Journalists should follow the lead of media scholars and look to the Internet as a rich source of data.

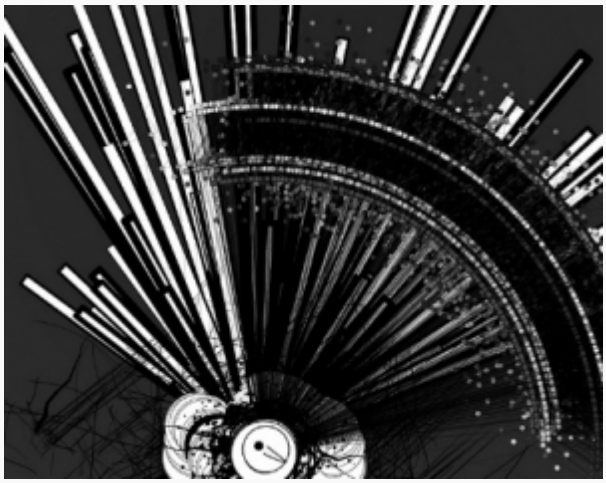
blogs.lse.ac.uk/impactofsocialsciences/2014/05/13/data-journalism-media-rich-data/

5/13/2014

Journalists rarely use the web as a source of data about the state of issues, debates and information flows in different societies. **Liliana Bounegru** *looks at how media scholars have leveraged digital data and algorithmic accountability. In times of shrinking news budgets and staff cuts journalists can turn to such readily available sources of data as a way to understand public engagement with major issues. Scholars can support this process by making the datasets, tools and protocols developed during their work available to others.*

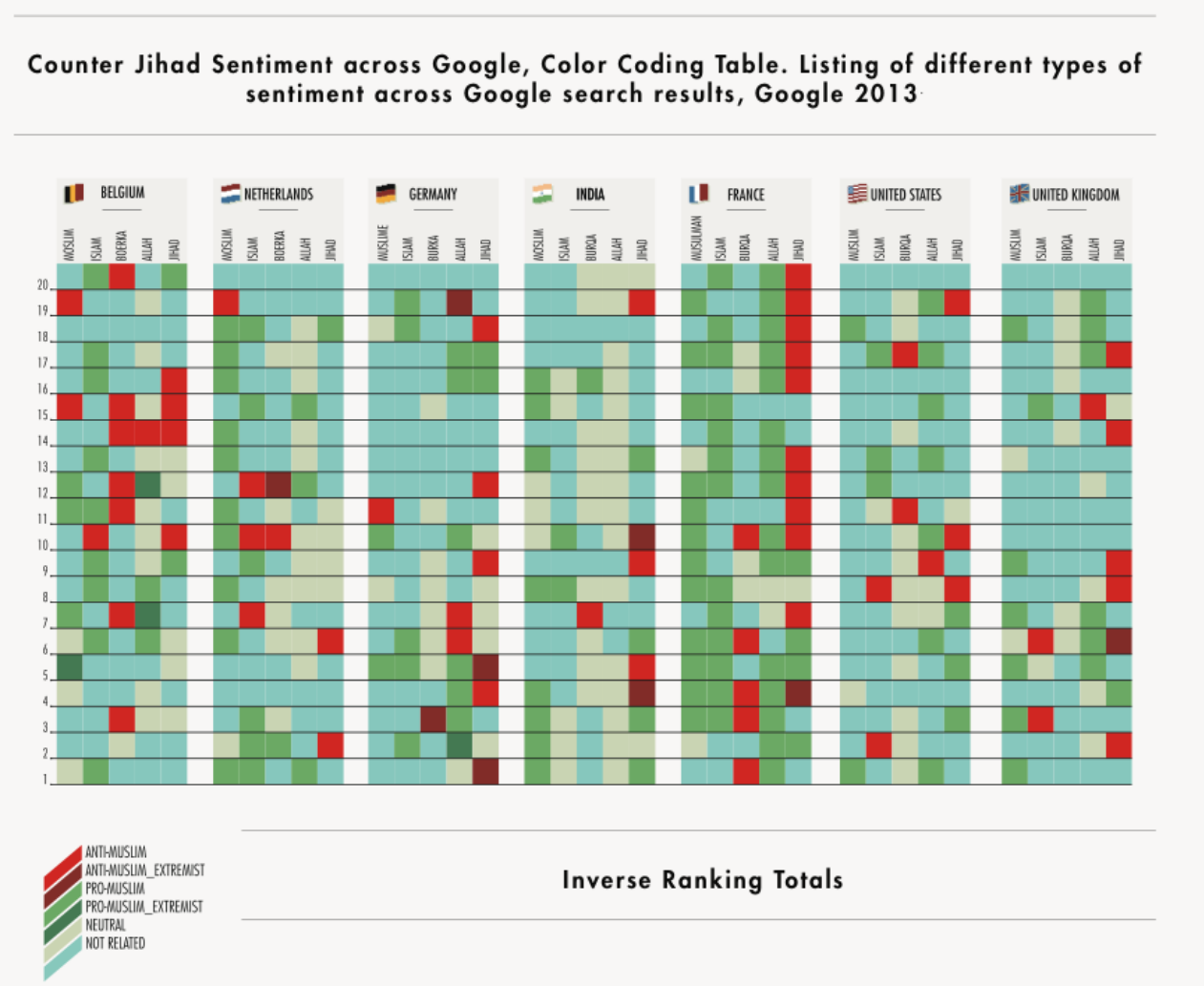
Earlier this year Alex Madrigal of Atlantic Tech published a fascinating piece that uncovered the mystery behind the very specific film genres that Netflix presents to its users. Netflix users are recommended categories of films such as "Tearjerkers from the 1970s" or "African-American Crime Documentaries" and there are 76,897 of them! What was really interesting about this piece was Madrigal's sourcing approach. To identify these genres Madrigal wrote a script that pulled the genres by navigating link by link (As he explains in the article, "http://movies.netflix.com/WiAltGenre?agid=*1* linked to 'African-American Crime Documentaries', http://movies.netflix.com/WiAltGenre?agid=*2* linked to 'Scary Cult Movies from the 1980s'. And so on.") In doing so he essentially used metadata generated by an online service and re-purposed it to start an investigation into our cinematic culture.

While this topic would typically be considered as 'soft news' in journalism, one can imagine how journalists can engage in scrutinising other algorithms and services that are central to our information acquisition habits such as Google or Wikipedia and their information politics. This is what Tow Center fellow Nick Diakopulos calls 'algorithmic accountability'. One approach to algorithmic accountability that we developed at the University of Amsterdam that could be valuable for journalists is measuring how key topics, views or actors rank in Google results, one of the main entry points to the web. This is a 21st century version of what media scholars traditionally do when they analyse how issues are prioritised in print and broadcast media, and how this contributes to the formation of public opinion.



Algorithms. Image credit: jm_escalante (CC BY-NC-SA)

For example, one of our studies mapping anti-Islamism or counter-Jihadism on the web examined how close or readily available hate content is to the every day search engine user. We looked at where this content is located in the top local Google domains of a number of countries: Belgium (google.be), the Netherlands (google.nl), Germany (google.de), India (google.co.in), France (google.fr), the United Kingdom (google.co.uk) and the "universal" Google (google.com) for keywords related to Islam such as "Muslim", "Allah", "burka" or "Jihad". We found that anti-Muslim content is indeed accessible in the top 20 Google results for these queries in each of the local domains – and in some cases extremist content is also present, as shown in the visualisation below from a forthcoming publication by Richard Rogers and the Digital Methods Initiative (Amsterdam):



Source: What Does the Internet Add? Studying Counter-Jihadism Online (forthcoming).Compiled by Richard Rogers and the Digital Methods Initiative, Amsterdam.

What these two examples have in common is the fact that they re-purpose information generated by online services, and treat it as a source of data, which we can use to understand how digital media can format the way we think and act in relation to issues. Whereas journalists typically use the web and social media to identify documents and human sources and to communicate with others, they rarely use social media and the web as a source of data about the state of issues, debates and information flows in different societies.

There have been a few notable exceptions to this such as the Guardian UK's 'Riot Rumours' project and the NRC Handelsblat's 'Opkomst en ondergang van extreemrechtse sites'. The award-winning 2011 'Riot Rumours' project of the Guardian UK analysed tweets produced around the London riots that year to identify the rise and fall of rumours on Twitter, thereby helping to shed light into information dynamics on this platform and how information can be verified in crucial situations like emergencies and disasters. An earlier example from the Dutch newspaper NRC Handelsblad (2007) shows how journalists used web data to investigate Dutch culture, and particularly the rise of extremist language in the Netherlands. Instead of interviewing subject-matter experts, the journalists mined hundreds of websites from the Internet Archive, dating back several years, and found that right-wing websites increasingly employed the language found on extremist sites, and concluded that right of centre Dutch political culture was shifting towards views traditionally associated with the far right.

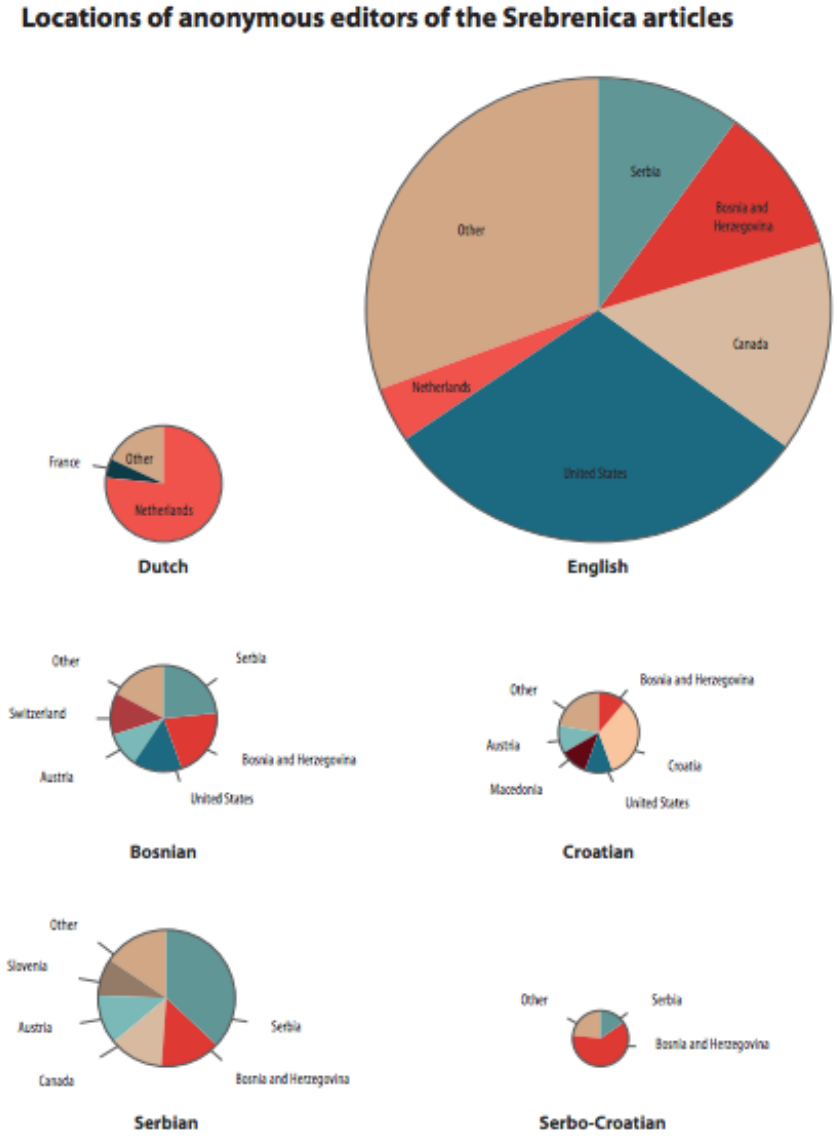
My experience leading a data journalism training programme and working with outlets pioneering the use of digital data in the service of journalism (such as the New York Times, the Guardian and ProPublica) showed me that

journalists tend to turn to governments, non-governmental organisations and companies for data to tell their stories. But how could using the internet as a source of data help us understand not only the policy-driven statistical evidence of key actors, but also the complex social lives of issues and controversies as they unfold, that mysterious beast called public opinion upon which democratic life depends? What can digital traces generated and captured in various online services add to our understanding of debates around climate change, inequality, ageing, global health, extremism, freedom of speech, migration and other major issues that we face?

This is where recent experimentation with new sources of data in the digital humanities and social sciences can contribute. At the University of Amsterdam we are currently working on a project led by the sociologist Bruno Latour, which aims to use digital data and digital methods to understand controversies around climate change. For example one of the things that we looked at is how “biodiversity hotspots” are treated by organisations working on climate change adaptation. Supporting biodiversity is a crucial part of many climate change adaptation strategies as it increases resilience in the face of climate change. While one might expect these vital hotspots to be treated equally by organisations who work on this topic, our research showed that some were neglected and others were given disproportionate attention.

Source: Mapping Climate Change Scepticism, Mitigation and Adaptation Online (fortcoming). Compiled by Richard Rogers and the Digital Methods Initiative, Amsterdam.

Another immensely rich source of data about the social lives of issues is Wikipedia. For example at the University of Amsterdam researchers undertook a comparative analysis of Wikipedia articles about the Srebrenica massacre that took place in July 1995. They found that across six different language versions of Wikipedia there are significant differences in how the events were portrayed. As well as analysing basic elements such as article titles and contents, the analysts also looked at less obvious things like the locations of anonymous editors (based on IP address), the discussion pages behind the articles, images and reference lists. By surfacing the differences between national views, this kind of approach indicates the potential for gaining a better empirical understanding of how different publics react to different issues and controversies, from Crimea to Snowden, the Eurozone to the financial crisis.



Source: Rogers, Richard (2013). Digital Methods. MIT Press.

In times of shrinking news budgets and staff cuts journalists can turn to these readily available sources of data to complement interviews or surveys as a way to understand public engagement with major issues of the day. Scholars can support this process by making the datasets, tools and protocols developed during their work available to others. Sciences Po and the University of Amsterdam are making their toolkits for Internet research available here and here.

I hope that these examples have shown that academic research in social sciences and digital humanities can help to open up the journalistic imagination regarding sourcing of stories and how digital traces left by users of various online services can become a rich source of evidence in understanding the formation of publics and opinions, a process central to our democracies.

Featured image credit: JM_Escalante (Flickr, CC BY-NC-SA)

Note: This article gives the views of the authors, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment

below.

About the Author

Liliana Bounegru *is leading the European Journalism Centre's Data Driven Journalism initiative, one of the main programmes for training, resources and networking in the area of data journalism. Liliana is also a new media researcher at the University of Amsterdam where she works on the Digital Methods Initiative and on the collaborative project EMAPS (Electronic Maps to Assist Public Science), led by the sociologist Bruno Latour. She is also PhD candidate studying data journalism at the University of Groningen (Netherlands) and the University of Ghent (Belgium). She blogs about her work at lilianabounegru.org and tweets at @bb_liliana.*