Bridging data gaps for policymaking: crowdsourcing and big data for development

Author : Anthony Swan

Date : July 8, 2016

Good data to inform policymaking, particularly in developing countries, is often scarce. The problem is in part due to supply issues – high costs, insufficient time, and low capacity – but also due to lack of demand: policies are rarely shown to be abject failures when there is no data to evaluate them. The wonderful phrase "policy-based evidence making" (the converse of "evidenced-based policy making") comes to mind when thinking about the latter. However, technological innovations are helping to bridge some of the data gaps. What are the innovations in data collection and what are the trade-offs being made when using them to inform policy?

By far the biggest innovation in data collection is the ability to access and analyse (in a meaningful way) user-generated data. This is data that is generated from forums, blogs, and social networking sites, where users purposefully contribute information and content in a public way, but also from everyday activities that inadvertently or passively provide data to those that are able to collect it.

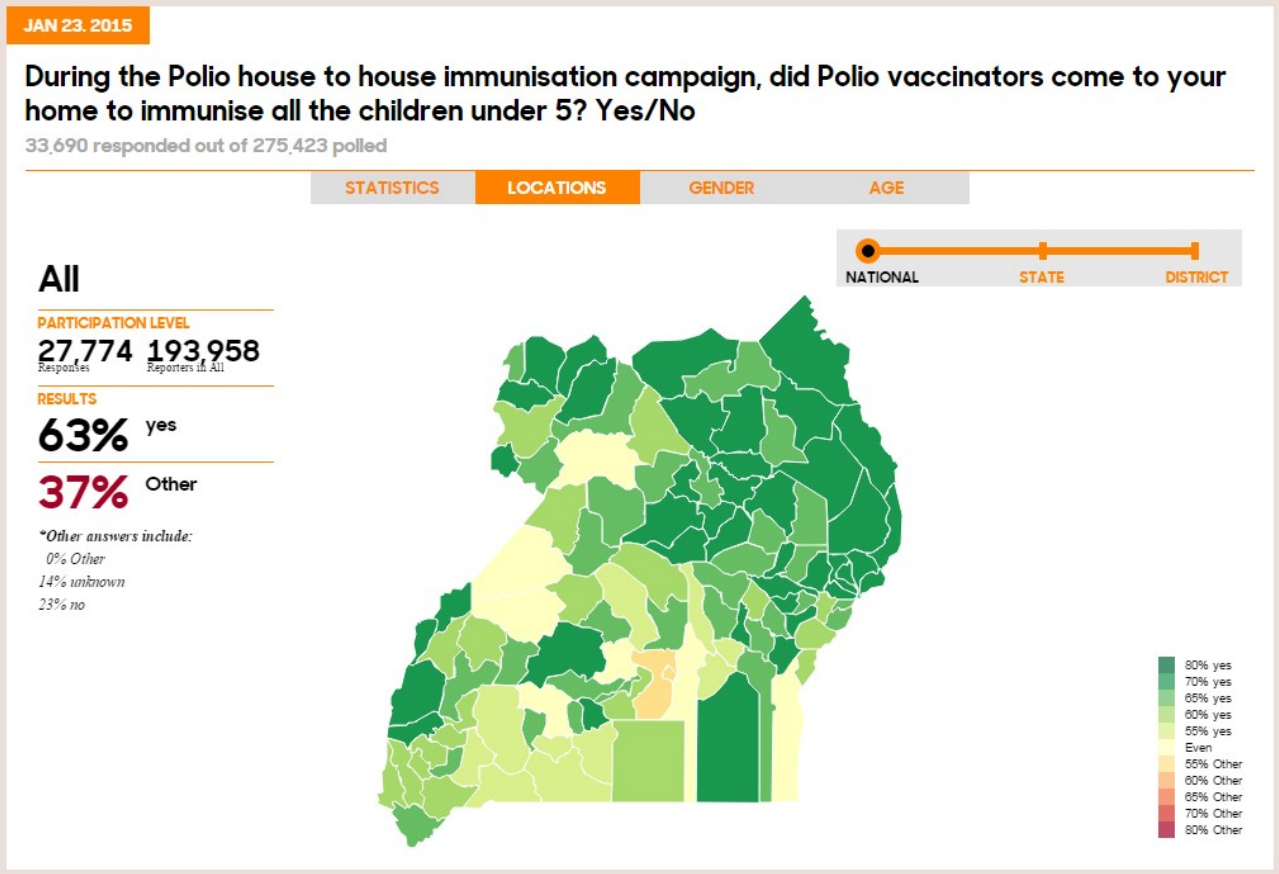
User-generated data can help identify user views and behaviour to inform policy in a timely way rather than just relying on traditional data collection techniques (census, household surveys, stakeholder forums, focus groups, etc.), which are often cumbersome, very costly, untimely, and in many cases require some form of approval or support by government.

It might seem at first that user-generated data has limited usefulness in a development context due to the importance of the internet in generating this data combined with limited internet availability in many places. However, U-Report is one example of being able to access user-generated data independent of the internet.

U-Report was initiated by UNICEF Uganda in 2011 and is a free SMS based platform where Ugandans are able to register as "U-Reporters" and on a weekly basis give their views on topical issues (mostly related to health, education, and access to social services) or participate in opinion polls. As an example, Figure 1 shows the result from a U-Report poll on whether polio vaccinators came to U-Reporter houses to immunise all children under 5 in Uganda, broken down by districts. Presently, there are more than 300,000 U-

Reporters in Uganda and more than one million U-Reporters across 24 countries that now have U-Report. As an indication of its potential impact on policymaking, UNICEF claims that every Member of Parliament in Uganda is signed up to receive U-Report statistics.

Figure 1: U-Report Uganda poll results



U-Report and other platforms such as Ushahidi (which supports, for example, I PAID A BRIBE, Watertracker, election monitoring, and crowdmapping) facilitate crowdsourcing of data where users contribute data for a specific purpose. In contrast, "big data" is a broader concept because the purpose of using the data is generally independent of the reasons why the data was generated in the first place.

Big data for development is a new phrase that we will probably hear a lot more (see here [pdf] and here). The United Nations Global Pulse, for example, supports a number of innovation labs which work on projects that aim to discover new ways in which data can help better decision-making. Many forms of "big data" are unstructured (free-form and text-based rather than table- or spreadsheet-based) and so a number of analytical techniques are required to make sense of the data before it can be used.

Measures of Twitter activity, for example, can be a real-time indicator of food price crises in Indonesia [pdf] (see Figure 2 below which shows the relationship between food-related tweet volume and food inflation: note that the large volume of tweets in the grey highlighted area is associated with policy debate on cutting the fuel subsidy rate) or provide a better understanding of the drivers of immunisation awareness. In these examples, researchers "text-mine" Twitter feeds by extracting tweets related to topics of interest and categorising text based on measures of sentiment (positive, negative, anger, joy, confusion, etc.) to better understand opinions and how they relate to the topic of interest. For example, Figure 3 shows the sentiment of tweets related to vaccination in Kenya over time and the dates of important vaccination related events.

Figure 2: Plot of monthly food-related tweet volume and official food price statistics

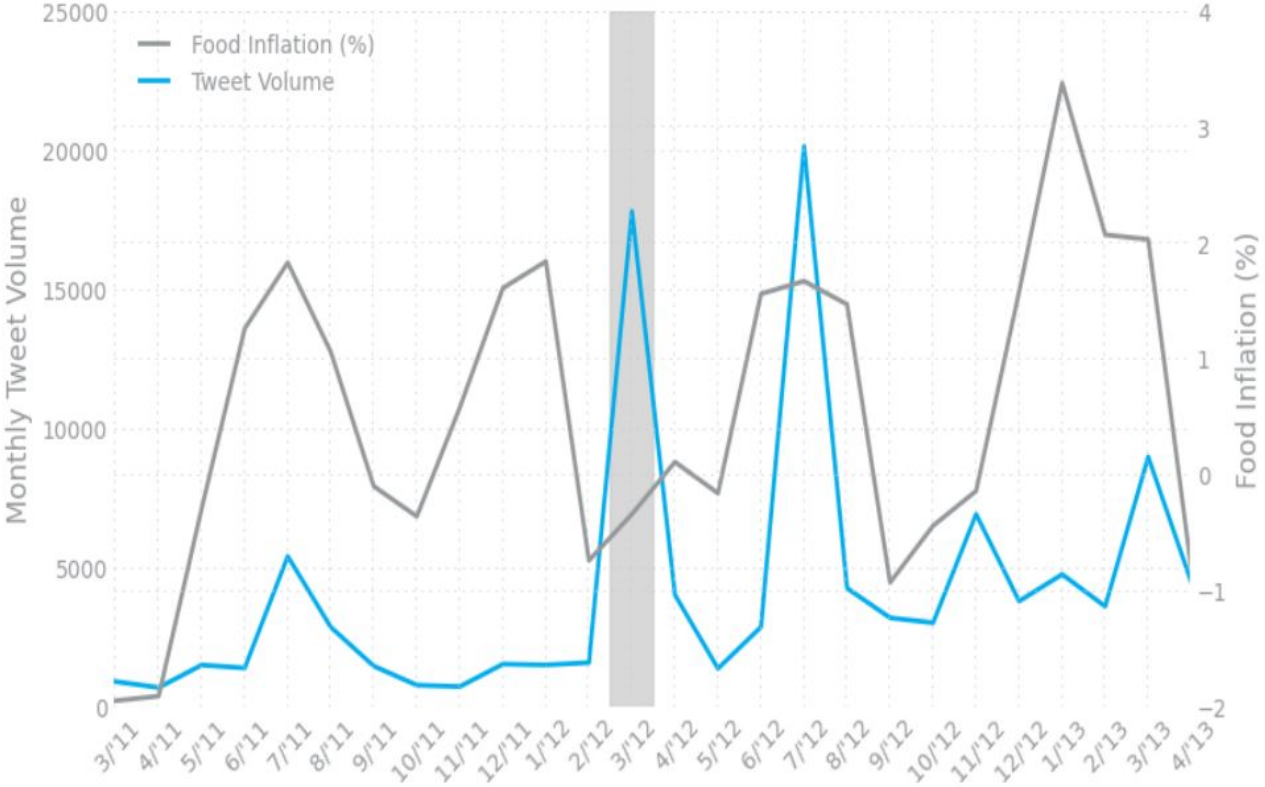
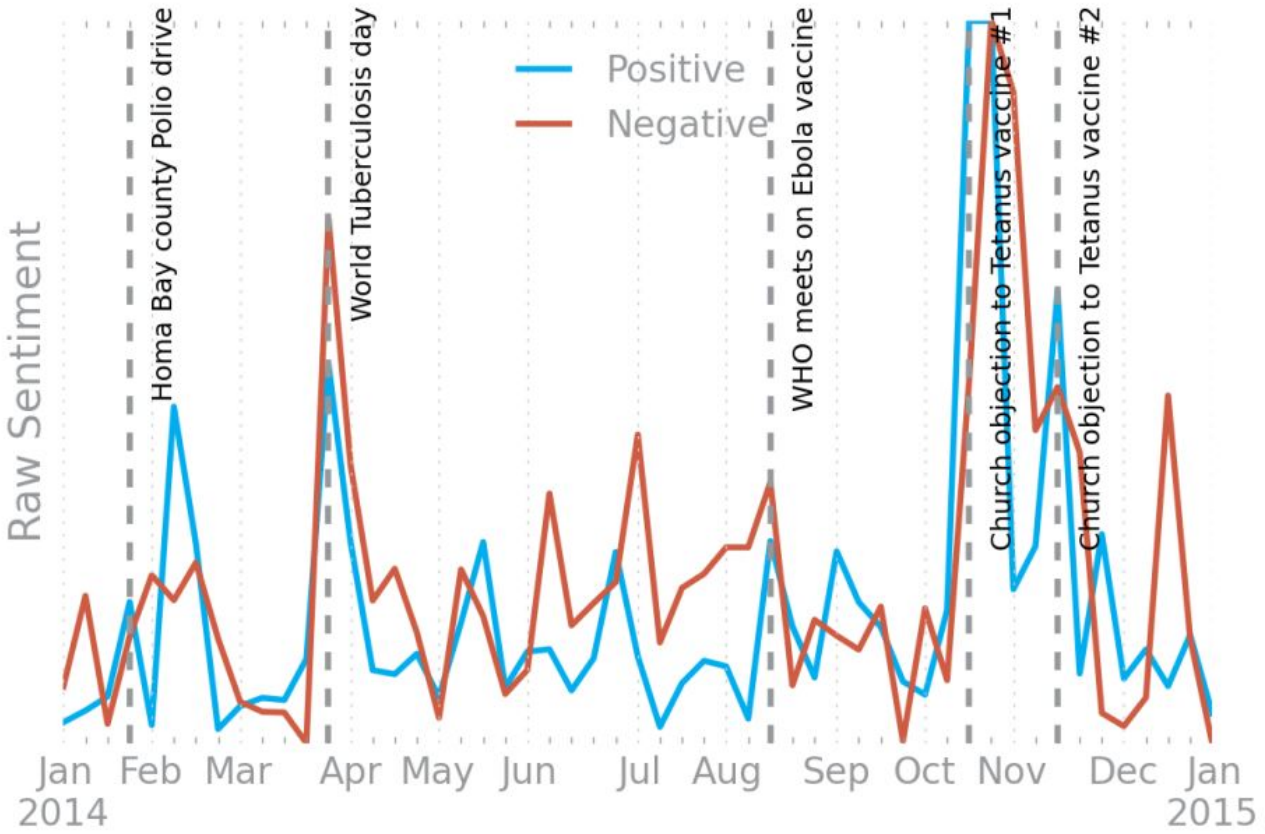


Figure 3: Sentiment of vaccine related tweets in Kenya



Another big data example is the use of mobile phone usage to monitor the movement of populations in Senegal in 2013. The data can help to identify changes in the mobility patterns of vulnerable population groups and thereby provide an early warning system to inform humanitarian response effort.

The development of mobile banking too offers the potential for the generation of a staggering amount of data relevant for development research and informing policy decisions. However, it also highlights the public good nature of data collected by public and private sector institutions and the reliance that researchers have on them to access the data. Building trust and a reputation for being able to manage privacy and commercial issues will be a major challenge for researchers in this regard.

Even when the data is made available, a difficulty with using statistics based on crowdsourcing and big data is working out who is being represented and how reliable the figures are. For example, were 63% of households in Uganda on average visited by polio vaccinators or were the U-Reporters for this poll unrepresentative of the general population, thereby giving a very biased result?

Indeed, it does seem that U-Reporters in Uganda represent particular parts of the population more so than others (e.g., two thirds are male, 90% under 35 years of age, and urban people are likely to be over-represented) so a lot of caution should be applied when interpreting U-Report statistics. However, it is possible to weight the data based on observed characteristics provided by users relative to the prevalence of those characteristics in census data, for example, in order to produce more representative results (Australia's Vote Compass uses this approach).

Clearly, statistics based on Twitter activity or other social media data are rarely going to represent broad populations of interest. However, this does not significantly reduce their value as indicators of change, which is useful for time critical analysis and response. Crowdsourcing also has benefits beyond generating statistics, such as giving voice and empowering individuals to take action (e.g., taking action against a rapist) as well as improving public awareness (such as uncovering the market price of bribes in India: see here and here).

Many researchers, myself among them, have lamented the lack of data in our region (especially Papua New Guinea and the Pacific). The challenge is to look beyond traditional sources of data, cultivate relationships with relevant public and private sector institutions, and apply some of the innovative data collection and analytical methods described here. In the case of PNG, for example, crowdsourced data collection on corruption in the public financial management system is already underway. Is anyone else interested in taking up this challenge in our region?

Anthony Swan is a Research Fellow at the Development Policy Centre.