ICA panel: Conducting Online Research - Methodological Challenges and Opportunities

Catching the User in Online Research: An Innovative Approach for Respondent Recruitment

Peter Mechant (iMinds-MICT-UGent) – peter.mechant@ugent.be
Pieter Verdegem (MICT-UGent) – pieter.verdegem@ugent.be

In addition to sensor data and transactional data being generated by e.g. Customer Relation Management (CRM) Systems, data available on online social media platforms such as Facebook, Twitter or YouTube is often labelled as 'Big Data'. After all, social media data sets can be indeed 'large'. Moreover, they are also 'big' in the sense that the data comes in different types and is generated (and updated) continuously at immense speeds (Dumbill, 2012). Most experts and analysts agree that analysing Big Data will help people be more agile and adaptive, and that Big Data has swept into every industry, business function and research discipline (Pew Research Center, 2012; McKinsey Global Institute, 2011). This presentation elaborates on working with 'big' social media data sets and focuses on how new possibilities emerge when combining user logging data ('what are people doing on social media?') with user survey data ('what do people say they are doing on social media?'). This contribution especially focuses on how to catch the user in online research.

In this context, we introduce the notion of 'APIs' (Application Programming Interfaces), which refers to a set of rules and specifications for interacting with websites as if it were a data server. When used in the context of the Web, an API is a defined set of request messages, along with a definition of the structure of response messages. It comprises both the grammar and vocabulary to 'talk' directly to the database 'behind' a website, and of course get a response. In non-technical terms, an API serves as both a gatekeeper at the back door and as a warehouse keeper who helps to get the right things if addressed in the appropriate language. APIs can be used in social science research to retrieve metadata on media objects, harvesting descriptive data (e.g. tags) and user interaction with the object (e.g. number of comments), or on media subjects (e.g. number of posts) as they enable instantaneous, automated and repeatable data collection.

However, our contribution specifically elaborates on another use of APIs; namely as a tool for the recruitment of respondents from online platforms. We will present two case studies in which APIs are used for recruiting respondents – i.e. people having an account on YouTube and Twitter – by inviting them to participate in an online survey about attitudes and uses of those social media platforms. In both case studies we will not digg into specific research questions and main findings but rather focus on a more general and thus meta description of both research projects, explaining how we used the APIs from the platforms YouTube and Twitter for recruiting survey respondents.

The first example elaborates on a study that focused on teenage YouTube uploaders' networked public expectancies when posting a video (Courtois et al., 2011). These expectancies allow uploaders to cope temporarily with the uncertainty of who exactly will view their video (imagined audiences or not, Marwick & boyd, 2010). Our results indicated that teenage uploaders strongly expect viewers that are situated close to them in both geographic and socio-demographic terms. An effect of the identified online public expectancy (viewers with a similar interest/activity) was found for the importance of feedback both on the platform (e.g., views, comments) and off the platform (e.g., interaction on a social-network site). The identified offline public expectancy (friends/family) affects the importance attributed to off-platform feedback. Surprisingly, no effect of the unidentified online public expectancy (the general public) was found on on-platform feedback. We also investigated the accuracy of the online networked public expectancies by testing their effects on the longitudinal growth of actual feedback (views, comments, and rates). The results provided modest evidence that teenage uploaders have accurate online public expectancies.
When covering this example we will explain how we administered a cross-sectional online survey, describing how YouTube uploaders from 12 to 18 years old were randomly selected from YouTube's Dutch ''Most Recent'' RSS feed. These selected respondents were then sent a comment on their freshly uploaded video, inviting them to participate to an online survey.

The second example shows how we used the Twitter API in an ongoing research project on mobile fitness apps (Stragier & Mechant, 2013). New media (tools) such as RunKeeper, Endomondo, Strava and MapMyRun might offer opportunities to encourage physical activity. Also, mobile fitness apps designed for smartphones experience an increasing popularity. They allow tracking and monitoring of activities as running, cycling and walking. More importantly, these activities can also be shared online on social network sites as Twitter and Facebook. In this research project we focused on the motivation of people for sharing their workouts with online peers. Results indicate that community identification, receiving positive feedback on activities and sharing information on activities are important predictors of a positive attitude towards sharing workouts, which leads to frequent sharing of those workouts.
For this research project we used a custom PHP-script that addressed the Twitter API, to collect a data set of 4556 random tweets with the hashtag #RunKeeper tweeted in a period of two months on Twitter. Next, this the sample of people who posted a tweet with the hashtag RunKeeper, was used to create a random subsample of 1,849 Twitter users who posted a tweet with the hashtag RunKeeper at least once in the past two months. Tweets reporting on walks, rides etc. and also retweets were excluded from the sample. We used the Twitter API in a second research phase to recruit respondents for an online survey from this random subsample of Twitter users. Using the Twitter REST API POST statuses/update-call, which updates the authenticating user's current status, also known as tweeting, we sent @messages to these 1,849 Twitter users, inviting them to navigate to a certain URL that corresponded to the start of an online survey. Simultaneously all the available metadata on these Twitter users was harvested using the Twitter REST API GET users/show-call and their Tweet-activity was followed daily using the Twitter REST API GET statuses/user_timeline-call. In this way, we could combine, supplement and confront the self-reported subjective data (survey data) with the 'pure' objective data captured by means of a (new) measurement system (API data).

Using the above-mentioned case studies on YouTube and Twitter as a springboard, we will point to the methodological opportunities and challenges, and also importantly, to some ethical considerations related to using APIs for enlisting respondents, such as the blurred distinction between private and public spaces, pointing to very concrete questions such as: What is the status of online 'public' data? Can it simply be used without permission? And what constitutes best ethical practices when approaching media users online?
Concerns might be raised whether the construction of a research tool in the shape of an automated multi-step protocol to monitor videos and video uploaders contributes to the enforcement of an online robust infrastructure of dataveillance, making us as researchers accomplices in marketers' and businesses' systematic monitoring of the actions of internet users through the application of information technology and the creation of 'data doubles'. After all, while scholars and researchers have the tools and the access, social media users as a whole have not, and researchers are rarely part of a user's imagined audience (boyd & Crawford, 2012). However, it could also be argued that users in our samples did not restrict access to their content (tweet, video clip…), and that, as they were treated as an aggregated and anonymised sample, there should be no ethical concerns whatsoever when collecting and analyzing data. Still, this ethical position bypasses the difference between 'being in public' and 'being public' (boyd & Crawford, 2011) and ignores the fact that publicly available content is not always meant to be aggregated or consumed by anyone. Although we could not ask consent from every YouTuber uploading a video (and hence automatically becoming part of our data sample) we would have welcomed an ethical framework with checks and balances for evaluating such research ethics.

On a broader level we will also reflect on working with data that has been harvested via social media APIs in general, addressing some of the problems we encountered. For example, there are issues with the representativeness of the sample that is created using APIs. Often, there is no central 'list' of users or content to draw a representative sample from (e.g. YouTube's API and its 'latest videos' RSS feed, by nature, only afford data harvesting from public videos and public user profiles). APIs are also most often of commercial or corporate nature forcing researchers, in a space which remains outside their control because any change in the functionality of APIs or in the data structures may jeopardize research goals or require extra work. Reliability may also be an issue as commercial data is subject to 'Karpf's Rule of Online Data' (Karpf, 2012), which describes the inverse relationship between on the one hand the reliability of an online metric and on the other hand its financial or political value. Importantly, there is also a question of skills needed to interact

with APIs. These skills are generally restricted to those with a computational background, setting up new hierarchies around 'who can read the numbers'. To put it harshly, if you are not a programmer, you are one of the programmed' (Rushkoff, 2011).

References

Boyd, D., & Crawford, A. (2011). Six Provocations for Big Data. Paper presented at the A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. Retrieved from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1926431

Boyd, D., & Crawford, K. (2012). Critical questions for big data. Information, *Communication & Society, 15*(5), 662-679.

Courtois, C., Mechant, P. & De Marez, L. (2011). Teenage uploaders on YouTube: networked public expectancies, online feedback preference, and received on-platform feedback. *Cyberpsychology Behavior and Social Networking, 14*(5), 315-322.

Dumbill, E. (2012). What is big data? An introduction to the big data landscape. Strata: making data work. Retrieved from: http://strata.oreilly.com/2012/01/what-is-big-data.html

Karpf, D. (2012). Social science research methods in internet time. Information, *Communication & Society, 15*(5), 639-661.

Marwick, A.E. & boyd, d. (2010). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*(1): 114-133.

McKinsey Global Institute. (2011). Big Data: the next frontier for innovation, competition and productivity. Retrieved from: http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

Pew Research Center. (2012). Big data. Retrieved from: http://pewinternet.org/~/media/Files/Reports/2012/PIP_Future_of_Internet_2012_Big_Data.pdf

Rushkoff, D. (2011). Program or be programmed: ten commands for a digital age. New York: OR Books.

Stragier, J.,& Mechant, P. (2013). Mobile fitness apps for promoting physical activity on Twitter: the #RunKeeper case. Proceedings of 'Etmaal van de communicatiewetenschappen', Rotterdam, 2013.