

# Digital Methods for Web Research

RICHARD ROGERS

## Abstract

Digital methods are techniques for the study of societal change and cultural condition with online data. They make use of available digital objects such as the hyperlink, tag, timestamp, like, share, retweet, and seek to learn from how the objects are treated by the methods built into the dominant devices online, such as Google Web Search and Facebook's Graph Search. They endeavor to repurpose the online methods and services with a social research outlook. Ultimately the question is the location of the baseline, and whether the findings made may be grounded online. Digital methods as a research practice is part of the computational turn in the humanities and social sciences, and as such may be situated alongside other recent approaches, such as cultural analytics, culturomics, and virtual methods, where distinctions may be made about the types of data employed (natively digital and digitized) as well as method (written for the medium, or migrated to it). The limitations of digital methods are also treated. Digital methods recognize the problems with web data, such as the impermanence of web services, and the instability of data streams, where, for example, APIs (application programming interfaces) are reconfigured or discontinued. They also grapple with the quality of web data, and the challenges of longitudinal study, where, for instance, all of Twitter's tweets may be archived by the Library of Congress, but new types of gaps emerge owing to changes over the years in the company's terms of service.

## DIGITAL METHODS FOR THE STUDY OF WEB DATA

*Digital methods* is a term that seeks to capture a recent development in Internet-related research, summarized as approaches to the web as data set. Joining a larger computational turn in the social sciences and the digital humanities, it asks a series of questions about the quality of web data, the productivity of online collection and analytical methods, and ultimately the prospects of having the web serve as a site for grounding findings. When may the web become the baseline for findings about social change?

When one raises the question of the web as a site for the study of social phenomena, a series of concerns arises. Web data are problematic. They have historical reputational issues, owing to the web's representation and study as a medium of self-publication as well as one of dubious repute, inhabited

by corsaires, pornographers, and conspiracy theorists (Dean, 1998). This was the cyberspace period, with an anything goes web, where it often was studied as a separate realm, even a “virtual society” (Woolgar, 2003). Later, the web came to be known perhaps more productively as an amateur production space for user-generated content (Jenkins, 2006). Nowadays the web is becoming a space for more than the study of online culture. Rather it has become a site to study a range of cultural and social issues, charting, for example, “concerns of the electorate” from the “searches they conduct,” and “the spread of arguments ( . . . ) about political and other issues,” among other questions concerning society at large (Lazer *et al.*, 2009, p. 722). More broadly, as Duncan Watts phrased it:

“If handled appropriately, data about Internet-based communication and interactivity could revolutionize our understanding of collective human behaviour”

(2007, p. 489)

As indicated the web, however, has had the general difficulty of meeting the standards of good data (Borgman, 2009). As such web data are also candidates for a shift, however slight, in methodological outlook. If web data are often considered dirty and poor, where could their value lie? The question could turn around. Where and how are web data handled routinely and deftly? Digital methods seek to learn from the so-called methods of the medium, that is, how online devices treat web data (Rogers, 2009a). Thus digital methods are, first, the study of the methods embedded in the devices treating online data (Rieder, 2012). How do search engines (as Google) treat hyperlinks, clicks, timestamps and other digital objects? How do platforms (as Facebook) treat profile interests as well as user interactions such as liking, sharing, commenting and liking comments?

Digital methods, however, seek to introduce a sociological imagination or a social research outlook to the study of online devices. Google Flu Trends is a good example, and serves as a case of how search engine queries may be employed to study social change (Ginsberg *et al.*, 2009). The location and intensity of flu and flu-related queries are used to chart the rising and falling incidence of flu in specific places. The places of flu is an imaginative use of web data for social research, breaking (albeit not in name) with the “trend” research that engines have been known for to date for marketing research under such names as Google Trends, Google Insights for Search, Yahoo Buzz Log, Yahoo Clues, Bing Webmaster Keyword Research, AOL Search Trends, YouTube Keyword Tool, YouTube Trends, and the Google AdWords Keyword Tool (Raehsler, 2012). It is also a case where the baseline is not web data or the web, but rather the (triangulated) findings from traditional flu surveillance techniques used by the Centers for Disease Control and Prevention in the

United States and its equivalents in other countries where Flu Trends as well as Google Dengue Trends have been instituted. Search engine query data are checked against the offline baseline of data from hospitals, clinics, laboratories, state agencies, and others. The offline becomes the check against which the quality of the online is measured.

For those seeking to employ web data to study social phenomena, the webometrician, Michael Thelwall, has suggested precisely the course of action taken by Google in the Flu Trends project: ground the findings offline. Given the so-called messiness of web data as well as the (historical) skepticism that accompanies its use in social research (as mentioned above), Thelwall relates the overall rationale for a research strategy that calls for offline correlation:

“One issue is the messiness of Web data and the need for data cleansing heuristics. The uncontrolled Web creates numerous problems in the interpretation of results (...). Indeed a skeptical researcher could claim the obstacles are so great that all Web analyses lack value. One response to this is to demonstrate that Web data correlate significantly with some non-Web data in order to prove that the Web data are not wholly random”

(Thelwall, Vaughn, & Björneborn, 2005, p. 81)

Digital methods raise the question of the prospects of online groundedness. When and under which conditions may findings be grounded with web data? One of the earlier cases that pointed up the prospects of web data as having a “say” or even a “great say” in the findings is journalistic and experimental. In the longform journalism in *NRC Handelsblad*, the Dutch quality newspaper, the journalist asked the question of whether Dutch culture was hardening, given the murders and the backlash to them of the populist politician, Pim Fortuyn, and the cultural critic, Theo van Gogh in the mid-2000s (Dohmen, 2007). By hardening is meant becoming less tolerant of others, with even a growing segment of radicalizing and more extremist individuals in society. The method employed is of interest to those considering web data as of some value. Instead of embedding oneself (e.g., among hooligans), studying pamphlets and other hard-copy ephemera, and surveying experts, the research turned to the web. Lists of right-wing and extremist web sites were curated, and the language on the two types of sites was compared over time, with the aid of the Wayback Machine of the Internet Archive. It was found that over time the language on the right-wing sites increasingly approximated that on the extremist sites. While journalistic the work provides a social research practice: chart change in language over time on the web, in order to study social change. (The article also was accompanied by the data set, which is unusual for newspapers, and heralded perhaps the rise of so-called data-driven journalism.) The journalist read the web sites, in a

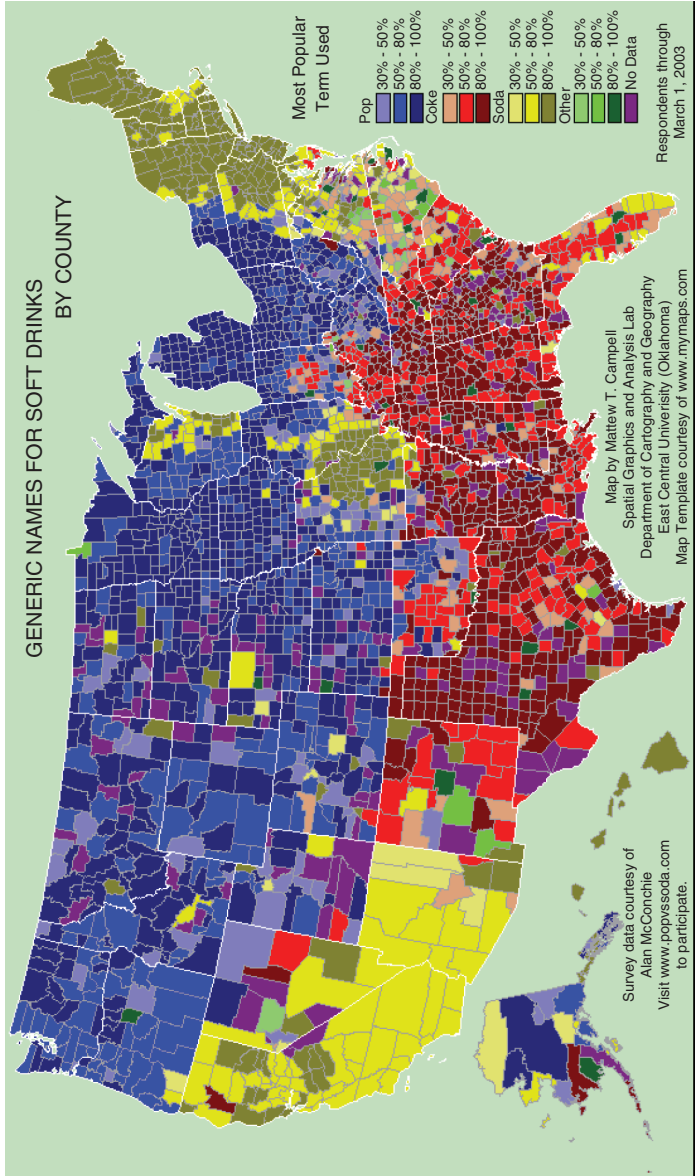
so-called close reading approach, yet one could imagine querying the sources as well in the distant reading approach which has come to be affiliated with the computational turn and big data studies more generally (Moretti, 2005; Boyd & Crawford, 2012).

Another project that is demonstrative of digital methods is the cartogram visualization of recipe queries, which appeared in the New York Times (Ericson & Cox, 2009). All the recipes (on allrecipes.com) queried the day before Thanksgiving, the American holiday and feast, were geo-located, showing in other words the locations from whence they came. The map is shaded according to frequency of queries (and is statistically normalized), where one notes differences in recipe queries, and perhaps food preference, across the United States. It presents, more broadly, a geography of taste. Here the question becomes how to ground the findings. Does one move offline with surveys or regional cookbooks, or seek more online data, such as food photos, tagged by location, and timestamped? Would Flickr or Instagram provide more grounding? Here the web becomes a candidate grounding site.

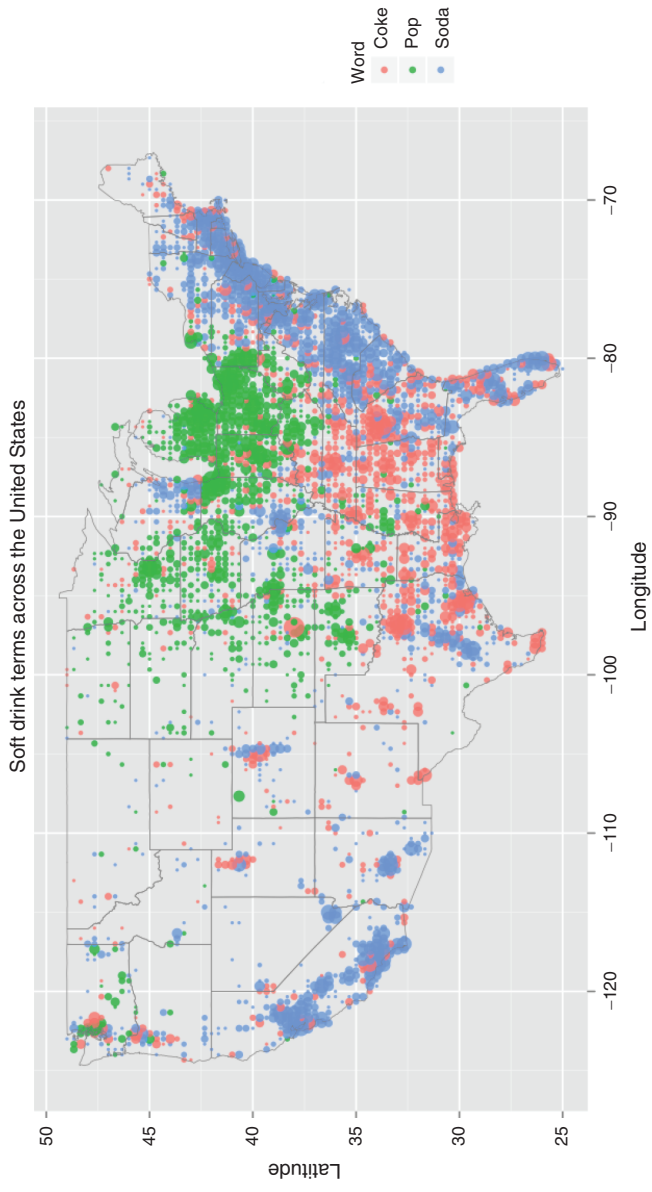
Online data has been employed to study further regional difference. One case in point is the classic discussion of language variation in the use of the terms, soda, pop and coke in the United States. Geo-tagged tweets with the words soda, pop or coke are captured, and plotted to a map, displaying a geography of word usage (Figure 1) (Chen, 2012). In the project the findings are compared to those made by another web data collection technique, albeit a method migrated online, also known as a *virtual method*, a term I will return to. A webpage serves as an online data collection vessel, where people are asked to choose their preferred term (soda, pop, coke, or other) and fill in their hometown, including state and zip code (Figure 2) (McConchie, 2002). The resultant map shows starker regional differentiation than the Twitter analysis (Campbell & Plumb, 2003). Chen, while not confirming the earlier findings, reports “similar patterns,” with pop mid-western, coke southern and soda northeastern and far western (2012).

#### SITUATING DIGITAL METHODS IN THE COMPUTATIONAL TURN

Digital methods may be situated as somewhat distinctive to other contemporary approaches within the computational turn in the social sciences and the digital humanities (Figure 3). First, it shares with other contemporary approaches in the study of digital data with methods based on queries, and has as a research practice what may be called *search as research*. It differs, however, from other approaches in that it relies on born-digital data, and online method as opposed to digitized data and migrated method.



**Figure 1** US map of self-reported usage of terms for soft drinks, 2003. Source: Alan McConchie, [popvssoda.com](http://popvssoda.com). Map by Campbell and Plumb, 2003.



**Figure 2** US map of found usage of terms for soft drinks, 2012. Source: Twitter. Map by Edwin Chen, 2012.

Data	Method	
	Digitized	Natively digital
Digitized	Culturomics* Cultural analytics*	
Natively Digital	Virtual methods webometrics	Digital methods

\*Culturomics and cultural analytics may have digitized qualitative method, but they employ digital methods in part because they use *search as research*.

**Figure 3** Situating digital methods among other approaches in the computational turn in the humanities and social sciences, according to their use of natively digital or digitized data and method.

Two approaches in the digital humanities that may be compared to digital methods are culturomics as well as cultural analytics. While digital methods studies web or natively digital data, culturomics and cultural analytics have as their corpi what one could call digitized materials, which then are searched for either words (in culturomics) or formal material properties (in cultural analytics). Culturomics queries Google books and performs longitudinal studies concerning the changes in use of language from the written word, inferring broader cultural trends. For example, American spelling is gradually supplanting British spelling, and celebrity or fame is increasingly more quickly gained and shorter-lived (Michel *et al.*, 2011). Cultural analytics is a research practice that also queries but at a lower level in a computing sense; it queries and seeks patterns and changes not to words but to formal properties of media, such as brightness and saturation (Manovich, 2007).

Digitized data are often considered better than web data. Both culturomics and cultural analytics have to their advantage the study of what has been described as good data. For culturomics the queries are made in a large set of historical books, which the researchers describe as the study of 5 million books, or approximately 4% of all books ever printed. For cultural analytics, the preferred corpus is the complete oeuvre of an artist (such as Mark Rothko) or the complete set of covers of a magazine (such as *Time*). In those cases, the data are good because they exist or have been captured from the beginning, cover long periods of time, and are complete, or rather so. One knows the percentage of missing data. With the web much data are from a recent past, cover a short period of time and are incomplete, where there is often a difficulty in grasping what complete data would be.

Indeed, the difficulties of moving method and collecting data online is the subject of a social science approach, in the computational turn, called *virtual methods*. While digital methods seek to make use of the methods of the medium, virtual methods migrate the social science instrumentarium online, such as online surveys. The transition of the methods online varies in smoothness. Online, net or virtual ethnography has been able to define communities, enter them and observe and participate (Hine, 2005). For other techniques virtual methods seek to overcome some difficulties that the web affords as a site of study, and data collection realm. When surveying, the question is how to find the respondents, and whether one knows (for example) the number of subscribers to a particular mailing list (and thus a response rate). For sampling, similarly, there are questions about whether one can estimate the population of web sites on a given topic. When interviewing there is the comfort level of being recorded on Skype compared to on a tape recorder or dictophone, and the additional question of whether that video will be put online, or perhaps has an expiration date, as online files may have when uploaded using a file-sharing service. Another aspect of interviewing is relevant in a discussion of the online as site for method debates. One looks up interviewees online, for example, through engine queries as well as social media profiles. It could be said that one not only prepares for an interview by checking background and references online. One may take notice of the person's klout or another social media metric, which is a ratings culture also shared by the online hospitality sector. However, perhaps one also checks the accounts of the interviewee after the fact by returning to the online. Indeed, one particular method of interest that has moved online is from journalism: fact-checking. Here the web, or the (political) blogosphere in certain countries, has taken on the mantle of fact-checkers, listing the points made in US presidential debates, for example, and subsequently putting data and published reports beside them (Annenberg School). It is instructive for the potential of the online to ground claims, or at least become the site of grounded claims. Here the online becomes the decisive source that provides the proof, so to speak, or the fact of the matter a posteriori. Finally, for user studies, how to migrate the methods online, or to the digital? How does one migrate to the medium the study of one's consumption of health information, for example (Mager, 2012)? May one consult people's browser histories? In all the migration of method online could be said to raise questions about the fit between the method and the medium.

Digital methods, contrariwise, strives to make use not only of born-digital data but also the methods that are native to the medium. "Native" is meant not in an ethnographic or anthropological sense. Rather it is meant in a computing sense of that which is written for a particular processor or operating



system, rather than simulated or emulated. By native here is meant as that written for the online medium, rather than migrated to it.

Recently a third type of “digital” object has been introduced, beyond the natively digital and the digitized. The reborn digital object is that which was once born in the medium, archived and reborn as an archived object in a digital library (Brügger, 2012). Thus the study of web archives would be not only the study of the natively digital materials, but also of the effects of the archiving as well as the archive as institution. For example, the Library of Congress’s Twitter collection, to contain all tweets of all time, is to be made accessible to bone fide researchers. The archive has Twitter user profiles only per September 2011, and researchers likely will have to take account of the fact that Twitter’s terms of service changed a number of times (Library of Congress, 2013).

Digital methods have a general research strategy, or set of moves, that have certain affinities with an online software project. First, stock is taken of the available digital objects, such as hyperlinks, tags, retweets, shortened URLs, Wikipedia edits, anonymous user IP addresses, timestamps, likes, shares, comments and others. Subsequently it is asked, how do the devices online handle these objects? How may we learn from online method? Here the sociological imagination or social research outlook enters the purview. How to repurpose the online methods and the devices so as to study not online culture or the virtual society, but cultural condition and societal change? At that point, the question of triangulation and benchmarking arises. How to ground the findings made with online data? Must we step offline to do so, or may we combine online and offline data and methods?

## DIGITAL METHODS AS A RESEARCH PRACTICE

Given certain devices or platforms (e.g., Internet Archive, Google Web Search, Wikipedia, Facebook, and Twitter), how may they be studied for social research purposes? It should be said at the outset that digital methods are often experimental and situational, because they developed in tandem with the medium conditions, and occasionally are built on top of other devices. They may be short-lived, as certain services are discontinued. They may fall victim to changes made by a platform, such as when a service is discontinued, advanced search in social media is removed, or if an API (application programming interface) is discontinued. When there are such changes research may be affected or perhaps discontinued, as was the case for many projects when Twitter changed its terms of service, and no longer allowed making tweet collections that could be stored and shared (Watters, 2011). Here, adding to Thelwall above, the researcher skeptical of

the value of web data becomes wary of the instability of the infrastructure that provides it.

In the following, the Internet Archive, Google Web Search, Wikipedia, Facebook, and Twitter is each taken in turn for the opportunities afforded for social research purposes, a la digital methods. For each the question is what digital objects are available, how are they handled by the device, and how can one learn from medium method, and repurpose it for social research. The interface on the Internet Archive, the Wayback Machine, has as its input a single URL. One is returned the stored pages of that URL since as far back as 1996. One also may have uniques returned. The research practice that has been developed follows from the Wayback Machines single-site focus, parlaying it into single-site histories. Changes to the interface of a homepage are captured, screengrabbed, placed in chronological order, and played back, in the style of time-lapse photography. A voiceover track is added, where the suggested approaches (among others) concern how the history of a single web site can tell the history of the web, the collision between old and new media (such as the history of an online newspaper), or the history of an institution (such as [whitehouse.gov](http://whitehouse.gov)). Making a single site history as a movie builds on particular, well-known screencast documentaries, especially "Heavy Metal Umlaut," made of the evolution of the Wikipedia article of that same name, in a sense telling the story of Wikipedia's editing culture (Udell, 2005). One example of such a single-site history screencast documentary, made from screenshots taken from the Wayback Machine of the Internet Archive, is "Google and the Politics of Tabs" (Govcom.org, 2008). By examining the changes to the search services privileged (as well as relegated) by Google.com on its interface over time tells the story of the demise of the human editors of the web (and the web directory), and the rise of the algorithm and the back-end taking over from the librarians.

Google web search has become so familiar that it requires some distancing efforts to consider its potential as a social research tool over its everyday value as a consumer information appliance. Google treats such digital objects as hyperlinks, clicks, and freshness, together with the neighborhood sites are considered to be in, in order to rank sites (Cutts, 2006). It is a ranking and also status-authoring machine for sources per key word, based on algorithmic notions of relevance. (Relevance increasingly relies on users' clicks and the page's freshness over how sites are linked, as in the past). Thus one could view the results of the query (climate change) as a list of web sites, mainly organizations, ranked according to relevance. Once one has a list of the "top sources" for climate change, one could query each source for the names of climate change skeptics, noting how close to the top of engine returns each appears (and with which frequency). "Source distance" is the name given to this two-step method which seeks to measure distance from the top of the

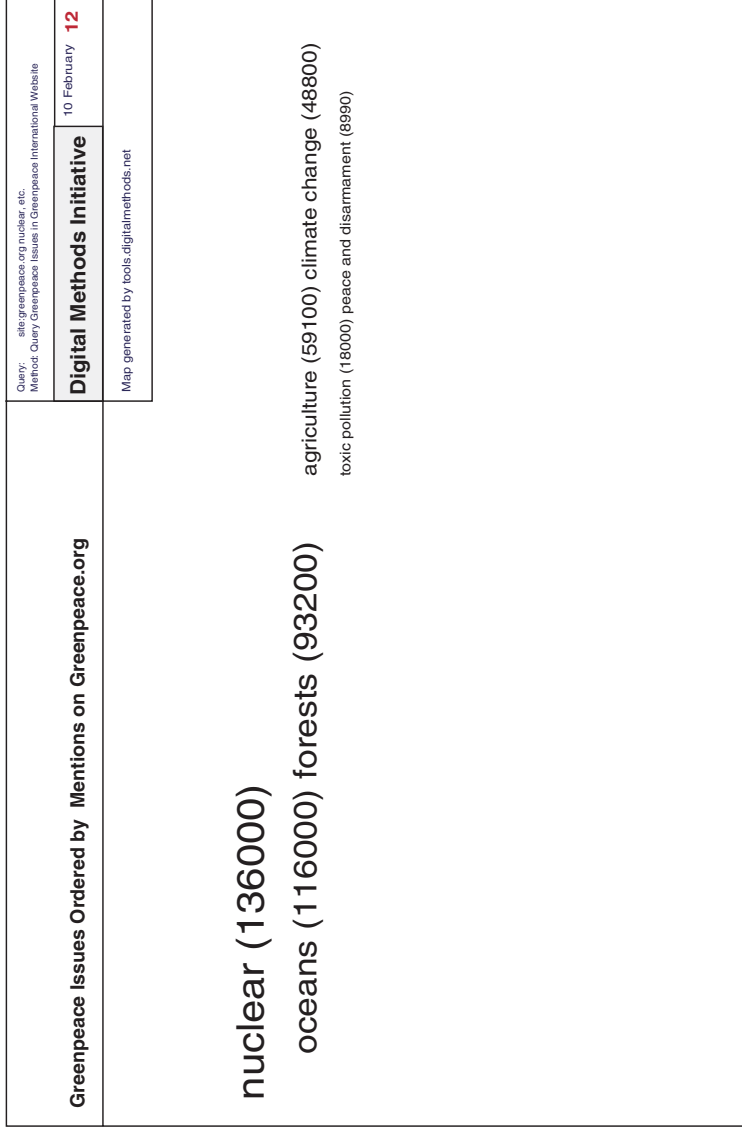
web for a given name or sub-issue, in a larger issue space (Rogers, 2013). It is the web equivalent of studying the top of the news (Figure 4).

Apart from a ranking machine, Google is also a massive indexing machine, meaning, for the user, that the contents of web sites may be queried, not only for single terms, but multiple ones, so as to gain a sense of which words appear more frequently than other ones. For example, Greenpeace.org is queried for all its campaigns, individually, to gain a sense of which campaigns have greater internal resonance than others, at least according to the number of mentions of its web site (Figure 5). One may also query multiple web sites for single terms, or for numerous terms. For example, one could query human rights web sites for different sorts of terms—such as campaigns and sub-issues—to gain a sense of the significance of each across the range of organizations. One could imagine seeking to begin the study of the agendas of the global human rights network in such a manner. This is precisely the purpose of the Lippmannian Device, which allows the user to create source clouds (which sources mention which issues) and issue clouds (which issues are mentioned by the given sources).

Wikipedia, the online encyclopedia, has a series of principles which its editors follow in order to have its articles achieve and retain “encyclopedia-ness,” namely, neutral point of view (NPOV), no original research and source verifiability. It is also routinely returned in the top results of Google for substantive queries (compared to navigational and transactional ones), making it a highly visible source of reference for its users. How would a digital methods researcher approach it? Wikipedia, however, also has language versions, and each article has links to its other language versions, so that the researcher can view the collection of articles on the one subject across the various Wikipedias. If the articles are not translated, then they are available for cross-cultural (or cross-linguistic) analysis. What may be compared? Each article has a series of digital objects such as anonymous edits with the IP address of those editors, whose location can be looked up. Thus one can study the places of edits. It also has a revision history and a discussion history, so one can study the intensity of editing as well as of debate. Furthermore, there are the article’s title, editors (including bots), table of contents, images and references. All may be compared. Projects such as Manypedia and Omnipedia have automated means of comparison of Wikipedia articles across language versions, which the former calls LPOV, or language points of view. Instead of a reference work, Wikipedia becomes the source of study for cultural reference, or even national point of view. One case in point is the Srebrenica massacre, which is how it is titled in the Serbian version, the Srebrenica genocide (Bosnian), and the Fall of Srebrenica (Dutch). A comparison of three significant parties to the events of July 1995, when some 6000–8000 (Serbian), 8000 (Bosnian), or



**Figure 4** Source cloud. Climate change skeptic's presence in the top Google results for the query (climate change), July 2007.  
Source: Alan McConchie, popvssoda.com. Map by Campbell and Plumb, 2003.

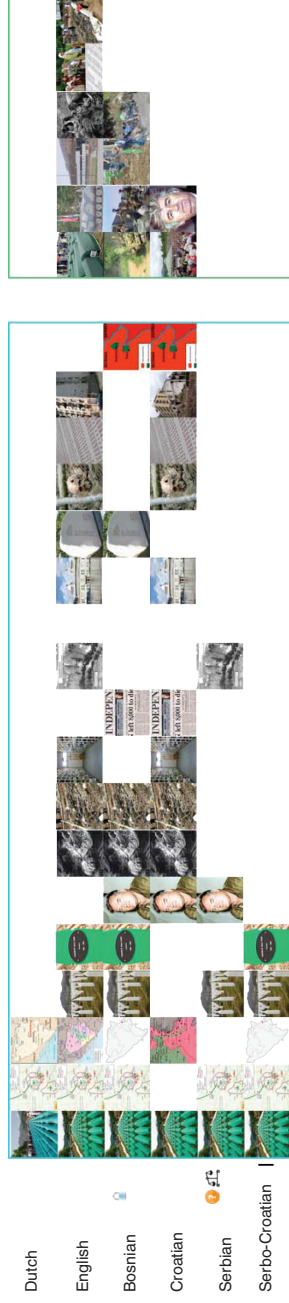


**Figure 5** Issue cloud. Greenpeace campaigns mentioned on Greenpeace.org's web site, February 2012. Source: Output by the Lippmannian Device, Digital Methods Initiative, Amsterdam.

7000–8000 (Dutch) Bosnians were killed. The Bosnian entry has distinctly different images, including a 13-year-old boy's grave, which given that he was not of fighting age would be evidence of genocide (Figure 6). The Dutch emphasize the military side of the story, and the Serbian, once similar in that respect to the Dutch, is alone in providing a section on the events according to the Republika Srpska, the part of Bosnia and Herzegovina where the town of Srebrenica is located. The articles also do not share references, or editors. The differences between the articles, not to mention the differences in locations of the edits as well as the activities of the editors, provide materials for the study of cultural memory as well as controversiality, which has prompted scholars to encourage home-grown articles over translations from the English-language Wikipedia (Callahan & Herring, 2011).

The digital objects much studied on Facebook are the friend as well as the privacy settings (Lewis, Kaufman, & Christakis, 2008; Lewis, Kaufman, Gonzalez, Wimmer, & Christakis, 2008). Using its API and the application Netvizz, for example, one could perform ego network research, pulling in the available data from yourself and your friends. Facebook's other digital objects include the profile, which provide the opportunity to study what I refer to as postdemographics—the media preferences and tastes of sets of social media users (Rogers, 2009b). In experimental work employing the advanced search of MySpace (which later was discontinued), compatibility comparisons were made of the interests of John McCain's friends and those of Barack Obama, prior to the 2008 US presidential elections where the two faced off (Figure 7). Here the profiles are repurposed to inquire into the so-called culture wars, considering the extent of the polarization between red (Republican) and blue (Democratic) supporters. On Facebook the operative digital objects for such analysis are the page and the group, together with what one may do there: like, share, comment and like a comment. Of interest would be the study of online networks and their content which are otherwise less in view, such as extremists. By liking a page, such as Stop Islamization of the World, one has access to data set of the likes, shares, comments and liked comments, also longitudinally. One is able to determine which content (and which content types) has elicited engagement (including which types of engagement). What is engaging to those who like, comment, and share on Stop Islamization of the World (Figure 8)? One also may join a group, and in doing so, gain access to its data.

In its early study tweets from Twitter were categorized as banal or having pass-along value, which eventually would be codified by its users as RT, or retweets, or those tweets of such interest that they should be tweeted again (Rogers, 2013b). The retweet was joined by other digital objects fashioned by its users, especially the hashtag, which would group content by subject, and particularly by event. Retweeted tweets per hashtag became a



**Figure 6** Wikipedia as the study of cultural point of view. Comparison of images present on the Srebrenica article in the Dutch, English, Bosnian, Croatian, Serbian, and Serbo-Croatian Wikipedia language versions, December 20, 2010. Source: Output by the Cross-Lingual Image Analysis Tool, Digital Methods Initiative, Amsterdam.

PROFILE BASED ON OBAMA	COMPATIBILITY	PROFILE BASED ON MCCAIN
<p><b>General</b></p> <p>. barack obama, reading, music, writing, history, politics, movies, traveling, friends, bob marley, peace, family guy, chicago, books, democrats, photography, harry potter, as a u, running, the producers, jon stewart, the daily show, art, senator, macs, napoleon dynamite, conan o'brien, sleeping, guinness, italy, the beatles, psychology, cats, dancing, shopping, summer, poetry, democracy, voting, gay rights, obama, mythology, coffee, borat, concerts, george orwell, david searls, icons, buddhism, ingrid chavez, pink floyd, piano</p> <p><b>Music</b></p> <p>radiohead, belle and sebastian, the beatles, the roots, pearl jam, sufjan stevens, bob dylan, swervedriver, the smiths, cat power, nina simone, amy winehouse</p> <p><b>Movies</b></p> <p>little miss sunshine, american beauty, goodfellas, alice in wonderland, amelie, a clockwork orange, archangel, secretary, memento, magnolia, lost in translation, rushmore</p> <p><b>Television</b></p> <p>the office, arrested development, weeds, the daily show, lost, heroes</p> <p><b>Books</b></p> <p>atlas shrugged, books, alice in wonderland, gone with the wind, harry potter, 1984, america: the book, catch-22, josef margale, gravity's rainbow, the baroque cycle (quicksilver, being written before obama decided on a political career it offers an honest introspective look that few other politicians could ever have offered:</p>	17%	<p><b>General</b></p> <p>yes, sometimes, edwin mccain, friends, reading, music, swimming, concerts, piercings, shopping, com, movies, metal, get paid to take surveys!, create your own!, (8), death metal, art, what?, sometimes, laguna beach, il mccain, john mccain, the beach, fugazi, freya, from autumn to ashes, full blown chaos, from first to last</p> <p><b>Music</b></p> <p>the shins, com, the bravery, coldplay, evanescence, guster, blink 182, dido, green day, taking back sunday, the used</p> <p><b>Movies</b></p> <p>love actually, old school, garden state, kill bill, pirates of the caribbean, napoleon dynamite, shawshank redemption, big fish, princess bride</p> <p><b>Television</b></p> <p>family guy, project runway, top chef, america's next top model, cs, desperate housewives, lust</p> <p><b>Books</b></p> <p>to kill a mockingbird, com, me talk pretty one day, "sneaking into the flying circus, it's just enough to get the blood pumping!"</p>
<p><b>Heroes</b></p> <p>my mom, , johnny cash</p>	0%	<p><b>Heroes</b></p> <p>haha, kat von d, john lennon, barry goldwater</p>

**Figure 7** Aggregated profiles of the interests of top 100 friends of Barack Obama and John McCain, MySpace.com, September 2008. Source: Analysis and output by Elfriendo.com, Govcom.org Foundation and Digital Methods, Initiative, Amsterdam.





**Figure 8** Most engaged with content, resized according to quantities of likes, comments and shares, on the Facebook page, Stop Islamization of the World, January, 2013. Source: Analysis by Ana Crisostomo, Despoina Mountanea, Juliana Marques, and Joe Mier. Digital Methods Initiative, Amsterdam.



means of studying significant tweets of the day, such as the Iran Elections and their aftermath in June 2009. How to repurpose the stream? In an effort “debanalize” Twitter, one digital methods approach has been to invert the reverse chronological order of Twitter, and place the most significant retweets per hashtag in chronological order, so as to tell the story of an event from Twitter, or the Twittiverse, as it sometimes referred to (Figure 9). Here the key question remains the relationship between what is happening on the ground, and in social media—a debate that has been led by Evgeny Morozov, who quotes the Al Jazeera’s head of new media as saying that during the Iran Election crisis there were perhaps six Twitter users tweeting from the ground in Tehran (Morozov, 2011).

Twitter, the company, began to recognize, as its cofounder Jack Dorsey related, that it does “well at natural disasters, man-made disasters, events, conferences, presidential elections” (Sarno, 2009). It changed its slogan in 2009 from “what are you doing?” to “what’s happening?,” indicating a shift from Twitter as a friend-following tool (for ambient intimacy) to a news medium for following events, especially elections and disasters. Here Twitter becomes a data set, not only of commercial but also historical value, indicated by the significance of Library of Congress’s embracing Twitter as a digital archival project. Similar to the Wayback Machine of the Internet Archive, it not only will differ in user experience from the live version, but also will have its own interface to be potentially repurposed for social research. Routines to build tweet collections and to output them as event chronologies may be among the scholarly uses.

## REFERENCES

- Borgman, C. (2009). The digital future is now: A call to action for the humanities. *Digital Humanities Quarterly*, 3(4), 1–30.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679.
- Brügger, N. (2012). When the present web is later the past: Web historiography, digital history, and Internet studies. *Historical Social Research*, 37(4), 102–117.
- Callahan, E. S., & Herring, S. C. (2011). Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology*, 62(10), 1899–1915.
- Campbell, M. T. & Plumb, G. (2003). *Generic names for soft drinks by county, The Pop vs. Soda Page*. Retrieved from <http://www.popvssoda.com/countystats/total-county.html>.
- Chen, E. (2012, July 6). *Soda vs. Pop, Edwin Chen’s Blog*. Retrieved from <http://blog.echen.me/2012/07/06/soda-vs-pop-with-twitter/>.

- Cutts, M. (2006, May 16). *Indexing timeline, Matt Cutts: Gadgets, Google, and SEO*. Retrieved from <http://www.mattcutts.com/blog/indexing-timeline/>.
- Dean, J. (1998). *Aliens in America: Conspiracy cultures from outerspace to cyberspace*. Ithaca, NY: Cornell University Press.
- Dohmen, J. (2007). Opkomst en ondergang van extreemrechtse sites, NRC Handelsblad, August 25.
- Ericson, M. & Cox, A. (2009). What's cooking on thanksgiving. New York Times, November 26.
- Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012–1014.
- Govcom.org (2008). *Google and the politics of tabs, movie*. Retrieved from <https://movies.issuecrawler.net/google.html>.
- Hine, C. (Ed.) (2005). *Virtual methods: Issues in social research on the Internet*. Oxford, England: Berg.
- Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York, NY: New York University Press.
- Lazer D., Pentland, A., Adamic, L., Aral, S., Barabasi, A-L., Brewer, D., ... , & Van Alstyne, M. (2009). Computational social science, *Science*, 323, 5915, 721–723.
- Lewis, K., Kaufman, J., & Christakis, N. (2008). The taste for privacy: An analysis of college student privacy settings in an online social network. *Journal of Computer-Mediated Communication*, 14(1), 79–100.
- Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30(4), 330–342.
- Library of Congress (2013). Update on the Twitter Archive at the Library of Congress (White Paper), January.
- Mager, A. (2012). Search engines matter: From educating users towards engaging with online health information practices. *Policy & Internet*, 4(2), 1–21.
- Manovich, L. (2007). *Cultural analytics: Analysis and visualization of large cultural data sets*. A proposal from Software Studies Initiative @ CALIT2, September 30.
- McConchie, A. (2002). *The Pop vs. Soda Page*. Retrieved from <http://www.popvsoda.com>.
- Michel, J-B, Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., ... , & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books, *Science*, 331, 6014, 176–182.
- Moretti, F. (2005). *Graphs, maps, trees: Abstract models for a literary history*. London, England: Verso.
- Morozov, E. (2011). *The net delusion: The dark side of Internet freedom*. New York, NY: Public Affairs.
- Raehsler, L. (2012, April 18). *What people search for—most popular keywords, search engine watch*. Retrieved from <http://searchenginewatch.com/article/2066257/What-People-Search-For-Most-Popular-Keywords>.
- Rieder, B. (2012). What is in PageRank? A historical and conceptual investigation of a recursive status index. *Computational Culture*, (2). [http://computationalculture.net/article/what\\_is\\_in\\_pagerank](http://computationalculture.net/article/what_is_in_pagerank)

- Rogers, R. (2009a). *The end of the virtual: Digital methods*. Amsterdam, The Netherlands: Amsterdam University Press.
- Rogers, R. (2009b). Post-demographic machines. In A. Dekker & A. Wolfsberger (Eds.), *Walled garden* (pp. 29–39). Amsterdam, The Netherlands: Virtueel Platform.
- Rogers, R. (2013a). *Digital methods*. Cambridge, MA: MIT Press.
- Rogers, R. (2013b). Debanalizing Twitter: The transformation of an object of study. In A. Bruns, J. Burgess, K. Weller, C. Puschmann & M. Mahrt (Eds.), *Twitter and society*. New York, NY: Peter Lang.
- Sarno, D. (2009). Jack Dorsey on the Twitter ecosystem, journalism and how to reduce reply spam. Part II, *Los Angeles Times*, February 19.
- Thelwall, M., Vaughn, L., & Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology*, 39, 81–135.
- Udell, J. (2005). *Heavy Metal Umlaut, movie*. Retrieved from <http://jonudell.net/udell/gems/umlaut/umlaut.html>.
- Watters, A. (2011, March 3). *How recent changes to Twitter's terms of service might hurt academic research*, ReadWriteSocial. Retrieved from [http://readwrite.com/2011/03/03/how\\_recent\\_changes\\_to\\_twitthers\\_terms\\_of\\_service\\_mi](http://readwrite.com/2011/03/03/how_recent_changes_to_twitthers_terms_of_service_mi).
- Watts, D. J. (2007). A Twenty-first century science. *Nature*, 445, 489.
- Woolgar, S. (Ed.) (2003). *Virtual society? Technology, cyberbole, reality*. New York, NY: Oxford University Press.

## FURTHER READING

- Borra, E. K., & Weber, I. (2012). Political Insights: Exploring partisanship in web search queries. *First Monday*, 17, 7.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5), 662–679.
- Marres, N., & Weltevrede, E. (2013). Scraping the social? Issues in real-time social research. *Journal of Cultural Economy*, 6(3), 313–335.
- Rieder, B., & Röhle, T. (2012). Digital methods: Five challenges. In D. M. Berry (Ed.), *Understanding digital humanities* (pp. 67–84). Palgrave Macmillan: Basingstoke, England.
- Rogers, R. (2010). Internet research: The question of method. *Journal of Information Technology and Politics*, 7(2/3), 241–260.

## RELATED BOOKS

- Digital Methods by Richard Rogers (MIT Press), <http://mitpress.mit.edu/books/digital-methods>
- Information Politics on the Web by Richard Rogers (MIT Press), <http://mitpress.mit.edu/books/information-politics-web>
- Related Tools and Methods Platforms
- Digital Methods Initiative, <http://www.digitalmethods.net/>

Govcom.org Foundation, <http://www.govcom.org/>  
Issue Crawler Mapping software, <http://www.issuecrawler.net/>  
Issue Mapping Online, <http://issuemapping.net/>  
Mapping Controversies, <http://mappingcontroversies.net/>

## RICHARD ROGERS SHORT BIOGRAPHY

**Richard Rogers** is University Professor and holds the Chair in New Media & Digital Culture at the University of Amsterdam, the Netherlands. He is Director of Govcom.org, the group responsible for the Issue Crawler and other info-political tools, and the Digital Methods Initiative, dedicated to developing methods and tools for online social research. Recently, Rogers was Annenberg Fellow at the Annenberg School for Communication at the University of Pennsylvania and Visiting Fellow at MIT. Over the past decade, Rogers has received research grants and tool development funding from the Open Society Institute, Ford Foundation, and MacArthur Foundation. Rogers is author of *Information Politics on the Web* (MIT Press, 2004) and *Digital Methods* (MIT Press, 2013).

## RELATED ESSAYS

To Flop Is Human: Inventing Better Scientific Approaches to Anticipating Failure (*Methods*), Robert Boruch and Alan Ruby  
Ambulatory Assessment: Methods for Studying Everyday Life (*Methods*), Tamlin S. Conner and Matthias R. Mehl  
Models of Nonlinear Growth (*Methods*), Patrick Coulombe and James P. Selig  
Quantile Regression Methods (*Methods*), Bernd Fitzenberger and Ralf Andreas Wilke  
Ethnography in the Digital Age (*Methods*), Alan Howard and Alexander Mawyer  
Participant Observation (*Methods*), Danny Jorgensen  
Structural Equation Modeling and Latent Variable Approaches (*Methods*), Alex Liu  
Data Mining (*Methods*), Gregg R. Murray and Anthony Scime  
Digital Methods for Web Research (*Methods*), Richard Rogers