collected with no quality control and frequently accompanied by no documentation or metadata (Miller and Goodchild, 2014).

CHARACTERISING VGI AS BIG DATA

We use OpenStreetMap, geolocated Twitter 'tweet' datasets and Foursquare Venue data as our three examples of VGI. These three sources are available openly and for free and have been used widely by researchers over the past number of years. We shall apply characterisations from Kitchin (2014) and Boyd and Crawford (2012) to assess these sources of VGI.

Kitchin (2014) characterises Big Data as being:

- **Voluminous:** consisting of terabytes or petabytes of data
- **High Velocity:** being created in or near real-time
- **Varied:** Being structured and unstructured in nature
- **Exhaustive:** In scope - striving to capture entire populations or systems
- **High Resolution:** Fine-grained and aiming to be as detailed as possible
- **Relational:** Containing common fields enabling the joining of different datasets
- **Flexible:** Having traits of being easily extended (adding new fields) and scalability (expand in size rapidly)

Boyd and Crawford (2012) characterises Big Data as having

- **Technology Requirements:** maximizing computation power and algorithmic accuracy to gather, analyze, link, and compare large data sets.
- **Analysis Possibilities**: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims.
- **Mythology:** the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy.

K denote characterisation from Kitchin (2014) and BC denotes from Boyd and Crawford (2012).

Dataset Characteristic	**OpenStreetMap**	**Foursquare**	**Twitter**
K:Voluminous	Whole world 36GB compressed XML (~500 GB Uncompressed). City and regional areas much smaller (several GB). A binary compressed format is also available. Difficult	50 million users have generated 6 billion "checkins". Approximately 60 million Foursquare Venues worldwide. Almost 2 million business listed	500 million tweets per day and around 200 billion tweets per year. Estimates vary from 1% to 5% on the quantity of tweets with explicit geographical information (geolocated). Estimates

	to estimate daily volume. Several hundreds of thousands polygons added per day		indicate about 12GB per day in tweet text (not including other message overheads)
K:High Velocity	Changes and edits are reflected quickly in OSM. Many services provide hourly updated downloads. Several APIs are available	Very high velocity. However to access ‘checkin’ data there are rate limits imposed which limit the number of API requests which can be processed in an hour by an application.	Very high velocity. Approximately 800 per second. Special access requirements for this stream of Tweets. There are rate limits imposed for free usage of API and Streaming
K:Varied	Variation in how tagging rules are implemented. Tags can contain structured and unstructured data	Data is returned from API calls in JSON format. This provides a robust data structure.	140 characters per tweet which includes multilingual free text, URLs, hashtags, twitter handles etc.
K:Exhaustive	Yes - data model flexibility allows any geographical feature to be included in OSM	Foursquare database of venues grows as businesses, venue owners and users add to the database.	Provides a communication medium for people. In this sense Twitter looks to connect a very large percentage of the world’s population.
K:High Resolution	Resolution can vary over features, device types used to capture the VGI, etc	Fine grained locational information is attached to venues (geographical coordinates and addresses)	Not directly relevant to datasets of Tweets. Very high temporal resolution. Due to small percentage of geolocated Tweets spatial resolution is difficult to quantify
K:Relational	Not directly - but mapping to other datasets have been performed	Yes - properties of the JSON responses can be linked to other datasets	Dependent upon the contents of the Tweets themselves.
K:Flexible	OSM’s flexibility, and also a cause of some QA/QC issues, is a flexible tagging/attribution model in combination with a reasonably simple data model.	Unclear. The data model appears to be fixed at present.	Twitter is open text limited to 140 characters.

BC: Technology Requirements:	Processing the entire OSM DB requires computing power and resources beyond that available on a desktop. Regions and subsets can be processed on standard desktop machines	Accessing Foursquare's API or Streaming processes requires programming and software knowledge. Storage of Foursquare data is not overly cumbersome. Rate limits means download of data may need to be spaced over a long time period.	Accessing Twitter's API or Streaming processes requires programming and software knowledge. Storage of Twitter data is not overly cumbersome. Analysis will require advanced string-based data mining algorithms.
BC: Analysis Possibilities	Analysis possibilities are beginning to emerge. Recent interest amongst research in the social construction of OSM on a regional and global basis.	Very wide range of possibilities as the Foursquare data can combine user movement patterns between venues over time. Venue data contains metadata about the venue itself. This offers great analysis possibilities	Large number of analysis studies have been produced. VGI type analysis is restricted by the low rate of geolocation in Tweets. The ability to link Tweets to user-profile and location offers significant analysis possibilities for researchers
BC: Mythology:	There is a belief amongst many OSM users that this VGI dataset could yield some very interesting social patterns and knowledge about the digital divide, socio-demographics online and spatial cognition	Some researchers suggest that Foursquare users use the service to document their social movement history and find venue information. This could provide a vast history of human movement and social patterns.	The entire dataset of Tweets has been called the largest dataset on human interaction ever created. A lucrative industry has emerged as being Twitter Content Partners involving the reselling, curation and analysis and business insight extraction of Twitter data for commercial partners.

CONCLUSIONS

In this abstract we have addressed the question 'Is VGI Big Data?'. Our analysis only considers VGI which can be accessed freely and openly. The three examples of VGI presented exhibit many of the characteristics of Kitchin and Boyd and Crawford's Big Data. These characteristics will exert different influences depending on the types of analysis or applications these VGI data are being used for. For example collecting one month of Twitter or Foursquare data for London will not present storage problems but significant computational resources may be required if highly complex spatial data mining algorithms are applied. Similarly this applies to OSM data which can be accommodated in any standard spatial database. However the type of analysis performed will greatly influence the resource requirements. For example Gao et al (2014) build a high-performance cloud-computing Hadoop-based geoprocessing MapReduce platform to facilitate gazetteer development using OSM and other georeferenced social media. This platform is implemented not because of characteristics of the datasets used but rather to speed-up the computation being performed.

Researchers accessing these sources of VGI are not necessarily working with 'Big Data'. The original creation of these data (in particular Foursquare and Twitter) and the potentially highly complex tools and skills required to analyse them exhibit Big Data characteristics. Many papers published on VGI, to date, have used VGI datasets for a specific set of locations over a specific time period. Few, if any, researchers are analysing VGI as it is produced in real-time. Instead the VGI is collected then analysed in the same way as the 'small-scale' studies of the past (Kitchin and Laurialt, 2014). We are presently witnessing a fast changing landscape with respect to geographical data. The types of data flows we are seeing from VGI and UGC are part of this changing landscape. Ubiquitous, ongoing, data flows are important because they allow us to capture spatio-temporal dynamics directly and at multiple scales (Mooney and Corcoran, 2014; Miller and Goodchild, 2014). Graham and Shelton (2013:255) write that while there has been significant discourse surrounding Big Data "there has yet to be a significant, sustained effort to understand its geographic relevance".

BIOGRAPHY

Peter Mooney is a Research Fellow with the Irish Environmental Protection Agency and the Department of Computer Science at Maynooth University. Adam Winstanley is a senior lecturer and Head of the Department of Computer Science at Maynooth University.

REFERENCES

Boyd, D. and Crawford, K. (2012) Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication & Society*. 15(5), 662-679. DOI: 10.1080/1369118X.2012.678878

Graham, Mark, and Taylor Shelton. 2013. "Geography and the Future of Big Data, Big Data and the Future of Geography." *Dialogues in Human Geography* 3 (3): 255–61.

Cinnamon, Jonathan, and Nadine Schuurman. 2013. "Confronting the Data-Divide in a Time of Spatial Turns and Volunteered Geographic Information." *GeoJournal* 78 (4): 657–74.

Elwood, Sarah, Michael F. Goodchild, and Daniel Z. Sui. 2012. "Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice." *Annals of the Association of American Geographers* 102 (3): 571–90.

Gao, Song, Linna Li, Wenwen Li, Krzysztof Janowicz, and Yue Zhang. "Constructing Gazetteers from Volunteered Big Geo-Data Based on Hadoop." *Computers, Environment and Urban Systems*, forthcoming.

Goodchild, M. F. 2013. “Quality of Big (Geo)Data.” *Dialogues Human Geography* 3(3): 280–84.

Graham, Mark, and Taylor Shelton. 2013. “Geography and the Future of Big Data, Big Data and the Future of Geography.” *Dialogues in Human Geography* 3 (3): 255–61.

Kitchin, Rob. 2013. “Big Data and Human Geography Opportunities, Challenges and Risks.” *Dialogues in Human Geography* 3 (3): 262–67.

Kitchin, Rob. (2014) Big Data, new epistemologies and paradigm shifts. *Big Data & Society*. 1(1), April 1, 2014. http://bds.sagepub.com/content/1/1/2053951714528481.abstract.

Kitchin, Rob and Lauriault, TraceyP. (2014) Small data in the era of big data. *GeoJournal*. 2014/10/11, 1-13. http://dx.doi.org/10.1007/s10708-014-9601-7.

Miller, HarveyJ and Goodchild, MichaelF. (2014) Data-driven geography. *GeoJournal*. 2014/10/10, 1-13. http://dx.doi.org/10.1007/s10708-014-9602-6.

Mooney, Peter, and Padraig Corcoran. 2014. “Analysis of Interaction and Co-Editing Patterns amongst OpenStreetMap Contributors.” *Transactions in GIS* 18 (5): 633–59.

Stefanidis, Anthony, Crooks, Andrew and Radzikowski, Jacek. (2013) Harvesting ambient geospatial information from social media feeds. *GeoJournal*. 78(2), 2013/04/01, 319-338. http://dx.doi.org/10.1007/s10708-011-9438-2..