Aleksandra K Krotoski

Data-driven research: open data opportunities for growing knowledge, and ethical issues that arise

Article (Published version) (Refereed)

Data-driven research: open data opportunities for growing knowledge, and ethical issues that arise

The Open Data Initiative in the UK offers incredible opportunities for researchers who seek to gain insight from the wealth of public and institutional data that is increasingly available from government sources – like NHS prescription and GP referral information – or the information we freely offer online. Coupled with digital technologies that can help teams generate connections and collaborations, these data sets can support large-scale innovation and insight. However, by looking at a comparable explosion in data-driven journalism, this article hopes to highlight some of the ethical questions that may arise from big data. The popularity of the social networking service Twitter to share information during the riots in London in August 2011 produced a real-time record of sense-making of enormous interest to academics, reporters and to Twitter users themselves; however, when analysed and published, academic and journalistic interpretations of aggregate content was transformed and individualized, with potential implications for a user-base that was unaware it was being observed. Similar issues arise in academic research with human subjects. Here, the questions of reflexivity in data design and research ethics are considered through a popular media frame.

In December 2011, the UK government outlined its plans to release patient information from the National Health Service (NHS). Over the next two years, the new Government Digital Service will make available GP prescribing data, references, electronic booking systems and demographics. The move was part of the next phase in data.gov.uk, an open data policy spearheaded by Sir Tim Berners-Lee and Professor Nigel Shadbolt of the University of Southampton that was initiated during the end of the Brown administration but was continued by Number 10 under David Cameron.

ALEKSANDRA K KROTOSKI
Visiting Fellow
Media@LSE
Research Associate
Oxford Internet Institute

It was lauded by open science advocates and researchers. These organizations, and research institutions across the country, foresaw the opportunities that the release of prescription and GP practice data afforded epidemiological and longitudinal studies, as well as other public health services, social science research questions and cross-disciplinary exchange. Yet the response by the UK population was less enthusiastic. Its reaction was reflected in news headlines, particularly concerned with privacy and the control of personal data which, if linked back to individuals, could have potential implications on their futures.

Although not all data released by the public sector is so contentious – for example, bus and rail timetables, environmental data and Ordinance Survey geographical data have been used to create services for the public and to generate new insight in these areas – there are similar challenges all open data users face when gaining access to data collected through public funds or released by private institutions. Most crucially, the data release must be accompanied by an assurance that the public is aware the data made open will be treated responsibly and ethically, and that no harm will come to the sources of the information. Traditionally, this issue is tackled by ethical protocols at research institutions, which continue to adapt guidelines to contend with the increasing amount of 'big' data sets that are being collected and are becoming available in digital formats.

"… data release must be accompanied by an assurance that the public is aware the data made open will be treated responsibly and ethically, and that no harm will come to the sources of the information."

There is another challenge that has emerged over the last two decades: the web offers a proliferation of *apparently* 'open' big data sets or opportunities to 'scrape', or automatically collect, data from large databases without the web service developer's knowledge or the web consumer's consent. These two aspects raise considerations alone, but there are also additional issues to be aware of due to the nature of the content that users publish online, and the subsequent harm that may come about as a result of the synthesis and analysis researchers and others subject it to.

For example, a few weeks after the government announced the NHS data schedule, I attended a meeting of the British Psychological Society's (BPS) Conducting Research on the Internet Working Party. This was a timely discussion; two days earlier a UK broadsheet had published an analysis of the sources and trajectories of rumours that diffused across the social networking service Twitter during the August 2011 riots in London. This incident resonated with the kind of work and analysis that psychology researchers increasingly face and that the BPS was hoping its internet guidelines would administer to. The Twitter riot data and analysis will be considered throughout the rest of this article as an example of open data made public on a mass media forum, which will serve to expose the issues raised in context.

Analysing the London riots: *The Guardian's* 'Riot Rumours' (7 December 2011)[1]

The Guardian newspaper has been an active leader in a new trend in 'data journalism', an investigative approach to generate news based on data analyses and visualizations. In December 2011, the paper began a series called 'Reading the Riots', in which they published intelligence gained by looking at data. One of their most popular pieces was an interdisciplinary project in association with several universities in the UK that analysed 2.6 million tweets in order to explore how rumours spread and were corrected in an unregulated forum. They published a dynamic social network visualization that linked rumour with username, identifying the original sources of false information by their handles, as well as those who passed the misinformation on.

There were many issues raised by the digital data, but amongst them were three specific areas relevant to the open data zeitgeist in academic communities:

1. Was permission to access the Twitter site (and data) gained by the researchers before they scraped the data?

2. Was permission gained from the individuals who were identified as hubs of misinformation by the researchers before they reported their analysis?

3. What potential harm might come to the individuals who held the named Twitter accounts based on their 'exposure' as sources of misinformation?

Relationship between researcher and service provider

In web-based scenarios, the relationship between the researcher and the software developer is multiple. However, two elements in particular stand out.

First, unless the web designer has decided that the contents should be private and have protected their site with a password or other mechanic to exclude non-members, the system is considered 'open' and individuals have the opportunity to 'lurk' unobtrusively. Further, if researchers have the tools, they can also collect data – interactions, clicks, posts, transcripts, etc. – according to their own research question specifications.

"… the relationship between the researcher and the software developer is multiple."

In the case of the Twitter riot data, the journalists were given 2.6 million tweets with usernames by the company, and were able to identify the pathways from which the information came and where it went afterwards. Yet it is possible to scrape this kind of data without a company's permission using commercial technologies developed for web crawling. Often, however, this activity falls outside a site's terms of service, as the company may be concerned about others' commercial gains from their system, the resilience of the community to scrutiny by individuals not invested in its aims, and the rights of their users as customers.

The second feature of the relationship between researcher and service provider is the context in which the site is used and built: from the meanings and etiquette ascribed to information and actions negotiated within the community, to the design decisions implemented by the developers to 'clean up' the 'messy' humans who will use the site according to its primary function.

Communities that emerge around a web service will often develop unspoken boundaries that determine what is and is not acceptable in terms of how content is consumed and shared, and these subtleties may not be apparent to the objective glance of the researcher. For example, a socially-generated context may have had an impact on how misinformation diffused through Twitter, through the user-generated trend of 'retweeting'. The significance of other community-evolved features, like Twitter's hashtag (#) followed by a user-determined subject word to collate subject-related content, may not be apparent in the data in its raw form.

These socially-derived boundaries may, in part, be determined by the ways the creators of the systems construct the software parameters; as Melvin Kranzberg opined, 'technology is neither good nor bad; nor is it neutral'.[2] It is therefore useful to look at the function the service seeks to provide and to extract the mechanics used to achieve this. At its most basic, Twitter's function is to provide a platform on which to share short messages with people who subscribe to a user's feed; a reciprocal relationship is not required. Contrast this with the relationship mechanic in another popular social network, Facebook: the connection implementation in that service has established a 'relationship' as existing between two consenting account holders. Therefore, although rumour will spread through a system like Facebook, it is likely that (mis)information would spread in a different pattern across Twitter than in Facebook as it is a service with less account verifiability and more heterogenous connections.

Relationship between researcher and individual

Most open data initiatives spearheaded by the public and university sectors in the UK have attempted to be as robust as possible in their data anonymization strategies, but it is still possible to identify an individual based on the information s/he proffers in the content of online interactions[3]. For example, the relationship between the offline and online self has been extensively studied[4], and so the connection with an individual and a pseudonym – or even a collection of data points – is a personal relationship regardless of identifiability.

Depending on the research question, most online research with human subjects includes an informed consent routine to give the participant the opportunity to opt out at any time. This is not always appropriate, as in observational studies in online forums, or when the data set is scraped. In these cases, consent is often assumed or waived.

In the commercial sphere, data content supplied to service developers usually falls under the ownership of the corporate entity and thus can be distributed to interested parties according to their internal commercial criteria. Using the Twitter-riots analysis as an example again, *The Guardian* received the tweets of service account holders *without* the consent of those people who were included in the data set.

When analysing data at an aggregate level, a waiver of consent can easily be argued, but when individuals are identified in the reports of the analysis – even by pseudonym only – it is at the discretion of the researcher whether or not to inform the account holders or to publish the pseudonym under a pseudonym.

Potential harm

The issue of consent is always related to the question of the degree of potential harm that may come to the individuals implicated in the data set. This is where the December 2011

NHS data became the subject of public debate. Analysis of data released through an openness agenda involving data that can be perceived as sensitive – whether data about prescriptions, GP practices, drug use, sexual activity, financial circumstances, political activity, etc. – is easily categorized as potentially harmful.

Data scraped from other online services may expose unexpectedly personal and potentially harmful information, which can be aggregated from various services to generate an identifiable profile of an individual. Open data sets that appear to be innocuous may thus in fact cause harm. In the case of the Twitter rumour diffusion generated by the riots, 'harm' is not clear-cut. It is only when direct association with misinformation or untruth and a user's online identity is made explicit in analysis that it is possible to see where potential harm may arise.

"Data scraped from other online services may expose unexpectedly personal and potentially harmful information ...Open data sets that appear to be innocuous may thus in fact cause harm."

It can be argued that by publishing the usernames of individuals who, according to the analysis, were the original sources of rumours about the London riots, the reputation of those individuals as sources of verifiable information can be called into question. The sanctity of the status and the reputation of the virtual identity in a community are paramount because they are the most important social currency in these environments[5]. This may also be the case for publishing the names of those who perpetuated, rather than started, the rumours: information from those user accounts may also be viewed in the future within the context of the misinformation that they were responsible for previously. The act of querying the data makes actions and networks explicit, which may result in, for example, reputational harm.

However, the individual may not perceive this as harmful, which is why guidelines for internet research like those published by the BPS, or the Association of Internet Researchers, recommend gaining informed consent pre-publication.

" …by publishing the usernames of individuals who, according to the analysis, were the original sources of rumours … the reputation of those individuals as sources of verifiable information can be called into question."

Conclusion

There are many opportunities in open data initiatives. The depth of insight gained from public sources of information can be enormously valuable in empirical and applied spheres. Coupled with technologies that can help render connections visible in big data sets and support cross-disciplinary and disproximate collaboration, the benefits for knowledge are enormous.

However, an 'open' data set, like any other data set, must be analysed in the context in which it was collected. Further, if the data refers to human subjects, it is possible to reverse engineer and identify anonymized data, thus introducing potential harm to the individuals behind the data points. Additionally, the analytic lenses we place on such data may also expose individuals to unanticipated outcomes that may cause perceived harm.

The computer technologies we use were built to connect people with information. What has happened is that the internet has connected people with people. As the greatest social experiment of our time, we must preserve the technologies that make this possible for the future.

References

1. Guardian Interactive team, Proctor, R, Vis, F and Voss, A, Riot rumours: how misinformation spread on Twitter during a time of crisis, *The Guardian*, 7 December 2011: http://guardian.co.uk/uk/interactive/2011/dec/07/london-riots-twitter (accessed 10 January 2011).
2. Kranzberg, M, Technology and history: 'Kranzberg's Laws', *Technology and culture*, 1986, 27(3), 544–ww560.
3. Jones, K C, Fallout from AOL's data leak just beginning, *InformationWeek*, 9 August 2006: http://informationweek.com/news/191900935 (accessed 10 January 2011).
4. Turkle, S, *Life on the Screen: Identity in the Age of the Internet*, 1995, Boston, Simon & Schuster.
5. Rosenberg, A, Virtual world research ethics and the private/public distinction, *International journal of Internet research ethics,* 2011, 3(1), 23–37.

Aleksandra K Krotoski, PhD, Visiting Fellow, Media@LSE and Research Associate, Oxford Internet Institute
E-mail: aleks@alekskrotoski.com | web: http://www.alekskrotoski.com

To cite this article:

Krotoski, A K, Data-driven research: open data opportunities for growing knowledge, and ethical issues that arise, *Insights*, 2012, 25(1), 28-32, doi: 10.1629/2048-7754.25.1.28

To link to this article:

http://dx.doi.org/10.1629/2048-7754.25.1.28